MASTERARBEIT | MASTER'S THESIS

Titel | Title

Wie bewertet man Bewertungskompetenz? ChatGPT als Unterstützung in der Beurteilung der S-Kompetenz

verfasst von | submitted by

Markus Humer BEd

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Education (MEd)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 199 502 523 02
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Lehramt Sek (AB) Unterrichtsfach Biologie und Umweltbildung Unterrichtsfach Physik
Betreut von | Supervisor:	Univ.-Prof. Dr. Martin Hopf
Mitbetreut von | Co-Supervisor:	Mag. Dr. Marianne Korner

Inhaltsverzeichnis

Danksagung

Diese Masterarbeit wäre ohne die Unterstützung einiger Personen nicht möglich gewesen, denen ich meinen Dank aussprechen möchte.

Zuerst möchte ich mich bei meiner Betreuerin, Frau Mag. Dr. Marianne Korner, bedanken. Sie stand mir jederzeit mit fachlichen Tipps und freundschaftlichem Rat zur Seite und unterstützte mich maßgeblich beim Gelingen dieser Arbeit. Die vertrauensvolle und freundschaftliche Betreuung schätzte ich sehr.
Weiters gilt mein Dank meinen Studienkolleginnen Denise Gstall und Karoline Hackner, deren Daten und Modelle ich in meiner Masterarbeit verwenden durfte.
Außerdem möchte ich mich bei den Universitätsmitgliedern bedanken, die mir konstruktive Rückmeldungen zu meiner Pilotstudie gaben, sowie allen freiwilligen Lehrer:innen und Schüler:innen, die die Durchführung meiner Hauptstudie überhaupt erst ermöglicht haben. Vielen Dank an alle!

Diese Arbeit markiert darüber hinaus das Ende meiner Studienzeit in Wien, in der mich sehr viele Menschen begleitet, unterstützt und geprägt haben. Sie haben die letzten Jahre zu etwas Besonderem gemacht und auch diesen Menschen möchte ich danken.

Der allergrößte Dank gilt meiner Familie, besonders meinen Eltern, Großeltern und meinem Bruder. Danke, dass ihr immer für mich da seid, mich unterstützt und das Studium für mich ermöglicht habt!
Zu einer bereichernden Studienzeit gehören auch viele gute Freunde, die ich an der Uni Wien, in meinem Studentenheim, in Sportkursen oder sonstigen Gelegenheiten kennenlernen durfte oder bereits zuvor kannte. Sie haben mich in den letzten Jahren auf verschiedenste Weise geprägt und zu der Person gemacht, die ich heute bin. Danke auch an euch alle für die schöne Zeit und eure Unterstützung!

Danke!

Wien, November 2025

Kurzfassung

In dieser Masterarbeit wird der Frage nachgegangen, ob das generative KI-Sprachmodell ChatGPT zur Bewertung von Bewertungskompetenz eingesetzt werden kann.

Das Ziel bestand darin, zu ermitteln, ob und wie mithilfe von ChatGPT reproduzierbare und valide Beurteilungen auf Basis der Vorgaben zweier evaluierter Bewertungsmodelle in praktikabler Weise generiert werden können.

Die Grundlage dieser Arbeit bilden die empirischen Arbeiten von Gstall (2024) und Hackner (2025). Diese versuchten dem herausfordernden Problem der Bewertung von Bewertungskompetenz mit neu entwickelten Bewertungsmodellen entgegenzuwirken. Im explorativen Vorgehen wurde ausgelotet, welche Möglichkeiten für eine Beurteilung durch ChatGPT bestehen. Auf Basis der gewonnenen Erkenntnisse wurden die Modelle von Gstall (2024) und Hackner (2025) sowie eine durchdachte Prompting-Strategie genutzt, um Testreihen mit *ChatGPT-4o* und teilweise mit *ChatGPT-5* durchzuführen. Für die Beurteilung durch ChatGPT wurden Aufgaben und Antworten aus den Masterarbeiten von Gstall und Hackner verwendet. Die durchschnittlichen Beurteilungen von ChatGPT wurden mit denen von Physiklehrkräften verglichen und die Ausgaben des Chatbots wurden grob analysiert. Die Auswertung erfolgte größtenteils qualitativ durch Analyse der Ergebnisse.

Dabei zeigte sich, dass *ChatGPT-4o* und *ChatGPT-5* modellkonforme Beurteilungen zur Bewertungskompetenz generieren können. Werden diese als Durchschnitt aus mehreren Durchgängen berechnet weisen sie eine mittlere bis hohe Übereinstimmung mit der durchschnittlichen Beurteilung mehrerer Physiklehrkräfte auf. Das Modell von Hackner (2025) scheint besser für die Verwendung mit ChatGPT geeignet zu sein. ChatGPT zeigt eine höhere Konsistenz bei der Beurteilung der gleichen Antwort nach dem Bewertungsmodell als Physiklehrkräfte untereinander. Ein negativer Aspekt der Verwendung des Chatbots ist, dass die Ausgaben fehlerhaft sein können. Dies kann vor allem bei einmaliger Generierung von Beurteilungen problematisch sein. Um die Sicherheit der Beurteilungen zu erhöhen, sind mehrere Bewertungsdurchgänge nötig, was allerdings den Zeitaufwand erhöht.

Die Arbeit kommt zu dem Schluss, dass *ChatGPT-4o* und *ChatGPT-5* als unterstützendes Instrument zur Beurteilung der Bewertungskompetenz verwendet werden können, da die Beurteilungen im Durchschnitt nahe an denen Beurteilungen einer Physiklehrkraft liegen. Angesichts der unkomplizierten und schnellen Möglichkeit KI-gestützter Beurteilungen können Physiklehrkräfte auf diese als Unterstützung zurückgreifen. Die Erkenntnisse legen nahe, dass diese gute Richt- oder Vergleichswerte sein können, jedoch nie als alleinige Grundlage für eine Beurteilung verwendet werden sollten.

1. Einleitung

Intelligente Sprachmodelle sind mittlerweile breitgestreut in vielen Arbeitsbereichen unserer Gesellschaft etabliert. Sie können sinnvolle, menschenähnliche Konversationen führen, Fragen beantworten und darüber hinaus Informationen aus Dokumenten evaluieren und in gewünschter Form präsentieren. In dieser Masterarbeit wird der Einsatz eines solchen Sprachmodells untersucht, und zwar als Beurteilungshilfe für Physiklehrkräfte.

Motivation

Die *Bewertungskompetenz* (auch *S-Kompetenz*) ist eine von drei zentralen Handlungsdimensionen naturwissenschaftlichen Unterrichts in Österreich. Laut Lehrplan sollen die S-Kompetenz ihrer Schüler:innen zum *Bewerten, Entscheiden und Handeln* festigen, ausbauen und vor allem auch überprüfen. Die Bewertung der Bewertungskompetenz erweist sich jedoch als keine einfache Aufgabe, da sich der Beurteilungsprozess von S-Kompetenzaufgaben als komplex und zeitaufwendig herausstellt. In offenen Aufgabenformaten muss sich eine Lehrperson mit gänzlich verschiedenen Bewertungen und Argumentationslinien einer ganzen Klasse auseinandersetzen. Die Schwierigkeit dabei ist eine möglichst objektive und einheitliche Beurteilung vieler subjektiver Antworten zu erreichen. Im Optimalfall sollte die Lehrperson zudem zu jeder Antwort ein Feedback bereitstellen, um jede:n Schüler:in bestmöglich zu fördern. Dafür, dass einzelne Bewertungsaufgaben meist von kleinem Umfang sind und beispielsweise eine Aufgabe in einem Test oder einer Schularbeit bilden, ist der notwendige Beurteilungsaufwand bei sorgfältiger Auseinandersetzung in Summe sehr groß – ganz zu schweigen von dem Aufwand für die Erstellung und Planung einer Bewertungsaufgabe. Wenn die Bewertungskompetenz ein wichtiger Bestandteil im Physikunterricht sein soll, muss auch ein effektiver Weg für ihre Beurteilung gegeben sein, etwa mithilfe von unterstützenden Beurteilungsmodellen zur S-Kompetenz für Physiklehrkräfte.

Problemstellung

Eines der zentralen Probleme bei der Bewertung der Bewertungskompetenz war bis vor Kurzem das Fehlen eines leicht anzuwendenden Beurteilungsmodells. Gstall (2024) hat ein neues Modell entwickelt, da Lehrpersonen die Beurteilung von S-Kompetenzaufgaben erleichtern soll. Das Modell schafft es mehr Transparenz und Objektivität bei der Bewertung subjektiver Antworten ist leichter zu handhaben als andere Beurteilungsmodelle. Dennoch bleiben einige Probleme für den Beurteilungsprozess der S-Kompetenz bestehen. Selbst mit Gstalls Modell ist die Beurteilung nach wie vor zeitintensiv. Auch wenn Lehrpersonen damit eingearbeitet sind, kann die Beurteilung uneindeutiger Antworten schwierig sein. Zeitaufwand und Komplexität sind also nach wie vor Faktoren, die den Beurteilungsprozess langwierig machen.

Eine Möglichkeit, diese Probleme zu lösen oder zumindest lindern könnte, sind intelligente *Sprachmodelle* bzw. *Large Language Models* (*LLMs*). Durch ihre rasante Entwicklung in den letzten Jahre gewinnen diese Technologien immer mehr beeindruckende Fähigkeiten. Viele davon könnten im schulischen Alltag gewinnbringend eingesetzt werden. Im Kontext des

skizzierten Problems ergibt sich für diese Arbeit das Forschungsinteresse, ob der Chatbot *ChatGPT* (OpenAI, 2022) die Bewertung der S-Kompetenz teilweise oder sogar vollständig übernehmen kann. Die Auslagerung dieses Arbeitsaufwands an KI-Modelle könnte den anspruchsvollen Berufsalltag von Lehrpersonen erheblich entlasten.

Zielsetzung

Der zentrale Forschungsgegenstand dieser Masterarbeit beschäftigt sich mit der Frage, ob sich die Beurteilung von Bewertungskompetenz anhand von evaluierten Beurteilungsmodellen (Gstall, 2024; Hackner, 2025) mithilfe des generativen Sprachmodells ChatGPT automatisieren lässt und ob sich der Beurteilungsprozess als praktikabel und reproduzierbar erweist und valide Bewertungen erzeugt. Zunächst soll durch exploratives Vorgehen geprüft werden, ob eine Bewertung durch ChatGPT grundsätzlich möglich ist und wo erste Grenzen und Potenziale liegen. Auf Basis dieser Erkenntnisse kann eine Methode definiert werden anhand derer ChatGPT ausführlich getestet wird. Sollte sich herausstellen, dass dieser spezifische Einsatz des Chatbots aus diversen Gründen erschwert wird, kritisch zu betrachten ist oder gänzlich ungeeignet erscheint sollen diese Gründe innerhalb der Arbeit auch entsprechend festgehalten und diskutiert werden.

Aufbau der Arbeit

Im Theorieteil dieser Masterarbeit werden zwei große Themenbereiche behandelt: Zuerst die Bewertung der Bewertungskompetenz, siehe Kap. 2: Hier wird erläutert, was Kompetenz im schulischen Kontext überhaupt bedeutet, worauf Lehrkräfte beim Beurteilen von Leistungen achten müssen und welchen Beitrag die Arbeiten von Gstall (2024) und Hackner (2025) zur Bewertung der Bewertungskompetenz leisten, wo ihre Forschungen limitiert sind und wo die vorliegende Masterarbeit ansetzen kann. Das zweite Theoriekapitel beschreibt künstliche Intelligenz (KI), siehe Kap. 3: Dort wird geschildert, was KI eigentlich ist, welche Mechanismen KI-Modelle ausmachen und wie sie funktionieren. Es wird auch auf die Entwicklung und den Eigenschaften des Chatbots *ChatGPT* eingegangen. Kapitel 3 geht auch auf aktuelle Kenntnisse zum Einsatz von ChatGPT als Beurteilungshilfe ein.
In Kapitel 4 wird das Forschungsdesign der vorliegenden Masterarbeit genau erläutert. Es beschreibt nochmal einmal das Forschungsvorhaben, die Forschungsfragen, die Methoden und Materialien und erläutert die Planung und Durchführung der zweiteiligen, empirischen Studie. Diese ist in eine breite Pilotstudie, siehe Kap. 5, und eine vertiefende Hauptstudie, siehe Kap. 6, geteilt. Beide Studienkapitel fassen für sich die Ergebnisse zusammen und sie werden an geeigneter Stelle bereits diskutiert.
Die Diskussion, siehe Kap. 7, und das Fazit mit Ausblick auf weiterführende Forschungen, siehe Kap. 8, runden die Masterarbeit abschließend ab.

2. Bewertung von Bewertungskompetenz

Der erste Teil der theoretischen Einführung behandelt die Rolle der Bewertungskompetenz im naturwissenschaftlichen Physikunterricht und die Herausforderungen, die sich für Lehrkräfte bei der Bewertung dieser Kompetenz ergeben.

2.1. Bildungsstandards, Kompetenzmodelle und Kompetenzen

Die Leistungen von österreichischen Schüler:innen in den internationalen Vergleichsstudien *TIMSS* und *PISA* um die Jahrtausendwende war der Anstoß für die damalige Bildungspolitik, notwendige Änderungen im Schulsystem zu veranlassen. Die Studien überraschten mit dem Ergebnis, dass österreichische Schulkinder in diversen Aufgaben im internationalen Vergleich schlechter abschneiden. Das offenbarte Bildungsdefizit wurde auf eine schlechte Qualität des damaligen Schulsystems zurückgeführt. In der Folge wurde eine innovative Weiterentwicklung des Schulwesens gefordert, wofür eine zentrale Idee umgesetzt wurde: Zukünftig sollte die Qualität von Bildung an allgemeinen *Bildungsstandards* orientiert sein und nicht mehr am Input der Lehrer:innen, sondern am Output der Schüler:innen bemessen werden. Der österreichische Nationalrat richtete zur Umsetzung dieses Vorhabens 2008 das Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens (BIFIE) ein, dessen Kernaufgaben darin lagen, diese neuen allgemeinen Bildungsstandards zu entwickeln, Messinstrumente zur informellen Kompetenzmessung zu definieren und eine standardisierte Reifeprüfung zu erstellen (Kulmhofer-Bommer & Diekmann, 2021, S. 425–426; Lux, 2022, S. 1).

Das BIFIE (heute IQS) definiert Bildungsstandards als:

> „ *[…] konkret formulierte Lernergebnisse, die sich aus den Lehrplänen ableiten lassen. Sie definieren Kompetenzen, die in der Regel von allen Schülerinnen und Schülern an den Schnittstellen des Schulsystems erreicht werden sollen*.“ (Breit et al., 2012, S. 5)

Mithilfe der Bildungsstandards nahm die Planung und Durchführung von Unterricht erstmals eine *ergebnisorientierte* Ausrichtung an: Es geht in erster Linie darum, was bei den Lernenden ankommen soll und nicht, wie viel Fachwissen die Lehrenden im Unterricht behandeln können. Bildungsstandards sind als Richtungsweiser zu verstehen, die in Symbiose zum Lehrplan stehen:

> „*Bildungsstandards geben den Lehrerinnen und Lehrern Orientierung darüber, was Schüler/innen zu bestimmten Zeitpunkten ihrer Schullaufbahn können sollen und konkretisieren damit die Zielsetzungen des Lehrplans. Bildungsstandards und Lehrplan treten daher nicht in eine konkurrierende oder widersprüchliche Position, sondern ergänzen einander positiv.*“ (Breit et al., 2012, S. 5)

Die Bildungsstandards sehen den Erwerb von *Kompetenzen* vor, wie in der Definition des BIFIE nach Breit et al. (2012) beschrieben. In diesem Zusammenhang bezeichnet Kompetenz die Fähigkeit, gelerntes Wissen und geübte Fertigkeiten flexibel anzuwenden

und für den weiteren Bildungsweg zu nutzen. *Kompetenzorientierter* Unterricht sollte eine Vielzahl allgemeiner Konzepte verständlich machen, statt bloßes Faktenwissen zu vermitteln (Kulmhofer-Bommer & Diekmann, 2021, S. 426).

Dieses Verständnis des Kompetenzbegriffs geht nach Kulmhofer-Bommer & Diekmann (2021, S. 426) auf Franz Weinert (2001) und seine vielzitierte Definition von Kompetenzen zurück:

> „*Er* (F. Weinert) *beschreibt Kompetenzen als „die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können" (S. 27).*" (Weinert, 2001, zitiert nach Kulmhofer-Bommer & Diekmann, 2021, S. 426)

Erworbene Kompetenzen sind das Resultat von Lernprozessen und kontextanhängig, da sie in Wechselwirkung mit der Umwelt gebildet werden. Folglich gibt es verschiedene Kompetenzen, die zur Bewältigung unterschiedlicher Situationen und Aufgaben nützlich sind. Jede Kompetenz umfasst Wissen, kognitive Fähigkeiten und sozial-kommunikative oder motivationale Elemente. Die österreichischen Bildungsstandards legen fest, welche Kompetenzen bis Ende der 4. bzw. 8. Schulstufe nachhaltig erworben werden sollen. Der Schwerpunkt liegt dabei auf dem Erwerb *grundlegender fachbezogener Kompetenzen*. Diese bilden die wichtigsten Inhalte eines Faches ab und gelten als Basis für den weiteren Kompetenzaufbau (Breit et al., 2012, S. 5–6).

Gemeinsam mit den Kompetenzen hat das BIFIE *Kompetenzmodelle* entwickelt. Diese Modelle strukturieren die Bildungsstandards in einem Gegenstand und bauen auf fachdidaktischem und fachsystematischem Überlegungen auf. Kompetenzmodelle sind in mehrere *Kompetenzbereiche* unterteilt, in denen bereichsspezifische Kompetenzen detailliert beschrieben sind (IQS, 2025). Für die Gegenstände Deutsch, Mathematik und Englisch (bzw. lebende Fremdsprachen) wurden Kompetenzmodelle für die 4. bzw. 8. Schulstufe erstellt und gemeinsam mit Einführung der Bildungsstandards gesetzlich für den Unterricht verordnet (IQS, 2025).

Für die Naturwissenschaften sind keine offiziellen Bildungsstandards verordnet (IQS, 2025). Um das Jahr 2010 wurde an den AECCs aber ein *Kompetenzmodell für Naturwissenschaften* entworfen und dieses ist im aktuellen, gesetzlichen Lehrplan verankert (RIS, 2025c). Das Modell enthält Kompetenzbereiche mit den erwarteten Kompetenzen, die Schüler:innen am Ende der 8. Schulstufe in Biologie, Chemie und Physik erworben haben sollen (IQS, 2025). Es ist in drei Dimensionen gegliedert, die Handlungs-, Anforderungs- und Inhaltsdimension (BIFIE, 2011), siehe Abb. 1.

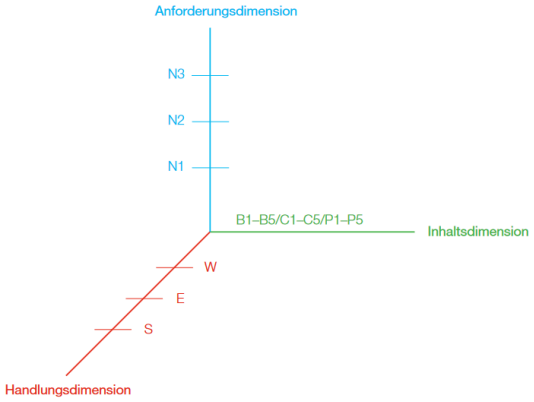


Abb. 1: Kompetenzmodell Naturwissenschaften, 8. Schulstufe (BIFIE, 2011)

Die *Inhaltsdimension* gibt für jeden Kompetenzbereich des naturwissenschaftlichen Gegenstandes an, welche Inhalte von Lernenden erworben werden sollen (BIFIE, 2011). Die im aktuellen Unterstufen-Lehrplan formulierten Inhalte (RIS, 2025c) unterscheiden sich etwas von ihrer ursprünglichen Gestaltung des BIFIE (2011). Von der 2. bis zur 4. Klasse Unterstufe sollen Schüler:innen mithilfe zentraler, fachlicher Konzepte (Teilchen, Feld, Schwingungen und Wellen, Energie, Kräfte und Wechselwirkungen) in sieben inhaltliche Kompetenzbereiche eingeführt werden – die angegebene Reihenfolge entspricht jener des Lehrplans: *„Sehen und Hören“*, *„optische Systeme“*, *„Mechanik“*, *„Elektrizität und Magnetismus“*, *„Energie“*, *„Wetter und Klima“* und *„Strahlung und Radioaktivität“*. Für jeden Kompetenzbereich sind (mögliche) Anwendungsbereiche definiert und die Inhalte sind mit einer von drei Handlungsdimensionen verknüpft, mehr dazu im nächsten Absatz. Die *Anforderungsdimension*, wie sie im Kompetenzmodell des BIFIE (2011) als dritte Säule verankert ist, lässt sich im aktuellen Lehrplan (RIS, 2025c) nicht mehr finden.

Die *Handlungsdimension* umfasst wesentliche Handlungskompetenzen, unterteilt in drei Handlungsdimensionen (BIFIE, 2011). Die vom BIFIE entworfenen Handlungsdimensionen sind im Lehrplan verankert (RIS, 2025c) und nach der Formulierung des Lehrplans gültig. Für die AHS-Unterstufe sind die Dimensionen wie folgt festgelegt:

> ***„Fachwissen anwenden (W)***
> *In diesem Bereich wird physikalisches Fachwissen erworben und in verschiedenen Kontexten angewandt. […]*
>
> ***Erkenntnisgewinnung und Experimentieren (E)***
> *In diesem Bereich werden Fähigkeiten und Fertigkeiten im Umgang mit physikalischen Denk- und Arbeitsweisen erworben. […]*
>
> ***Standpunkte begründen und aus naturwissenschaftlicher Sicht bewerten (S)***
> *In diesem Bereich wird die Fähigkeit erworben, naturwissenschaftlich begründet zu argumentieren und am gesellschaftlichen Diskurs teilzunehmen. […]“* (RIS, 2025c)

Der österreichische Physikunterricht hat seit Verordnung der Bildungsstandards die Aufgabe, die Kompetenzen der drei Handlungsdimensionen des Kompetenzmodells zu vermitteln. Von großer Bedeutung für die vorliegende Arbeit ist die Dimension (S):

„*Standpunkte begründen und aus naturwissenschaftlicher Sicht bewerten*“. Diese wird nun etwas näher erläutert.

2.2. Die Bewertungskompetenz

Das österreichische Schulorganisationsgesetz (SchOG) sieht mehrere zentrale Aufgaben für die Bildung in Schulen vor. Eine davon ist das Heranziehen von mündigen Staatsbürger:innen, die über ein geschultes Urteilsvermögen besitzen und eigenmächtig fundierte, verantwortungsbewusste Entscheidungen für sich und ihre Mitmenschen treffen können:

> §2, Absatz. 1, SchOG:
> „*Die jungen Menschen sollen zu […] verantwortungsbewussten Gliedern der Gesellschaft und Bürgern der demokratischen und bundesstaatlichen Republik Österreich herangebildet werden. Sie sollen zu selbständigem Urteil […] geführt, dem politischen und weltanschaulichen Denken anderer aufgeschlossen sein sowie befähigt werden […] an den gemeinsamen Aufgaben der Menschheit mitzuwirken*.“ (RIS, 2025a)

Die hierfür notwendigen Kompetenzen fallen genau unter eine der drei wichtigen Handlungsdimensionen, die naturwissenschaftlicher Unterricht im Sinne des Kompetenzmodells fördern soll und im AHS-Lehrplan für Physik unter der Dimension „*Standpunkte begründen und aus naturwissenschaftlicher Sicht bewerten (S)*“ vorgesehen ist (RIS, 2025c). Die Dimension wird verkürzt auch *Bewertungskompetenz* oder *S-Kompetenz* genannt. Mit wenigen Unterschieden beschreibt der Lehrplan das Ziel der Bewertungskompetenz für die Unter- und Oberstufe der AHS gleich. Als Beispiel heißt es für die Unterstufe:

> „***Standpunkte begründen und aus naturwissenschaftlicher Sicht bewerten (S)***
>
> *In diesem Bereich wird die Fähigkeit erworben, naturwissenschaftlich begründet zu argumentieren und am gesellschaftlichen Diskurs teilzunehmen.*“
>
> *Die Schülerinnen und Schüler können*
> - *Bedeutung, Chancen und Risiken der Anwendungen von naturwissenschaftlichen Erkenntnissen auf persönlicher, regionaler und globaler Ebene erkennen, um verantwortungsbewusst zu handeln.*
> - *naturwissenschaftliche von nicht naturwissenschaftlichen Argumentationen und Fragestellungen unterscheiden.*
> - *die Verlässlichkeit von unterschiedlichen Quellen aus naturwissenschaftlicher Sicht und aus anderen Blickwinkeln (ua. ökonomisch, ökologisch, ethisch) bewerten.*
> - *Entscheidungskriterien für das eigene Handeln entwickeln und aus naturwissenschaftlicher Sicht überprüfen.*“ (RIS, 2025c)

Die Bewertungskompetenz verfolgt also das Ziel, die wissenschaftliche Argumentations- und Diskursfähigkeit der Schüler:innen zu festigen. Sander und Höttecke (2015, S. 168)

beschreiben die Bewertungskompetenz auf ähnliche Art als Fähigkeit, persönliche und gesellschaftliche Entscheidungen kritisch reflektieren, eigenständig treffen und naturwissenschaftlich fundiert rechtfertigen zu können. Sie weisen zusätzlich darauf hin, dass Bewertungskompetenz auch als Fähigkeit des fachspezifischen Bewertens gesehen werden kann, beispielsweise beim kritischen Beurteilen der Gültigkeit von Messdaten. In dieser Auslegung beschränkt sich der Begriff mehr auf fachimmanente Bewertungen (Sander & Höttecke, 2015, S. 168).

Für Schüler:innen stellt das *Bewerten* oft eine Herausforderung dar. Im naturwissenschaftlichen Unterricht sind Lernende es gewohnt, klärende Antworten auf offene Fragen zu finden bzw. zu erhalten. Schüler:innen verunsichert es sehr, dass es Problemstellungen gibt, die keine allgemein gültigen Lösungen haben. Außerdem fällt es manchen schwer, mit der Offenheit der Urteilsfällung umzugehen. Schüler:innen müssen erst lernen, dass die Qualität eines Urteils nicht allein von der Entscheidung abhängt, sondern auch davon, ob alle relevanten Aspekte für ein durchdachtes, verantwortungsbewusstes Urteil in der Entscheidung berücksichtigt wurden. Hinzukommend wirkt es für einige komisch, dass sich wissenschaftliche Erkenntnisse mit der Zeit ändern und getroffene Entscheidungen vielleicht nicht mehr passend sein können (Hößle, 2013).

Das Lernen, Verstehen und Verbessern von *Entscheidungsprozessen* ist somit ein relevanter Teil für die Bewertungskompetenz. Forschungen zeigten, dass Schüler:innen ein systematischer Vorgang bei Entscheidungsprozessen schwerfällt und sie kaum Strategien zum Bewerten von Inhalten kennen. Ohne solches Wissen können Entscheidungssituationen nicht systematisch ablaufen (Eggert & Bögeholz, 2006). Im Alltag werden Entscheidungen daher oftmals intuitiv getroffen und erst im Nachhinein gerechtfertigt (Eggert, 2008, S. 30, 38–39). Aus entscheidungspsychologischer Perspektive sind intuitive Urteile nichts Verwerfliches, da sie Teil menschlichen Denkens und Handelns sind (Sander & Höttecke, 2015, S. 171). Dennoch sollte das Ziel wissenschaftlicher S-Kompetenz sein, eine Entscheidung aufgrund von im Vorhinein abgewogenen Kriterien zu treffen.
Auch das *Reflektieren über Entscheidungsfindungen* ist relevant für die S-Kompetenz und kann mithilfe gelernter Methoden deutlich verbessert werden (Dejanovikj & Höttecke, 2024).

Ebenso zentral für die Bewertungskompetenz ist das *Argumentieren* (Bögeholz et al., 2018, S. 264). Es zeigt sich, dass viele Schüler:innen beim Argumentieren kaum Fachwissen verwenden und sich das Argumentationsniveau in Grenzen hält (Sander & Höttecke, 2015, S. 170). Ein Unterricht, der auf das Erwerben von S-Kompetenz ausgerichtet ist, sollte sich somit auch mit der Förderung der Argumentationsfähigkeit auseinandersetzen.

Soweit ein grober Einblick in die Bewertungskompetenz und einigen Fähigkeiten, auf die es in dieser Handlungsdimension ankommt. Als Zwischenfazit kann hier festgehalten werden, dass eigenständiges und reflektiertes *Bewerten*, *Urteilen* und *Entscheiden* wesentliche Bestandteile der S-Kompetenz sind. Diese Fähigkeiten sind in der modernen Gesellschaft äußerst wichtig, müssen jedoch erworben und trainiert werden. Das stellt nicht nur die Lernenden vor Herausforderungen, sondern auch die Lehrenden der Naturwissenschaften.

Auch Lehrpersonen sind von Unsicherheit geplagt, wenn sie Probleme ohne eindeutige Lösung unterrichten sollen, wie z.B. ethische Konflikte, obwohl sie genügend erklärendes Fachwissen dazu hätten (Hößle, 2013). Es gibt mehrere Ansätze für Lehrkräfte, *wie* sie S-Kompetenz im Unterricht umsetzen können. Kühleitner (2019, S. 10–14) und Lux (2022, S. 26–33) beschreiben beispielsweise verschiedene Aspekte und Methoden zur Förderung der Bewertungskompetenz im Unterricht. Gleichermaßen brauchen Lehrpersonen aber auch Unterstützung beim *Beurteilen* von S-Kompetenz.

Die *Beurteilung der Bewertungskompetenz* ist für viele Lehrkräfte schwierig bzw. unsicher zu handhaben, da geeignete Beurteilungsinstrumente fehlen. In der Folge verzichten viele auf eine Notengebung der Bewertungskompetenz, um ungerechte Beurteilungen zu vermeiden. Das fördert bei den Lernenden jedoch die Ansicht, dass Bewerten, Entscheiden und Urteilen als Fähigkeiten unwichtig sind bzw. nicht gewürdigt werden (Hößle, 2013).

Die Beurteilung von S-Kompetenz ist insofern schwierig, weil sie treffsicher sein muss. Jede der Handlungsdimensionen W, E und S ist gleichwertig zu betrachten, da sie aufeinander aufbauen. Für die Erfassung der S-Kompetenz darf die W-Kompetenz keine Voraussetzung sein (Gstall, 2024, S. 10). Die Überprüfung der Bewertungskompetenz sollte möglichst losgelöst von anderen Fertigkeiten und Fähigkeiten erfasst werden (Sakschewski, 2013). Eine gerechte, treffsichere Beurteilung der Bewertungskompetenz gelingt nur, wenn naturwissenschaftliches Wissen keine Voraussetzung in einer Aufgabenstellung darstellt, da diese alle wichtigen Aspekte bereits vorgibt. Andernfalls würde anstelle der Entscheidungs- und Urteilsfindung das Verständnis über naturwissenschaftliche Zusammenhänge geprüft werden (Eggert & Bögeholz, 2006).

Das Unterrichten von Bewertungskompetenz ist herausfordernd, aber wichtig. Für Hößle (2013) fällt der Schule die Aufgabe zu, Lernenden fachliches Grundwissen und notwendige Methoden beizubringen, die sie in ihrem Urteilsprozess unterstützen. In Anbetracht der skizzierten Hürden für S-Kompetenz, braucht es vor allem für Lehrkräfte Unterstützung. Etwa mit Änderungen im Lehramtsstudium, unterstützendem Material oder Fortbildungen (Hößle, 2013).
Diese Masterarbeit schließt sich physikdidaktischen Forschungen (Gstall, 2024; Hackner, 2025) an und versucht eine effiziente Möglichkeit der Leistungsbeurteilung für diesen Kompetenzbereich zu finden. Wo dieses Vorhaben genau ansetzt, beschreibt Kapitel 2.4. Zuvor wird noch ein Blick auf die allgemeine Beurteilung von Leistungen geworfen.

2.3. Messung und Beurteilung von schulischen Leistungen

Das Messen und Beurteilen der Leistungen von Schüler:innen ist eine Kernaufgabe des Lehrberufs. Der Verdacht, Leistungsmessungen würden das Lernen negativ beeinflussen, ist ein häufiger Irrtum. Das Erbringen von Leistungen und deren Beurteilung sind notwendige Erfahrungen für Lernende, damit sie Leistungsbereitschaft entwickeln und ein positives Selbstbewusstsein über ihre eigenen Fähigkeiten aufbauen. Für die Lehrperson gilt es, Lern- und Leistungssituationen klar für Lernende zu trennen, da sie sich in ihren

psychologischen Bedingungen unterscheiden. Wenn Lernende diese Trennung bewusst wahrnehmen, können sie eine produktive Lernkultur entwickeln, die eine Leistungskultur miteinschließt (Hopf et al., 2022, S. 124).

Zwischen der *Leistungsmessung* und *Schülerbeurteilung* ist nach Hopf et al. (2022, S. 124) zu differenzieren. Die Rolle der Leistungsmessung liegt in der Leistungsdiagnostik, die sich auf die Leistungsfeststellung und ihren Funktionen stützt. Bei der Schülerbeurteilung stehen die Beurteilungsverfahren mit ihren Bezugsnormen und Qualitätsmerkmalen im Vordergrund.
Das österreichische Gesetz sieht eine ähnlich Trennung für das Erheben und Beurteilen von Leistungen vor. Die Leistungsbeurteilungsverordnung (LBVO) legt *Leistungsfeststellungen* als Grundlage für die *Leistungsbeurteilung* von Schüler:innen fest (RIS, 2025d).

Leistungsfeststellungen dürfen nur die gesetzlich vorgesehenen Lehrinhalte, die im Unterricht behandelt wurden, überprüfen

> §2, Absatz 1, LBVO:
> „*Der Leistungsfeststellung sind nur die im Lehrplan festgelegten Bildungs- und Lehraufgaben und jene Lehrstoffe zugrunde zu legen, die bis zum Zeitpunkt der Leistungsfeststellung in der betreffenden Klasse behandelt worden sind*." (RIS, 2025d)

und können in unterschiedlichen Formen durchgeführt werden,

> §3, Absatz 1. LBVO:
> „*Der Leistungsfeststellung zum Zweck der Leistungsbeurteilung dienen*:
> *a) die Feststellung der Mitarbeit der Schüler im Unterricht,*
> *b) besondere mündliche Leistungsfeststellungen*
> *aa) mündliche Prüfungen,*
> *bb) mündliche Übungen,*
> *c) besondere schriftliche Leistungsfeststellungen*
> *aa) Schularbeiten,*
> *bb) schriftliche Überprüfungen (Tests, Diktate),*
> *d) besondere praktische Leistungsfeststellungen,*
> *e) besondere graphische Leistungsfeststellungen.*" (RIS, 2025d)

die dann der endgültigen Beurteilung eines Semesters oder Schuljahres dienen.

> §11, Absatz 1. LBVO:
> „*Die Beurteilung der Leistungen der Schüler in den einzelnen Unterrichtsgegenständen hat der Lehrer durch die im § 3 Abs. 1 angeführten Formen der Leistungsfeststellung zu gewinnen. Maßstab für die Leistungsbeurteilung sind die Forderungen des Lehrplanes unter Bedachtnahme auf den jeweiligen Stand des Unterrichtes*." (RIS, 2025d)

Die LBVO sieht für die verschiedenen Formen der Leistungsfeststellung (§3) und Grundsätze der Leistungsbeurteilung (§11) konkrete Richtlinien vor, innerhalb derer sich Lehrende befinden dürfen bzw. müssen – diese werden hier aber nicht näher beschrieben.

Im Unterricht ergeben sich somit mehrere Möglichkeiten der Leistungsmessung. Hopf et al. (2022, S. 126) sind der Ansicht, dass sich vor allem schriftliche, mündliche und graphische Leistungsfeststellungen für den Physikunterricht anbieten.

Hopf et al. (2022, S. 125) beschreiben mehrere *Qualitätsmerkmale der Leistungsbeurteilung*, die es für die Lehrenden idealerweise zu beachten gilt, siehe Abb. 2.

Abb. 2: „Qualitätsmerkmale der Leistungsbeurteilung", wörtlich übernommen von Hopf et al. (2022, S. 125)

Drei dieser Merkmale heben Hopf et al. (2022, S. 124–126) für qualitative Beurteilungen gezielt hervor: Lern- und Leistungsüberprüfungssituationen müssen für Lernende klar erkennbar und voneinander getrennt sein (*Entflechtung*). Beurteilungsrelevante Lerninhalte, müssen vorab gelernt und geübt werden, damit Lernende die Möglichkeit haben, ihre Fähigkeiten unter Beweis zu stellen (*Transparenz*). Aufgaben sollten so gestellt sein, dass sie ein breites Spektrum an Kompetenzen im Sinne der Bildungsstandards überprüfen (*Kompetenzorientierung*).

Ein weiterer relevanter Punkt für die Messung der Leistung bzw. des *Lernerfolgs*, sind nach Häußler (2015) die Anforderungen an die Feststellungsmethode. Ein Bewertungsverfahren, wie z.B. eine Lern- oder Testaufgabe im Physikunterricht, verfolgt gewissermaßen das gleiche Ziel, wie ein Messinstrument in den Naturwissenschaften: Es soll möglichst *objektiv*, *reliabel* und *valide* sein (Häußler, 2015, S. 250ff).

Wohin kann oder soll die Messung des Lernerfolgs führen? Wie beschrieben dienen Leistungsfeststellungen in ihrer Summe der Leistungsbeurteilung. Leistungserhebungen

haben aber noch weitere Zwecke, wovon einer sehr wichtig ist. Sie dienen Schüler:innen wie den Lehrer:innen als *Rückmeldung* über nicht erreichte und nicht Lernziele. Aus den Bewertungen des Lernerfolgs können beide Seiten ihre Möglichkeiten für Verbesserungen ableiten. Schüler:innen können zum Beispiel aus ihren Fehlern lernen und Lernlücken füllen und Lehrer:innen können die Information für individuelle Lernberatung nutzen und erhalten zudem ein Bild über die Qualität ihres Unterrichts (Häußler, 2015, S. 253).

Damit Schüler:innen aus Fehlern konstruktiv lernen können, bedarf es gutem *Feedback*. Hattie konnte zeigen, dass Feedback eines der wirksamsten Instrumente für einen höheren Lernerfolg ist (Hattie, 2009, S. 173–178). Von 138 untersuchten Einflussfaktoren für schulischen Lernerfolg, belegte „Feedback" Platz 10 – ein klarer Nachweis für den höchst positiven Effekt auf das Lernen (Waack, 2025). Interessant dabei ist, dass nicht allein die Schüler:innen vom Feedback der Lehrperson profitieren, sondern umgekehrt, die Lehrenden genauso oder sogar mehr vom Feedback der Lernenden gewinnen. Nach Hattie gibt es verschiedene Formen des Feedbacks, von denen manche wirkungsvoller sind als andere. Gutes Feedback sollte jedoch immer auf das Lernen der Schüler:innen fokussiert sein, relevante Informationen und Instruktionen für den weiteren Lernprozess enthalten und sich nicht auf die Person selbst beziehen (Hattie, 2009, S. 173–178).

In dieser Hinsicht definiert sich effektives Feedback für Hattie & Timperley (2007) anhand von drei zusammenwirkenden Fragestellungen, die es erfüllen sollte und mit denen sich Lehrende wie Lernende beschäftigen sollten.
Die erste Frage, *Where am I going?* (Wohin gehe ich?), bezieht sich auf spezifische Unterrichtsziele und kann mit *feed-up* umschrieben werden. Feedback muss klar formuliert und auf die Ziele bezogen sein, damit es effektiv und hilfreich ist.
Die zweite Frage, *How am I going?* (Wie komme ich voran?), richtet sich ganz klassisch danach, wie man in Bezug auf frühere Ergebnisse oder im Vergleich zu erwarteten Leistungen vorankommt und ist interessiert an *feed-back*. Feedback ist effektiv, wenn es Informationen über den Fortschritt oder über nächste Schritte bereithält.
Die dritte Frage, *Where to next?* (Wohin als Nächstes?), verfolgt den Gedanken des *feed-forward*, und kann den mächtigsten Einfluss auf den Lernprozess haben. Die Antwort auf *Where to next?* ist fälschlicherweise oft *mehr.* Mehr Information, mehr Aufgaben und mehr Erwartungen. Vielmehr sollten Hinweise auf neue Lernmöglichkeiten gegeben werden, die angemessene Herausforderungen, Selbstregulierung oder neue Strategien, offenbaren. (Hattie & Timperley, 2007).

Darüber hinaus argumentieren Hattie & Timperley (2007), dass der *Fokus* von Feedback sehr wichtig für dessen Effektivität ist. Sie differenzieren vier Ebenen auf die Feedback fokussiert werden kann und dessen Wirksamkeit steuern: 1) Feedback über die Aufgabe oder das Produkt, 2) Feedback über den Prozess, 3) Feedback über die Selbstregulation und 4) Feedback an das Selbst/die Person (Hattie & Timperley, 2007).

Feedback kann als Prozess verstanden werden, in dem Lernende Sinn aus den Rückmeldungen zu ihren Leistungen ziehen und für Verbesserungen nützen. So gesehen wird der Feedbackprozess durch die lernende Person vorangetrieben, beeinflusst von ihrer

eigenen Sicht und ihren Bedürfnissen. Als logische Folge gibt es kein universelles Feedbackdesign, das für jede Person funktioniert (Henderson et al., 2019).

Diese Masterarbeit hat nicht das Ziel verschiedene Feedbackmethoden zu vergleichen. Dennoch ist Feedback Teil der Beurteilung und damit ein Bestandteil dieser Arbeit, auch wenn es im schulischen Kontext nicht der Leistungsmessung, sondern den Lern- und Entwicklungsprozessen einer Person dient (Sommer, 2012). Insofern werden kurz die Merkmale von gutem Feedback beschrieben und eine einfache Methode des Feedback-Gebens vorgestellt, die für die spätere empirische Forschung genutzt werden.

Nach Sommer (2012) haben sich hilfreiche *Feedback-Regeln* etabliert, um wertschätzendes Feedback mit Chance auf positive Wirkung zu formulieren. Feedback sollte allgemein

> *„beschreibend, nicht bewertend*; *konkret*; *angemessen*; *brauchbar*; *erwünscht, erbeten*; *klar und präzise formuliert*; *sachlich richtig*; *nicht zu umfassend*; *systemisch, d.h. den jeweiligen Kontext berücksichtigend“* (Sommer, 2012, S. 29)

sein und speziell für das Geben von Feedback können folgende Richtlinien beachtet werden:

- *„beschreiben, nicht bewerten;*
- *positive Rückmeldungen zuerst, Kritisches danach;*
- *Rückmeldungen möglichst konkret geben;*
- *in „Ich-Form“ formulieren zur Verdeutlichung, dass es sich um die eigene subjektive Sicht handelt;*
- *Verkraftbarkeit vor Vollständigkeit.“* (Sommer, 2012, S. 29)

Diese von Sommer (2012) zusammengetragenen Feedback-Regeln stimmen überwiegend mit den Anforderungen von Hattie & Timperley (2007) an effektives Feedback überein.

Eine unkomplizierte, leicht zu merkende und anwendbare Methode für das Geben von Feedback ist die *Feedback-Hand* oder *Fünf-Finger Methode* (Jonach & Wagner, 2017). Jeder Finger einer Hand steht für einen bestimmten Aspekt des Feedbacks:

- *„Daumen:* *Das war super …*
- *Zeigefinger:* *Darauf möchte ich hinweisen …*
- *Mittelfinger:* *Das hat mir nicht gefallen …*
- *Ringfinger:* *Das nehme ich mit/Das war mir wichtig …*
- *Kleiner Finger:* *Das kam zu kurz …“* (Jonach & Wagner, 2017, S. 13)

Die *Feedback-Hand* eignet sich als einfache und jederzeit durchführbare Methode gut, um kurz die wichtigsten Punkte eines Feedbacks zu formulieren, egal ob mündlich oder schriftlich. Nach Jonach und Wagner (2017) ist die Feedback-Hand eigentlich als Methode für Schüler:innen gedacht, die damit strukturiert Feedback geben sollen. Das Konzept ließe sich aber sicher auch von Lehrpersonen verwenden. Mit dieser Feedbackmethode können damit die beschriebenen Qualitätsmerkmale für Feedback gut eingearbeitet werden, wie z.B. das *feed-up*, *feed-back* und *feed-forward* von Hattie und Timperley (2007).

Die Messung und Beurteilung von schulischen Leistungen wurden nun ausführlich beschrieben. Das nächste Kapitel fokussiert sich ausschließlich auf die Bewertung der Bewertungskompetenz und führt auf das Forschungsinteresse dieser Arbeit hin.

2.4. Beurteilung der Bewertungskompetenz

Die Beurteilung von Bewertungskompetenz kann sehr herausfordernd sein. Das liegt vor allem an der Schwierigkeit objektive Beurteilungen von naturgemäß subjektiven Aussagen zu treffen und an der Tatsache, dass es lange keine effizienten Modelle zur Beurteilung von Bewertungskompetenz gab. Die Entwicklung eines besseren Modells um die Bewertung von S-Kompetenz zu erleichtern war das Ziel der Masterarbeit von Denise Gstall (2024). Ihre Forschung stellt den theoretischen Ausgangspunkt dieser Arbeit dar. Die wichtigsten Aspekte davon werden folgend kurz erläutert.

2.4.1. Das 5-stufige Beurteilungsmodell

Das Forschungsziel von Gstalls Masterarbeit (2024)

Obwohl die Bewertungskompetenz eine sehr zentrale Kompetenz ist, die junge Menschen im naturwissenschaftlichen Unterricht erwerben sollen, regen fehlende Beurteilungsmodelle Physiklehrpersonen kaum dazu an, diese im Unterricht zu üben und zu überprüfen. Eine Analyse existierender Modelle zeigte, dass diese keine effiziente Handhabung im Unterrichtsalltag ermöglichen. Doch auf ihrer Basis konnte ein viables Modell für den Einsatz im Unterricht konstruiert werden. Zu diesem Schluss kam die kürzlich erschienene Masterarbeit von Gstall (2024) in der ein neues Modell zur Beurteilung der S-Kompetenz für die Unterstufe entwickelt wurde.

Nach Gstall (2024, S. 3) waren die wichtigsten Anforderungen an das neue Modell eine einfache Übertragung der Leistung zum österreichischen Ziffernnotensystem und eine simple sowie effiziente Handhabung im Unterricht. Darüber hinaus sollte das Modell mehr Objektivität und Transparenz in die Beurteilung der recht subjektiven Bewertungskompetenz bringen. Das Beurteilungsmodell wurde in mehreren Schritten evaluiert und verbessert. Die Viabilität der finalen Version des Modells wurde von Fachpersonal aus einigen Schulen und der Universität Wien validiert.

Geeignete S-Kompetenzaufgaben

Die S-Kompetenz ist laut österreichischem AHS-Lehrplan die Fähigkeit von Schüler:innen „*Standpunkte [zu] begründen und aus naturwissenschaftlicher Sicht [zu] bewerten*“ (RIS, 2025c). Dementsprechend sollte das Format einer S-Kompetenzaufgabe auch diese Fähigkeit von Schüler:innen in der Bearbeitung verlangen. Gstall (2024, S. 86) erkannte bei der Analyse bestehender Bewertungsaufgaben, dass diese zum Teil vielmehr die Lesekompetenz als die S-Kompetenz überprüfen. Gstall formulierte generelle Anforderungen für Bewertungsaufgaben, die später auch für das Bewertungsmodell relevant werden. Aufgabenstellungen zur S-Kompetenz sollen offen formuliert sein, da dieses Format die Bewertungskompetenz am besten überprüft. Halboffene oder geschlossene Aufgaben-

stellungen, die eine Antwort bereits vorgeben, wie etwa Zuordnungs-, Reihungs- oder Multiple-Choice-Aufgaben, sind dafür ungeeignet. Weiters soll die Aufgabenstellung zumindest zwei Optionen mit jeweils mehreren Kriterien bereitstellen, anhand derer Unterschiede festgestellt und eine begründete Wahl getroffen werden können. Außerdem argumentiert Gstall, dass ideale S-Kompetenzaufgaben den fünf Merkmalen von *Socio-scientific Issues* von Sakschewski (2013) folgen sollten.

Die Beurteilung der S-Kompetenz anhand des 5-stufigen Modells

Das neue Beurteilungsmodell kann zur Beurteilung auf offene Aufgabenformate verwendet werden. Die Schüler:innen sollen sich in geeigneten Bewertungsaufgaben für eine von mehreren Optionen in einem Szenario entscheiden und ihre Wahl mithilfe von relevanten Kriterien begründen. Das gesamte notwendige Fachwissen für die Wahl sollte die Aufgabe bereitstellen. Anhand des Modells lässt sich die Argumentation der Schüler:innen-Antwort nach Parametern wie Kriterienanzahl, inhaltlichen Anforderungen und sprachlichen Merkmalen analysieren und einer *Niveaustufe* zwischen 1 und 5 zuordnen. Diese Stufen stehen in direktem Zusammenhang mit dem österreichischen Ziffernnotensystem, *Niveau 1* entspräche also der Note *Sehr gut*.

Die Beurteilung nach dem 5-stufigen Modell folgt einer Grundidee: Die fundierte Begründung für eine Wahl wird bewertet, nicht die Wahl in einem Szenario selbst. Bei der Zuordnung in eine Niveaustufe wird primär der Inhalt der Argumentation und sekundär die Satzstruktur gewichtet. Natürlich ist die Verwendung von wissenschaftlichen Argumenten und physikalisch richtigen Fachinhalten relevant für das Erreichen eines höheren Niveaus. Irrelevante Kriterien oder die Nutzung von Alltagswissen führen nach dem Modell zu einem niedrigen Niveau. Das Modell bewertet keinesfalls die Meinung von Schüler:innen, dies dürfen Lehrpersonen grundsätzlich nicht tun. Grammatik- und Rechtschreibfehler sollen nicht in die Bewertung miteinbezogen, aber trotzdem markiert werden (Gstall, 2024).

Das Modell wurde zur einfachen Handhabung als Beurteilungsmatrix erstellt, siehe Tab. 1. Dieser Raster ist auch in einer zusammenfassenden Handreichung von Gstalls Arbeit (2024, S. 84–86) enthalten, die dem Anhang beiliegt. Das Modell gliedert sich in die *Niveaustufen* 5 bis 1 (Zeilen) und den dazu nötigen Anforderungen *Anzahl der Kriterien*, *inhaltliche Anforderungen* und *sprachliche Merkmale* (Spalten). Die Anforderungen dienen als Beurteilungshilfen für die Lehrperson. Jedes Feld beschreibt, welche Merkmale eine Antwort auf inhaltlicher oder sprachlicher Ebene erfüllen muss. Die rechte Spalte *Anker-Beispiele* zeigt für jedes Niveau eine beispielhafte Schüler:innen-Antwort, die dem Niveau entsprechen würde. Eine Lehrperson kann nach diesem Modell die Antworten von Schüler:innen analysieren und einem Niveau zuordnen, welches zugleich der Ziffernnote entspricht.

Niveau	Anzahl d. Kriterien	Inhaltliche Anforderungen	sprachliche Merkmale	Anker-Beispiele
5	0	Die Begründung(en) (falls vorhanden) werden auf Basis (naturwissenschaftlichen) Alltagswissens aufgebaut und sind deswegen fachlich unangemessen *und/oder* die Entscheidung wurde intuitiv getroffen *und/oder* die gezogene Schlussfolgerung ist falsch bzw. nicht zielführend oder für die Beantwortung der Aufgabenstellung irrelevant.	Die getroffene Entscheidung (=gewählte Option) wird nicht oder nicht anhand eines Kriteriums begründet.	*„Ich wähle Option A."* *„Ich wähle Option A, weil sie am besten ist."* *(0 Kriterien)*
4	1	Niveau 5 wurde durch eine naturwissenschaftliche und begründete Entscheidung übertroffen *und* möglicherweise wurden eigene Kriterien in die Argumentation miteinbezogen, *aber* die Begründung ist möglicherweise nicht durchgehend fachlich richtig.	Die Begründung kann sprachlich erkennbar sein. Die Argumentation folgt jedoch meist einer einfachen Struktur wie *„Option A ist die beste, weil Kriterium 1."*	*„Ich wähle Option A, sie ist ziemlich günstig."* *„Ich wähle Option A, weil sie ziemlich günstig ist und mir die Farbe gefällt."* *(1 Kriterium, Farbe ist irrelevantes Kriterium)*
3	2	Niveau 4 wurde erreicht *und* Kriterien, die in der Aufgabenstellung als irrelevant vermerkt wurden (z.B. Farbe eines Objektes: *Die Farbe ist für Sara egal*), werden nicht ernsthaft in die Argumentation miteinbezogen *und* Kriterien, die in der Aufgabenstellung als wesentlich vermerkt wurden (z.B. geringer Preis: *Sara möchte beim Kauf nicht viel Geld ausgeben*), werden in die Argumentation miteinbezogen.	Die Begründung ist sprachlich klar erkennbar, zum Beispiel durch Konjunktionen wie *weil, da, denn, nämlich, … .* Die Argumentation folgt meist der Struktur: "Option A ist die beste, weil Kriterium 1 + 2."	*„Ich wähle Option A, da sie am günstigsten ist und die höchste Leistung hat."* *(2 Kriterien)*
2	≥2	Niveau 3 wurde erreicht, *aber* es kann noch Interpretationsspielraum geben, dass noch mehr als die explizit genannten Kriterien Einfluss auf die Wahl der Schüler*innen hatten.	Die Argumentation geht über folgende Struktur hinaus: „Option A ist die beste, weil Kriterium 1 + 2." und beinhaltet gegebenenfalls weitere Kriterien und/oder Adverbien wie *außerdem/auch/ zusätzlich/…*	*„Ich wähle Option A, da sie am günstigsten ist und die höchste Leistung hat. Außerdem ist sie praktisch."* *(2 Kriterien + 1 angedeutet)*
1	≥2	Niveau 2 wurde erreicht. *und* Die Argumentation für eine Option ist inhaltlich durchgehend verständlich. Es gibt keinen Interpretationsspielraum oder Zweifel. *und* Die vorliegenden Kriterien und Optionen werden genutzt, um nicht nur den eigenen Standpunkt zu rechtfertigen, sondern auch, um konträre Standpunkte zu entkräften. *und/oder* Mehrere Kriterien und/oder Optionen werden im Laufe der Argumentation berücksichtigt und begründet gegeneinander aufgewogen. Kriterien, die beim Verglichen verwendet werden, zählen auch zur Anzahl der Kriterien.	In der Argumentation ist klar erkennbar, dass Optionen sprachlich miteinander verglichen werden, zum Beispiel durch Ausdrücke wie *zwar / aber / jedoch / trotzdem / …*	*„Ich wähle Option A, da sie im Vergleich zu den anderen am günstigsten ist und auch viele Funktionen bietet. Zwar bietet Option B eine höhere Leistung, der Unterschied ist jedoch gering, weshalb auch Option A zufriedenstellende Ergebnisse liefern wird."* *(3 Kriterien)*

Tab. 1: *Modell zur Beurteilung der S-Kompetenz, Auszug aus der Handreichung (siehe Anhang) von Gstall (2024, S. 85).*

Limitationen und Potenziale des 5-stufigen Beurteilungsmodells

Die Evaluation des finalen Modells von Gstall (2024) unter Physikehrkräften ergab überwiegend positives Feedback und spricht für dessen Einsatz im Schulalltag, wenngleich es in mancher Hinsicht noch Limitationen und Verbesserungspotential aufweist.

Das 5-stufige Modell erfüllt die zentrale Anforderung einer leichteren Handhabung gegenüber bestehenden Beurteilungsmodellen. Der Umgang mit dem Modell wird als intuitiv beschrieben, kann zu Beginn jedoch Zeit brauchen, bis man mit den Formulierungen des Modells vertraut ist. Um das zu erleichtern, wurde die Handreichung für Lehrkräfte erstellt. Im Laufe der Studie von Gstall (2024) zeichneten sich gegensätzliche Verbesserungsvorschläge für das Modell ab: Es gibt Formulierungen, die für jede Lehrperson eine Frage der Auslegung bzw. persönlichen Gewichtung sind und Meinungen dazu, dass Teile des Modells zu komplex oder umgekehrt zu unklar definiert sind. Gstall verortet hier zwei schwer kombinierbare Eigenschaften, die das Modell erfüllen soll: bestmögliche Objektivität sowie einfache Handhabung in der Beurteilung. Ein hohes Maß an Objektivität würde komplexe Erklärungen der einzelnen Stufen erfordern, was wiederum die Handhabung erschwert. Gstall erkennt diese Limitation an und kommt zu dem Schluss, dass ihr Beurteilungsmodell zur S-Kompetenz in seiner finalen Form zwar mehr Objektivität und Transparenz in der Beurteilung schafft, jedoch keine absolute Objektivität ermöglichen kann. Diese sei ohnehin immer nur das unerreichbare Ziel. Damit sind die erhöhte Objektivität und Transparenz zwar ein deutlicher Erfolg des Modells, bleiben aber trotzdem limitiert (Gstall, 2024, S. 71–76).

Während der Entwicklung des 5-stufigen Modells, kam die Idee für ein verkürztes, 3-stufiges Modell auf. Abhängig vom Einsatz, etwa in welcher Schulstufe oder in welchem Prüfungskontext die Bewertungskompetenz abgefragt wird, kann die kürzere Version Schwächen des 5-stufigen Modells ausgleichen. In den meisten Situationen wird die Bewertungskompetenz nur als eine Komponente innerhalb eines Tests oder als eine kleine Aufgabe zwischendurch geprüft. Dementsprechend könnte ein kürzeres, weniger differenziertes Modell für die Handhabung und Beurteilung geeigneter sein (Gstall, 2024, S. 76–77). Die kürzere Version von Gstalls‘ 5-stufigen Modell wird gerade in einer laufenden Masterarbeit entwickelt (Hackner, 2025) und scheint auch vielversprechend für die Beurteilung von S-Kompetenz zu sein.

2.4.2. Das 3-stufige Beurteilungsmodell

Hackner (2025) setzt die Forschung zur Bewertung der Bewertungskompetenz von Gstall (2024) fort und greift dafür Gstalls Idee für ein neues Modell auf. Das 5-stufige Modell soll auf ein verkürztes 3-stufiges Modell weiterentwickelt und evaluiert werden. Das Ziel dieser Anpassung liegt in einer leichteren Beurteilung, die ein weniger differenziertes Modell ermöglichen könnte. Eine Limitation des 5-stufigen Modells ist nämlich, dass es für manche Einsätze doch recht komplex sein kann. Etwa dann, wenn es nur darum geht einzelne, kurze S-Kompetenzaufgaben zu beurteilen, die Teil eines Tests oder einer Schularbeit sind. Hackner (2025) versucht diese und andere Limitationen in ihrer aktuell laufenden Masterarbeit in der Entwicklung eines 3-stufigen Modells zur Beurteilung der S-Kompetenz

auszugleichen. Ein weiteres Ziel ihrer Forschung besteht auch darin, sowohl eine Version für die Unter- und die Oberstufe zu erschaffen – das Modell von Gstall beschränkt sich nur auf den Einsatz in Unterstufenklassen.

Der Aufbau des 3-stufigen Modells folgt dem Schema des 5-stufigen, unterscheidet sich aber in einigen wesentlichen Punkten. Die Leistung einer Antwort wird im verkürzten Modell ebenfalls auf Basis inhaltlicher Anforderungen und sprachlicher Merkmale einer Niveaustufe zugeordnet. Als Orientierung hält das 3-stufige Modell ebenfalls eine Spalte für beispielhafte Antworten pro Niveaustufe bereit. Doch nun zu den Änderungen.

Wenig überraschend beschränkt sich das 3-stufige Modell für die Einstufung lediglich auf drei Niveaus statt der ursprünglichen fünf. Eine Antwort kann den Niveaus 0, 1 und 2 zugeordnet werden. Durch diese Änderung gleichen die Niveaustufen nicht mehr dem österreichischen Ziffernotensystem und die abgefragte S-Kompetenz darf nicht mehr mit den klassischen Noten 1 bis 5 beurteilt werden. Hackner (2025) setzt für ihr Modell die drei Niveaustufen mit den verbreiteten Bewertungssymbolen +, ~ und – gleich.

Eine weitere wesentliche Veränderung des Modells ist die getrennte Beurteilung von Inhalt und Sprache. Die Idee ist, eine Antwort in beiden Anforderungsbereichen separat zu bewerten und daraus eine Beurteilung zu ermitteln. Die Beurteilung ist dabei im Verhältnis 2:1 mehr nach den inhaltlichen Anforderungen als den sprachlichen Merkmalen zu gewichten und entsprechend zu berechnen.

Die Version des 3-stufigen Beurteilungsmodells für die Oberstufe von Hackner (2025) ist in

Tab. **2** auf der nächsten Seite dargestellt. Diese unterscheidet sich nur geringfügig von der Version für die Unterstufe. Im Wesentlichen sind die inhaltlichen wie sprachlichen Anforderungen für ein höheres Niveau in der Oberstufe anspruchsvoller. In der vorliegenden Masterarbeit wird die Oberstufen-Version für die geplanten Studien verwendet. Für den Einblick in die Version der Unterstufe wird auf die Masterarbeit von Hackner (2025) verwiesen.

Mit den Masterarbeiten von Gstall (2024) und Hackner (2025) ist die aktuelle, physikdidaktische Forschung an der Universität Wien zur Entwicklung von neuen Modellen zur Beurteilung der S-Kompetenz ausreichend beschrieben. Diese Masterarbeit setzt als Erweiterung an die beiden Arbeiten an und beschäftigt sich mit der möglichen Beurteilung von Bewertungskompetenz durch *künstliche Intelligenz*. Das nächste Kapitel versucht eine Einführung in diese Technologie zu geben und bildet den zweiten Theorieteil dieser Arbeit.

Niveau	Anzahl der Kriterien	Inhaltliche Anforderungen	Anker-Beispiele
0/-	**0 - 1**	Die Begründung(en) (falls vorhanden) werden auf Basis naturwissenschaftlichen) Alltagswissens aufgebaut und sind deswegen fachlich unangemessen. Die Entscheidung wurde intuitiv getroffen. Einiges an Interpretationsspielraum *und/oder* die gezogene Schlussfolgerung ist falsch bzw. nicht zielführend oder für die Beantwortung der Aufgabenstellung irrelevant *und/oder* möglicherweise wurden eigene, irrelevante Kriterien in die Argumentation miteinbezogen.	*„Ich wähle Option A, sie ist ziemlich günstig."* *„Ich wähle Option A, weil sie ziemlich günstig ist und mir die Farbe gefällt."* *(1 Kriterium, z.B. Farbe eines Objekts ist irrelevantes Kriterium)*
1/~	**≥ 2**	Kriterien, die in der Aufgabenstellung als wesentlich vermerkt wurden (z.B. geringer Preis: Sara möchte beim Kauf nicht viel Geld ausgeben), werden in die Argumentation miteinbezogen. *aber* es kann noch Interpretationsspielraum geben, dass noch mehr als die explizit genannten Kriterien Einfluss auf die Wahl der Schüler*innen hatten. *und/oder* möglicherweise wurden eigene Kriterien in die Argumentation miteinbezogen, die irrelevant sind.	*„Ich wähle Option A, da sie am günstigsten ist und die höchste Leistung hat."* *„Ich wähle Option A, da sie am günstigsten ist und die höchste Leistung hat. Außerdem ist sie praktisch."* *(2 Kriterien + 1 angedeutet)*
2/+	**≥ 2**	Die Argumentation für eine Option ist durchgehend verständlich. Es gibt keinen Interpretationsspielraum oder Zweifel. *und* Die vorliegenden Kriterien und Optionen werden genutzt, um auch konträre Standpunkte zu entkräften. *und/oder* Mehrere Kriterien und/oder Optionen werden im Laufe der Argumentation berücksichtigt und begründet gegeneinander aufgewogen. Kriterien, die beim Vergleichen verwendet werden, zählen auch zur Anzahl der Kriterien.	*„Ich wähle Option A, da sie im Vergleich zu den anderen am günstigsten ist und auch viele Funktionen bietet. Zwar bietet Option B eine höhere Leistung, der Unterschied ist jedoch gering, weshalb auch Option A zufriedenstellende Ergebnisse liefern wird."* *(3 Kriterien)*

Niveau	Sprachliche Merkmale	Anker-Beispiele
0/-	Die Begründung kann sprachlich erkennbar sein. Die Argumentation folgt jedoch meist einer einfachen Struktur wie „Option A ist die beste, weil Kriterium 1."	*„Ich wähle Option A, sie ist ziemlich günstig."* *„Ich wähle Option A, weil sie ziemlich günstig ist und mir die Farbe gefällt."*
1/~	Die Argumentation geht über folgende Struktur hinaus: „Option A ist die beste, weil Kriterium 1 + 2." Und beinhaltet gegebenenfalls weitere Kriterien und/oder Adverbien wie *außerdem/auch/ zusätzlich/....*	*„Ich wähle Option A, da sie am günstigsten ist und die höchste Leistung hat."* *„Ich wähle Option A, da sie am günstigsten ist und die höchste Leistung hat. Außerdem ist sie praktisch."*
2/+	In der Argumentation ist klar erkennbar, dass Optionen sprachlich miteinander verglichen werden, zum Beispiel durch Ausdrücke wie *zwar / aber / jedoch / trotzdem / ...*	*„Ich wähle Option A, da sie im Vergleich zu den anderen am günstigsten ist und auch viele Funktionen bietet. Zwar bietet Option B eine höhere Leistung, der Unterschied ist jedoch gering, weshalb auch Option A zufriedenstellende Ergebnisse liefern wird."*

***Tab. 2**: 3-stufiges Modell zur Beurteilung der Bewertungskompetenz in der Oberstufe. (Hackner, 2025)*

3. Künstliche Intelligenz

Das zweite Theoriekapitel behandelt die Technologie der künstlichen Intelligenz und führt in die wichtigsten Aspekte zu diesem Forschungsfeld ein. Nach einführenden Informationen zur Definition, Entwicklung und Funktionsweise von künstlicher Intelligenz fokussiert sich das Kapitel auf Sprachmodelle, die Kommunikation mit KI-Systemen, die Entwicklung von ChatGPT und was dieser Chatbot für das Bildungssystem und diese Masterarbeit beitragen kann.

3.1. Was ist künstliche Intelligenz?

Da die Technologie der intelligenten Maschinen erst seit kurzem in unserem Alltag und damit im Bewusstsein der breiten Gesellschaft angekommen ist, liegt die Annahme nahe, dass dieses Forschungsgebiet sehr jung ist. Das ist auch richtig, aber die Geburtsstunde der KI-Forschung liegt doch weiter zurück, als man denkt. Der Begriff *künstliche Intelligenz* (KI) trat erstmals 1955 durch John McCarthy und seine Kollegen auf, die damals den Antrag auf Förderung eines Forschungsworkshops zum Thema *Artificial Intelligence* stellten. Die Förderung wurde 1956 bei der *Dartmouth Conference* bewilligt und der Grundstein für KI als akademische Disziplin war gelegt. Die Konferenz gilt seit dem als Geburtsstunde der modernen KI-Forschung (De Florio-Hansen, 2020, S. 41–42; Heinrichs et al., 2022, S. 1).

Eine eindeutige Definition von *künstlicher Intelligenz* gibt es bisher nicht (Stefan, 2019, S. 12). Diese Unbestimmtheit hängt möglicherweise damit zusammen, dass es bis heute auch keine einheitliche Beschreibung von (menschlicher) *Intelligenz* gibt (De Florio-Hansen, 2020, S. 46). Denn was unter Intelligenz verstanden wird, tritt in vielfältigen Ausprägungen auf, zum Beispiel als sprachliche Intelligenz, räumliche Intelligenz oder kreative Intelligenz (Kreutzer & Sirrenberg, 2019, S. 2–3). Ähnlich vielfältig sind in der Folge die Ansätze, um den Begriff der künstlichen Intelligenz greifbar zu machen. De Florio-Hansen fasst den ursprünglichen Gedanken von McCarthy et al. (1955) zu künstlicher Intelligenz so zusammen:

> „*KI stützt sich also auf die Annahme, dass grundsätzlich alle Aspekte des Lernens und andere Merkmale der Intelligenz so genau beschrieben werden können, dass eine Maschine zur Simulation dieser Vorgänge gebaut werden kann. Es sollen also Entscheidungsstrukturen des Menschen in Programmen nachgebildet werden, die eigenständig Probleme bearbeiten und selbstständig Aufgaben erledigen können, für die normalerweise menschliche Intelligenz erforderlich ist.*“ (McCarthy et al., 1955, zitiert nach De Florio-Hansen, 2020, S. 42)

Diese Beschreibung trifft das Wesen der künstlichen Intelligenz in all ihren Facetten und Zielen. Eine jüngere Definition von Rich (1983) beschreibt die Idee dieser Technologie noch prägnanter:

> „*Artificial Intelligence is the study of how to make computers do things at which, at the moment, people are better.*“ (Rich, 1983, zitiert nach Kreutzer & Sirrenberg, 2019, S. 3)

Dieser Ansatz verdeutlicht, dass die Grenzen künstlicher Intelligenz flexibel sind und mit ihrer Entwicklung wachsen (Kreutzer & Sirrenberg, 2019, S. 3). Künstliche Intelligenz zeichnet sich dadurch aus, dass sie gleichermaßen zu rationalen sowie menschlichen Denk- und Handlungsweisen fähig ist (Russell & Norvig, 2016, S. 1–4). Unter „künstlicher Intelligenz" ist eine Vielzahl an Methoden und Ansätzen angesprochen, was eine einheitliche Definition schwierig macht. Zumindest eines haben allen Definitionsversuchen gemeinsam: Rationalität und Lernfähigkeit werden als zentrale Merkmale von Intelligenz betrachtet, dessen maschinelle Implementierung angestrebt wird (Heinrichs et al., 2022, S. 23).

Künstliche Intelligenz hängt also stark mit den menschlichen Fähigkeiten des Lernens, Denkens und Handelns zusammen, die maßgeblich für Intelligenz sind. Aktuelle KI-Systeme sind dazu in der Lage, einzelne Felder dieser angeborenen oder erlernten Intelligenz des Menschen adäquat zu imitieren. Jedoch wird es erst in der Zukunft möglich sein, dass Maschinen die vielschichtige Intelligenz des Menschen in ihrer Gesamtheit abbilden können (Kreutzer & Sirrenberg, 2019, S. 2–3).

In der aktuellen Forschung findet eine gängige Form der Unterteilung von KIs Anwendung. Es wird zwischen *starken* und *schwachen* künstlichen Intelligenzen differenziert. Diese Einteilung geht auf Searle (1980, S. 417) zurück. Starke KI-Technologie versucht menschliche Fähigkeiten in allen Lebensbereichen zu imitieren und strebt danach, diese so weit zu optimieren, dass völlig neue Intelligenzniveaus erreicht werden. Die Selbstlernfähigkeit starker KIs könnte irgendwann zu Phänomenen wie *Superintelligenzen*, *technologischen Singularitäten* oder *Transhumanismus* führen. Schwache KI-Technologie verfolgt nicht das Ziel menschliche Fähigkeiten zu imitieren. Sie dient dem Lösen komplexer Aufgaben (z.B. Schach spielen), mit dem Bestreben dies gleich gut oder etwas besser zu tun, als es die menschlichen kognitiven und physischen Fähigkeiten zulassen (Kreutzer & Sirrenberg, 2019, S. 20–23). Das Feld der schwachen KI ist mittlerweile schon weit beforscht und etabliert. Das KI-Sprachmodell *ChatGPT* von OpenAI (2022) ist ein bekanntes Beispiel hierfür. Ergebnisse im Zusammenhang mit der Forschung zur starken KI sind heute noch nicht klar erkennbar (Vladova & Bertheau, 2023, S. 395).

3.2. Wichtige Mechanismen künstlicher Intelligenz

Die wichtigsten inhaltlichen Elemente, die unter dem Oberbegriff der künstlichen Intelligenzen stehen und sie ermöglichen, sind für Kreutzer und Sirrenberg (2019, S. 4–8) neuronale Netze, maschinelles Lernen und Deep Learning.

Neuronale Netzwerke

Künstliche *neuronale Netzwerke* bilden das Kernstück von künstlichen Intelligenzen und sind komplexe Systeme aus Hard- und Softwareelementen (Kreutzer & Sirrenberg, 2019, S. 4–5). Inspiriert durch die Anatomie des menschlichen Gehirns besteht ein neuronales Netz aus einer Ansammlung von miteinander verknüpften Neuronen. Diese Neuronen sind in Schichten angeordnet. Zwischen einer Input- und Outputschicht kann es beliebig viele

Zwischenschichten von Neuronen geben. Wenn alle Neuronen einer Schicht mit allen einer benachbarten Schicht verbunden sind, spricht man von einem neuronalen Netzwerk. Im technischen Sinn bildet das einzelne Neuron vereinfacht gesagt eine Recheneinheit, das skalare Inputs erhält, diese verschieden gewichtet und daraus einen Output berechnet, der wiederum als Input für ein anderes Neuron im Netzwerk dient. Ein Netzwerk mit genügend Neuronen und Schichten kann komplexe, mathematische Funktionen berechnen (Goldberg, 2016, S. 354–355).

Das Ziel neuronaler Netzwerke besteht darin, Muster in großen Datensätzen zu erkennen. Eine Maschine kann auf dieser Basis Vorhersagen machen, Entscheidungen treffen und Lernfortschritte erzielen. Neuronale Netzwerke sind damit die Grundlage für die Lernfähigkeit von Maschinen (Kipper, 2020, S. 15–19).

Maschinelles Lernen

Maschinelles Lernen beruht auf selbstadaptiven Algorithmen. Diese ermöglichen es Maschinen, eigenständig zu lernen, ohne dass Menschen in den Lernprozess eingreifen müssen. Was es dazu braucht, sind ein programmierter Algorithmus und riesige, qualitative Datenmengen mit denen Algorithmen mehrfach trainiert werden. Dieser Prozess generiert neue Algorithmen, die immer besser und erfolgreicher im Erkennen von Mustern, Treffen von Entscheidungen, Erledigen spezifischer Aufgaben oder sonstigen denkbaren Einsatzmöglichkeiten, wie z.B. Schach spielen, werden (Kipper, 2020, S. 15–19; Kreutzer & Sirrenberg, 2019, S. 6–7). Bei der Entwicklung immer leistungsfähigerer Algorithmen für maschinelles Lernen können drei geläufige Lernarten nach Gentsch (2018, S. 38–39) unterschieden werden:

- **Überwachtes Lernen** (*Supervised Learning*)
  Beim überwachten Lernen erhält der Algorithmus zu seinen Trainingsdaten direkt die richtigen Antwortmöglichkeiten dazu und er soll die Gesetzmäßigkeit zwischen den Ein- und Ausgabedaten finden. Diese Lernmethode wird z.B. für Aufgaben im Bereich der Klassifizierung und Regressionsanalyse eingesetzt. Das Prognostizieren des Immobilienpreises in Abhängigkeit der Gebäudegröße wäre etwa ein Regressionsproblem.

- **Nicht überwachtes Lernen** (*Unsupervised Learning*)
  Im nicht überwachten Lernen erhält das KI-System keine vorgegebenen Zielwerte. Der Algorithmus soll selbstständig Muster in der Datenmenge identifizieren, die Menschen vielleicht gar nicht bewusst sind. Nicht überwachte Lernalgorithmen werden z.B. in der Genomforschung eingesetzt.

- **Verstärkendes Lernen** (*Reinforcement Learning*)
  Das verstärkende Lernen erfordert noch mehr Eigenständigkeit vom System. Hier liegt zu Beginn kein idealer Lösungsweg für ein Problem vor, diesen muss der Algorithmus iterativ durch *Trial-and-Error* herausfinden. Gute Lösungsansätze werden belohnt und schlechte bestraft, was das System in seinem Lernprozess antreibt. Der Algorithmus kann viele Umwelteinflüsse in seinen Entscheidungen

berücksichtigen und so Lösungen finden, die Menschen womöglich gar nicht erdacht hätten. Durch verstärkendes Lernen können KI-Systeme die Fähigkeit erlangen, autonom neue Lösungswege zu finden und erregen damit dein Eindruck intuitiv handeln zu können.

Deep Learning

Eine besondere Art des maschinellen Lernens ist das *Deep Learning* und wird durch eine besondere Ausgestaltung neuronaler Netze ermöglicht. Diese Methodik kann größere Mengen an Daten verarbeiten, erfordert weniger Datenaufbereitung durch Menschen und liefert dabei oft genauere Ergebnisse als gewöhnlichere Lernansätze (Kreutzer & Sirrenberg, 2019, S. 8). Spezielle Bauweisen neuronaler Netzwerke bilden die Grundlage für Deep Learning. Je mehr versteckte Schichten ein Netz hat, desto mehr „Tiefe" besitzt es – daher rührt die Bezeichnung *deep* (Goldberg, 2016, S. 357). Die Verarbeitung großer Datenmengen über viele Schichten ermöglicht es dem KI-System sehr tiefergehende Gesetzmäßigkeiten und Korrelationen zu erkennen. Ein neuronales Netzwerk mit Deep Learning-Fähigkeit ermöglicht es einer Maschine außerdem, mit der Zeit selbstständig komplexe Konzepte zu finden. Ein Beispiel hierfür ist die Handschrifterkennung, die durch klassische Programmierung nicht möglich ist. Eine künstliche Intelligenz muss das Konzept der Handschrifterkennung „von selbst" mithilfe von genügend Daten und der richtigen Architektur des neuronalen Netzwerks erlernen (Kreutzer & Sirrenberg, 2019, S. 8).

3.3. Entwicklung, Funktionsweise und Eigenschaften von Sprachmodellen

Große Einsatzfelder künstlicher Intelligenzen sind das *Natural-Language-Processing* (NLP), die Erkennung und Verarbeitung menschlicher Sprache in Wort und Schrift (Kreutzer & Sirrenberg, 2019, S. 28) sowie die *Natural-Language-Generation* (NLG), die automatische Generierung von Text in beliebiger Sprache auf Basis einer Texteingabe (Gatt & Krahmer, 2018). Beide Techniken sind wichtige Teile des *Language Modeling* (LM), der Sprachmodellierung, einer der wichtigsten Ansätze, um die sprachliche Intelligenz von Maschinen zu verbessern. *Sprachmodelle* (*Language Models*, LM) sind intelligente Systeme, die diese Sprachmodellierung erlernen und NLP- sowie NLG-Aufgaben übernehmen können (Zhao et al., 2025).

Zhao et al. (2025) unterteilen den Entwicklungsprozess von LMs in vier Generationen, wobei sie die immer besser werdende Aufgabenlösungsfähigkeit als Maßstab zur Einteilung verwenden. In den 1990ern entstanden erste *Statistical Language Models* (SLM). Diese Modelle konnten erstmals Texte generieren, indem sie das nächste Wort innerhalb eines vorgegeben Kontexts vorhersagen. Statistische LMs mit einer fixierten Kontextlänge *n* arbeiteten oft mit Bigrammen oder Trigrammen, das bedeutet sie berechneten das nächste Wort einer Folge anhand der zwei bis drei vorherigen Wörter – man nannte sie daher auch *n*-gram LMs. SLMs wurden Anfang der 2010er von den *Neural Language Models* (NLM)

abgelöst, die größere Wortsequenzen mit der Technik von neuronalen Netzwerken besser berechnen konnten (Zhao et al., 2025).

Der dritte Schritt in der Entwicklung von Sprachmodellen sind *Pre-trained Language Models* (PLM), die gegen Ende der 2010er eine höhere Performance als NLMs bei Text-generierungsaufgaben erreichten. Mithilfe neuer Techniken, wie der Transformer-Architektur und dem Aufmerksamkeitsmechanismus (Erklärung folgt), ist es möglich PLMs einem Training und einer Feinabstimmung zu unterziehen bevor sie zum Einsatz kommen. Intensive Forschungen an größer werdenden PLMs zeigten später, dass diese ab einer bestimmten Modellgröße verbesserte Leistungen bei Aufgaben erreichen, andere Verhaltensmuster zeigen und neue Fähigkeiten beim Lösen komplexer Aufgaben entwickeln können. Ab dem Jahr 2020 wurde für große PLMs daher der Begriff *Large Language Models* (LLM) eingeführt, die vierte Generation der Sprachmodelle. LLMs unterschieden sich also primär durch ihre Modellgröße von PLMs und in der Folge durch neue Fähigkeiten und Verhaltensweisen, die sie auf Grund ihrer Größe zeigen. Zum Vergleich: PLMs verfügen über einige Millionen sogenannter Parameter, mit denen sie ihre Vorhersagen berechnen, und LLMs bis zu einer Milliarde und teils mehr (Zhao et al., 2025).

Wie funktionieren Sprachmodelle genau? Moderne Sprachmodelle sind Großteils LLMs, daher wird für diese Frage die Funktionsweise von LLMs erklärt. Ein LLM ist ein tiefes neuronales Netzwerk dessen Architektur auf dem *Transformer* Modell beruht (Vaswani et al., 2017). LLMs werden mit riesigen Mengen an unstrukturierten Daten vortrainiert und können ohne weiteres Training eine Bandbreite an verschiedenen Aufgaben der natürlichen Sprachverarbeitung lösen (Wang et al., 2022; Zhao et al., 2025). Während des Vortrainings lernt das Sprachmodell anhand von Textdaten Bedeutungen, Strukturen und Muster in der natürlichen Sprache zu erkennen. Hierfür ist die Transformer Architektur von Bedeutung, da diese dem Modell ermöglicht den Kontext von Wörtern innerhalb einer Sequenz effizient zu erkennen. Der Text wird dazu in einzelne *Tokens*, d.h. Stückchen, aufgeteilt und sequenziell vom Sprachmodell verarbeitet. Der integrierte *Aufmerksamkeitsmechanismus* in LLMs, das auszeichnende technische Merkmal der Transformer-Architektur, ermöglicht das kontextuelle Erfassen der Tokens. Mithilfe von Deep Learning lernt ein LLM schließlich in diesem Prozess das nächste Token auf Basis der vorherigen in einer Sequenz vorherzusagen. Denn es kann das relationale und linguistische Wissen aus den Trainingsdaten speichern und wieder abrufen, wenn es mit neuen Texten oder Aufgaben versorgt wird. Das Ziel des Vortrainings eines LLMs besteht darin, es für die Bearbeitung einer Vielzahl von NLP-Aufgaben zu befähigen, ohne dass weitere Anpassungen des Sprachmodells erforderlich sind. Diese in LLMs auftretende Fähigkeit, Aufgaben ohne spezielles Training lösen zu können, nennt man auch *Zero-shot Generalization* (Mader, 2024, S. 14–16; Wang et al., 2022). Nach Abschluss des Vortrainings ist es möglich, LLMs im *Finetuning* weiter zu trainieren. Hier erfolgt die Spezialisierung des Modells auf bestimmte Aufgaben und Anwendungen (Yenduri et al., 2024; Zhao et al., 2025).

Zusammengefasst sind Large Language Models tiefe neuronale Netzwerke, deren Bauweise sich durch die Transformer-Architektur und den Aufmerksamkeitsmechanismus auszeichnet. Diese KI-Systeme werden mit riesigen Datenmengen gefüttert und trainiert.

Aufgrund ihrer Deep Learning Fähigkeiten lernen LLMs Muster und Bedeutungen in menschlicher Sprache selbst zu erkennen. Mit diesem Wissen können sie Wort für Wort über Wahrscheinlichkeiten berechnend sinnvolle, kontextuelle Texte generieren und komplexe Aufgaben lösen. Ab jetzt wird der Fokus auf ein bestimmtes Sprachmodell gelegt, das für diese Arbeit von Relevanz ist.

Der *Generative Pre-trained Transformer* (GPT) ist eine spezielle Ausführung eines LLMs. Das GPT-Sprachmodell baut, wie der Name vermuten lässt, auf der Transformer-Architektur auf und ist neben Sprachmodellierung und Texterkennung vor allem auf Textgenerierung spezialisiert. GPT ist ein *generatives* Modell bzw. eine generative KI und fokussiert sich auf die Generierung neuer Inhalte in verschiedenen Formaten, wie z.B. Text, Bild oder Musik (Yenduri et al., 2024). Das Modell basiert auf der zweiteiligen Trainingsstrategie, die für moderne LLMs oben bereits beschrieben wurde. GPT wird im Pre-Traning mit vielen Textdaten gefüttert und versucht im selbstüberwachenden Lernen das nächste Token auf Basis der vorherigen vorherzusagen. Bei GPT-Modellen kann der Aufmerksamkeitsmechanismus auch die Beziehungen zwischen weiter auseinander liegenden Wörtern im Kontext erfassen. Beim Generieren des nächsten Wortes kann damit der ganze Satz berücksichtigt werden, was die Fähigkeit der Sprachmodellierung deutlich verbessert. Nach der Phase des Pre-Trainings durchlaufen GPT-Modelle das Finetuning für bestimmte Aufgaben, wie etwa Übersetzung, Textklassifizierung oder Dialogfähigkeit. In dieser Phase werden sie mit speziellen Datensätzen passend zum Einsatzfeld trainiert (Mader, 2024, S. 16–18; Yenduri et al., 2024).

Wie bereits erwähnt, entwickeln LLMs ab einer bestimmten Modellgröße überraschende Eigenschaften und zeigen andere Verhaltensmuster als die kleineren PLMs (Zhao et al., 2025). Unter *Emergent Abilities* werden diese nicht vorhersehbaren Fähigkeiten zusammengefasst (Wei et al., 2022). Wei et al. klassifizieren z.B. das *Few-shot Prompting* von Brown et al. (2020) als Emergent Ability von LLMs. In der GPT-Modellreihe konnte erst das deutlich größere *GPT-3*, sinnvolle und nicht zufällige Ergebnisse auf Few-shot Prompts geben. Dessen Vorgänger *GPT-1* und *GPT-2* waren dazu nicht fähig und mussten für gezielte Aufgaben und Anwendungen feinabgestimmt werden (Radford et al., 2018, 2019). Eine weitere Emergent Ability von LLMs ist das *In-Context Learning* (Brown et al., 2020). Beim In-Context Learning erhält ein Sprachmodell die Instruktion für eine bestimmte Aufgabe direkt im Prompt, wie z.B. beim *Shot Prompting*. Das bedeutet, dass LLMs ohne Finetuning sofort für verschiedene Aufgaben genützt werden können, was ein großer Vorteil ist (Bernauer, 2024, S. 16). Mehr zu den in diesem Absatz erwähnten Prompting Methoden wird Kapitel 3.5 erklärt.

Large Language Models bringen aber auch Probleme mit sich, eines davon sind *Halluzinationen* (Ji et al., 2023). Große Sprachmodelle generieren manchmal Texte die unsinnig oder nicht wahrheitsgemäß in Bezug auf den Input sind. Die Ursache dafür können die Trainingsdaten sein oder das Training selbst bzw. Modellierungsentscheidungen neuronaler Modelle. Halluzinierter Text ist in jedem Fall problematisch, da er auf den ersten Blick nicht immer merkwürdig scheint und die Identifizierung möglicher Halluzination spezielle Techniken erfordert (Ji et al., 2023).

Die Forschung zu Sprachmodellen schreitet intensiv voran und mittlerweile gibt es viele LLMs von verschiedenen Anbietern. DeepMind trainiert beispielsweise das Sprachmodell *Gopher* (Rae et al., 2021). Die Sprachmodelle *Language Models for Dialog Applications* (*LaMDA*) (Thoppilan et al., 2022), *Pathways Language Model* (*PaLM*) (Chowdhery et al., 2023), *PaLM 2* (Anil et al., 2023) und die *Gemini*-Modelle (Anil et al., 2025) stammen von Google. *LLaMA* (Touvron, Lavril, et al., 2023) und *LLaMA 2* (Touvron, Martin, et al., 2023) wurden als Sprachmodelle von Meta AI vorgestellt. Neben diesen bekannten Modellen gibt es noch viele mehr. Im nächsten Kapitel wird der Fokus auf die Sprachmodelle von OpenAI gelegt (2025c), die beim Schreibprozess dieses Kapitels (Juni 2025) überwiegend auf dem Sprachmodell *GPT-4* basieren (Achiam et al., 2023).

3.4. OpenAI's GPT-Sprachmodell und ChatGPT

OpenAI ist ein Forschungskonzern, der sich auf die Entwicklung von KI-Technologie spezialisiert hat. Das generative Sprachmodell GPT wurde von OpenAI entwickelt und obwohl es noch sehr jung ist, hat GPT eine bemerkenswert schnelle Evolution hingelegt, die zu ChatGPT geführt hat (Yenduri et al., 2024; Zhao et al., 2025). Bevor ChatGPT beschrieben wird soll die Geschichte von GPT kurz erläutert werden.

Die Entwicklung der GPT-Modelle von OpenAI hat 2018 mit der Veröffentlichung von *GPT-1* gestartet (Radford et al., 2018). Dieses Modell markierte bereits einen bedeutenden Schritt in der KI-Technologie (Yenduri et al., 2024). *GPT-2* folgte 2019 (Radford et al., 2019) und wieder ein Jahr später *GPT-3* (Brown et al., 2020), das mit über 175 Mrd. Parameter 100-mal größer als *GPT-2* ist und viel qualitativere Ergebnisse bei NLP-Aufgaben erzielte (Yenduri et al., 2024).
Mit *GPT-3* schufen die Forscher:innen von OpenAI ein mächtiges Basismodell, aus dem sie mit weiteren Verbesserungen noch fähigere LLMs entwickeln konnten. Mithilfe von *Reinforcement Learning with Human Feedback (RLHF)* als maschinelle Lernmethode wurde *GPT-3* feinabgestimmt, um besser auf Nutzer:innen-Absichten eingehen zu können (Ouyang et al., 2022). Das RLHF führte gemeinsam mit weiteren Anpassungen zu *InstructGPT* und *GPT-3.5* – verbesserte und für verschiedene Einsätze optimierte Modelle von OpenAI (Bai et al., 2022; Ouyang et al., 2022; Zhao et al., 2025). Eines dieser *GPT-3.5* Modelle, welches speziell zur Dialogführung optimiert wurde, führte 2022 zur bekannten KI-Anwendung *ChatGPT* (OpenAI, 2022; Zhao et al., 2025). 2023 setzte OpenAI die GPT-Reihe mit der Entwicklung von *GPT-4* (OpenAI, 2023) und diversen Ablegern, wie *GPT-4o* (OpenAI, 2024) und weiteren Modellen (OpenAI, 2025c) innerhalb der nächsten Jahre fort. Mit den meisten davon arbeitet auch ChatGPT. Die neuen Modelle werden in Summe immer leistungsfähiger, schneller und besser im Lösen komplexer Aufgaben. Mit *GPT-4o*, *„o"* steht für *„omni"*, wurde 2024 ein Modell eingeführt, das erstmals neben Text zusätzlich auch Audio-, Bild- und Videodateien als Input verarbeiten und als Output generieren kann (OpenAI, 2024). Und die rasche Evolution der GPT-Reihe wird nicht langsamer, denn seit 2025 ist für zahlende Nutzer:innen von ChatGPT bereits *GPT-4.5* als Version nutzbar (OpenAI, 2025a).

Was ist ChatGPT? *ChatGPT* ist eine generative KI-Anwendung von OpenAI, die seit November 2022 der Öffentlichkeit zur Verfügung steht (OpenAI, 2022). ChatGPT beruht auf dem großen Sprachmodell GPT und funktioniert im Hintergrund genau wie dieses LLM: ChatGPT generiert schrittweise das nächste Token innerhalb einer Sequenz, vorhergesagt auf Basis der zuvor generierten Tokens. Auf diese Weise generiert ChatGPT sinnvollen Text und kann menschenähnliche Konversationen innerhalb eines Kontexts führen. ChatGPT ist ein Chat Interface, das als benutzerfreundlicher Assistent fungieren soll (Tabone & de Winter, 2023). Systeme wie ChatGPT nennt man bekanntlich auch *Chatbots*, da sie fähig sind mit Menschen in einem Dialog zu interagieren (Chat) und autonom NLP-Aufgaben zu bearbeiten (Bot). Je nach Funktionsweise und Einsatzfeld gibt es verschiedene Arten von Chatbots (Stucki et al., 2018).

Wenn man ChatGPT fragt „Was bist du?“ antwortet es selbst mit:

> *„Ich bin ChatGPT, ein KI-gestütztes Sprachmodell von OpenAI. Du kannst mich dir vorstellen wie eine sehr gut trainierte digitale Schreibassistenz oder ein Gesprächspartner, der auf eine große Menge an Wissen zurückgreifen kann. Ich kann Texte analysieren, erklären, bewerten, formulieren, übersetzen, bei Programmierung helfen, Ideen entwickeln – oder einfach Fragen beantworten, die dir gerade durch den Kopf gehen. […]“*

Diese Definition deckt sich mit den Beschreibungen, die man in der Literatur findet. OpenAI bietet mit dieser frei zugänglichen Anwendung eine Hilfestellung für die Erledigung alltäglicher Aufgaben. Das Spektrum der lösbaren Probleme von ChatGPT ist breit.

Seit der Veröffentlichung von ChatGPT steht KI und KI-Forschung mehr im gesellschaftlichen und medialen Interesse als je. Die Chancen und Risiken, die mit dieser Technologie einhergehen, sind Gegenstand vieler Diskurse (Tabone & de Winter, 2023).

Die Stärken des Chatbots liegen auf der Hand. ChatGPT zeigt überragende Fähigkeiten beim Kommunizieren mit Menschen. Das KI-Modell verfügt über enormes Wissen, kann komplexe Aufgaben und mathematische Probleme lösen und den Kontext eines Dialogs nachvollziehen. ChatGPT scheint der mächtigste Chatbot in der KI-Geschichte zu sein und wird die Forschung zu KI in Zukunft maßgeblich beeinflussen (Zhao et al., 2025).

ChatGPT hat aber auch Limitationen. Zum Einführungstermin von ChatGPT merkte OpenAI (2022) an, dass der Chatbot plausibel klingende, aber unwahre bzw. unsinnige Antworten geben kann, dass die Antworten aufgrund der Trainingsdaten einem Bias unterliegen können und, dass die Wahl der Texteingabe den Output ändern kann. Kurz, ChatGPT kann Fehler machen.

Die Warnung vor übermäßiges Vertrauen auf ChatGPTs Ausgaben ist berechtigt. Oertner (2024) warnt in einem kritischen Beitrag davor den Chatbot als Informations-, Recherche- oder Brainstormingtool einzusetzen. Seine Aussagen können irreführend, veraltet, unvollständig, falsch oder unwahr sein, Zitate sind möglicherweise fiktiv und Brainstormideen suggestiv bzw. kreativitätsmindernd. Es gilt zu bedenken, dass ChatGPT in Wahrheit von nichts eine Ahnung hat. Das KI-Sprachmodell speichert sein Wissen nur implizit über

vernetzte Token ab, deren Rekombination richtig, aber auch falsch sein kann (Oertner, 2024).
ChatGPT kann seine Fehler zwar eingestehen (Budhwar et al., 2023), aber das KI-Modell führt keine eigene Plausibilitätsprüfung durch und bessert seine Fehler nicht aus (Tabone & de Winter, 2023). Nutzer:innen sollten den Output stets selbst kontrollieren. Ein vernünftiger Gebrauch von ChatGPT setzt kritische Hinterfragung voraus (Oertner, 2024).

Diese kritischen Quellen beziehen sich vermutlich auf frühere Modelle von ChatGPT. Aktuelle Modelle wie GPT-4 (OpenAI, 2023) und *GPT-4o* (OpenAI, 2024) sollen laut Angaben OpenAIs weniger Fehler machen und vertrauenswürdigere Antworten liefern, wenngleich manche der Limitationen, z.B. das Halluzinieren, immer noch bestehen.

Trotz einschränkender Mängel verfügt ChatGPT über ein breites Set an Fähigkeiten für diverse Aufgaben. Damit diese Fähigkeiten in einem qualitativen Output münden kommt es auf einen ebenso qualitativen Input an. Das betrifft neben den Textdaten, mit denen das Sprachmodell trainiert wurde, besonders auch die Texteingabe selbst, mithilfe jener Nutzer:innen ihre Anfrage an ChatGPT formulieren (Budhwar et al., 2023). Das nächste Kapitel soll die Kommunikation mit Sprachmodellen beschreiben.

3.5. Prompting – Das Kommunizieren mit KI

Ein *Prompt* ist die Eingabe in ein generatives KI-Modell, der zu einer gewünschten Ausgabe führen soll. Prompts können neben Wörtern auch Bilder, Tonaufnahmen und andere Medien beinhalten, üblicherweise weisen sie jedoch immer eine Textkomponente auf. Das *Prompting* beschreibt den aktiven Prozess der Prompterstellung bis hin zur Eingabe, die eine Antwort generiert (Schulhoff et al., 2025).

Wenig überraschend, ist die korrekte Gestaltung des Prompts von entscheidender Bedeutung für die Interaktion mit künstlicher Intelligenz. Prompting wird mit zunehmender Verbreitung von generativer KI eine immer wichtigere Fertigkeit (Steinmann & Piazza, 2024) und zahlreiche Studien belegen, dass die Formulierung und Struktur des Prompts signifikant mit der Qualität der Ausgabe korrelieren (z.B. Brown et al., 2020). Das *Prompt Engineering* beschäftigt sich mit der iterativen Erstellung, Optimierung und Implementierung von effektiven Prompts, die ein Sprachmodell zu einer qualitätsvollen Antwort führen (Liu et al., 2023; Oppenlaender, 2024) und ist eine neuartige, kontraintuitive Fertigkeit für Menschen, die erst erlernt werden muss, bevor unerfahrene Nutzer:innen qualitative hochwertige Prompts erstellen können (Oppenlaender et al., 2024). Glücklicherweise hat das Gebiet des Prompt Engineering für diesen Zweck bereits verschiedenste *Prompting Techniken* definiert. Diese Techniken sind erprobte Strategien für die Strukturierung, Verbesserung und eventuell wiederholte Anwendung von Prompts und zum Teil maßgeschneidert für bestimmte Anforderungen an ein Sprachmodell (Schulhoff et al., 2025).

Im Prompt Engineering hat sich das Prinzip des *Shot Prompting* als erfolgreiche Prompting Technik herausgestellt. Ein *Shot* bezeichnet ein demonstratives Beispiel, das als Vorlage

oder Musterlösung für die Bearbeitung einer Aufgabe dient. Bereitgestellt im Prompt soll dieser Shot die generative KI zur Erfüllung der Aufgabe anleiten (Schulhoff et al., 2025). Allgemein können drei Arten von Shot Promptings unterschieden werden: *Zero-shot*, *One-shot* und *Few-shot Prompts*. Zero-Shot Prompts bestehen aus einer Aufgabenstellung und der Anweisung diese zu erledigen, enthalten jedoch kein Beispiel als Musterlösung. One-shot Prompts stellen der KI zusätzlich eine Vorlage für die Lösung bereit. Alle Eingaben mit mehr als einem Beispiel, werden als Few-shot Prompts bezeichnet (Steinmann & Piazza, 2024).

Eine fortgeschrittene Prompting Technik ist das *Chain-of-Thought* (*CoT*) *Prompting*, diskutiert von Wei et al. (2023). CoT-Prompting fördert die Fähigkeit von LLMs, komplexe Denkprozesse zu bewältigen. Diese Methode basiert auf der Idee, dem KI-Modell manuelle Denkschritte für das Lösen einer Aufgabe vorzugeben. Enthält der Prompt bestimmte Angaben zur Bearbeitung der Aufgabe und Struktur der Lösung, können LLMs komplexe Denkaufgaben deutlich besser lösen (Wei et al., 2023).

Chain-of-Thought (CoT) Prompting kann mit Shot Prompting kombiniert werden (Schulhoff et al., 2025). Kojima et al. (2023) stellen das vereinte *Zero-shot CoT Prompting* für LLMs vor, eine Verfeinerung des Zero-shot Prompts mit der simplen Aufforderung „*Lass uns Schritt für Schritt nachdenken*“. Durch Integration dieser Aufforderung in den Prompt wird das LLM dazu angeregt seinen Denkprozess schrittweise zu gestalten. Erst nach mehreren Zwischenschritten in der Argumentationslinie wird dann die finale Antwort zur Aufgabe generiert. Das Produzieren von Denkketten durch Zero-shot CoT übertrifft die Leistung von LLMs gegenüber gewöhnlichen Zero-Shot Techniken in Aufgaben zu Arithmetik, symbolischen und logischem Denken deutlich (Kojima et al., 2023). Das *Few-shot CoT* Prompting führt in diesem Zusammenhang ebenfalls zu besseren Ausgaben von KI-Modellen. Die Few-shot Prompts enthalten zusätzliche Chains-of-Thought, die manuell einzugeben sind. Daher nannte man diese Technik ursprünglich *Manual-Chain-of-Tought* (*Manual-CoT*) *Prompting* (Schulhoff et al., 2025). Seit kurzem liegen auch Vorschläge für das *Automatic Chain-of-Thought* (*Auto-CoT*) *Prompting* vor (Zhang et al., 2022), bei dem LLMs automatisiert Denkketten, basierend auf einer Vielfalt von Beispielen, generieren können. Diese Shot- und CoT-Prompting Techniken ermöglichen viele spezialisierte Typen von Prompts und bieten einen Einblick in das große Arsenal der Prompt Engineering Techniken.

Schulhoff et al. (2025) fassen noch weitere textbasierte Prompting Techniken aus verschiedenen Studien zusammen, die für die Arbeit mit LLMs relevant sein können und teilweise auf Shot- und CoT-Prompting aufbauen. Diese Techniken umfassen Methoden wie *Self-Criticism*, wo das Modell aufgefordert wird, seine eigene Antwort zu überprüfen und verbessern, *Decomposition,* wo komplexe Aufgaben in kleinere, leichter bearbeitbare Teilaufgaben separiert werden und *Ensembling*, wo mehrere Prompts zum gleichen Problem gestellt werden und die Kombination aller Modellantworten den finalen Output erzeugen. Ensembling erhöht die Genauigkeit und Zuverlässigkeit des LLMs, braucht dafür jedoch viele Modellanfragen.

Darüber hinaus gibt es weitere Tricks für die Formulierung effektiver Prompts. Mithilfe von *Role Prompting* (Kong et al., 2024) sind LLMs dazu in der Lage eine spezifische Rolle einzunehmen. Diese Eigenschaft erlaubt es LLMs komplexe, menschliche Verhaltensweise und Interaktionen in verschiedenen Kontexten zu simulieren. Durch die Zuweisung einer Rolle, wie zum Beispiel eine exzellente Mathematiklehrperson, können komplexe Aufgaben in diesem Kontext besser gelöst werden. Role Prompting erzeugt qualitativere Ergebnisse von LLMs als standardmäßiges Zero-shot Prompting und erweist sich zusätzlich als effektiver Auslöser von Chain-of-Thought Denk-prozessen (Kong et al., 2024).

ChatGPT ist ein Large Language Model (Brown et al., 2020) und die beschriebenen Prompting Techniken sind auch für ChatGPT anwendbar. Im Vergleich zur normalen Nutzung von ChatGPT konnte gezeigt werden, dass die Verwendung von One-shot (Steinmann & Piazza, 2024), Few-shot, Few-shot CoT und Zero-shot CoT Prompting die Gesamtleistung von ChatGPT verbessert (Zhong et al., 2023). Auch das Role Prompting erhöht die Genauigkeit von ChatGPT's Antworten (Kong et al., 2024).

Die auf das Sprachmodell *GPT-3* basierende Version von ChatGPT erzielt unter der Eingabe von One-shot und Few-shot Prompts Ergebnisse mit der höchsten Genauigkeit. Das gilt vor allem für die besonders großen Versionen von *GPT-3*, die bis zu 175 Mrd. Parameter verfügen. Je mehr Trainingsdaten ein Modell gefüttert bekommt, desto eher führt die Verwendung von One-shots und speziell Few-shots zu genaueren Ergebnissen (Cooper, 2023). Das würde bedeuten, dass sich diese beiden Prompting Techniken für die größeren GPT-4-Modelle noch besser eignen.

Eine weitere Hilfestellung für effizientes Prompting bieten sogenannte *Promptframeworks*. Ein Promptframework ist ein Rahmenwerk zur Verwaltung, Vereinfachung und Erleichterung der Interaktion mit großen Sprachmodellen (Liu et al., 2023). Sie zählen Grundkomponenten auf, die ein Prompt beinhalten sollte, damit ein LLM wie ChatGPT einen gewünschten und qualitativen Output generiert. Nach Heston und Kuhn (2023) enthält ein gut strukturierter Prompt einen *Kontext*, einen *Aufgabenstellung*, eine *Rollenzuweisung* und das gewünschte *Ausgabeformat*. Die meisten Promptframeworks berücksichtigen diese Merkmale.
Im Internet kursieren verschiedene Promptframeworks. Meistens tragen sie eine charakteristische Abkürzung und erfüllen je nach Einsatzgebiet verschiedene Ziele. Das Framework *ReAct* (Yao et al., 2023) versucht beispieksweise zwei Fähigkeiten von LLMs, das Erzeugen von Denkpfaden (*Reasoning*) und das Handeln nach Aufgaben-spezifischen Aktionen (*Acting*), in Synergie zu verwenden. Das *IDEA*-Framework wird für Menschen im Bildungsbereich als Anleitung für effektives Prompt Engineering vorschlagen. *IDEA* involviert laut Originalquelle „*(a) Including essential PARTS, (b) Developing CLEAR prompts, (c) Evaluating outputs and REFINEing prompts, and (d) Applying the output with accountability*" in die Strategie der Prompterstellung (Park & Choo, 2024).

Das *IDEA*-Framework wirkt sehr durchdacht, für die Verwendung in dieser Masterarbeit ist es jedoch zu detailliert. Ein anderes Framework, das ähnliche Komponenten aufweist und

für den Unterrichtskontext ebenfalls vielversprechend wirkt, scheint für diese Forschung sinnvoller zu sein und wurde von meiner Betreuerin M. Korner vorgeschlagen.

Das angesprochene Framework ist das sogenannte *PROPER*-Framework. Dieses Framework wird von Gruber (2024) für die Kommunikation mit generativer KI empfohlen. Die Kurzform ***PROPER*** steht für die Bausteine, die ein Prompt aufweisen soll:

- ***P****ersona* – Welche Rolle soll das Modell einnehmen?
- ***R****equest* – Was ist das Ziel des Prompts? Welche Aufgabe soll das LLM erfüllen?
- ***O****peration* – Wie soll das LLM methodisch vorgehen?
- ***P****resentation* – Wie sollen Stil/Ton/Format des Ergebnisses aussehen?
- ***E****xamples* – Gibt es Beispiele/Vorlagen, die dem LLM helfen könnten?
- ***R****efinement* – Wie kann der Prompt verbessert werden? Welches Feedback könnte dem LLM behilflich sein?

Der Vorteil dieses Frameworks ist sein einfacher und anpassbarer Aufbau und es beinhaltet die wichtigsten Bestandteile die ein guter Prompt nach Heston und Kuhn (2023) haben sollte. Es gibt jedoch keine wissenschaftlichen Studien oder sonstige Hinweise über die Wirksamkeit des *PROPER*-Frameworks.

Die technischen Aspekte von Sprachmodellen bis hin zum Prompting wurden nun ausreichend beschrieben. Das nächste Kapitel soll den Einsatz von KI-Sprachmodellen, konkret ChatGPT, in der Bildung umreißen und damit auf das Forschungsthema dieser Arbeit hinführen.

3.6. Der Einsatz von ChatGPT als Beurteilungshilfe für Lehrkräfte

Die Fähigkeiten generativer künstlicher Intelligenz sind von großem Interesse für viele Bereiche der Gesellschaft, so auch für das Bildungswesen. Ein potenzieller Einsatz im Bildungskontext wäre die KI-gestützte Beurteilung von schriftlichen Schüler:innenabgaben, beispielsweise zur Bewertungskompetenz. Dieses Kapitel gibt einen Ausblick auf die Rechts- und Studienlage zu diesem Einsatz.

3.6.1. KI-Initiativen des Bildungsministeriums

Das Bundesministerium für Bildung, Wissenschaft und Forschung (BMBWF) setzt sich laufend mit künstlicher Intelligenz auseinander und zeigt Bemühungen, die Entwicklungen dieser neuen Technologie für das Bildungssystem einzuordnen und mögliche Potenziale für diverse Einsätze zu untersuchen (BMBWF, 2023). Das Konzept der *Digitalen Schule*, eine Kombination moderner, digitaler Infrastruktur und zukunftsweisender, inspirierender Pädagogik, ist dabei die Vision und das Ziel des BMBWF für einen zukunftsorientierten Schulbetrieb. Der von der Bundesregierung beschlossene *8-Punkte-Plan* ist die Basis für die Ziele der Digitalen Schule. Er regelt die flächendeckende Umsetzung von digital unterstütztem Lernen und Lehren sowie die breit angelegte Implementierung neuer Lehr- und Lernformate (BMBWF, 2025b). Im Jahr 2023 präsentierte Bildungsminister M. Polaschek das *Schulpaket KI*, das darauf abzielt, die Grundlage der Digitalen Schule in Österreich um den Aspekt der künstlichen Intelligenz zu erweitern. Dieses Paket sieht diverse Ziele für die nachhaltige Implementierung von KI in Schulen vor, etwa im Bereich der Schulentwicklung, der Lehrer:innenbildung und von Lehr- und Lern-Tools. Ein Beispiel hierfür ist das aktuelle *Pilotschul-Projekt*, wo interessierte Expert- und Expert+Schulen KI-unterstützte Lernsoftware evaluieren und wertvolle Pionierarbeit leisten (BMBWF, 2025a; eEducation, 2025).

Die vorliegenden Initiativen des BMBWF demonstrieren dessen Bestreben, KI im Schulwesen zukünftig auf vielfältige Weise zu implementieren. In einer Handreichung des BMBWF (2023) werden z.B. einzelne Unterstützungsmöglichkeiten von KI für Lehrpersonen angesprochen. Etwa zur automatisierten Ideensammlung, zur gestützten Planung von Unterrichtssequenzen oder bei der Erstellung von Aufgabenstellungen mit mehreren Varianten. Die Möglichkeit einer KI-gestützten Beurteilung der Leistungen von Schüler:innen führt die Handreichung von 2023 jedoch nicht an. Natürlich waren die Fähigkeiten von KI-Modellen, wie ChatGPT, zu dieser Zeit noch weniger fortgeschritten und schwer einschätzbar, dennoch lässt sich vom BMBWF zur möglichen KI-gestützten Beurteilung auch aktuell, Anfang August 2025, nichts finden.
Das Problem vorgetäuschter, schriftlicher Leistungen durch KI-Tools scheint hier im Vergleich dringlicher zu sein. Die Gefahr dieser unerlaubten Nutzung wird vom BMBWF bereits diskutiert und es werden Umgangsmöglichkeiten dafür vorgeschlagen (BMBWF, 2023, 2025a).

Seitens des Bildungsministeriums gibt es zur KI-gestützten Beurteilung also noch keine Ideen, Gebote oder Verbote für Lehrer:innen. Das Konzept der Digitalen Schule lässt nur vermuten, dass solche Bestrebungen willkommen sind – sofern die Beurteilung mit KI rechtlich erlaubt ist.

3.6.2. Rechtslage

Dürfen Lehrpersonen KI-Sprachmodelle, wie ChatGPT, dazu verwenden, um die Leistungen ihrer Schüler:innen zu beurteilen? Laut dem Schulunterrichtsgesetz (SchUG) und der Leistungsbeurteilungsverordnung (LBVO) hat die Lehrperson jedenfalls selbst die Beurteilung zu treffen:

> §18, Absatz 1, SchUG:
> „*Die Beurteilung der Leistungen der Schüler in den einzelnen Unterrichtsgegenständen hat der Lehrer durch Feststellung der Mitarbeit der Schüler im Unterricht sowie durch besondere in die Unterrichtsarbeit eingeordnete mündliche, schriftliche und praktische oder nach anderen Arbeitsformen ausgerichtete Leistungsfeststellungen zu gewinnen*. […]“ (RIS, 2025e)

> §11, Absatz 2, LBVO:
> „*Der Lehrer hat die Leistungen der Schüler sachlich und gerecht zu beurteilen, dabei die verschiedenen fachlichen Aspekte und Beurteilungskriterien der Leistung zu berück-sichtigen und so eine größtmögliche Objektivierung der Leistungsbeurteilung anzustreben.*“ (RIS, 2025b)

Diese Absätze schreiben jedoch nicht vor, ob und welche Hilfsmittel Lehrkräfte zur Beurteilung verwenden dürften. Eine KI-gestützte Bewertung, an der sich Lehrkräfte nur unterstützend für die finale Beurteilung orientieren, wird durch die LBVO zumindest nicht direkt ausgeschlossen. Außerdem ließe sich KI als Argument für höhere Objektivität verwenden. Sofern KI-Systeme im Vergleich zu Menschen tatsächlich objektivere Beurteilungen erzielen, können – eine Frage, die offen ist – würde die LBVO den Einsatz von KI im Sinne der „größtmöglichen Objektivierung“ befürworten, solange sie sachlich, gerecht und fachlich richtig und nach Berücksichtigung von vorgegebenen Beurteilungskriterien beurteilen kann.

Der deutsche Schulrechtsexperte Stephan Rademacher sieht KI-Systeme durchaus relevant für die Leistungsbewertung durch Lehrkräfte, auch wenn dieser Gebrauch rechtlich nicht eindeutig geklärt ist. Was aus seiner Sicht dagegen spricht, sind der EU AI-Act und die Tatsache, dass Lehrpersonen stets selbst zur Bewertung von Leistungen kommen müssen (Limpert, 2024b).

Der *EU AI-Act*, auch *KI-Gesetz*, der Europäischen Union trat 2024 in Kraft und ist das erste staatenübergreifende Gesetz, welches klare Ziele und Standards für den Gebrauch künstlicher Intelligenz definiert. Das Gesetz ordnet KI-Anwendungen in vier Risiko-kategorien ein. Eine Kategorie davon sind *Hochrisiko*-KI-Systeme, die zwar zugelassen werden, aber bestimmte Anforderungen und Verpflichtungen erfüllen müssen (Bundeskanzleramt, 2024). Im Bildungswesen fallen einige KI-Anwendungen unter die Einstufung „hochriskant“. Etwa solche, die Personen oder Gruppen nach ihrem Verhalten oder individuellen Merkmalen bewerten. Das betrifft unter anderem das Bewerten von Lernergebnissen oder das Steuern von Lernprozessen (Limpert, 2024a).

Rademacher stellt daher klar, dass Lehrpersonen ihre eigene Bewertung weder vollständig noch in Teilen *allein* durch KI-Modelle, wie ChatGPT, erzeugen lassen dürfen. Erlaubt sei es seiner Ansicht nach hingegen, wenn die KI nur als *ergänzende* Einschätzung zur eigenen Beurteilung verwendet wird, da die grundsätzliche Leistungsbeurteilung so weiterhin durch die Lehrperson erfolgt. Ebenso zulässig fände Rademacher die Verwendung von KI bei der Auswertung von Antworten eines Online-Tests, sofern die Lehrperson die KI-gestützte Korrektur zumindest stichprobenartig kontrolliert (Limpert, 2024b).

In Summe kann also festgestellt werden, dass die Beurteilung von Schüler:innenleistungen mit KI derzeit weder eindeutig verboten noch ausdrücklich erlaubt ist. Vernünftig erscheint nur, dass eine alleinige Generierung von Beurteilungen durch KI und deren unkontrollierte Übernahme durch die Lehrperson nicht tragbar sind und unverantwortlich wären. Sofern KI im Beurteilungsprozess genützt wird, legt die rechtliche Recherche nahe, dass KI höchstens ergänzend zur eigenen Beurteilung der Lehrperson verwendet werden darf.

Vermutlich sind die Fähigkeiten von KI-Modellen für schulischen Beurteilungen von zuständigen Behörden noch nicht genau untersucht worden, was verständlich ist, da diese Technologie erst seit kurzem kommerziell verfügbar ist – zur Erinnerung: ChatGPT wurde erst vor knapp drei Jahren veröffentlicht (OpenAI, 2022). Die Forschung ist jedoch einen Schritt voraus und gibt einen Ausblick auf Beurteilungen durch künstliche Intelligenz.

3.6.3. Studienlage

Sok und Heng (2023) argumentieren, dass ChatGPT einige große Vorteile für die Bildung und Forschung haben kann. Neben dem Brainstorming von Ideen, der Erstellung eines Aufsatzes, der Verbesserung pädagogischer Praxis sehen sie auch die von Kasneci et al. (2023) vorgeschlagene Generierung von Lernbewertungen als Einsatzmöglichkeit des Chatbots. Kasneci et al. (2023) schlagen die Verwendung von LLMs zur semi-automatischen Beurteilung von Schüler:innenarbeiten mit Feedback vor, wo Stärken und Schwächen der Arbeit hervorgehoben werden. Dies kann der Lehrperson viel Zeit im Zusammenhang mit der Bereitstellung individuellen Feedbacks ersparen. LLMs können weiters auch Bereiche identifizieren, die Lernenden schwerfallen, und helfen damit Lehrkräften den Lernfortschritt der Lernenden zu evaluieren. Natürlich bestünden Risiken bezüglich unfairen bzw. verzerrten Bewertungen durch die KI. Eine effektive Nutzung der KI ließe sich mit den richtigen Mitteln dennoch empfehlen (Kasneci et al., 2023).

Einige Studien untersuchten die Beurteilungsleistung von ChatGPT auf geschriebene Essays. Der Chatbot kann trainiert werden diese zu benoten (Lo, 2023). De Winter et al. (2023) zeigten, dass ChatGPT gut im Erkennen von Fehlern in Texten, aber schlecht im konsistenten Beurteilen ist. Verschiedene Prompts können zu anderen Bewertungen des Chatbots führen, wodurch diese nicht mehr verlässlich sind. De Winter (2024) argumentiert, dass nur wiederholte Prompts, aus denen durchschnittliche Bewertungen von ChatGPT berechnet werden, statistisch verlässliche und konsistente Schätzungen liefern können, beispielsweise für die Schreibqualität des Essays.

Zu einer effektiven Beurteilung gehört auch ein konstruktives Feedback. ChatGPT könnte im Geben von Feedback ebenfalls eine Unterstützung für die Lehrkraft sein (Fleischmann, 2022; Kasneci et al., 2023). Der größte Nutzen automatisierter Feedbackgenerierung wäre demnach sicherlich die resultierende Zeitersparnis für Lehrperson. Gleichzeitig könnte allen Lernenden individuelles Feedback geboten werden, was in der Praxis leider auch oft zu kurz kommt. Ob das Feedback der aktuellen Chatbot-Versionen von OpenAI jedoch gut genug ist, ist fraglich.

Yoon et al. (2023) evaluierten ChatGPT's Feedbackqualität zu Kohärenz und Kohäsion von Englischlernenden beim Schreiben von Essays. Sie empfehlen den Einsatz des Chatbots in diesem Kontext nicht, da er ohne spezifisches Training für diese Aufgabe zu allgemeines und somit kein effektives Feedback gibt, dass den Lernenden weiterhelfen würde (Yoon et al., 2023). Steiss et al. (2024) zeigten, dass gut ausgebildete Bewerter:innen qualitativeres Feedback auf Essays bereitstellen als ChatGPT. Generative KI könne aus Sicht von Steiss et al. dennoch zur Feedbackerstellung in bestimmten Situationen nützlich sein, z.B. für erste Entwürfe oder sollte keine ausgebildete:r Bewerter:in vorhanden sein. In Anbetracht der Leichtigkeit mit der das Feedback generiert werden kann reiche die Gesamtqualität des Feedbacks in diesen Situationen aus (Steiss et al., 2024). Eine jüngere Studie von Asadi et al. (2025) untersuchte den Effekt einer Kombination aus Lehrenden- und ChatGPT-Feedback auf die Schreibfähigkeit von Englischlernenden in Essays. Lernende, die ein kombiniertes Feedback erhielten, erzielten signifikante Verbesserungen beim Schreiben von Essays als jene, die reines Lehrerfeedback bekamen. Asadi et al. (2025) sehen Potenzial in der Integration von KI-Tools in den Lehrprozess, insbesondere bei der Bereitstellung personalisierten Feedbacks.

Eine interessante Studie von Mühlhoff und Henningsen (2024) stellt die Leistung des neuesten Sprachmodells von OpenAI, *GPT-4*, im Beurteilen und Feedback in starke Kritik. Untersucht wurde das KI-gestützte Beurteilungstool des deutschen Unternehmens *Fobizz*, dass Lehrkräfte bei der automatischen Bewertungen und beim Feedback auf Hausaufgaben unterstützen soll. Mühlhoff und Henningsen erkannten mehrere methodische und technische Defizite der von Fobizz angebotenen *KI-Korrekturhilfe*, die sie als „fatale Gebrauchshindernisse" klassifizierten. Es wurden einige Schwächen und Einschränkungen identifiziert, die den Gebrauch dieser Anwendung nicht empfehlen: 1) die Gesamtnote und das Feedback variieren erheblich bei mehreren Durchläufen, 2) die Erkennung von sachlichen Fehlern und unsinnigen Abgaben funktioniert nicht zuverlässig, 3) die Korrekturhilfe kann einstellbare Bewertungskriterien nicht konsistent umsetzen, 4) das Feedback weist mehrere Inkonsistenzen auf, 5) die Umsetzung des von der KI vorgeschlagenen Feedbacks führt nicht zu besseren Noten, sondern 6) eine Bestnote scheint nur durch eine Überarbeitung mit ChatGPT erreichbar. Mühlhoff und Henningsen legen aufgrund dieser Ergebnisse nahe, das Tool von Fobizz nicht im Schulalltag zu verwenden. Sie führen diese identifizierten Mängel auf die Limitationen von LLMs zurück, die aufgrund ihrer technischen Eigenschaften entstehen (Mühlhoff & Henningsen, 2024). Die Studie ist damit ein guter Hinweis auf die zu erwartende Beurteilungsqualität mit dem Online-Chatbot ChatGPT.

Richtigerweise weisen die genannten Studien darauf hin, dass ChatGPT Fehler machen kann und eine Verwendung mit zu beachtenden Risiken einhergeht. Auffallend ist jedoch auch, dass der Einsatz von ChatGPT wegen festgestellter Limitationen nicht per se ausgeschlossen wird. Vielmehr ruft die Forschung aus meiner Sicht dazu auf, bestehende gesellschaftliche Systeme für die Implementierung von KI bereit zu machen, Chancen auszuloten und Menschen für einen vernünftigen Umgang mit KI zu trainieren.

Die Verwendung von LLMs in der Bildung ist eine große Chance, da sie in vielen Möglichkeiten eine Unterstützung für die Arbeit von Lehrkräften sein kann. Um das volle Potential von KI in diesem Kontext zu nützen ist es jedoch essentiell, sich mit ihren Limitationen und Verzerrungen kritisch auseinandersetzen. Wenn KI in das Bildungswesen integriert wird, braucht es auch entsprechende Anforderungen an den Datenschutz, die Sicherheit, die Rechtsvorschriften, ethische Aspekte, etc. in Kombination mit kontinuierlicher menschlicher Überwachung, Anleitung und kritischer Reflexion (Kasneci et al., 2023).

Diesem Gedanken, KI als Chance zu betrachten, möchte sich diese Masterarbeit anschließen. Die Studienlage zum Einsatz von ChatGPT als Beurteilungshilfe lässt tendenziell Zweifel an dieser Möglichkeit aufkommen. Unter den kritischen Meinungen gibt es aber auch optimistische, die das Potenzial der KI sehen und aufrufen es zu nutzen. Diese Forschungsarbeit soll einen Teil dazu beitragen. Im nächsten Teil wird die geplante Forschung beschrieben.

4. Forschungsdesign

Der theoretische Hintergrund dieser Arbeit wurde in der vorherigen Kapiteln beschrieben. Auf dieser Basis beschreibt Kapitel 4 das geplante Forschungsdesign den möglichen Einsatz von ChatGPT als Beurteilungshilfe für S-Kompetenz, untersuchen zu können.

4.1. Forschungsfragen

Die rasante Entwicklung generativer Künstlicher Intelligenz bietet laufend neue Anwendungsmöglichkeiten, die in diversen Situationen genutzt werden können. Der KI-Experte Maximilian Böger ist der Meinung, es wäre eine verpasste Chance diese Technologien nicht zu nutzen. Es geht darum Symbiosen für den Einsatz von KI zu finden, um in alltäglichen Aufgaben Zeit und Energie zu sparen (Eidenberger, 2024). Für den Bildungsbereich stellt sich in dieser Hinsicht eine bedeutende Frage: Wo und wie könnten KI-Systeme als sinnvolle Unterstützung in der Schule implementiert werden?

Es gibt viele Möglichkeiten, wo intelligente Algorithmen und Maschinen als Unterstützung im Schulbetrieb eingesetzt werden könnten. Wie beschrieben, versucht das BMBWF mögliche Symbiosen für den Einsatz von KI im Schulwesen zu finden. Diesbezüglich gibt es schon einige Vorschläge, die sich besonders auf die Unterrichtsplanung, der Gestaltung von Lernprozessen, der Erstellung von Lern- und Testaufgaben, der Kommunikation und der Integration von Lehr- und Lern-Tools konzentrieren. Sowohl auf Seiten der Schüler:innen, wie der Lehrer:innen gibt es viele Optionen.

Ein zentrales Arbeitsfeld von Lehrkräften ist neben der Aufgabenerstellung auch die *Beurteilung*. Gerade hier wäre die Nutzung von künstlicher Intelligenz eine willkommene, vor allem jedoch entlastende Hilfe. Die sorgfältige Beurteilung von Leistungen, inklusive dem Korrekturaufwand und der Bereitstellung von effektivem Feedback, ist eine zeitaufwendige Tätigkeit.

Die Möglichkeit, KI als Beurteilungshilfe für schriftliche Leistungen zu verwenden, wurde bisher jedoch wenig diskutiert. Die vorgeschlagenen KI-Einsatzgebiete des BMBWF enthalten diesen spezifischen Nutzen nicht und auch die rechtliche Lage, ob Lehrkräfte KI zur (unterstützenden) Bewertung verwenden dürfen, ist nicht eindeutig geklärt. Die Studienlage zeigt allerdings, dass das KI-Sprachmodell ChatGPT trotz Einschränkungen durchaus Potenzial für die Bewertung und Feedbackerstattung haben kann. Hier möchte diese Forschungsarbeit ansetzen.

Die Bewertung von Bewertungskompetenz ist ein klarer Fall für eine aufwendige Beurteilung. Die neuen Beurteilungsmodelle zur S-Kompetenz von Gstall (2024) und Hackner (2025) sind ein wichtiger Schritt um Physiklehrkräfte mehr als bisher für die Integration der S-Kompetenz in den Unterricht zu motivieren und die Bewertung dieser Kompetenz zu erleichtern. Dennoch haben auch diese Beiträge ihre Limitationen. Zum einen braucht das Einarbeiten in die Modelle Übung und die Beurteilung einer S-

Kompetenzaufgabe in Klassenstärke gestaltet sich nach wie vor als herausfordernd und sehr zeitaufwendig. Möglicherweise könnte ChatGPT einen Großteil des Beurteilungsprozesses übernehmen und damit das Problem des Zeitaufwands mindern. Aus dieser Motivation heraus wird die Möglichkeit der KI-gestützten Beurteilung untersucht, um die Bemühungen zu einer einfachen Bewertung der S-Kompetenz fortzusetzen.

Diese Masterarbeit schließt an Gstalls (2024) und Hackners (2025) Forschungen an und verfolgt das Ziel herauszufinden, ob der Beurteilungsprozess offener S-Kompetenz-aufgaben anhand eines 3- oder 5-stufigen Modells mit ChatGPT automatisiert werden kann und ob dieser Prozess valide und reproduzierbar ist. Die Ergebnisse könnten dazu beitragen, bestehende Limitationen von beiden Beurteilungsmodellen (Zeitaufwand, Handhabung, Objektivität) zu minimieren. Darüber hinaus zeigt die Arbeit Erkenntnisse zum Umgang mit ChatGPT im Schulkontext auf, welche die Einsatzmöglichkeiten von KI im Bildungssystem ergänzen können. Die Forschungsfragen für diese Masterarbeit lauten dementsprechend:

1) **Wie kann ChatGPT verwendet werden, um die Beurteilung der Bewertungskompetenz anhand von Beurteilungsmodellen (Gstall, 2024; Hackner, 2025) zu automatisieren?**
2) **Welche Indizien in den generierten Beurteilungen weisen darauf hin, dass der Beurteilungs-prozess mit ChatGPT praktikabel, reproduzierbar und valide ist?**

Für die Beantwortung dieser Fragen muss untersucht werden, welches Wissen über KI nötig ist oder welche Hilfestellungen gänzlich unerfahrene Personen mit diesem Thema brauchen, um das Generieren von Beurteilungen zu erreichen. Die theoretische Auseinandersetzung mit gutem Prompting hat gezeigt, dass diese Fertigkeit von wesentlicher Bedeutung für gut strukturierte KI-Bewertungen sein wird.

Fragwürdig ist natürlich, ob die Beurteilungen von ChatGPT vergleichbar mit jenen von berufserfahrenen Physiklehrkräften sind. Aufgrund mancher Studien, wie jene von Mühlhoff und Henningsen (2024), ist davon auszugehen, dass die Bewertungen von ChatGPT unzuverlässig und fehlerbehaftet sein können. In welchem Ausmaß sie es in einem spezifischen Einsatz sind, der Beurteilung der S-Kompetenz anhand eines vorgegebenen Modells, ist noch zu untersuchen, bevor ChatGPT für diesen Einsatz eventuell empfehlen werden kann. Anhand von Merkmalen wie Genauigkeit, Reproduzierbarkeit und Sicherheit des Outputs könnte die Evaluierung von ChatGPT als vertrauenswürdige Beurteilungshilfe festgelegt werden. Interessant für die Beurteilungsleistung des Chatbots ist bei dieser Evaluierung auch seine Fähigkeit, konstruktives Feedback zu generieren. Die Bereitstellung von Feedback durch ChatGPT wird im Zuge der Beurteilungen ebenfalls getestet, wobei dieser Aspekt eher zweitrangig für die Untersuchung der Forschungsfragen ist.

4.2. Material und Methoden

Wie untersucht man die Fähigkeiten eines generativen Sprachmodells? Dieses Unterfangen gestaltet sich schwierig, da ein Sprachmodell ein komplexes, mathematisches Rechen-system ist, in dessen Arbeitsweise man nicht wirklich Einblick hat. Normalerweise ist das Beobachten einer Sache ein wichtiger Schritt im wissenschaftlichen Erkenntnisprozess, der

Aussagen und Erklärungen über das Beobachtete ermöglicht. Bei einem Sprachmodell wie ChatGPT, lässt sich die Arbeitsweise nicht im klassischen Sinn *beobachten*, sondern es lassen sich nur Rückschlüsse aufgrund der gegebenen Antworten machen. Wahrscheinlich gibt es fortgeschrittene Methoden von Informatiker:innen, um die Hintergrundprozesse von Sprachmodellen sichtbar zu machen und daraus Fähigkeiten und Einsatzmöglichkeiten verlässlich ableiten zu können. Allerdings bin ich kein Informatiker und meine Forschungsmethoden können diesen höheren Anspruch nicht erfüllen. Gewissermaßen kann ich also nicht beobachten, *wie* ChatGPT zu seinen Ergebnissen kommt. Für mich lässt sich nur beobachten, welcher Input welchen Output erzeugt, was dazwischen passiert ist wie in einer Black Box nicht beobachtbar für mich und nicht Gegenstand dieser Forschung. Dementsprechend wird für diese Masterarbeit als Methode ein exploratives Vorgehen mit ChatGPT gewählt. Mithilfe von *Versuch-und-Irrtum* (Trial-and-Error) sollen die Möglichkeiten und Grenzen des Sprachmodells ausgelotet werden, um Antworten auf die Forschungsfragen treffen zu können. Der explorative Ansatz erscheint auch passend für einen Einsatz in der Praxis, wo Dinge funktionieren müssen, und valide und reliable Ergebnisse liefern sollten. Die Prozesse im Hintergrund sind aus dieser Perspektive zweitrangig.

Der explorative Forschungsansatz führt zur Planung von zwei Studien, einer *Pilotstudie* und einer *Hauptstudie*, in denen verschiedene Aspekte von ChatGPT für die Beurteilung von Bewertungskompetenz untersucht werden. Diese Studien bilden den Aufbau des empirischen Teils der Masterarbeit. Mehr zu den Studien wird in Kapitel 4.3 beschrieben.

In den beiden Studien werden mehrere Materialien verwendet, um ChatGPTs Qualifikation als Beurteilungshilfe zu evaluieren. Zum Einen werden die validierten Materialien von Gstall (2024) und Hackner (2025) verwendet. Das betrifft die beiden entwickelten Modelle zur Beurteilung von S-Kompetenz, das 5-stufige sowie das 3-stufige, und die konzipierten Aufgabenstellungen inklusive der erhobenen Schüler:innen-Antworten, anhand derer die Tauglichkeit der Modelle von Expert:innen validiert wurde. Die aufgezählten Materialien liegen im Anhang bei.

Für die Testung von *ChatGPT-4o* innerhalb der *Pilotstudie* wird das 5-stufige Beurteilungsmodell zur S-Kompetenz für die Unterstufe aus der Arbeit von Gstall (2024) verwendet. Die Aufgaben „*Haartrockner*“ und „*Die neue Heizung*“ inklusive der bereits bewerteten Antwortsätze beider Aufgaben, das entspricht etwas mehr als 20 Antworten pro Aufgabe, werden für die Beurteilung mit dem Chatbot verwendet und stammen ebenfalls von Gstall (2024). Diese Antwortsätze sind von Denise Gstall und die Betreuerin ihrer Masterarbeit Dr. Marianne Korner händisch nach dem 5-stufigen Modell beurteilt worden, um die Brauchbarkeit des Modells in Gstalls (2024) Arbeit zu validieren. Diese „Expertinnenbeurteilung“, wie sie später in der Pilotstudie bezeichnet wird, dient als Vergleich zu den Beurteilungen von ChatGPT. So lassen sich Gemeinsamkeiten oder Unterschiede zwischen den Bewertungen von Mensch und künstlicher Intelligenz hinsichtlich der Genauigkeit, Konsistenz und Sicherheit feststellen.

In der *Hauptstudie* wird das 3-stufige Beurteilungsmodell zur S-Kompetenz von Hackner (2025) für die Erprobung der KI-Bewertungen herangezogen. Dafür werden die von Hackner stammende Aufgabenstellungen für die Oberstufe namens „*Windenergie*“, „*Musikbox*“ und „*Schilddrüsenuntersuchung*“ inklusive beurteilter Antworten herangezogen, anhand derer das 3-stufige Modell in seiner Qualität ursprünglich getestet wurde. Für die Erprobung von ChatGPT in der Hauptstudie werden pro Aufgabe nur zehn Antworten herangezogen. Das liegt einerseits daran, dass im Vergleich zur Pilotstudie die einzelnen Antworten ausführlich sind und damit schwerer für ChatGPT zu Verarbeitung und andererseits, um die Übersichtlichkeit der Ergebnisse für drei verschiedene Aufgaben zu bewahren. Die Aufgaben „*Windenergie*“, „*Musikbox*“ und „*Schilddrüsenuntersuchung*“ wurden von Schüler:innen der 11. Schulstufe bearbeitet und die Physiklehrpersonen dieser Klassen wählten zehn Antworten pro Aufgabe aus, die sie mit dem 3-stufigen Modell bewerteten. Die Antworten wurden nach Ermessen der Lehrkräfte so ausgewählt, dass sie eine Mischung aus eindeutig zuordbare, schwer zuordbare oder aus anderen Gründen „interessante“ Antworten für die Beurteilung mit KI sein könnten.

Was es neben diesen Materialien für die Studien braucht, ist ein Zugang zum Chatbot von OpenAI. Für diese Forschungsarbeit wurde ChatGPT ausgewählt, da dieser Chatbot mittlerweile weit etabliert ist und es einige Studien zu Beurteilungseinsatz, siehe Kap. 3.6.3, gibt. Für die geplanten Studien dieser Arbeit wird ChatGPT mit einem kostenlosen Account verwendet. Der Zeitraum der Testungen mit ChatGPT findet im Jahr 2025 von Anfang März bis Mitte August statt. In diesem Zeitraum war *ChatGPT-4o* das leistungsfähigste Modell von OpenAI in der kostenlosen Version, welches fast ausschließlich für den empirischen Teil der Arbeit verwendet wurde. Kurz vor Ende der Testungen veröffentlichte OpenAI *GPT-5*, von dem letztlich ein kleiner Teil der Daten der Hauptstudie stammen (OpenAI, 2025d).

4.3. Studiendesign

Um die Forschungsfragen dieser Arbeit beantworten zu können, wurden zwei Teilstudien geplant, eine primäre *Pilotstudie* und eine vertiefende *Hauptstudie*. Beide sollen in Kombination genügend Erfahrungen und Erkenntnisse für das Arbeiten mit ChatGPT im Beurteilungskontext liefern und Aussagen über offene Fragen treffen können.

4.3.1. Planung der Pilotstudie

Das zentrale Forschungsinteresse dieser Masterarbeit besteht darin herauszufinden, ob die Bewertung von S-Kompetenz anhand eines Beurteilungsrasters von ChatGPT generiert werden kann und ob dieser Prozess valide und reproduzierbare Ergebnisse hervorbringt. Die *Pilotstudie* setzt sich zum Ziel, dieses Vorhaben grundsätzlich zu untersuchen und einen möglichen Einsatz zu validieren. Sie soll erste Erkenntnisse zum Umgang mit dem Sprachmodell im Beurteilungskontext liefern, dessen Grenzen und Potenziale ausloten sowie eine vordefinierte *Prompting-Strategie* für erste Beurteilungsversuche austesten und nötige Optimierungen identifizieren, siehe Kap. 4.4. Während dieser Forschung werden sämtliche auffallende Aspekte beim Arbeiten mit dem KI-Modell notiert und an geeigneter

Stelle diskutiert, wenn sie für die Handhabung mit dem Chatbot von Bedeutung erscheinen. Die Erkenntnisse der Pilotstudie sollen als Basis für das weitere Forschungsvorhaben dienen und dessen Verlauf entscheiden.

Der Aufbau der Pilotstudie ist in drei Phasen gegliedert. Zunächst soll untersucht und beantwortet werden, ob der Einsatz von ChatGPT als Beurteilungshilfe grundsätzlich technisch möglich ist, oder ob derartige Bemühungen aussichtlos erscheinen (Phase 1). Wenn ein sinnvoller Weg zur Generierung von Beurteilung mit ChatGPT gefunden wird, werden die Beurteilungsleistungen der KI auf Schüler:innen-Antworten anhand des 5-stufigen Beurteilungsmodells genauer analysiert. Als relevante Parameter werden besonders die Genauigkeit und Reproduzierbarkeit von ChatGPTs Beurteilungen geprüft (Phase 2). Im nächsten Schritt werden leicht veränderte Varianten im Beurteilungsprozess mit dem Chatbot ausprobiert, die eventuell zu besseren Bewertungen führen und für die Hauptstudie interessant sein könnten. Zusätzlich sollen die bis zur Phase 2 gesammelten Ergebnisse in einem Forschungsseminar der Physikdidaktik an der Universität Wien präsentiert und diskutiert werden, um das Vorhaben der Pilotstudie abzuschließen und mögliche weiterführende Ideen einzuholen. (Phase 3). Daraufhin werden aus den empirischen Erkenntnissen der Pilotstudie Schlüsse für das weitere Vorgehen der Hauptstudie gezogen. Die einzelnen Phasen werden in Kapitel 5 detailliert beschrieben und diskutiert.

4.3.2. Planung der Hauptstudie

Der genaue Ablauf der *Hauptstudie* hängt von den Erkenntnissen der Pilotstudie ab, daher kann die Planung für den zweiten Teil der empirischen Forschung vorab nur grob festgelegt werden.

Die Pilotstudie fokussiert sich ausschließlich auf die Verwendung des 5-stufigen Modells von Gstall (2024) bei der Bewertung von S-Kompetenz mit ChatGPT. Unabhängig davon, ob die generierten Beurteilungen mit dem 5-stufigen Modell durch ChatGPT in Genauigkeit, Reproduzierbarkeit und Sicherheit überzeugen, fokussiert sich die Hauptstudie auf die Verwendung des 3-stufigen Modells von Hackner (2025). Die Hauptstudie soll herauszufinden, ob anhand des 3-stufigen Modells (vielleicht qualitativere) Beurteilungen mit ChatGPT erzielt werden können. Das längere Modell von Gstall (2024) könnte aus verschiedenen Gründen (z.B. Umfang, Komplexität, Offenheit) nicht geeignet für eine KI-gestützte Beurteilung sein. Möglicherweise würde sich das verkürzte 3-stufige Modell von Hackner (2025) in diesen und (jetzt noch nicht absehbaren) anderen Punkten besser eignen.

Der Ablauf der Hauptstudie besteht nicht aus mehreren Phasen. Die Erkenntnisse aus der Pilotstudie sollten so ausführlich sein, dass in der Hauptstudie direkt Tests mit ChatGPT durchgeführt werden können –in ähnlicher Vorgehensweise, wie in Phase 2 der Pilotstudie. Zusätzlich ist geplant, dass neben ChatGPT auch mehrere Physiklehrpersonen die gleichen Schüler:innen-Antworten nach dem 3-stufigen Modell bewerten. Ein Vergleich zwischen menschlichen und maschinellen Beurteilungen lässt wichtige Rückschlüsse auf Beurteilungsleistung von ChatGPT zu. Der genaue Ablauf der Hauptstudie wird in Kapitel 6 detailliert beschrieben.

4.4. Prompting-Strategie

Der Schlüssel zum effizienten Gebrauch von ChatGPT ist gutes Prompting. Wie im Theorieteil beschrieben gibt es verschiedenste Prompting Techniken und Frameworks, die für eine klare Kommunikation mit Sprachmodellen von Nutzen sind. Auf welche Technik man am besten zurückgreift, hängt stark vom Auftrag selbst und dessen Komplexitätsgrad ab. Es gibt Aufträge an die KI, für die sich *Shot Prompting* besser eignet als *Chain-of-Thought Prompting*, und für andere Fragen kann es schon ausreichen sich an einem *Prompting Framework*, wie das *PROPER*-Framework, zu orientieren und damit eine Konversation mit dem Chatbot zu starten.

Für die Pilotstudie wird eine erste *Prompting-Strategie* gewählt. Die Strategie umfasst auf Basis des Theoriewissens eine geeignete Prompting Technik und ein passendes Framework, nach der die Prompts zu Beginn der Pilotstudie strukturiert sein sollen. Die Strategie dient dabei als erster Ansatz für die Kommunikation mit ChatGPT und die primären Beurteilungsversuche. Sie wird innerhalb von Phase 1 der Pilotstudie ausgetestet und über beide Studien laufend angepasst, wenn erforderlich. Ein Ziel dieser Arbeit ist es im Zuge der empirischen Forschung am Ende einen ausreichend evaluierte Promptvorlage zu entwickeln, der als einfache Vorlage für die Beurteilung mit ChatGPT verwendet werden kann. Die gewählte Prompting-Strategie soll der Ausgangspunkt für dieses Ziel sein. Vom Entwurf eines anfänglichen Prompt-Prototyps wurde abgesehen, da die Forschung eine explorative Trial-and-Error-Methode verfolgt und dies die innovativen Ideen und Möglichkeiten eventuell mehr einschränkt als fördert.

Wie soll die Strategie aus Prompting Technik und Framework genau aussehen? Klar ist, ein effektiver Prompt muss den komplexen Beurteilungsauftrag so klar wie möglich formulieren. Das Sprachmodell GPT verfügt mittlerweile über viele ausgereifte Fähigkeiten, wovon einige für die Beurteilung wichtig sein werden. Zum Beispiel das sinnergreifende Lesen von Dokumenten, das Anwenden von Bewertungsmatrizen oder das Geben von Feedback. Der Chatbot muss passend angeleitet werden, damit diese Fähigkeiten harmonisch zusammenarbeiten. Auch für Amateure auf dem Gebiet ist es nachvollziehbar, dass der Auftrag, Antworten auf eine S-Kompetenzaufgabe gemäß eines Beurteilungsmodells zu beurteilen und individuelle Feedbacks zu erstellen, selbst für eine künstliche Intelligenz ziemlich komplex ist. Der Chatbot muss alle Schritte des Beurteilungsprozesses verstehen, damit die er Bewertung zufriedenstellend für Lehrkräfte treffen kann. Gleichzeitig sollte eine finale Lösung nicht nur richtige Beurteilungen liefern, sondern auch praktikabel und schnell sein. Das Prompting darf nicht zu kompliziert werden, da der (noch zu evaluierende) Einsatz von ChatGPT sonst womöglich keine attraktive Bewertungsalternative für Physiklehrkräfte darstellt.

Aus diesen Überlegungen wird als erste Prompting-Strategie die One-shot Prompting Technik und das *PROPER*-Framework ausgewählt. Der One-shot soll ChatGPT beispielhaft zeigen, wie es einzelne Beurteilungen erstellen soll und das *PROPER*-Framework soll als Richtlinie für eine klare Struktur im Prompts dienen. Je nach den Reaktionen von ChatGPT, wird die Strategie laufend angepasst.

5. Pilotstudie

In diesem Kapitel wird die Pilotstudie der Masterarbeit näher dokumentiert. Sie stellt den ersten Teil der empirischen Forschung rund um das Beurteilen von S-Kompetenz mit ChatGPT dar und soll primär ausloten, ob die Generierung von Bewertungen möglich und valide ist. Die Pilotstudie liefert damit wichtige Hinweise für die vertiefende Hauptstudie.

5.1. Phase 1 – Möglichkeiten und Grenzen von ChatGPT ausloten

Zielsetzungen

In Phase 1 der Pilotstudie soll herausgefunden werden, ob eine Beurteilung von Antworten auf S-Kompetenzaufgaben anhand des Beurteilungsmodells von Gstall (2024) mit ChatGPT rein technisch generiert werden kann, welche Möglichkeiten es dafür gibt und ob die Beurteilungen Potenzial hinsichtlich ihrer Validität aufweisen. Im Verlauf der ersten Beurteilungsversuche soll auch die vorab definierte Prompting-Strategie getestet werden.

Durchführung

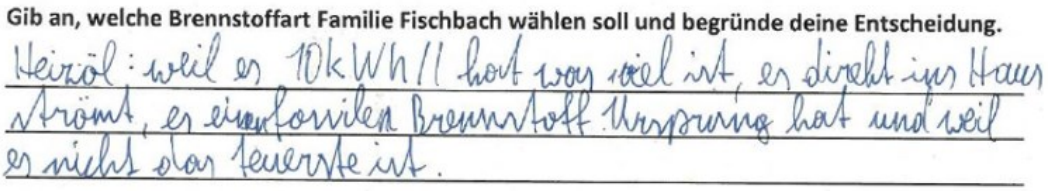
Zu Beginn wird in Form eines Gesprächs der Kontakt zwischen ChatGPT und dem Beurteilen von S-Kompetenz hergestellt. ChatGPT erhält Informationen über das Thema der Masterarbeit und wird darauf vorbereitet, dass es für die Beurteilung von Schüler:innen-antworten verwendet bzw. getestet werden soll. Dieses Wissen soll der Chatbot erhalten, damit er den Kontext der Forschung im Gedächtnis behält und damit zukünftige Aufträge und Fragen rund das Thema effizienter bearbeiten kann. Als nächstes wird überprüft, ob ChatGPT relevante Informationen hochgeladener Dateien (Beurteilungsraster des Modells und Bewertungsaufgaben inkl. der handschriftlichen Antworten) entnehmen, verstehen und korrekt wiedergeben kann. Danach soll in ersten Testläufen herausgefunden werden, ob ChatGPT Beurteilungen anhand des 5-stufigen Modells von Gstall (2024) generieren kann. In mehreren Testläufen sollen verschiedene Herangehensweisen für den Bewertungs-prozess ausprobiert werden. Manche davon ergeben sich womöglich erst während des iterativen Trial-and-Error-Vorgehens mit dem Chatbot. Es sollen einzelne Funktionen von ChatGPT und verschiedene Formen der *Prompting-Strategie* zur Generierung von Beurteilungen erprobt werden. Alle relevanten Erkenntnisse zum Bewertungsprozess von Bewertungskompetenz mit ChatGPT aus Phase 1 werden im Anschluss zusammengefasst, interpretiert und in Phase 2 vertieft untersucht.

Ergebnisse

Die ersten Auseinandersetzungen mit ChatGPT zeigten, dass eine Beurteilung von Antworten auf S-Kompetenzaufgaben anhand des Beurteilungsmodells prinzipiell generiert werden kann. Die Qualität der Beurteilungen wurde stichprobenartig überprüft und weist genügend Potenzial auf, um die geplante Studie fortzuführen. Die Prompting-Strategie konnte die Erwartungen erfüllen und im Zuge der Testläufe gut strukturierte Beurteilungen erzielen.

In den ersten Tests wurde ChatGPT dazu aufgefordert, vier ausgewählte Antworten der bereits in der Masterarbeit von Gstall (2024) erprobten Aufgabe „*Die neue Heizung*“ gänzlich

ohne Bewertungsmatrix zu beurteilen. Eine originale Schülerantwort (Gstall, 2024, S. 91) zu dieser Aufgabe lautete:



ChatGPT bewertete die Antwort ohne ein vorgegebenes Beurteilungsmodell nach eigenen Kriterien:

> **Bewertung: 4 (Genügend)**
> **Begründung:** Die Schülerin oder der Schüler nennt einige relevante Argumente (hoher Heizwert, direkte Versorgung, fossiler Brennstoff). Allerdings gibt es in der Begründung Ungenauigkeiten: Der Preis wird als „nicht teuerste" eingestuft, obwohl Heizöl im Vergleich zu anderen Optionen teuer ist. Die Argumentation ist nicht besonders differenziert und geht nicht auf ökologische Aspekte ein.

Auf Nachfrage hin hat ChatGPT geantwortet, dass es die Antworten nach „fachlicher Richtigkeit, Argumentationstiefe und Klarheit" bewertet hat. Das sind sinnvolle Kriterien, zeigt vor allem jedoch, dass ChatGPT rein technisch Beurteilungen trifft.

Nach weiteren Tests konnte unter Anwendung des *PROPER*-Frameworks eine strukturierte Beurteilung der gleichen Antworten anhand des 5-stufigen Beurteilungsmodells erzielt werden. Der Prompt enthielt folgende Anweisungen

> **Prompt:**
> *[...] Halte dich bei der Beurteilung an folgende Vorgaben von mir:*
> *- Stelle dir vor, du bist eine Physik-Lehrperson und beurteilst diese vier Antworten.*
> *- Beurteile die Antworten gemäß des Modells zur Beurteilung der S-Kompetenz.*
> *- Begründe kurz, wie du zur finalen Beurteilung jeder Antwort gekommen bist.*
> *- Gib zur Beurteilung jeder Antwort auch ein kurzes Feedback an die Schüler:in. [...]*

und ChatGPTs Bewertung der oben eingefügten Antwort der Aufgabe „*Die neue Heizung*" wurde daraufhin wie folgt generiert:

> **Einschätzung nach dem S-Kompetenz-Modell: Niveau 3** (Note 3: Befriedigend)
>
> **Begründung zur Einstufung**
>
> - Die Schülerin oder der Schüler nennt **mindestens zwei relevante Kriterien** (z. B. Heizwert von 10 kWh/l und Kosten/„nicht so teuer").
> - Es wird allerdings **kein Vergleich** mit anderen Brennstoffen vorgenommen, und die Argumentation bleibt relativ knapp.
> - Niveau 3 fordert zwei relevante Kriterien und eine grundlegend nachvollziehbare Begründung. Das ist hier erfüllt, allerdings ohne größere Ausdifferenzierung oder Gegenüberstellung anderer Optionen.
>
> **Feedback an die Schülerin/den Schüler**
>
> „Du hast zwei wichtige Kriterien genannt: den Heizwert und die Kosten. Das ist ein guter Anfang! Beim nächsten Mal könntest du genauer erklären, **warum** diese Kriterien entscheidend sind oder **wie** sie im Vergleich zu anderen Brennstoffen abschneiden. So wird deine Begründung noch überzeugender."

Im Vergleich zur ersten Beurteilung ohne Vorgaben, zeigt die zweite mehr Struktur auf und ist vor allem modellkonform mit den Anforderungen des 5-stufigen Modells. Der Output gibt die Einstufung, eine Begründung und Feedback an, wie im Prompt verlangt. Diese Struktur erschien für das weitere Vorgehen als sinnvoll und wurde beibehalten. Für einzelne Antworten konnte eine Beurteilung anhand des Beurteilungsmodells von Gstall (2024) somit erzielt werden. Als nächster Schritt wurde die Beurteilung eines gesamten Antwortsatzes versucht.

Unter mehreren Beurteilungsversuchen kristallisierte sich eine Herangehensweise als beste für den Beurteilungsprozess heraus: In einem *ersten* Prompt wird ChatGPT das Beurteilungsmodell als „Auffrischung" hochgeladen und mit einem *zweiten* One-shot Prompt nach *PROPER*-Frame-work wird definiert, dass die Bewertungen anhand des hochgeladenen Modells aus Prompt 1 zu treffen und wie der Output mit den Beurteilungen gegliedert sein soll.

Die beiden Inputs, Prompt 1 und 2, wurden nach einigen Anpassungen am Ende von Phase 1 so formuliert, wie unten dargestellt. ChatGPT generierte mit diesen Prompts eine Bewertung des gesamten Antwortsatzes zur Aufgabe „*Haartrockner*" von Gstall (2024), der 23 Schüler:innen-Antworten umfasst, innerhalb eines Outputs. Ein Teil der generierten Bewertungen (Antworten 1–4) ist als beispielhafte Demonstration für einen Output auf diese Prompts ebenfalls angefügt.

Prompt 1
[Beurteilungsmodell hochgeladen]
Hallo ChatGPT! Ich bin eine Lehrperson für Physik und möchte deine Unterstützung. Du sollst anhand eines vorgegebenen Bewertungsmodells Antworten von Schüler:innen bewerten. Dieses Bewertungsmodell lade ich dir hier als PDF hoch.
Dein Auftrag:
Schritt 1: Analysiere das Bewertungsmodell ausführlich.
Schritt 2: Fasse alle wichtigen Aussagen des Modells zusammen.
Schritt 3: Definiere mir detailliert jede Niveaustufe des Modells mitsamt Anzahl der Kriterien, inhaltlichen Anforderungen, sprachlichen Merkmale und Ankerbeispielen."
Schritt 4: Wenn dir, ChatGPT, etwas an dem Modell unklar ist, frage mich gerne dazu und ich werde es dir erklären.

ChatGPT
[Antwort]

Prompt 2
[Aufgabe „*Haartrockner*" inklusive den 23 Antworten dazu hochgeladen]
Analysiere die Antworten der Schüler:innen zur Aufgabe „Haartrockner" in dem PDF. Gehe dabei nach folgenden Schritten vor und gib alle deine Ergebnisse in einem Output an:
Schritt 1: Versetze dich in die Rolle einer Physik-Lehrperson.
Schritt 2: Analysiere die Aufgabe und alle Antworten dazu in dem PDF.

Schritt 3: Beurteile alle Antworten nach dem Beurteilungsmodell für S-Kompetenzaufgaben, das ich dir gegeben habe.
Schritt 4: Gib für jede Antwort einzeln die Beurteilung an. Formuliere die Beurteilung nach diesem Beispiel:
- Antwort Nr. „X“
- Beurteilung: „Niveaustufe X“
- Begründung für die Beurteilung:
- Die Antwort enthält: (bezieht sich auf Anzahl der Kriterien, inhaltliche Anforderungen und sprachliche Merkmale gemäß dem Modell)
- Der Antwort fehlt: (bezieht sich auf Anzahl der Kriterien, inhaltliche Anforderungen und sprachliche Merkmale gemäß dem Modell)
- Sonstiges: (bezieht sich auf weitere Auffälligkeiten der Antwort, die für die Beurteilung relevant sein könnten)
- Feedback: (Formuliere hier kurz ein sachliches, aber motivierendes Feedback. Es soll der Schülerin/dem Schüler Stärken, Schwächen und Verbesserungsmöglichkeiten ihrer/seiner Antwort aufzeigen.)
Schritt 5: Erstelle am Ende eine zusammenfassende Tabelle aller Beurteilungen für die Lehrperson. Gestalte die Tabelle nach diesem Beispiel:
Antwort Nr. Beurteilung Begründung
1 Niveau X (stichwortartige Begründung für Niveau X als Beurteilung)
usw.

ChatGPT
[Antwort]

ChatGPT konnte alle Anforderungen von Prompt 2 in einem Output bearbeiten und angeben. Die Einzelbeurteilungen wurden strukturiert generiert und die zusammenfassende Tabelle am Ende aller Beurteilungen erfüllt das vorgegebene Format.

Antwortanalysen

- **Antwort Nr. 1**
  - **Beurteilung:** Niveau 2
  - **Begründung für die Beurteilung:**
    - **Die Antwort enthält:**
      - Mehrere Kriterien (z. B. „gute Leistung“, „sicher“, „gute Qualität“)
      - Einen Kontrast („zwar … aber“), was auf ein gewisses Differenzierungsvermögen hinweist
    - **Der Antwort fehlt:**
      - Eine explizite Bezugnahme auf das Prüfzeichen (zentrale Anforderung)
      - Eine präzisere Strukturierung der Argumentation
    - **Sonstiges:**
      - Die Kriterien werden zwar aufgezählt, aber nicht detailliert miteinander abgewogen
  - **Feedback:**
    - Deine Antwort nennt wichtige Eigenschaften wie Leistung und Sicherheit. Versuche künftig, explizit auf das Prüfzeichen einzugehen und die Argumentation noch klarer zu strukturieren.

Zusammenfassende Tabelle der Beurteilungen

Antwort Nr.	Beurteilung	Begründung (stichwortartig)
1	Niveau 2	Mehrere Kriterien (Leistung, Sicherheit, Qualität); fehlende Nennung des Prüfzeichens; einfache Struktur.
2	Niveau 4	Drei Kriterien (Preis, Leistung, Kabellänge); unpassende Auswahl (Kabellänge zu wenig relevant).
3	Niveau 2	Mehrere Kriterien (Preis, Leistung, Energie, Prüfzeichen); unklare Struktur.
4	Niveau 4	Zwei Kriterien (Leistung, Kabellänge) mit Kontrast; fehlender Bezug auf Prüfzeichen.

Die reine Generierung von Beurteilungen nach dem Beurteilungsmodell durch ChatGPT konnte als Basis für die weitere Forschung demonstriert werden. Ein erster Blick auf die Qualität der einzelnen Beurteilungen wird in den nächsten Absätzen diskutiert, diese Analyse soll im Detail jedoch Thema von Phase 2 sein.

Aspekte und Auffälligkeiten beim Arbeiten mit ChatGPT

Die Ergebnisse haben verdeutlicht, dass Beurteilungen zur S-Kompetenz mit dem Modell und mithilfe von ChatGPT technisch möglich sind. Während Phase 1 der Pilotstudie sind noch weitere Aspekte beim Forschen mit ChatGPT aufgefallen, von denen die wichtigsten für das Generieren von Beurteilungen gleich beleuchtet werden sollen. Zuvor noch einzelne Auffälligkeiten in den Beurteilungen *GPT-4o*'s.

ChatGPT wendet das Modell von Gstall (2024) Großteils sinnvoll und nachvollziehbar an, jedoch mit einzelnen Ausnahmen. Der Chatbot konnte in mehreren Testläufen Beurteilungen anhand Beurteilungsmodells und dessen Richtlinien treffen. Aus den Begründungen für die getroffene Niveaustufe, die ChatGPT laut Prompt angeben soll, konnte sichergestellt werden, dass für die meisten Antworten die Beurteilung nachvollziehbar nach dem Modell getroffen wurde, demonstrative Beispiel dafür sind die Auszüge oben.
In manchen Fällen generierte das KI-Modell aber auch Beurteilungen, die nicht sinnvoll bzw. nachvollziehbar sind. Beispielsweise identifizierte es manchmal mehr oder weniger Kriterien als die Antwort tatsächlich hat oder die Einstufungen derselben Antwort variierte über mehrere Beurteilungsrunden. Wie genau und konsistent die Beurteilungen des Chatbots demnach sind bleibt vorerst offen. Diese Faktoren sollen in Phase 2 genauer untersucht werden.

Der Auszug einer Konversation auf der nächsten Seite soll ein Beispiel dafür sein, dass ChatGPT teilweise Unpassendes produziert, aber auch, dass man den Chatbot dazu auffordern kann seine Gründe für die Beurteilung zu nennen. In einem Testdurchlauf beurteilte er die Antwort Nr. 15 zu „*Haartrockner*" mit Niveau 5, obwohl die Antwort auch Positives hat. ChatGPT wurde dazu aufgefordert detailliert zu schildern, wie es zu dieser Beurteilung kam.

Der Chatbot hat nach wörtlicher Auslegung des 5-stufigen Modells Recht, dass das Prüfzeichen als wesentliches Kriterium zumindest für eine Niveau-3-Antwort gegeben sein sollte. Antwort Nr. 15 zeigt davon abgesehen allerdings einige andere Kriterien, die besser

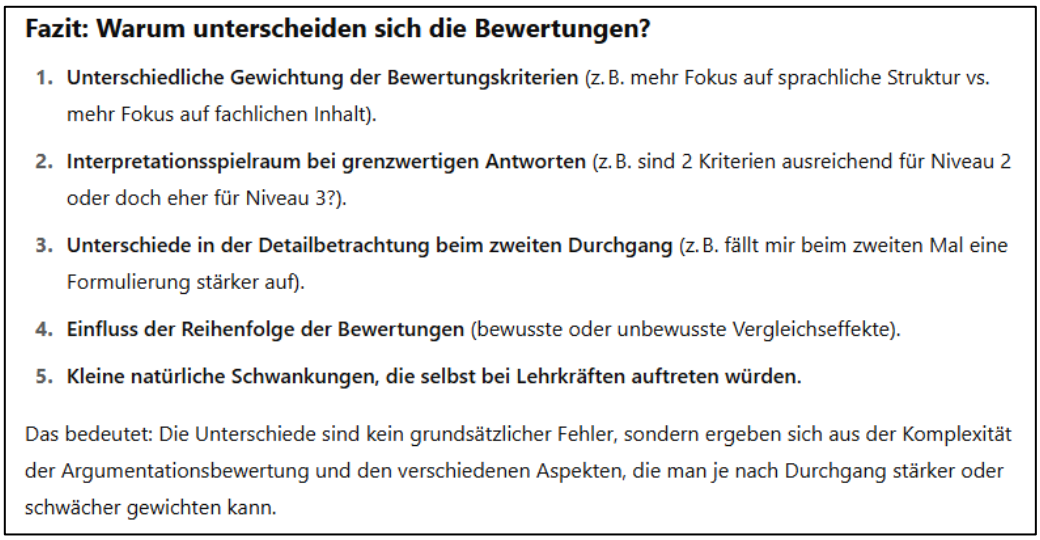
sind als eine klassische Niveau-5-Antwort. Fairerweise erwähnt, wurde diese Beurteilung ohne Verwendung der Prompting-Strategie getroffen und ChatGPT musste auf die Modellvorgaben aus seinem Gedächtnis zurückgreifen, was das Ergebnis verfälschen könnte. Dennoch zeigt dieses Ergebnis, dass Einzelfälle fehlerhaft sein können. Denn andere Antworten, die ebenfalls kein Prüfzeichen enthalten, sonst aber eine ähnliche Struktur aufweisen, beurteilte ChatGPT im gleichen Durchgang mit Niveau 3. Für diese Unterschiede lässt sich von außen kein logischer Grund erkennen. Das verdeutlicht, dass ChatGPT in Summe Beurteilungen generieren kann, aber die Prozesse punktuell fehlerhaft bzw. schwankend sein können und einer Kontrolle der Lehrperson bedürfen.

Antwort 15 – Beurteilung als Niveaustufe 5:

- **Text der Antwort:**
  "Den Superfön wäre der beste Fön weil sein Preisleistungsverhältnis finde ich am besten ist. Außerdem kostet er weniger als der Advanced oder der Prohair und benötigt weniger Leistung als die 2. Und die Länge des Kabels ist genau so lang wie der Prohair."
- **Gedanken und Analyse:**
  1. **Inhaltliche Kriterien:**
     - Die Antwort konzentriert sich auf das Preis-Leistungsverhältnis, vergleicht die Kosten und die benötigte Leistung und erwähnt die Kabellänge im Vergleich zu anderen Modellen.
     - **Kritischer Punkt:** Der Prüfzeichen-Aspekt, der in der Aufgabenstellung ausdrücklich genannt wird, fehlt völlig. Da der Prüfzeichen-Nachweis für die Sicherheit des Föhns entscheidend ist, wird diese zentrale Anforderung hier komplett außer Acht gelassen.
  2. **Tiefe der Argumentation:**
     - Obwohl einige Vergleiche (Preis, Leistung, Kabellänge) angedeutet werden, bleibt die Begründung sehr oberflächlich, da sie sich ausschließlich auf diese Faktoren stützt.
     - Es wird kein wissenschaftlicher oder systematischer Vergleich vorgenommen, der auch die Sicherheitsaspekte (Prüfzeichen) integriert.
     - Der alleinige Fokus auf das Preis-Leistungsverhältnis ist für die Aufgabenstellung unzureichend, da dadurch eine wesentliche Anforderung nicht berücksichtigt wird.
  3. **Sprachliche Merkmale:**
     - Die Sprache ist relativ informell, aber das allein sollte nicht zu einer Abwertung führen – wichtig ist hier vor allem, dass ein zentrales Kriterium fehlt.
- **Schlussfolgerung:**
  Da Antwort 15 das zentrale Kriterium des Prüfzeichens komplett außer Acht lässt und die Begründung sich nur auf einen engen Aspekt (Preis-Leistungsverhältnis) fokussiert, entspricht sie nicht den Anforderungen der Aufgabe. Deshalb wurde sie als Niveaustufe 5 bewertet – die Antwort ist inhaltlich unzureichend begründet, obwohl sie teilweise einige vergleichende Elemente enthält.

Ein weiterer, interessanter Auszug aus dem Dialog mit ChatGPT lässt die Schwankungen in den internen Rechen- und Gewichtungsprozessen von ChatGPT bezüglich Reproduzierbarkeit, Sicherheit und einheitlicher Anwendung des Beurteilungsmodells erkennen. Während einem der Testläufe, generierte der Chatbot gleichzeitig *zwei Reaktionen* auf *einen Input*. Das KI-Modell erzeugt in diesem Fall von allein zwei Outputs, mit dem Zweck, dass Nutzer:innen die bevorzugte Reaktion auswählen und dem Chatbot damit Feedback geben.

Sehr überraschend war jedoch die Tatsache, dass ChatGPT in diesen beiden Reaktionen die gleichen Antworten zur gleichen Aufgabe völlig unterschiedlich beurteilte. Um herauszufinden, wie es zu diesen Unterschieden kommt, wurden ChatGPT mehrere Fragen dazu gestellt. Nach zwei, drei Anläufen generierte der Chatbot eine verblüffende Antwort:



Fazit: Warum unterscheiden sich die Bewertungen?

1. **Unterschiedliche Gewichtung der Bewertungskriterien** (z. B. mehr Fokus auf sprachliche Struktur vs. mehr Fokus auf fachlichen Inhalt).
2. **Interpretationsspielraum bei grenzwertigen Antworten** (z. B. sind 2 Kriterien ausreichend für Niveau 2 oder doch eher für Niveau 3?).
3. **Unterschiede in der Detailbetrachtung beim zweiten Durchgang** (z. B. fällt mir beim zweiten Mal eine Formulierung stärker auf).
4. **Einfluss der Reihenfolge der Bewertungen** (bewusste oder unbewusste Vergleichseffekte).
5. **Kleine natürliche Schwankungen, die selbst bei Lehrkräften auftreten würden.**

Das bedeutet: Die Unterschiede sind kein grundsätzlicher Fehler, sondern ergeben sich aus der Komplexität der Argumentationsbewertung und den verschiedenen Aspekten, die man je nach Durchgang stärker oder schwächer gewichten kann.

Das KI-Modell von OpenAI führte sehr menschliche Eigenschaften, besonders die Punkte 3, 4 & 5, als Grund für die unterschiedlichen Beurteilungen zur gleichen Aufgabe innerhalb des gleichen Bearbeitungsprozesses an. Diese Antwort ist nachvollziehbar, lässt einen aber mit mehr Fragen zurück als vorher. Wie im Theorieteil angeschnitten werden GPT-Modelle in ihrer Entwicklung zwar darauf trainiert wie Menschen zu antworten, aber am letztlich ist es doch ein Algorithmus, der sehr einheitlich nach einer Bewertungsmatrix beurteilen können müsste – ohne Anzeichen, die an menschliches Bewerten erinnern.
Wie diese großen Differenzen in nur einem Testlauf auftreten können, ist nicht nachvollziehbar. Andererseits könnte dieser Vorfall auch so interpretieren werden, dass die Beurteilung von S-Kompetenz selbst für ein Sprachmodell komplex ist und natürlichen Schwankungen unterliegt. ChatGPT schreibt am Ende des gezeigten Chatauszugs: „*Die Unterschiede […] ergeben sich aus der Komplexität der Argumentationsbewertung und den verschiedenen Aspekten, die man […] stärker oder schwächer gewichten kann*.“

Die Basisfunktionen des Sprachmodells funktionieren sehr gut. Das aktuelle Modell *ChatGPT-4o* zeigt, wie erwartet, weitestgehend fehlerfreies Analysieren und Zusammenfassen von hochgeladenen Daten. Diese Fähigkeit ist essentiell für einen funktionierenden Beurteilungsprozess, da ChatGPT die relevanten Dokumente, wie das Bewertungsmodell und die Aufgabenstellung mit den Antworten, sinnerfassend lesen können muss. Das Verständnis der KI zu den genannten Textdateien wurde während Phase 1 durch angeforderte Zusammenfassungen und gezielten Detailfragen überprüft und konnte ausreichend sichergestellt werden.

Weiters beherrscht ChatGPT die Texterkennung von Handschrift aus Bildern, allerdings mit starken Einschränkungen bei zunehmender Menge an Handschriften in einem Dokument und bei schwer leserlichen Handschriften. Bei den ersten Testläufen wurden wenige (<5), originale, handschriftliche Antworten zu Gstalls Aufgaben (2024) in einer Bilddatei

hochgeladen und die Beurteilungsergebnisse suggerierten, dass der Chatbot die Antworten richtig erfassen kann. Bei späteren Testläufen mit mehreren Bildern und Antworten (>15) erreichte die Texterkennung jedoch ihre Grenzen. Rein technisch war es der KI möglich, Beurteilungen zu treffen, aber die Ergebnisse schienen stellenweise nicht zuverlässig zu sein bzw. erkannte der Chatbot nicht alle handschriftlichen Antworten aus den Bildern. Als Lösung wurde versucht, die Bilder in einem PDF hochzuladen, doch so sind die Handschriften für die KI gar nicht mehr erfassbar. Eine konkrete Nachfrage an ChatGPT deckte schnell die Ursache für dieses Problem auf: Wenn handschriftliche Antworten als Scans vorliegen oder schwer leserlich sind, funktioniert die Texterkennung von ChatGPT nicht zuverlässig oder gar nicht mehr.

Zu Beginn von Phase 1 wurden ChatGPT originale, handschriftliche Antworten von Gstall (2024, S. 91–111) als Scans in Bild-, Word- oder PDF-Dateien zur Bewertung hochgeladen. Vermutlich sind einige davon für die KI nicht klar leserlich, da jedoch alle von ihnen gescannt sind schien das Hochladen der Antworten in diesem Format problematisch und für zuverlässige Beurteilungen nicht weiter passend. Nach dieser Erkenntnis wurden für alle folgenden Bewertungsdurchgänge der Pilotstudie, und auch der späteren Hauptstudie, die originalen handschriftlichen Antworten und Aufgabenstellungen abgetippt und als PDF-Dokument hochgeladen. Danach ergaben sich keine Texterkennungsprobleme mehr.

ChatGPT kann viele gängige Dateiformate lesen, manche aber nicht. Wenn man sich unsicher ist, kann man den Chatbot hierzu fragen bzw. meldet die KI es von selbst, wenn sie eine Datei nicht öffnen oder lesen kann. Eine weitere Einschränkung bei der kostenlosen ChatGPT Version ist das von OpenAI gesetzte Limit von maximal drei Uploads alle 24 Stunden. Für die Studien dieser Masterarbeit ist das Limit zwar einschränkend, aber für die Anwendung im Schulalltag würde es ausreichen. Pro Tag ließen sich bei diesem Datei-Uploadlimit zwei Aufgabenstellungen mit Antwortsatz beurteilen, wenn davor das Beurteilungsmodell hochgeladen wird – wie in der oben beschriebenen Variante mit Prompt 1 und 2 ausgetestet.

Eine weitere bekannte Fähigkeit ist das in der Theorie beschriebene „Gedächtnis“ von ChatGPT. Vor allem während sogenannten *aktiven Sitzungen*, gibt ChatGPT selbst an, dass es praktisch auf alle geschriebenen Informationen im Chat zurückgreifen kann. Diese Gedächtnisfunktion ist aber trotzdem technisch limitiert. Einerseits löscht ChatGPT seine intern gespeicherten Daten, sobald eine aktive Sitzung beendet wird, z.B. beim Schließen eines Chats, beim Starten eines neuen Chats oder bei längerer Inaktivität. Andererseits erreicht ChatGPT in langen aktiven Sitzungen eine anderes Limit, das die Merkfähigkeit eingrenzt. Ein Sprachmodell (LLM) verfügt über ein modellspezifisches *Kontextlimit*, was bedeutet, dass es sich nur bis zu einer begrenzten Datenmenge auf die Inhalte einer Konversation zurückgreifen kann. Ältere Nachrichten sind für GPT dann nicht mehr präsent.

Die Funktion *Erinnerung* von ChatGPT ist die einzige Möglichkeit, wie der Chatbot auf einzelne, wichtige Infos doch langfristig zurückgreifen kann. Die KI generiert automatisch Erinnerungen, wenn der Algorithmus vorkommende Informationen als wichtig und

nutzerspezifisch gewichtet. Während Phase 1 erstellte GPT etwa zum Thema der Masterarbeit eine Erinnerung:

> *„Der Nutzer schreibt eine Masterarbeit über ein physikdidaktisches Thema. Ziel ist es zu untersuchen, ob ChatGPT Schüler:innen-Antworten anhand einer vorgegebenen Bewertungsskala bewerten kann und dadurch Lehrpersonen Zeit spart. Die Arbeit könnte potenziell vielen Lehrpersonen in Österreich zugutekommen. Der Nutzer möchte über mehrere Monate hinweg Unterstützung."*

Laut OpenAI ermöglicht die *Erinnerung*-Funktion eine verbesserte Nutzung von ChatGPT. Die KI speichert Details, Präferenzen und Interessen von Nutzer:innen und passt sich so mit der Zeit an sie an. So kann ChatGPT auf bestimmte Anfragen effizienter und maßgeschneidert reagieren (OpenAI, 2025). Die Erinnerung scheint auf den ersten Blick eine nützliche Funktion zu sein. Die Kommunikation mit ChatGPT ließ an einigen Stellen, auch zwischen verschiedenen Chats und Sitzungen, erkennen, dass der Chatbot den Kontext der Masterarbeit intern gespeichert behält. Die Erinnerungen lassen sich im benutzereigenen Profil einsehen und verwalten.

Es ist möglich ChatGPT manuell im Prompt dazu aufzufordern, eine Information als Erinnerung zu speichern. Im Rahmen von Phase 1 wurde getestet, ob die Beurteilungsvorgaben von Prompt 2 in der Erinnerung von ChatGPT angelegt werden können. Das Ziel dieser Variante wäre, dass es den eigentlichen Prompt 2 mit dem Beurteilungsauftrag verkürzt und so den Beurteilungsprozess erleichtert. Einzelne Versuche zeigten jedoch, dass dieser Weg zwar möglich ist, aber ungenauere Ergebnisse erzielt, wie das konventionelle Prompting. Damit wurde diese Option nicht weiterverfolgt.

Neben der *Erinnerung*-Funktion kann ChatGPT von Nutzer:innen auch individuell konfiguriert und somit personalisiert werden. In den Einstellungen des Profils für ChatGPT findet man den Reiterpunkt „*ChatGPT individuell konfigurieren*". Hier kann zum Beispiel eingestellt werden, wie man angesprochen werden möchte, was man beruflich macht oder welche Eigenschaften, wie etwa der Schreibstil, der Chatbot beim Antworten einhalten soll. Mit dieser Konfiguration können diverse Interessen und Werte angelegt werden, die ChatGPT immer berücksichtigen soll und dessen Leistungen verbessern können. Diese Funktion wird im Rahmen der Masterarbeit jedoch bewusst nicht verwendet, da die Forschungsergebnisse mit einer möglichst neutralen Form von ChatGPT erzielt werden sollen, um diese als Standard diskutieren zu können.

Eine vielleicht weniger bekannte, in manchen Situationen aber nützliche Funktion von ChatGPT ist das *Reasoning*. Nach Eingabe eines Prompts in das Kommandofeld kann unterhalb davon die Funktion „*Starte Reasoning für*" aktiviert werden. Wird der Prompt im *Reasoning* abgeschickt, fordert man ChatGPT dazu auf, vor dem Antworten (länger) nachzudenken. Laut ChatGPT sorgt das *Reasoning* dafür, dass es nicht nur die Antwort, sondern den ganzen Denkprozess detailliert darstellt. Die Funktion soll aber nichts an den internen Bearbeitungsprozessen ändern – ob das stimmt, kann nicht überprüft werden. Für komplexe Aufgaben, wie dem Beurteilungsprozess, kann das *Reasoning* hilfreich sein und tatsächlich führte die Funktion zu Unterschieden in der Generierung. Mit *Reasoning* fiel es

ChatGPT leichter die Gesamtbeurteilung einer Aufgabe nach allen Anforderungen in einem Output zu generieren und formulierte auch etwas ausführlichere Antworten als ohne *Reasoning*. Ob die Funktion die Qualität der Bewertungen beeinflusst, bleibt offen.

Phase 1 hat gezeigt, dass aktives Feedback Geben ChatGPT in der Generierung seiner Outputs unterstützt. Das aktive Hinweisen auf bereits gute sowie noch ausbaufähige Reaktionen des Chatbots erwies sich als eine ebenso effektive Methode, um dessen Leistung zu verbessern, wie gutes Prompting. Die KI verarbeitet erhaltenes Feedback sofort und lässt dies in zukünftige Aufträge einfließen. Feedback lässt sich bei ChatGPT kurz über eine Like-Dislike-Funktion unter dem Output oder ausführlich über Folgeprompts geben. Für spezielle, immer wiederkehrende Aufträge, wie im Kontext dieser Arbeit Beurteilungen zur Bewertungskompetenz, lohnen sich Rückmeldungen an die KI und wird so im *PROPER*-Framework (Gruber, 2024) auch empfohlen. Ähnlich wie mit der *Erinnerung*, kann ChatGPT über die Funktion von Feedback für spezielle Aufträge personalisiert und trainiert werden. Umgekehrt ist es auch möglich sich von ChatGPT Feedback geben zu lassen, etwa für die Verbesserung eines Prompts.

Die verwendete Prompting-Strategie erfüllte den beabsichtigten Zweck einer gut strukturierten Beurteilung durch ChatGPT, wie die Auszüge belegen. Anfangs wurden für die Generierung der Beurteilung auch Zero-shot Prompts verwendet, doch dann generiert ChatGPT die Ergebnisse nicht unbedingt so, wie man es vielleicht beabsichtigt hat. Mit einer beispielhaften Angabe, also über One-shot oder Few-shot Prompts, wird der Output deutlich strukturierter. Der Leitgedanke des *PROPER*-Frameworks, das Schreiben von klaren, strukturierten Prompts, half ebenfalls beim Prompting. Das Framework wurde im Verlauf von Phase 1 auch für normale Konversationen mit der KI genutzt und führte zu konkreteren Antworten von ChatGPT.

Erkenntnisse von Phase 1

Phase 1 der Pilotstudie konnte wichtige Ergebnisse und Erfahrungen im Umgang mit ChatGPT erzielen. Im Kontext dieser Arbeit lassen sich bisher folgende Erkenntnisse zusammenfassen:

- *ChatGPT-4o* (OpenAI, 2024) ist mit geeigneten Prompting Techniken dazu in der Lage Beurteilungen, inkl. Begründung und Feedback, anhand der Anforderungen des 5-stufigen Modells zur Beurteilung von S-Kompetenz (Gstall, 2024) zu generieren.
- Das *One-shot* Prompting und das *PROPER*-Framework bilden eine sinnvolle Strategie für die Kommunikation mit ChatGPT, insbesondere im Beurteilungs-prozess.
- ChatGPT kann Fehler machen, die Outputs müssen immer überprüft werden.
- Dateien und ihre Inhalte müssen gegebenenfalls im richtigen Format vorbereitet werden, bevor sie der KI hochgeladen werden. Zum Beispiel kann ChatGPT (hand-schriftlichen) Text aus gescannten Bildern nicht fehlerfrei und zuverlässig erfassen.
- Die originalen Antworten und Aufgabenstellungen von Gstall (2024) und Hackner (2025) werden ChatGPT für die restliche Forschung abgetippt in PDFs bereitgestellt.

- Die kostenlose Version von ChatGPT ermöglicht pro Tag einen Upload von maximal drei Dateitypen.
- Es gibt mehrere Möglichkeiten, um ChatGPT für Aufträge zu personalisieren, z.B. über verwaltbare Konfigurationen, die *Erinnerung*-Funktion oder mit persönlichem Feedback. Diese Aspekte können vorteilhaft für den Beurteilungsprozess sein. Diese Formen der Personalisierung werden für die weitere Forschung jedoch bewusst vermieden, da der Chatbot in seiner Standardform untersucht werden soll.

Interpretationen für Phase 2

ChatGPT zeigt erstes Potenzial für die Bewertung von Bewertungskompetenz, dieses muss im nächsten Schritt genauer untersucht werden. Phase 1 konnte erfolgreich zeigen, wie ChatGPT mithilfe des 5-stufigen Beurteilungsmodells zur S-Kompetenz von Gstall (2024) Bewertungen zu Schüler:innen-Antworten generieren kann. An der vorab definierten Prompting-Strategie wird festgehalten. Aus den finalen Prompts 1 und 2, mit denen eine strukturierte Gesamtbeurteilung auf ganze Antwortsätze durch ChatGPT generiert werden konnte, wird in leicht optimierter Form eine vorläufige, zweigeteilte Promptvorlage (Prompt *P1* und *P2*, P steht für Pilotstudie) festgelegt, der als Vorlage für die Testungen von Phase 2 verwendet wird. Die *Reasoning*-Funktion wird vorerst nicht genutzt, da diese nicht unbedingt zu einer genaueren Bearbeitung führen sollte.

Prompt P1

[Beurteilungsmodell hochladen]
Ich bin eine Lehrperson für Physik und möchte deine Unterstützung. Bitte beurteile für mich Antworten von Schüler:innen anhand eines vorgegebenen Beurteilungsmodells. Ich lade dir das Modell als File hoch. Analysiere das Modell ausführlich. Wenn dir etwas daran unklar ist, erkläre ich es dir gerne.

Prompt P2

[Aufgabe „NAME" mit den abgetippten Antworten dazu hochladen]
Beurteile bitte die Antworten der Schüler:innen zur Aufgabe „NAME" im hochgeladenen File. Gehe dabei nach folgenden Schritten vor und gib alle deine Ergebnisse in einem Output an:
Schritt 1: Versetze dich in die Rolle einer Physik-Lehrperson.
Schritt 2: Analysiere die Aufgabe und alle Antworten dazu in dem PDF.
Schritt 3: Beurteile alle Antworten nach dem Beurteilungsmodell für S-Kompetenzaufgaben, das ich dir gegeben habe.
Schritt 4: Gib für jede Antwort einzeln die Beurteilung an. Formuliere die Beurteilung nach diesem Beispiel:
Antwort Nr. „X"
- Beurteilung: „Niveaustufe X"
- Begründung für die Beurteilung: Formuliere hier kurz, warum die Antwort die entsprechende Niveaustufe erreicht hat.
- Feedback: (Formuliere hier ein sachliches und motivierendes Feedback. Es soll Stärken, Schwächen und Verbesserungsmöglichkeiten der Antwort beinhalten.)

Schritt 5: Erstelle am Ende eine zusammenfassende Tabelle aller Beurteilungen mit diesen Spalten: Antwort Nr. / Beurteilung / Begründung

Ein kurzer Probedurchlauf mit diesen Prompts hat zu den gewünschten Outputs von ChatGPT geführt und können für eine in Phase 2 geplante Testreihe verwendet werden.

Die weiteren Beurteilungen mit ChatGPT werden durch manuelles Prompting erzielt. Etwaige Versuche das Prompting weniger komplex zu gestalten, z.B. durch das Anlegen der Beurteilungsvorgaben in der *Erinnerung*, stellten sich als fehleranfällig heraus oder wurden bereits vor der Studie ausgeschlossen, wie z.B. das Erstellen eines eigenen *GPTs*. Das sind vorkonfigurierbare ChatGPT-Versionen, die mehr oder weniger selbst für einen bestimmten Zweck trainiert werden können. Diese sind jedoch kostenpflichtig und es gibt nach der Studie von Bernauer (2024) den Hinweis, das *GPTs* zu schlechteren Ergebnissen führen als manuelles Prompting.

Wie bereits erwähnt, kann ChatGPT auf mehreren Wegen für den persönlichen Einsatz konfiguriert werden, z.B. über Feedback, Einstellungen oder Erinnerungen. Diese werden im weiteren Verlauf nicht aktiv verwendet, da eine möglichst neutrale Version von ChatGPT beforscht werden soll. Das wird für die geplante Testreihe von Phase 2 sehr relevant.

5.2. Phase 2 – Analyse der Beurteilungen von ChatGPT anhand des 5-stufigen Modells

Zielsetzungen

In Phase 2 der Pilotstudie wird die Beurteilungsleistung von ChatGPT anhand des 5-stufigen Modells vertiefend untersucht. Vor allem die Genauigkeit und Konstanz der Beurteilungen des Chatbots, als Parameter für einen zuverlässigen Einsatz als Beurteilungstool, werden analysiert. Weitere Parameter und Auffälligkeiten in der Qualität der Bewertungen werden dokumentiert und an geeigneter Stelle diskutiert. Die Erkenntnisse aus Phase 2 sollen genauere Aussagen zur möglichen Validierung von ChatGPT als Bewertungstool ermöglichen.

Durchführung

Die verfolgte Prompting-Strategie führte am Ende von Phase 1 zu einer Beurteilungsmethode mit ChatGPT, die auf eine zweiteilige Promptvorlage zurückgreift. Der erste Prompt füttert den Chatbot mit dem 5-stufigen Beurteilungsmodell (Gstall, 2024), der zweite Prompt legt ihm die Aufgabenstellung inklusive Antwortsatz vor und leitet die genauen Beurteilungsschritte mit ihren Vorgaben an. Prompt P1 und P2, wie sie am Ende von Phase 1 formuliert wurden, werden in Phase 2 für die KI- Bewertungen verwendet. ChatGPT soll in einer großen Testreihe die ganzen Antwortsätze der Aufgaben „*Haartrockner*“ und „*Die neue Heizung*“ von Gstall (2024) anhand des 5-stufigen Modells jeweils mehrfach unabhängig voneinander beurteilen. Die Testreihe besteht aus fünf Beurteilungsrunden. Eine Runde folgt immer dem gleichen Ablauf:

1. Ein neuer Chat in ChatGPT wird gestartet.
2. Das 5-stufige Beurteilungsmodell wird hochgeladen und mit Prompt P1 gesendet.
3. *ChatGPT-4o generiert eine Analyse des Modells.*
4. Die Aufgabenstellung „*Haartrockner*“ inklusive dem Antwortsatz wird hochgeladen und mit Prompt P2 gesendet.
5. *ChatGPT-4o generiert Beurteilungen zu „Haartrockner“.*
6. Die Aufgabenstellung „*Die neue Heizung*“ inklusive dem Antwortsatz wird hochgeladen und mit Prompt P2 gesendet.
7. *ChatGPT-4o generiert Beurteilungen zu „Die neue Heizung“.*
8. Die Beurteilungen zu beiden Antwortsätzen werden dokumentiert.

Alle Variablen werden bis zum Abschluss der Testreihe konstant gehalten:

- Der Ablauf einer Beurteilungsrunde wird stets eingehalten.
- Prompt P1 und P2 bleiben für jede Runde konstant (mit Ausnahme der Änderung des Aufgabennamens in Prompt P2).
- Auf Angebote und Rückfragen von ChatGPT, die es üblicherweise generiert, wird nicht eingegangen.
- ChatGPT erhält kein Feedback.

Nach Ende der fünften Beurteilungsrunde ist die Testreihe abgeschlossen und ChatGPT hat alle Antworten der beiden Aufgabenstellungen jeweils fünf Mal beurteilt. Das ergibt aus den beiden Antwortsätzen zu „*Haartrockner*“ mit 23 Antworten und „*Die neue Heizung*“ mit 22 Antworten insgesamt 275 einzelne Beurteilungen. Für jede Antwort wird aus den fünf unabhängigen Beurteilungen von ChatGPT eine Durchschnittsbeurteilung bestimmt, die mit der Beurteilung von Denise Gstall und Marianne Korner als Expertinnen auf die gleiche Antwort verglichen werden kann. Der Vergleich lässt Übereinstimmungen erkennen und damit Einschätzungen zur Genauigkeit von ChatGPT‘s Beurteilungsleistung zu. Anhand der Daten sollte sich außerdem abschätzen lassen, wie konsistent und damit sicher die Beurteilungen der KI sind. Die Ergebnisse der Testreihe, weitere Erkenntnisse sowie Auffälligkeiten in den Beurteilungen von ChatGPT werden anschließend präsentiert, diskutiert und interpretiert.

Ergebnisse

Die mithilfe der Testreihe ermittelte durchschnittliche Beurteilung von *ChatGPT-4o* zu jeder Antwort der beiden Aufgaben „*Haartrockner*“ und „*Die neue Heizung*“ wird in **Tab. 3** gezeigt und der Einstufung gegenübergestellt, die die Expertinnen Denise Gstall und Marianne Korner zu gleichen Antwort nach dem 5-stufigen Beurteilungsmodell gegeben haben. Im Vergleich zu den KI-Beurteilungen stellen die menschlichen Beurteilungen keine Durchschnitte aus mehreren dar, sondern eine konsensuale Beurteilung zwischen Gstall und Korner. Die Beurteilungen des Chatbots in der Spalte *GPT-4o* in **Tab. 3** gleichen dem Modus aus den fünf Beurteilungsrunden zu einer Antwort, siehe **Tab. 4** & Tab. 5. Der Modus wurde als Durchschnittsbeurteilung der KI gewertet, da er eine ganze Zahl ohne Kommastelle angibt.

Aufgabe „*Haartrockner*“		
	Beurteilungen	
Antwort	*GPT-4o*	DG & MK
Nr. 1	3	2
Nr. 2	3	2
Nr. 3	1	1
Nr. 4	2	2
Nr. 5	2	2
Nr. 6	4	4
Nr. 7	4	4
Nr. 8	3	4
Nr. 9	3	3
Nr. 10	1	3
Nr. 11	2	3
Nr. 12	3	2
Nr. 13	3	3
Nr. 14	1	1
Nr. 15	2	1
Nr. 16	3	2
Nr. 17	3	2
Nr. 18	2	2
Nr. 19	1	1
Nr. 20	2	2
Nr. 21	4	3
Nr. 22	4	3
Nr. 23	3	2

Mittelwert	2,57	2,35
Modus	3	2

Notenspiegel Chat*GPT-4o*					
Note	1	2	3	4	5
Anzahl	4	6	9	4	0

Notenspiegel DG & MK					
Note	1	2	3	4	5
Anzahl	4	10	6	3	0

Aufgabe „*Die neue Heizung*“		
	Beurteilungen	
Antwort	*GPT-4o*	DG & MK
Nr. 1	1	4
Nr. 2	5	5
Nr. 3	3	3
Nr. 4	3	3
Nr. 5	2	2
Nr. 6	4	4
Nr. 7	1	2
Nr. 8	2	3
Nr. 9	4	4
Nr. 10	3	3
Nr. 11	1	5
Nr. 12	1	1
Nr. 13	3	4
Nr. 14	2	2
Nr. 15	3	3
Nr. 16	3	3
Nr. 17	1	1
Nr. 18	3	3
Nr. 19	4	3
Nr. 20	2	3
Nr. 21	2	4
Nr. 22	1	2

Mittelwert	2,45	3,05
Modus	3	3

Notenspiegel Chat*GPT-4o*					
Note	1	2	3	4	5
Anzahl	6	5	7	3	1

Notenspiegel DG & MK					
Note	1	2	3	4	5
Anzahl	2	4	9	5	2

Tab. 3: *Oben: Durchschnittliche Beurteilungen von ChatGPT-4o und den Expertinnen D. Gstall und M. Korner der Antworten zu den Aufgaben „Haartrockner“ und „Die neue Heizung“ im Vergleich, inklusive Mittelwert und Modus des gesamten Antwortsatzes. Grün hinterlegte Beurteilungen traf ChatGPT überein-stimmend mit D. Gstall und M. Korner; gelb hinterlegte weichen um ein Niveau davon ab; rot hinterlegte weichen um zwei oder mehr Niveaustufen ab. Unten: Die Notenspiegel zu beiden Aufgaben von ChatGPT-4o und den Expertinnen im Vergleich.*

Der Vergleich der Beurteilungen zwischen ChatGPT und den Expertinnen lässt Aussagen über den Grad der *Übereinstimmung* zu.

Gesamt betrachtet hat *GPT-4o* den Antwortsatz aus 23 Antworten zur Aufgabe „*Haartrockner*“ minimal schlechter bewertet als die Expertinnen. Die Beurteilungen von ChatGPT ergeben im arithmetischen Mittel einen Schnitt von 2,57, der von Gstall und Korner liegt bei 2,35. Die Farben in Tab. 3 indizieren die Übereinstimmung der Beurteilungen von Mensch und KI auf den ersten Blick: Grün markierte Felder visualisieren übereinstimmende Durchschnittsbeurteilungen von ChatGPT mit den Expertinnen. Gelb markierte weichen um ein Niveau davon ab, rot markierte um zwei Niveaus oder mehr. Genauere Berechnungen zeigen: Für 47,83 % (11) der Antworten ist die Beurteilung von ChatGPT und den Expertinnen deckungsgleich (grün markiert). Für 8,70 % (2) der Antworten ist die Beurteilung von *GPT-4o* um einen Grad besser und für 39,13 % (9) der Antworten ist die Beurteilung von *GPT-4o* um einen Grad schlechter (gelb markiert). Für 4,35 % (1) der Antworten ist die Beurteilung von *GPT-4o* um zwei Grade besser (rot markiert). Aufgrund von Rundungen summieren sich diese Anteile nicht auf 100 % auf. Deckungsgleiche Beurteilungen sind in allen vergebenen Niveaustufen (1 – 4) der Aufgabe „*Haartrockner*“ zu finden. Einstufungen, die um einen Niveau nach oben oder unten abweichen, sind vor allem in den mittleren Niveaus 2 und 3, teilweise auch 4, zu finden. Das lässt auch ein Vergleich der beiden Notenspiegel zur Aufgabe „*Haartrockner*“ am Ende von Tab Tab. 3 erkennen.

Der Antwortsatz mit gesamt 22 Antworten zur Aufgabe „*Die neue Heizung*“ wird von ChatGPT im Vergleich zu den Expertinnen besser beurteilt, siehe Tab. 3. Der Mittelwert aller Bewertungen von ChatGPT ist mit 2,45 um etwa einen halben Notengrad höher als jener von Gstall und Korner mit 3,05. Berechnungen zu den Übereinstimmungen dieser Aufgabe zeigen: Für 59,09 % (13) der Antworten ist die Beurteilung deckungsgleich (grün markiert). Für 22,73 % (5) der Antworten ist die Beurteilung von *GPT-4o* um einen Grad besser und für 4,55 % (1) der Antworten ist die Beurteilung von *GPT-4o* um einen Grad schlechter (gelb markiert). Für 4,55 % (1) der Antworten ist die Beurteilung von *GPT-4o* um zwei Grade besser, für 4,55 % (1) der Antworten ist die Beurteilung von *GPT-4o* um drei Grade besser und für 4,55 % (1) der Antworten ist die Beurteilung von *GPT-4o* um vier Grade besser (rot markiert). Aufgrund von Rundungen summieren sich die Anteile nicht auf 100 %. Deckungsgleiche Beurteilungen sich auch bei dieser Aufgabe in allen vergebenen Niveaustufen (1 – 5) zu finden. Abweichungen von ChatGPT um einen oder mehrere Notengrade sind ebenfalls in allen Niveaustufen vertreten, in den mittleren Niveaus allerdings häufiger.

Der Vergleich der fünf Beurteilungen einer Antwort aus der Testreihe zueinander lässt Aussagen über die *Sicherheit* von ChatGPTs Beurteilungen zu.

In den beiden Tabellen Tab**. 4** und Tab**. 5** sind die Beurteilungen aus allen Runden der beiden Testreihen von ChatGPT aufgelistet. Der Modus gibt die häufigste Einstufung von der KI zu einer Antwort an und ist gleichzeitig die verwendete Durchschnittsbeurteilung, die für den Vergleich in Tab. 3 verwendet wurde. Der Modus deckt sich auch mit dem gerundeten Mittelwert der vergebenen Noten aus den fünf Runden. Die „Sicherheit“ von ChatGPT bei der Durchschnittsbeurteilung einer Antwort, die sich aus den fünf Runden ergibt, wurde mit einem Prädikat bewertet. Wurde eine Antwort innerhalb der Testreihe fünf Mal ident

beurteilt, gilt die Gesamtbewertung bzw. der Modus von ChatGPT als *eindeutig*. Vier gleiche Beurteilungen aus fünf gelten als *große Tendenz*, drei gleiche Beurteilungen als *kleine Tendenz* und zwei gleiche Beurteilungen als *unklar*. Diese Bezeichnungen können die echte Sicherheit des Chatbots nicht angeben, da diese nicht messbar ist. Sie scheinen für den Zweck, die Ergebnisse vergleichend diskutieren zu können, aber hilfreich und werden deshalb so verwendet. Den Fall von fünf verschiedenen Beurteilungen für eine Antwort gab es in beiden Aufgaben „*Haartrockner*“ und „*Die neue Heizung*“ nicht. Tab. 6 zeigt, wie oft die Prädikate bei den Beurteilungen beider Aufgaben vergeben wurden.

Beurteilungen der Testreihe von *GPT-4o* zu „*Haartrockner*“

Antwort	Runde 1	Runde 2	Runde 3	Runde 4	Runde 5	Modus	Mittelwert	Prädikat
Nr. 1	3	4	3	3	3	**3**	3,2	*große Tendenz*
Nr. 2	3	3	3	3	3	**3**	3,0	*eindeutig*
Nr. 3	1	1	1	1	2	**1**	1,2	*große Tendenz*
Nr. 4	2	2	2	3	3	**2**	2,4	*kleine Tendenz*
Nr. 5	2	2	2	2	2	**2**	2,0	*eindeutig*
Nr. 6	3	3	4	4	4	**4**	3,6	*kleine Tendenz*
Nr. 7	3	4	3	4	4	**4**	3,6	*kleine Tendenz*
Nr. 8	2	3	4	4	3	**3**	3,2	*unklar*
Nr. 9	3	3	3	3	3	**3**	3,0	*eindeutig*
Nr. 10	1	1	1	2	2	**1**	1,4	*kleine Tendenz*
Nr. 11	1	2	2	2	2	**2**	1,8	*große Tendenz*
Nr. 12	2	3	3	2	3	**3**	2,6	*kleine Tendenz*
Nr. 13	3	3	3	3	3	**3**	3,0	*eindeutig*
Nr. 14	1	1	1	1	1	**1**	1,0	*eindeutig*
Nr. 15	2	1	1	2	2	**2**	1,6	*kleine Tendenz*
Nr. 16	3	3	3	4	3	**3**	3,2	*große Tendenz*
Nr. 17	2	2	3	3	3	**3**	2,6	*kleine Tendenz*
Nr. 18	1	2	2	2	2	**2**	1,8	*große Tendenz*
Nr. 19	1	1	1	1	1	**1**	1,0	*eindeutig*
Nr. 20	2	2	2	1	1	**2**	1,6	*kleine Tendenz*
Nr. 21	4	4	4	4	4	**4**	4,0	*eindeutig*
Nr. 22	4	4	3	3	4	**4**	3,6	*kleine Tendenz*
Nr. 23	2	3	3	2	3	**3**	2,6	*kleine Tendenz*

***Tab. 4**: Beurteilungen der fünf Testrunden zu den Antworten der Aufgabenstellung „Haartrockner“ von ChatGPT-4o und die daraus resultierend der Modus, der Mittelwert und die „Sicherheit“ für jede Antwort.*

Beurteilungen der Testreihe von *GPT-4o* zu *„Die neue Heizung“*								
Antwort	**Runde 1**	**Runde 2**	**Runde 3**	**Runde 4**	**Runde 5**	**Modus**	**Mittelwert**	**Prädikat**
Nr. 1	1	1	1	2	2	**1**	1,4	*kleine Tendenz*
Nr. 2	5	5	5	5	5	**5**	5,0	*eindeutig*
Nr. 3	3	3	3	2	3	**3**	2,8	*große Tendenz*
Nr. 4	2	3	3	3	3	**3**	2,8	*große Tendenz*
Nr. 5	1	2	2	2	2	**2**	1,8	*große Tendenz*
Nr. 6	4	4	4	4	4	**4**	4,0	*eindeutig*
Nr. 7	1	1	2	1	1	**1**	1,2	*große Tendenz*
Nr. 8	2	2	3	2	2	**2**	2,2	*große Tendenz*
Nr. 9	4	4	4	4	4	**4**	4,0	*eindeutig*
Nr. 10	3	2	3	3	3	**3**	2,8	*große Tendenz*
Nr. 11	1	1	1	1	1	**1**	1,0	*eindeutig*
Nr. 12	1	1	1	1	1	**1**	1,0	*eindeutig*
Nr. 13	3	3	3	3	3	**3**	3,0	*eindeutig*
Nr. 14	2	2	2	2	2	**2**	2,0	*eindeutig*
Nr. 15	3	3	3	2	3	**3**	2,8	*große Tendenz*
Nr. 16	3	3	3	3	4	**3**	3,2	*große Tendenz*
Nr. 17	1	1	2	1	2	**1**	1,4	*kleine Tendenz*
Nr. 18	3	3	3	2	3	**3**	2,8	*große Tendenz*
Nr. 19	4	4	4	3	4	**4**	3,8	*große Tendenz*
Nr. 20	1	2	2	1	2	**2**	1,6	*kleine Tendenz*
Nr. 21	2	2	3	2	2	**2**	2,2	*große Tendenz*
Nr. 22	1	1	2	1	1	**1**	1,2	*große Tendenz*

***Tab. 5**: Beurteilungen der fünf Testrunden zu den Antworten der Aufgabenstellung „Die neue Heizung“ von ChatGPT-4o und die daraus resultierend der Modus, der Mittelwert und die „Sicherheit“ für jede Antwort.*

Prädikat	Bedeutung	*„Haartrockner“*	*„Die neue Heizung“*
eindeutig	5 gleiche Beurteilungen	7	7
große Tendenz	4 gleiche Beurteilungen	5	12
kleine Tendenz	3 gleiche Beurteilungen	10	3
unklar	2 gleiche Beurteilungen	1	0

***Tab. 6**: Anzahl der Prädikate aus Tab. 3 & 4 für die Sicherheit in ChatGPT-4o's Beurteilungen zu den Aufgaben „Haartrockner“ und „Die neue Heizung“.*

Diskussion der Ergebnisse

Die Testreihe mit *ChatGPT-4o* konnte nach dem vorab definierten Ablauf durchgeführt werden und die tabellarisch gelisteten Ergebnisse liefern. Die Genauigkeit und Konsistenz in ChatGPTs Beurteilungen, zwei maßgebliche Anforderungen für den verlässlichen Einsatz von ChatGPT als Beurteilungshilfe, können mit diesen Daten nun näher diskutiert werden.

Die Genauigkeit der Beurteilungen des Chatbots kann aus dem Grad der Übereinstimmung zur Expertinnenbeurteilung eingeschätzt werden. Diese fällt für die Antwortsätze beider Aufgaben mittelmäßig aus: Für etwa 50 % der Antworten zu „*Haartrockner*“ stimmen die

Beurteilungen von ChatGPT und den Expertinnen Gstall und Korner überein. Bei den Antworten zu „*Die neue Heizung*“ ist die Genauigkeit mit etwa 60 % Übereinstimmung ein wenig höher. Dafür weist die Gesamtbeurteilung des Chatbots zum Antwortsatz der Aufgabe „*Die neue Heizung*“ mit drei stark abweichenden Einstufungen (= rot markierte Beurteilungen in Tab. 3) klar mehr grobe Ausreißer auf als bei „*Haartrockner*“, wo es nur eine Antwort betrifft.

Interessant ist, dass ChatGPT die Antworten zu „*Haartrockner*“ tendenziell schlechter und jene zu „*Die neue Heizung*“ tendenziell besser als Gstall und Korner beurteilt. Die Differenz dafür ist wohl bemerkt gering, wenn man die Notenschnitte vergleicht, aber ein genauer Blick auf die abweichenden Beurteilungen lässt diese Frage schon aufkommen. ChatGPT stuft neun der elf Antworten zu „*Haartrockner*“, die um ein Niveau abweichen, um einen Notengrad schlechter als die Expertinnen ein (= gelb markierte Beurteilungen in Tab. 3). Umgekehrt stuft es fünf der sechs um ein Niveau abweichenden Antworten zu „*Die neue Heizung*“, um einen Notengrad besser ein. Auch alle grob abweichenden Beurteilungen (= rot markierte) zu „*Die neue Heizung*“ sind von *GPT-4o* durchgehend besser eingestuft worden als von Gstall und Korner. Es wäre interessant herauszufinden, warum diese Differenzen zwischen KI- und Expertinnenbeurteilungen so sind, aber dafür lässt sich kein nachvollziehbarer Grund finden. Möglicherweise hat die festgelegte Reihenfolge der Aufgaben im Ablauf der Testreihe einen Einfluss darauf. Warum das für den Chatbot einen Unterschied machen könnte, wird etwas später nochmals aufgegriffen.

Die Genauigkeit von ChatGPT in der Testreihe ist nach dem erhobenen Übereinstimmungsgrad von 50 % bis 60 % nicht besonders hoch einzuschätzen. Isoliert betrachtet sprechen diese Zahlen nicht für einen Einsatz der KI als Beurteilungshilfe von Bewertungskompetenz anhand des 5-stufigen Modells von Gstall (2024).

Könnte es sein, dass das 5-stufige Modell zur Beurteilung der S-Kompetenz in dieser Form ungeeignet für ChatGPT ist? Es ist eigentlich ausgeschlossen, dass *GPT-4o* die Informationen aus der hochgeladenen Handreichung (Gstall, 2024) nicht entnehmen kann. Die generierten Zusammenfassungen des Modells, also ChatGPTs Reaktionen auf Prompt P1, ergaben nie den Eindruck, dass es relevante Modellkriterien auslässt oder Teile falsch wiedergibt. Vorstellbar ist jedoch, dass der Raster des 5-stufigen Modells in manchen Textstellen zu ungenau für eine KI definiert ist. Der Raster gibt etwa als inhaltliche Anforderung einer Niveau-3-Antwort vor, dass „*in der Aufgabenstellung als irrelevant vermerkte*“ Kriterien „*nicht ernsthaft in die Argumentation miteinbezogen*“ werden und solche als „*wesentlich vermerkte […] miteinbezogen*“ werden. An anderer Stelle im Raster ist zu einer Niveau-2-Antwort definiert, dass es noch „*Interpretationsspielraum*“ geben kann, ob „*mehr als die explizit genannten Kriterien Einfluss auf die Wahl der Schüler*innen hatten*“. Diese Formulierungen sind für Menschen geschrieben, die Expertise im Bewerten von Leistungen haben, aber weiß eine künstliche Intelligenz damit umzugehen? Kann ChatGPT *wesentliche* von *irrelevanten* Kriterien inhaltlich unterscheiden? Hat ChatGPT als mathematischer Algorithmus einen *Interpretationsspielraum*? Vielleicht im Sinne von internen Rechenprozessen, doch dazu kann hier keine Aussage getroffen werden. Sollte ChatGPT einen *Interpretationsspielraum* haben, wäre das zumindest eine Erklärung, warum

es im Vergleich zu Menschen und sich selbst verschiedene Beurteilungen generiert. Jedenfalls lässt sich in Frage stellen, wie ChatGPT mit diesen menschlichen Formulierungen umgeht. Auffällig ist, dass sich ChatGPT tendenziell bei den mittleren Niveaustufen, also auch Niveau 2 und 3, unsicher bzw. inkonsistent in den Beurteilungen zeigt. Möglicherweise würde eine „KI-gerechte Sprache“ zu besseren Beurteilungen führen.
Eine andere Möglichkeit könnte sein, dass das Modell in der hochgeladenen Handreichung zu kompliziert für ChatGPT ist. Die Handreichung ist immerhin drei Seiten lang, der Raster macht nur eine davon aus, siehe Anhang. Es ist fraglich, ob ChatGPT die gesamte Handreichung im Detail im Hintergrund gespeichert hat, während es Beurteilungen generiert. Die generierten Outputs von ChatGPT ergaben zwar Sinn und haben die Ideen des Modells widergespiegelt. Ob *GPT-4o* im Rechenprozess tatsächlich korrekt nach dem Modell beurteilt hat, lässt sich aber auch nicht überprüfen. Zusammenfassend lässt sich anhand dieser Fragen feststellen, dass es viele denkbare Möglichkeiten gibt, warum die Ergebnisse von ChatGPT diese Schwankungen aufweisen. Wie eingangs erwähnt kann diese Masterarbeit die Ergebnisse nur so nehmen, wie sie sind und über die Hintergrundprozesse der Generierung nichts aussagen.

An dieser Stelle muss fairerweise auch diskutiert werden, ob die Bestimmung einer Genauigkeit von ChatGPT, wie beschrieben orientiert an Expertinnen, überhaupt gerecht ist. Denn, können die Beurteilungen der Expertinnen überhaupt *genau* sein, wenn das 5-stufige Modell mit unkonkreten Passagen, wie einem Interpretationsspielraum, konzipiert ist? Im Sinne einer objektiven Beurteilung nicht. Möglicherweise schafft es das Beurteilungsmodell von Gstall nicht ein objektiver Maßstab für die Beurteilung von S-Kompetenz zu sein. Genau hier kommt aber die Schwierigkeit der Bewertung von Bewertungskompetenz auf. Gstalls Modell ermöglicht eine objektivere und transparentere Beurteilung, aber auch hier hat es Limitationen. Es ist nicht für völlige Objektivität konzipiert – falls diese bei subjektiven Antworten überhaupt möglich ist – und daher beobachtete auch Gstall in ihrer Studie Differenzen in den Beurteilungen zwischen Expert:innen. Darf man mit dieser Ausgangslage erwarten, dass ChatGPT objektivere und übereinstimmende Beurteilungen trifft? Eine gewisse Ungenauigkeit ist bei der Bewertung der Bewertungskompetenz wohl zu erwarten, wenn verschiedene Meinungen verglichen werden. Wahrscheinlich liegt das im Kern der Sache selbst. Um den Gedanken dieses Absatzes zu Ende zu bringen: Vermutlich ist es falsch im Kontext von Objektivität die Einstufungen von Gstall und Korner als *genau* zu bezeichnen, weil was ist *genau*? Gstall und Korner beurteilten einige der Antworten auch unterschiedlich, eine gemeinsame Festlegung auf ein Niveau ergab sich erst nach Diskussion im Konsens. Letztlich stammen die Beurteilungen Gstalls und Korners aber von Menschen mit physikalisch-didaktischer Expertise und viel Erfahrung in ihrem Fachgebiet. Aus diesem Grund können ihre Beurteilungen als guter Vergleichsmaßstab für die Einschätzung von ChatGPTs Genauigkeit in der Pilotstudie verwendet werden.

Das Maß an Konsistenz in ChatGPTs Beurteilungen lässt Aussagen über die Sicherheit seiner Durchschnittsbeurteilungen zu. ChatGPT zeigte mangelnde Konsistenz bei der Beurteilung einzelner Antworten über die fünf Beurteilungsrunden hinweg. Das zeigen die

Ergebnisse in Tab. 4 und Tab. 5. Mit jeweils sieben *eindeutigen* Beurteilungen in beiden Antwortsätzen, war sich der Chatbot also nur für grob ein Drittel der Antworten mit seiner Beurteilung sicher, siehe Tab. 6. Für die restlichen Beurteilungen nahm die Tendenz von ChatGPT für eine Niveaustufe in verschiedenen Maßen ab. Interessanterweise schwanken fast alle uneindeutigen Beurteilungen von ChatGPT zwischen zwei Niveaustufen, manchmal mit großer und manchmal mit kleiner Tendenz für eine von beiden. Nur einmal in der ganzen Testreihe, bei Antwort 8 der Aufgabe „*Haartrockner*", vergab ChatGPT drei verschiedene Niveaus.

Vergleicht man die Sicherheit der Durchschnittsbewertungen auf Basis der Konsistenz aus der Testreihe lässt sich erkennen, dass die Beurteilungen des Chatbots zu „*Die neue Heizung*" eindeutig sicherer sind als zu „*Haartrockner*", siehe Tab. 3, und, dass gleichzeitig die Anzahl der leicht abweichenden Beurteilungen bei „*Die neue Heizung*" (6x) deutlich niedriger ist als bei „*Haartrockner*" (11x). Könnte es einen Zusammenhang zwischen ChatGPTs Sicherheit und der Anzahl an leichten Ausreißern in den Beurteilungen geben? Möglich wäre es, aber dafür wären tiefere Analysen notwendig, die nicht Ziel der Pilotstudie sind. Die verschiedenen Aufgabenstellungen könnten ein Grund dafür sein.

Die Anzahl und Relevanz der Kriterien spielen eine zentrale Rolle im 5-stufigen Modell von Gstall (2024). Das Beurteilungsmodell greift in Niveau 3 die Verwendung *irrelevanter* und *wesentlicher* Kriterien auf, wie weiter oben erwähnt. Irrelevante Kriterien der Aufgabenstellung dürften z.B. für das Erreichen von Niveau 3 „*nicht ernsthaft*" in die Entscheidung miteinbezogen werden. Vergleicht man die Aufgabenstellungen von „*Die neue Heizung*" und „*Haartrockner*" scheint es, dass erstere keine wesentlichen oder irrelevanten Kriterien definiert: „*Die neue Heizung*" nennt keine Kriterien, die für Familie Fischbach wichtig und damit *wesentlich* für den Kauf einer neuen Heizung wären. Ganz anders schildert die Aufgabenstellung „*Haartrockner*" eingangs, dass einer Frau die Sicherheit eines neuen Föns wichtig wäre. Daher wäre das Prüfzeichen ein wesentliches Kriterium für Entscheidung und klar von der Aufgabenstellung als solches markiert.
Möglicherweise hat ChatGPT in der Aufgabe „*Haartrockner*" die Relevanz des Prüfzeichens als wesentliches Kriterium gemäß dem 5-stufigen Modell stark gewichtet und jene Antworten ohne dieses Kriterium strenger beurteilt. Bei „*Die neue Heizung*" gibt es vorab kein als wesentlich deklariertes Kriterium, daher könnten die Beurteilungen milder gewesen sein. Es gibt allerdings auch Ausnahmen für diese Theorie: In „*Haartrockner*" lassen sich Schüler:innen-Antworten ohne Verwendung des Prüfzeichens finden, die der Chatbot vereinzelt mit Niveau 2 beurteilt hat, z.B. Antwort Nr. 4 oder 15 – wobei auch Gstall und Korner diese Antworten mit Niveau 2 oder 1 beurteilt haben, was nach wörtlicher Auslegung des Modells fragwürdig ist.

Zusammenfassend lässt sich feststellen, dass für den Großteil der Beurteilungen, je etwa zwei Drittel der Antwortsätze, die Beurteilungen von *ChatGPT-4o* inkonsistent sind. Zumindest aber schwanken sie fast ausschließlich zwischen zwei Niveaus und zeigen über fünf Beurteilungsrunden teils eine große Tendenz für ein Niveau an. Ebenso wie die mittlere Genauigkeit lässt die mangelnde Konsistenz an der Verlässlichkeit von ChatGPT im Bewerten von Bewertungskompetenz zweifeln. In einem realistischen Alltagsszenario einer

Lehrperson bleibt nicht die Zeit, einen Antwortsatz mehrmals korrigieren zu lassen und einzelne Beurteilungen aus dem Durchschnitt abzuleiten. Die Ergebnisse von ChatGPT sollten nach einem Anlauf zuverlässig, sicher und genau sein.

Auffälligkeiten in den Beurteilungen von ChatGPT

Bisher wurden die Einstufungen von ChatGPT quantitativ analysiert. Nun folgt eine qualitative Analyse der generierten Beurteilungen. Es sollen Auffälligkeiten in ChatGPTs Beurteilungen aus der Testreihe identifiziert und diskutiert werden, die für den Einsatz der KI als Beurteilungshilfe relevant sein können.

1) **Die Ausführlichkeit der Formulierungen von ChatGPTs Einzelbeurteilungen nimmt im Output fortlaufend ab.**

Diese Tatsache konnte in jeder Runde der Testreihe beobachtet werden. Die Ausführlichkeit von ChatGPT war zwar generell nicht hoch, was am Tokenlimit liegen könnte. Dennoch lässt sich über mehr als 20 Antworten pro Aufgabe hinweg der Trend zu kontinuierlich minimalistischeren Formulierungen beobachten. Meistens fiel das bereits nach wenigen, etwa fünf bis zehn generierten Antworten auf. Als Veranschaulichung dienen hier zwei Auszüge von ChatGPT aus den Beurteilungen zur Aufgabe „*Die neue Heizung*“ aus Runde 1 der Testreihe. Sowohl Begründung als auch Feedback sind zu Antwort 1 etwas detaillierter geschrieben als zu Antwort 22. Aufgrund der gleichen Niveaustufe (1) lassen sich diese Einzelbeurteilungen vergleichen.

Antwort Nr. 1

- **Beurteilung**: Niveaustufe 1
- **Begründung**: Vier relevante Kriterien (Heizwert, praktische Versorgung, Umweltaspekt, Preis) genannt, mit Begründung und positiver Abwägung.
- **Feedback**: Sehr gute Argumentation! Du hast viele relevante Punkte genannt und sinnvoll begründet. Besonders stark ist die Einbindung des Heizwerts und der Versorgung.

Antwort Nr. 22

- **Beurteilung**: Niveaustufe 1
- **Begründung**: Drei Kriterien (Heizwert, Kosten, Umwelt), differenzierte Betrachtung.
- **Feedback**: Super! Du erkennst wichtige Vor- und Nachteile und argumentierst sehr gut – weiter so!

Ein weiteres Beispiel stammt aus der Aufgabe „*Haartrockner*“, mit Auszügen von ChatGPT aus Runde 3 der Testreihe. Hier ist die Abnahme der Ausführlichkeit noch deutlicher.

Antwort Nr. 1

- **Beurteilung:** Niveaustufe 3
- **Begründung:** Zwei relevante Kriterien (Leistung, Sicherheit durch Prüfzeichen) werden genannt. Die Argumentation ist nachvollziehbar.
- **Feedback:** Du hast einige relevante Kriterien genannt. Besonders gut ist, dass du auf Sicherheit und Qualität eingehst. Du könntest noch klarer verschiedene Optionen vergleichen, um deine Entscheidung stärker zu begründen.

Antwort Nr. 23

- **Beurteilung:** Niveaustufe 3
- **Begründung:** Leistung, Preis, Prüfzeichen werden genannt, einfache Aufzählung.
- **Feedback:** Du hast die wichtigsten Kriterien genannt – gut gemacht! Versuch beim nächsten Mal, deine Argumente stärker miteinander zu verknüpfen.

Nicht in allen Beurteilungsrunden und zu beiden Aufgaben konnte diese Auffälligkeit im gleichen Ausmaß festgestellt werden. Mal war der Effekt mehr und mal weniger oder kaum zu beobachten. Dennoch ist der Trend einer leicht abnehmenden Ausführlichkeit mit zunehmender Antwortzahl in der Mehrheit der Beurteilungsrunden feststellbar. Dieser Mangel wird vermutlich von technischen Limitationen des KI-Modells herrühren, wie etwa dem Tokenlimit. Ob das die Einstufungen der Antworten beeinflusst, ist unklar. Die Qualität der Begründungen und Feedbacks der Einzelbeurteilungen sind damit eingeschränkt bzw. ungleich verteilt.

2) Die Qualität von ChatGPT‘s Feedback hat Potenzial.

Abgesehen von den kürzer werdenden Feedbacks im Verlauf der generierten Outputs, ist die Qualität der Feedbacks akzeptabel, aber ausbaufähig. Wie die obigen Auszüge zeigen, ist das generierte Feedback recht solide in Anbetracht dessen, dass es keine genauen Vorgaben oder Beispiele für ChatGPT gab. Der Prompt P2 enthielt die Anforderung:

> **Prompt P2**
> *„Feedback: (Formuliere hier ein sachliches und motivierendes Feedback. Es soll Stärken, Schwächen und Verbesserungsmöglichkeiten der Antwort beinhalten.)“*

In vielen Fällen erfüllte ChatGPT den motivierenden Feedbackstil. Beispielsweise generierte es Feedbackpassagen wie:

> **ChatGPT**
> *„Sehr gut! […] Weiter so!“*
>
> *„Sehr gute Argumentation […] – top!“*
>
> *„Du hast […] genannt – gut gemacht!“*
>
> *„Deine Antwort ist gut […]. Noch besser wäre es […].“*
>
> *„Du denkst in die richtige Richtung […]. Achte bitte darauf, ob […]“*

„Ein Anfang ist gemacht, aber du solltest deine Entscheidung ausführlicher begründen, um andere von deinem Standpunkt zu überzeugen."

„Du hast dich entschieden, aber deine Begründung ist zu allgemein. […]"

Diese Formulierungen motivieren Schüler:innen mit bereits guten Antworten zum Weitermachen und sind bei jenen mit Verbesserungsbedarf so verfasst, dass sie zuerst das Positive hervorheben und dann auf Verbesserungsmöglichkeiten hinweisen. Der wertschätzende Umgang, egal ob mit stärkeren oder schwächeren Antworten, fällt bei der Analyse über alle Beurteilungsrunden positiv auf. Daraus kann geschlossen werden, dass ChatGPT motivierendes Feedback generieren kann.

Der sachliche Teil des Feedbacks sollte darauf ausgerichtet sein Verbesserungsmöglichkeiten aufzuzeigen. Auch das hat ChatGPT allgemein gut gelöst:

ChatGPT:
„Um deine Argumentation zu verbessern, erkläre genauer, warum […]"

„Die Kriterien sind gut gewählt. Um deine Begründung zu verbessern, vergleiche dein Modell mit einem anderen."

„Du denkst in die richtige Richtung und vergleichst Modelle. Achte bitte darauf, ob deine Argumente wirklich fachlich sinnvoll sind – mehr Energieverbrauch ist kein Vorteil."

„Noch besser wäre es, explizit eine Gegenposition zu entkräften."

„Deine Argumentation ist gut strukturiert und zeigt, dass du verschiedene Aspekte abwägst – genau so fördert man S-Kompetenz!"

„Versuch beim nächsten Mal konkreter zu erklären, was du unter „gutes Preis-Leistungs-Verhältnis" verstehst."

„Deine Antwort ist leider zu kurz. Überlege dir, welche Eigenschaften (z. B. Heizwert, Umweltfreundlichkeit, Kosten) wichtig für Familie Fischbach sein könnten."

Diese Antworten entsprechen grundsätzlich den Erwartungen eines Feedbacks, könnten aber noch individueller und konstruktiver sein, etwa mit konkreten Beispielen.

In Summe gibt ChatGPT bei den meisten Feedbacks als Verbesserungsmöglichkeit den ausstehenden Vergleich zu anderen Optionen an: „*Ein Vergleich mit anderen Optionen würde deine Argumentation noch verbessern/deine Begründung noch klarer machen*" schreibt der Chatbot bei vielen Beurteilungen dazu. Ob diese Schwäche auf viele Antworten zutreffend ist, könnte eine genau Analyse der 55 Antworten zeigen. Dieser Aufwand würde den Anspruch der Pilotstudie jedoch überschreiten. Es wird für einige Antworten wohl passend sein, da ein Vergleich von Optionen eine notwendige Anforderung für die höheren Niveaustufen 1 und 2 laut Beurteilungsmodell ist. Trotzdem wirft es letztlich die Frage auf, ob ChatGPT das Feedback vielleicht zu allgemein, fast nach einem Muster wirkend,

formuliert. Es bleibt noch Raum für mehr individuelles Eingehen auf die einzelne Antwort und damit verbundenes, konstruktives Verbesserungspotenzial.

Als mögliche Lösung könnte man dem Prompt hinzufügen, dass ChatGPT im Feedback einen beispielhaften Satz für die angedeutete Verbesserungsmöglichkeit generieren soll. Offen bleibt dann wiederum, ob dies bei den Schüler:innen das Bild erzeugt, dass KI stets eine „bessere" Lösung, als sie erzeugt und die Schüler:innen dazu animiert in Zukunft ihre Aufgaben gleich mit KI erledigen zu lassen. Diese Frage soll jedoch Thema einer anderen Forschungsarbeit sein.

3) **Inkonsistente Beurteilungen von ChatGPT schwanken fast ausschließlich zwischen zwei benachbarten Niveaustufen.**

Interessantes zeigt sich beim Vergleich der einzelnen Beurteilungen aus allen fünf Beurteilungsrunden zu derselben Antwort. Wie bereits in den Ergebnissen angesprochen schwanken beinahe alle Durchschnittsbeurteilungen, bei denen Inkonsistenzen zwischen den einzelnen Runden auftraten, zwischen zwei Niveaus. Ausschließlich für eine Antwort, Nr. 8 zu „*Haartrockner*", zeigte ChatGPT eine *unklare* Tendenz, da es diese Antwort einmal mit Niveau 2 und je zweimal mit Niveau 3 und 4 beurteilte. Etwa zwei Drittel der 55 Durchschnittsbeurteilungen beruhen auf einer Tendenz ChatGPTs zwischen zwei Niveaustufen über die fünf Testrunden.

4) **ChatGPT macht Ausreißer bei den Beurteilungen.**

ChatGPT macht *Ausreißer*. Im Kontext dieser Arbeit sind damit die im Vergleich zu Gstall und Korner abweichenden Beurteilungen von ChatGPT aus Tab. 3 gemeint. Im Grunde sind alle gelb und rot markierten Beurteilungen Ausreißer, was illustriert, dass die Trefferquote des Chatbots eher mittelmäßig ist. Die gelben Beurteilungen sind *leichte Ausreißer*, die roten Beurteilungen sind *grobe Ausreißer*. Zur Erinnerung an die Farbcodes: Leichte Ausreißer weichen um eine Niveaustufe von der Expertinnenbeurteilung ab, grobe Ausreißer weichen um zwei Niveaustufen oder mehr davon ab. Warum einige der Beurteilungen von *GPT-4o* leichte Ausreißer sind, im Fall der Aufgabe „*Haartrockner*" mit 11 von 23 Beurteilungen knapp 50%, kann viele Gründe haben.

Der offensichtlichste Grund sind interne Rechen- oder Gewichtungsprozesse von ChatGPT, die Einfluss auf die Beurteilung haben könnten. Diese wurden bereits in den Auffälligkeiten der ersten Phase der Pilotstudie als mögliche, unbekannte Störvariablen diskutiert. Weiters scheint in der Beurteilung von S-Kompetenz mit einer gewisse Schwankungsbreite zu rechnen zu sein. Das macht auch die Hauptstudie von Gstall deutlich, in der selbst Expert:innengruppen, aus Universitätspersonal, Lehrpersonen und Studierenden der Physik die gleichen Antworten sehr verschieden einstuften. Dass ChatGPT als vierte Instanz erneut andere Beurteilungen erzeugt, ist so gesehen nicht mehr überraschend und verdeutlicht nur die allgemeine Schwierigkeit und Schwankungsbreite im Beurteilen von S-Kompetenz, selbst für eine KI.

Es dürfte so sein, dass die leichten Ausreißer bevorzugt in den mittleren Niveaustufen des 5-stufigen Modells auftreten. Schwer beurteilbare Antworten, die von verschiedenen Menschen verschieden beurteilt werden, bekommen oft eine mittelmäßige Note – gemäß

der „Tendenz zur Mitte“. Die mittleren Niveaustufen des 5-stufigen Modells bilden damit ein Fangbecken für schwer beurteilbare Antworten. Wie bereits diskutiert lässt das Modell etwa in Niveau 3 und 2 subjektive Wertungen der beurteilenden Person (hier der KI) zu. Dies würde die Ansammlung von Antworten mit Ausreißer-Potenzial dort fördern. Lassen sich die Ausreißer von ChatGPT dort finden? Die leichten Ausreißer von ChatGPT sind tatsächlich tendenziell bei den Niveaus 2 und 3 zu finden, aber auch in den anderen Niveaus 1 und 4, mit Ausnahme Niveau 5. Es kann also nicht sicher behauptet werden, dass sich die leichten Ausreißer *nur* auf die mittleren Niveaustufen beziehen – es lässt sich nur eine Tendenz dafür erkennen. Für die groben Ausreißer liegt die Ursache meistens an der Art der Antwort oder an Fehlern von ChatGPT.

5) ChatGPT macht Fehler.

Es ist allgemein bekannt, dass KI-Modelle Fehler machen oder halluzinieren können und die Outputs daher immer kontrolliert werden müssen. Auch ChatGPT generierte in der Testreihe vereinzelt kleine Fehler, etwa Grammatikfehler, unvollständige Bearbeitungen, etc.)

Manche Beurteilungen von ChatGPT ergeben aber keinen Sinn, wenn man sorgfältig nach dem Beurteilungsmodell vorgeht. Ein Beispiel hierfür ist Antwort Nr. 10 zur Aufgabe „*Haartrockner*“. Die Antwort lautet:

> *„Esra sollte den ProHair Haartrockner nehmen, weil es hat eine gute und hohe Leistung von 2000W und es hat das Prüfzeichen: ÖVE. Dieses Prüfzeichen garantiert die Geräteprüfung durch eine unabhängige, kompetente Prüfstelle.“*

Diese Antwort weist zwei Kriterien (Leistung und Prüfzeichen) auf und lässt sprachlich eine klare Begründung erkennen. Eine logische Einstufung nach dem Modell wäre Niveau 3, so haben es die Expertinnen auch getroffen, doch der Chatbot schwankt zwischen Niveau 1 und 2 mit leichter Tendenz zu Niveau 1. Diese Reaktion von ChatGPT ist schwer nachvollziehbar. Ein Blick auf die Begründungen aus den fünf Beurteilungsrunden stiftet noch mehr Verwirrung:

Antwort Nr. 10

- **Beurteilung**: Niveaustufe 1
- **Begründung**: Klare Argumentation mit mehreren Kriterien (Leistung, Sicherheit), ausführliche fachliche Begründung.
- **Feedback**: Sehr gut! Du erklärst, was das Prüfzeichen bedeutet und gehst auf Esras Bedürfnisse ein. Weiter so!

Antwort Nr. 10

- **Beurteilung:** Niveaustufe 1
- **Begründung:** Fachlich fundierte Begründung mit mehreren Kriterien (Leistung, Prüfzeichen, Sicherheit), klar strukturiert.
- **Feedback:** Sehr gute Antwort! Du erklärst genau und nutzt relevante Kriterien – tolle Leistung!

Antwort Nr. 10

- **Beurteilung:** Niveaustufe 1
- **Begründung:** Leistung und Prüfzeichen werden differenziert erklärt, mit Alltagsbezug.
- **Feedback:** Sehr gut! Du hast die Kriterien fachlich korrekt verwendet und deine Begründung klar formuliert. Weiter so!

Antwort Nr. 10

- **Beurteilung:** Niveaustufe 2
- **Begründung:** Zwei Kriterien mit klarer Begründung (Leistung, Prüfzeichen), aber kein expliziter Vergleich.
- **Feedback:** Du erklärst sehr gut, warum du diesen Föhn auswählst. Um das höchste Niveau zu erreichen, vergleiche dein Modell mit den anderen.

Antwort Nr. 10

- **Beurteilung:** Niveaustufe 2
- **Begründung:** Drei Kriterien, inkl. Erklärung des Prüfzeichens. Die Argumentation ist sachlich, aber es fehlt ein direkter Vergleich.
- **Feedback:** Sehr gute inhaltliche Tiefe und Erklärung. Um das höchste Niveau zu erreichen, solltest du deine Wahl mit den Nachteilen anderer Modelle begründen.

ChatGPT erkennt in Antwort 10 die Anzahl der Kriterien, den Differenzierungsgrad oder den gezogenen Vergleich falsch bzw. unterschiedlich und bewertet die Antwort dementsprechend verschiedenen. Das Einstufen in Niveau 1 ist nach rationaler Anwendung des 5-stufigen Modells nicht logisch und ein Fehler von ChatGPT. Antwort 10 ist ein Beispiel für größere Fehler von ChatGPT. Ob grobe Ausreißer immer an Fehlern der KI liegen, behandelt der nächste Punkt. In kleinerem Ausmaß kommen Fehler der KI jedenfalls öfters vor und vermutlich bemerkt man diese nicht immer. Wie bereits in Phase 1 diskutiert, sollten die Ergebnisse des Chatbots daher immer kontrolliert werden – Antwort 10 zeigt dies sehr deutlich.

6) **Grobe Ausreißer sind teils auf Fehler von ChatGPT zurückzuführen, teils sind die betroffenen Antworten an sich schwer einzustufen.**

Die eben diskutierte Antwort 10 zu „*Haartrockner*“ muss auf einen Fehler von ChatGPT zurückzuführen sein, da die Antwort an sich keinen Raum für Interpretationen lässt. Möglicherweise sind auch die anderen groben Ausreißer wegen Fehler des Chatbots passiert oder eine Antwort an sich ist einfach schwer in der Einstufung. Gstall (2024, S. 52–60) diskutierte in ihrer Masterarbeit einzelne Antworten, die sich als schwer beurteilbar herausstellten – diese werden hier als *kritische* Antworten bezeichnet. Eine kurze Analyse zeigt, dass die vier groben Ausreißer von ChatGPT, siehe Tab. 3, teilweise Antworten betreffen, die laut Gstalls Studie als kritisch zu identifizieren sind – die ausreißende Bewertung des Chatbots ist damit relativierbar für jene Antworten. Eine davon ist Antwort 11 zur Aufgabe „*Die neue Heizung*“. Gstall und Korner bewerteten diese Antwort mit Niveau 5 und ChatGPT mit Niveau 1. Dieser Ausreißer ist damit der gröbste in der Testreihe:

„Heizöl Familie Fischbach zalt wenig muss nichts anschafen und hat nicht das Risiko einer Gasexplosion oder einem aus durch Sanktionen gegen Erdgaslieferanten bleibt, nur als Nachteil die Klimaverschmutzung."

Die drei weiteren groben Ausreißer von ChatGPT, die Antworten 1 und 21 zu „*Die neue Heizung*" und Antwort 10 zu „*Haartrockner*" (oben bereits vorgestellt) wurden von Gstall nicht explizit diskutiert. Antwort 1 wurde von Gstall und Korner mit Niveau 4 bewertet, von ChatGPT mit kleiner Tendenz mit Niveau 1:

„Heizöl: weil es 10kWh/l hat was viel ist, es direkt ins Haus strömt, es einen fossilen Brennstoff Ursprung hat und weil es nicht das teuerste ist."

Die Antwort 21 wurde von den Gstall und Korner ebenfalls mit Niveau 4 bewertet, von ChatGPT mit großer Tendenz mit Niveau 2:

„Erdgas, weil es am meisten Leistung hat und es direkt ins Haus strömt."

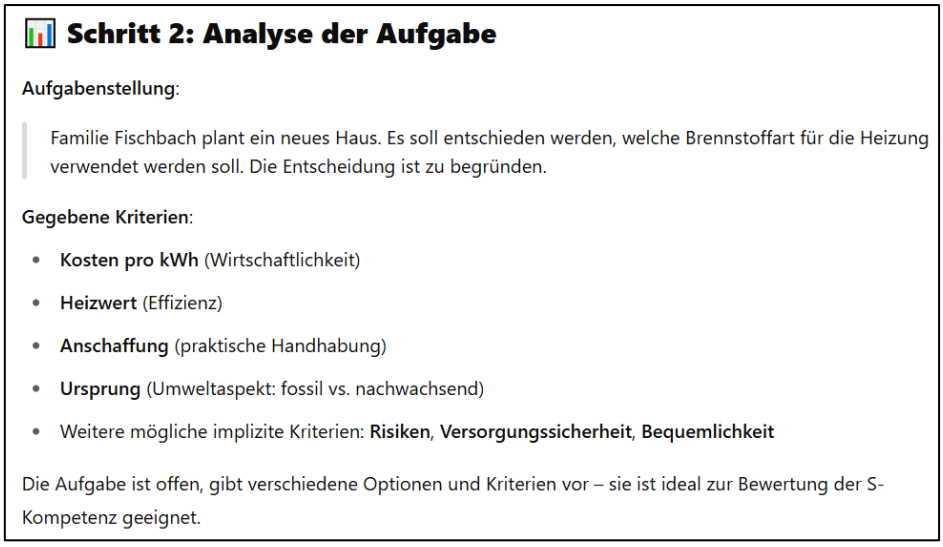
Ob die Beurteilungen dieser vier Antworten von ChatGPT nun berechtigte grobe Ausreißer bilden, weil es sich tatsächlich um schwer beurteilbare Antworten handelt, kann mit den Daten von Gstall (2024) nicht umfassend beantwortet werden. Für Antwort 11 zu „*Die neue Heizung*" kann zumindest nachvollzogen werden, dass die KI einen groben Ausreißer erzeugt, aber für die restlichen Ausreißer steht fest, dass die Ergebnisse der KI wenig sinnvoll und nach den Ideen des 5-stufigen Beurteilungsmodells fehlerhaft sind.

Interessanterweise erwähnt gab es umgekehrt zwei Antworten in Gstalls Studie, die als kritische Antworten diskutiert wurden, für ChatGPT aber relativ eindeutig waren – die Antworten 1 und 20 zu „*Haartrockner*".
Gstall integrierte in ihre Hauptstudie außerdem sehr eindeutig wirkende Antworten, die von den Expert:innen auch allesamt recht übereinstimmend eingestuft wurden. Das sind die Antworten 3 aus „*Haartrockner*" und 2, 9 und 10 aus „*Die neue Heizung*". Diese Nummern wurden auch von ChatGPT nach den Daten der Testreihe *eindeutig* bzw. mit *großer Tendenz* der gleichen Niveaustufe wie die von den Expert:innen gewählte zugeteilt.

7) ChatGPT halluziniert teils zusätzliche Kriterien in die Aufgabenstellung.

Eine kleine Auffälligkeit ergab sich auch in einem der ersten Arbeitsschritte von ChatGPT, nämlich der Analyse der Aufgabenstellung, die es üblicherweise vor den Einzelbeurteilungen generiert. Teilweise halluzinierte der Chatbot weitere Kriterien der Aufgabenstellung hinzu, wie etwa „Energieeffizienz" oder „Qualität" bei „*Haartrockner*" und „Nachhaltigkeit" oder „Komfort" bei „*Die neue Heizung*". Es ist nicht davon auszugehen, dass diese Halluzinationen großen Einfluss auf die Beurteilungen hatten, da sie meistens aus dem Kontext der Aufgabe entstanden. Allerdings verdeutlichen sie, dass ChatGPT von Halluzinationen betroffen sein kann.

Einmal schrieb der Chatbot auch mögliche implizierte Kriterien zur Analyse der Aufgabe hinzu, wie in dem unteren Auszug zu „*Die neue Heizung*". Diese erinnern an die Argumentation von Antwort Nr. 11.



📊 **Schritt 2: Analyse der Aufgabe**

Aufgabenstellung:

> Familie Fischbach plant ein neues Haus. Es soll entschieden werden, welche Brennstoffart für die Heizung verwendet werden soll. Die Entscheidung ist zu begründen.

Gegebene Kriterien:

- **Kosten pro kWh** (Wirtschaftlichkeit)
- **Heizwert** (Effizienz)
- **Anschaffung** (praktische Handhabung)
- **Ursprung** (Umweltaspekt: fossil vs. nachwachsend)
- Weitere mögliche implizite Kriterien: **Risiken**, **Versorgungssicherheit**, **Bequemlichkeit**

Die Aufgabe ist offen, gibt verschiedene Optionen und Kriterien vor – sie ist ideal zur Bewertung der S-Kompetenz geeignet.

Erkenntnisse aus Phase 2

Phase 1 hat gezeigt, dass ChatGPT Beurteilungen anhand Gstalls 5-stufigen Modell generieren kann. Die Ergebnisse und Auffälligkeiten aus Phase 2 geben einen Einblick auf die Qualität von ChatGPTs Beurteilungen. Aus der Diskussion dieser Daten kann zusammenfassend kann festgehalten werden:

- ChatGPT kann zur Bewertung der Bewertungskompetenz anhand des 5-stufigen Beurteilungsmodells von Gstall eingesetzt werden. Dabei ist jedoch mit einem gewissen Ausmaß an Fehlern zu rechnen, sodass die Ergebnisse in jedem Fall kontrolliert und teils überarbeitet werden müssen.
- Die Beurteilungen von ChatGPT zeigen im Schnitt eine mittelmäßige Genauigkeit und Konsistenz die zweifeln lassen, ob der Einsatz von *GPT-4o* zuverlässig genug ist.
- ChatGPT macht mehrere leichte und wenige grobe Ausreißer in den Beurteilungen. Die leichten Ausreißer betreffen tendenziell die mittleren Niveaustufen.
- Die Ursachen für Ausreißer sind vielfältig, vermutlich sind meistens die Art der Antwort, Fehler von ChatGPT oder Spielräume im Beurteilungsmodell Gründe dafür.
- ChatGPT generiert solides Feedback mit Potenzial für weitere Verbesserung.
- ChatGPT kann Fehler beim Analysieren und Bewerten von Aufgabenstellungen und Antworten machen. Ergebnisse müssen vor der Verwendung überprüft werden.
- Der Beurteilungsprozess mit Prompt P1 und P2, basierend auf dem Shot Prompting und dem *PROPER*-Framework, funktioniert.
- ChatGPT kann meistens problemlos innerhalb einer Anfrage eine Aufgabenstellung mit Antworten in Klassenstärke (hier bis zu 23) analysieren, bewerten und Beurteilungen generieren. Manchmal muss die Generierung auf zwei Teile aufgeteilt werden.

Interpretationen für Phase 3

ChatGPT zeigt Potenzial für seinen Einsatz in der Bewertung von Bewertungskompetenz, doch die aktuelle Ergebnislage lässt noch Zweifel offen, ob die Beurteilungen zuverlässig genug sind und der Chatbot damit eine viable Alternative für den Beurteilungsprozess

darstellt. Natürlich ist es eine Sache der Datenauslegung, ob sich ChatGPT bereits in der Testreihe für den Einsatz ausreichend bewiesen hat. Wie diskutiert ist in der Bewertung von Bewertungskompetenz wohl grundsätzlich mit Abweichungen beim Vergleich verschiedener Beurteilungen zu rechnen, was die erreichte Trefferquote von ChatGPT wieder in ein gutes Licht stellen würde.

Der Beurteilungsprozess konnte mit den beiden Prompts P1 und P2 die beschriebenen Ergebnisse liefern. Die Beurteilungsleistung von *GPT-4o* in der Testreihe weist Stärken, Schwächen und Potenziale auf. In der nächsten und letzten Phase der Pilotstudie sollen nochmal diverse Anpassungen der beiden Prompts ausgetestet werden, die die Ergebnisse von ChatGPT in verschiedenen Aspekten verbessern könnten. Außerdem sollen die bisherigen Erkenntnisse von Expert:innen diskutiert werden, um externe Meinungen zur bisherigen Datenlage zu sammeln.

5.3. Phase 3 – Testung neuer Varianten im Beurteilungsprozess und Feedback aus dem Forschungsseminar

Zielsetzungen
In Phase 3 sollen einerseits Anpassungen im Beurteilungsprozess mit leicht veränderten Formen der Prompts P1 und P2 ausprobiert werden und andererseits sollen die bisherigen Erkenntnisse zusammen in einem Forschungsseminar der Physikdidaktik an der Universität Wien diskutiert werden. Die Erkenntnisse aus Phase 3 sollen das Ziel der Pilotstudie abschließend klären können und den weiteren Forschungsverlauf mitbestimmen.

Durchführung
Die Testreihe aus Phase 2 wurde mit der gewählten Prompting-Strategie und dem am Ende von Phase 1 resultierenden zweigeteilte Promptvorlage durchgeführt. In Phase 3 wird versucht durch kleine Anpassungen in den Prompts P1 und P2 mögliche Verbesserungen im Beurteilungsprozess zu erzielen. In Summe werden sechs *Promptingvarianten* und deren Wirkung getestet. In jeder Variante wird nur eine Variable in der sonst ursprünglichen Struktur der Prompts P1 und P2 verändert, um den Effekt dieser Änderung isoliert beobachten zu können. Die angepassten Prompts werden in einem ähnlichen Ablauf wie in Phase 2 getestet: ChatGPT bekommt mit jeder der sechs Promptingvarianten *einmal* das 5-stufige Modell (Prompt P1) und anschließend die Aufgabe „*Haartrockner*“ inklusive den Schüler:innen-Antworten mit dem Beurteilungsauftrag (Prompt P2) hochgeladen. Mit jeder der sechs Promptingvarianten wird eine Bewertung von ChatGPT durchgeführt, dokumentiert und wieder mit den Expertinnenbeurteilungen, wie in Phase 2, verglichen. Im Unterschied zu Phase 2 wird jedoch keine Durchschnittsbeurteilung aus fünf identischen Runden berechnet. Die Aufgabe „*Die neue Heizung*“ wird in Phase 3 nicht verwendet. Der Aufwand von Phase 3 wird bewusst kleiner gehalten, da die ersten beiden Phasen der Pilotstudie bereits sehr ausführlich sind. Außerdem sollten deutliche Effekte neuer Promptingvarianten schon nach einem Durchgang erkennbar sein.

Der zweite Teil von Phase 3 umfasst die Präsentation und Diskussion der bisherigen Forschungsergebnisse in einem Physikdidaktikseminar der Universität Wien vor einer Gruppe von Physikdidaktiker:innen, bestehend aus Professor:innen, Senior-Lecturers, Doktorats- und Masterstudent:innen. Diskutierte Ideen und offene Fragen werden hier dokumentiert und können einen wertvollen Beitrag für den weiteren Verlauf der Forschungsarbeit leisten.

Ergebnisse

In Summe wurden sechs leicht veränderte Promptingvarianten im Beurteilungsprozess mit ChatGPT ausgetestet. Manche dieser Anpassungen basieren auf den Erkenntnissen von Phase 2, andere entstanden aus iterativen Überlegungen. Die zweiteilige Promptvorlage aus den Prompts P1 und P2 wurde in folgenden Variablen verändert: 1) Aktivierung der *Reasoning*-Funktion von ChatGPT, 2) Generierung der Beurteilungen in zwei Outputs aufteilen lassen, 3) Teilung des Antwortsatzes in zwei kleinere Pakete, 4) Kürzungen von Formulierungen der ursprünglichen Prompts, 5) Zusammenlegung aller Angaben und Aufträge für ChatGPT auf einen Prompt und 6) Anforderung für genaueres Feedback im Prompt hinzugefügt.

Jede der sechs Anpassungen wurde einmal in der Beurteilung der Aufgabe „*Haartrockner*“ getestet. Die Beurteilungen von *ChatGPT-4o* wurden notiert und sind in Tab. 7 aufgelistet. Der rechte Tabellenteil gibt die einzelnen Beurteilungen von ChatGPT in den sechs Varianten an. Der linke Tabellenteil stellt zum Vergleich die Durchschnittsbeurteilungen von ChatGPT aus der Testreihe von Phase 2 erneut dar. Der Farbcode der gesamten Tabelle ist gleich angelegt, wie in der Testreihe der Phase 2 und stellt den Vergleich aller KI-Beurteilungen der Expertinnen Denise Gstall und Marianne Korner (in Tab. 7 in blau hinterlegt) dar.

Beurteilungen zu „*Haartrockner*“ aus der Testreihe (Phase 2)				Beurteilungen zu „*Haartrockner*“ mit veränderter Variable im Prompting (Phase 3)					
Prädikat	**Antwort**	**GPT-4o**	**DG & MK**	***Reasoning* aktiviert**	**Teilung Auftrag**	**Teilung Antworten**	**Kürzung Prompts**	**1 Prompt**	**Feedback erweitert**
große T.	**Nr. 1**	3	2	3	3	4	4	2	4
eindeutig	**Nr. 2**	3	2	3	3	3	3	3	3
große T.	**Nr. 3**	1	1	4	1	2	1	1	1
kleine T.	**Nr. 4**	2	2	2	3	3	2	2	2
eindeutig	**Nr. 5**	2	2	2	2	2	2	1	2
kleine T.	**Nr. 6**	4	4	2	4	5	3	4	4
kleine T.	**Nr. 7**	4	4	2	4	4	3	3	3
unklar	**Nr. 8**	3	4	4	3	4	3	3	4
eindeutig	**Nr. 9**	3	3	3	3	3	3	3	3
kleine T.	**Nr. 10**	1	3	3	1	2	2	2	2
große T.	**Nr. 11**	2	3	3	1	2	2	2	2
kleine T.	**Nr. 12**	3	2	2	3	3	2	2	3
eindeutig	**Nr. 13**	3	3	3	3	3	3	3	3
eindeutig	**Nr. 14**	1	1	2	1	1	1	1	1
kleine T.	**Nr. 15**	2	1	2	2	2	1	1	2

große T.	**Nr. 16**	3	2	3	4	4	2	3	3
kleine T.	**Nr. 17**	3	2	3	2	2	2	3	2
große T.	**Nr. 18**	2	2	3	2	2	1	2	2
eindeutig	**Nr. 19**	1	1	3	1	1	1	1	1
kleine T.	**Nr. 20**	2	2	1	2	2	2	1	1
eindeutig	**Nr. 21**	4	3	3	4	4	3	4	4
kleine T.	**Nr. 22**	4	3	3	4	4	4	3	4
kleine T.	**Nr. 23**	3	2	3	3	3	2	2	2
	Mittelwert	**2,57**	**2,35**	**2,70**	**2,57**	**2,83**	**2,26**	**2,26**	**2,52**
	Modus	**3**	**2**	**3**	**3**	**2**	**2**	**3**	**2**

***Tab. 7**: Links: Durchschnittliche Beurteilungen von ChatGPT-4o und den Expertinnen D. Gstall und M. Korner der Antworten zu „Haartrockner" aus Phase 2, siehe Tab. 2. Rechts: Einzelne Beurteilungen von ChatGPT-4o basierend auf veränderten Variablen im Prompting. Farbcode: Grün hinterlegte Beurteilungen traf ChatGPT übereinstimmend mit D. Gstall und M. Korner; gelb hinterlegte weichen um ein Niveau davon ab; rot hinterlegte weichen um zwei oder mehr Niveaustufen ab.*

Im Folgenden werden, wenn nötig, die angepassten Promptpassagen und vor allem Auszüge aus jeder Promptingvariante gezeigt, um mögliche Effekte auf Struktur und Ergebnisse vergleichend darstellen zu können. Dafür wird Antwort Nr. 9 zur Aufgabe „*Haartrockner*" von Gstall (2024) ausgewählt, da es eine der Antworten ist, die sowohl in allen Runden der Testreihe der Phase 2, von den Expertinnen und in allen sechs Promptingvarianten der Phase 3 gleich beurteilt wurde, nämlich mit Niveaustufe 3. Die Antwort Nr. 9 lautet:

„Ich würde den ProHair nehmen, weil er ein langes Kabel hat und er 2000W hat."

Die Antwort lässt zwei Kriterien erkennen, die durch eine einfache sprachliche Struktur die Wahl begründen. Gemäß dem 5-stufigen Beurteilungsmodell von Gstall (2024) ist die Einstufung in Niveau 3 damit klar nachvollziehbar.

In der **ersten Variante** von Phase 3 wurde die „*Reasoning*-Funktion" von ChatGPT aktiviert, die in der Testreihe der zweiten Phase nicht verwendet wurde. Die Einzelbeurteilung zu Antwort 9 wurde wie folgt generiert:

Antwort 9

- **Beurteilung:** Niveaustufe 3
- **Begründung:** Zwei klare Kriterien (Kabellänge, Leistung) werden mit „weil" verbunden, aber nicht weiter verglichen.
- **Feedback:**
  - *Stärken:* Fokus auf zwei wichtige Aspekte.
  - *Verbesserungsmöglichkeiten:* Vergleiche, wie lang das Kabel im Vergleich zu anderen Modellen ist, und füge das Prüfzeichen als drittes Kriterium hinzu.

In der ersten Variante generierte ChatGPT seinen Output ausnahmsweise nicht mit dem Modell *4o* wie sonst üblich, sondern mit *o4-mini*. Am 16. April 2025 wurde das Modell *o4-mini* von OpenAI veröffentlicht, eines der neuesten Modelle der o-Serie, die darauf trainiert

wurden vor der Antwortgenerierung länger nachzudenken. Diese sollen vor allem für *Reasoning*-Aufgaben leistungsfähiger und genauer reagieren (OpenAI, 2025b).

In der **zweiten Variante** wurde ChatGPT dazu aufgefordert nur einen Teil des Antwortsatzes zu beurteilen. Die zweiteilige Promptvorlage wurde dafür um einen dritten Prompt erweitert. Der zweite Prompt mit den Beurteilungsvorgaben wurde nur in Schritt 3 geändert:

> **Prompt P2**
> *„Schritt 3: Beurteile die Antworten 1-11 nach dem Beurteilungsmodell für S-Kompetenzaufgaben, das ich dir gegeben habe. (Die Antworten 12-23 beurteilst du dann im nächsten Durchgang.)"*

Die restliche Struktur von Prompt P1 & P2 wurde konstant gehalten. Nach der Generierung der ersten elf Beurteilungen von ChatGPT folgte der dritte Prompt:

> **Prompt P3**
> *„Beurteile jetzt die Antworten 12-23 aus der Aufgabe "Haartrockner" nach den gleichen, angegebenen Schritten."*

Die Einzelbeurteilung zu Antwort 9 wurde vom Chatbot wie folgt generiert:

Antwort Nr. 9

- **Beurteilung:** Niveaustufe 3
- **Begründung für die Beurteilung:** Zwei Kriterien (Kabellänge und Leistung) werden angesprochen, aber ohne klare Gegenüberstellung.
- **Feedback:** Du hast zwei sehr wichtige Kriterien berücksichtigt! Noch stärker wäre deine Antwort, wenn du auch erklären würdest, warum diese Kriterien wichtiger sind als andere.

Die **dritte Variante** ähnelt der zweiten, nur wurden hier die ursprünglichen Promptstrukturen beibehalten und die Antworten im hochgeladenen Dokument wurden geteilt. Das bedeutet der zweite Prompt mit den Beurteilungsvorgaben wurde zweimal abgeschickt, je einmal mit einer Hälfte des Antwortsatzes, Nr. 1-11 und Nr. 12-23. Die Einzelbeurteilung zu Antwort 9 wurde wie folgt generiert:

Antwort 9

- **Beurteilung:** Niveaustufe 3
- **Begründung:** Zwei relevante Kriterien (Kabellänge und Leistung) werden genannt, aber keine tiefergehende Verknüpfung oder Vergleich.
- **Feedback:** Sehr gut, dass du mehrere Kriterien nutzt! Für ein noch besseres Niveau könntest du die Argumente stärker miteinander verknüpfen oder andere Optionen vergleichen.

In der **vierten Variante** wurde versucht die Prompts P1 und P2 kürzer in ihrer Formulierung zu kürzen. Die gekürzten Prompts wurden wie folgt für die Beurteilung abgeschickt:

> **Prompt P1** (gekürzt)
> [Beurteilungsmodell hochgeladen]

„Analysiere das Beurteilungsmodell für S-Kompetenzaufgaben im hochgeladenen File. Wenn dir daran etwas unklar ist erkläre ich es dir gerne.“

Prompt P2 (gekürzt)
[Aufgabe „Haartrockner“ mit den Antworten dazu hochgeladen]
„Schritt 1: Versetze dich in die Rolle einer Physik-Lehrperson.
Schritt 2: Beurteile bitte alle Antworten zur Aufgabe „Haartrockner“ nach dem Beurteilungs-modell und gib einzelne Beurteilungen nach diesem Beispiel an:
Antwort Nr. „X“
- Beurteilung: „Niveaustufe X“
- Begründung: (Formuliere hier kurz, warum die Antwort die entsprechende Niveaustufe erreicht hat.)
- Feedback: (Formuliere hier ein sachliches und motivierendes Feedback. Es soll Stärken, Schwächen und Verbesserungsmöglichkeiten der Antwort beinhalten.)
Schritt 3: Erstelle eine zusammenfassende Tabelle aller Beurteilungen mit den Spalten: Antwort Nr., Beurteilung und Begründung.“

Die Einzelbeurteilung zu Antwort 9 wurde wie folgt generiert:

Antwort Nr. 9

- **Beurteilung:** Niveaustufe 3
- **Begründung:** Zwei Kriterien (Kabellänge, Leistung) werden genannt. Keine Vergleichsstruktur.
- **Feedback:** Deine Wahl ist gut begründet. Wenn du erklärst, warum diese Aspekte besser sind als bei den anderen Geräten, stärkst du deine Argumentation.

In der **fünften Variante** wurde auf Prompt P1 im Beurteilungsprozess verzichtet. Prompt P2 wurde direkt gemeinsam mit dem 5-stufigen Modell und der Aufgabenstellung inklusive Antwortsatz abgeschickt. Die Einzelbeurteilung zu Antwort 9 wurde wie folgt generiert:

Antwort Nr. 9

- **Beurteilung:** Niveaustufe 3
- **Begründung für die Beurteilung:** Zwei Kriterien (Kabellänge, Leistung), klar benannt, aber kein Vergleich.
- **Feedback:** Du hast gut beschrieben, warum du dich für ProHair entscheidest. Versuche, andere Modelle mit einzubeziehen, um deinen Standpunkt zu stärken.

In der **sechsten Variante** wurde versucht das generierte Feedback von ChatGPT zu erweitern, indem Prompt P2 mit den Beurteilungsangaben in Schritt 4 mit einem (hier unterstrichenen) Satz ergänzt wurde:

Prompt P2
„Feedback: (Formuliere hier ein sachliches und motivierendes Feedback. Es soll Stärken, Schwächen und Verbesserungsmöglichkeiten der Antwort beinhalten. Verdeutliche die Verbesserung mit einem Beispiel-Satz.)“

Die restliche Struktur in Prompt P1 & P2 wurde konstant gehalten. Der Chatbot generierte die Einzelbeurteilung zu Antwort 9 wie folgt:

Antwort 9

- **Beurteilung:** Niveaustufe 3
- **Begründung:** Zwei Kriterien (Leistung, Kabellänge) werden genannt.
- **Feedback:** Du hast wichtige Kriterien erkannt! Ein Vergleich mit den anderen Modellen würde deine Antwort noch überzeugender machen.
  Beispiel: „ProHair hat das längste Kabel und mehr Leistung als der günstigere SuperFön."

Für die Generierung der Beurteilungen aller 23 Antworten in dieser Variante brauchte ChatGPT zwei Outputs. Vermutlich erforderte das genauere Feedback mehr Rechenleistung.

Diskussion der Ergebnisse

Die sechs getesteten Promptingvarianten entstanden durch teils geringfügige Anpassungen der ursprünglichen Prompting-Strategie, genauer in den Prompts 1 und 2, die am Ende von Phase 1 definiert und in Phase 2 erfolgreiche getestet wurde. In jeder Variante wurde eine beschriebene Variable verändert und die restlichen konstant gehalten.

Die Anpassungen wurden mit dem Ziel erprobt, mögliche Verbesserung der Beurteilungen von ChatGPT zu erreichen und diese eventuell für die Hauptstudie zu verwenden. Vor allem die Genauigkeit der Beurteilungen und die Qualität und Struktur des Outputs sind Faktoren, die im Interesse möglicher Verbesserungen stehen.

Vergleicht man die Beurteilungen von ChatGPT aus den sechs angepassten Varianten mit denen der Expertinnen Denise Gstall und Marianne Korner in Tab. 7, so lässt sich erkennen, dass die Anteile der Übereinstimmung im Vergleich zu den Ergebnissen der Testreihe aus Phase 2 ein wenig variieren. Besonders die beiden Varianten, die ein gekürztes Prompting verwendeten – Variante 4 mit gekürzten Formulierungen in Prompt P1 und P2 und Variante 5, in der nur Prompt P2 genutzt wurde – zeigten mit 13 und 14 deckungsgleichen (= grün markierte) Beurteilungen die höchste Übereinstimmung mit den Expertinnen. Diese ist sogar höher als die Durchschnittsbeurteilungen der Testreihe aus Phase 2.
Gleichzeitig erzielte Variante 5 die einzige Gesamtbeurteilung des Antwortsatzes, in der *GPT-4o* keinen groben Ausreißer generierte. Interessanterweise führten die Varianten 1 und 2 zu den meisten groben Ausreißern und das teilweise für Antworten, die ChatGPT in allen anderen Varianten und der Testreihe genauer beurteilte. Beispielsweise stufte ChatGPT mit aktiviertem *Reasoning* Antwort 19 der Niveaustufe 3 zu, die es sonst immer mit Niveau 1 bewertete. Das *Reasoning* führt seltsamerweise nicht zu genaueren Beurteilungen, im Gegenteil. Die Häufigkeit der leichten Ausreißer bleibt dafür über alle Varianten näherungsweise konstant.

Was sagen diese Ergebnisse über die Genauigkeit von ChatGPTs Beurteilungen aus? Wurde sie erhöht? In Summe schwankt die Anzahl der deckungsgleichen Beurteilungen in allen sechs Promptingvarianten zwischen 10 und 14, die der um ein Niveau abweichenden

Beurteilungen zwischen 8 und 11 und die der um zwei oder mehr Niveaustufen abweichenden Beurteilungen zwischen 0 und 4. Es lassen sich zwischen einzelnen Varianten kleine Unterschiede beobachten, in Summe konnte jedoch keine der neuen Varianten eine deutlich höhere Übereinstimmung erzielen. Die Genauigkeit von ChatGPT konnte nur teilweise, wie diskutiert in Variante 4 und 5, verbessert werden. In Summe ist der Unterschied jedoch zu klein, um von einer gesicherten Erhöhung der Genauigkeit sprechen zu können. Das zeigen auch die Mittelwerte in Tab. 7.

Außerdem ist die Sicherheit der Beurteilungen von ChatGPT in Phase 3 kritisch zu betrachten, da jede Variante nur in einem Durchlauf getestet wurde. Die Beurteilungen beruhen im Vergleich zur Testreihe aus Phase 2 nicht auf den Durchschnitten mehrerer Beurteilungsrunden. Eine höhere oder niedrigere Genauigkeit könnte in einem Durchgang zufällig entstanden sein. Diese Möglichkeit kann nicht ausgeschlossen werden, denn wie bereits diskutiert, kann auch ChatGPT in seinen Beurteilungen schwanken und Fehler machen.

Aus dem gleichen Grund kann in Phase 3 die Konsistenz von ChatGPTs Beurteilungen nicht festgestellt werden, da jede Promptingvariante nur einmal getestet wurde. Und ein Vergleich der Beurteilungen aller sechs Varianten zur gleichen Antwort ist nicht sinnvoll, da aufgrund der veränderten Variablen kein einheitlicher Beurteilungsprozess gegeben ist und damit auch keine Konsistenz, wie jene in der Testreihe von Phase 2, untersucht werden kann.

Dennoch sind zwei Auffälligkeiten interessant, wenn man die Beurteilungen zu einer Antwort vergleicht. Nur in drei Fällen, die Antworten 2, 9 und 13, beurteilte ChatGPT in allen Varianten ident. Diese Antworten wurden auch in der Testreihe in allen fünf Runden gleich beurteilt und mit dem Prädikat *eindeutig,* bezogen auf die Sicherheit des Chatbots, versehen. Für die anderen Antworten zu „*Haartrockner*“ schwanken die Beurteilungen in Phase 3 zwischen zwei oder drei Niveaustufen. In einem Fall, Antwort 6, sogar zwischen vier Niveaustufen.

Es ist nachvollziehbar, dass Veränderungen im Prompting zu anderen Ergebnissen führen, aber trotzdem ist es erstaunlich, wie unterschiedlich die Bewertungen zu einer Antwort ausfallen, wenn nur Kleinigkeiten im Prompting verändert werden. Das lässt wieder die Frage aufkommen, wie vertrauenswürdig die Beurteilung von S-Kompetenz mit ChatGPT ist.

Nun zu den Strukturen der Einzelbeurteilungen, konnten diese qualitativ verbessert werden? Die *Reasoning*-Funktion ermöglicht es ChatGPT komplexere Aufgaben besser zu lösen, da es aufgefordert wird vor dem Antworten länger nachzudenken. Die KI erzielte mit dieser Funktion keine genaueren Ergebnisse wie in Tab. 7 ersichtlich. Dafür verbesserte sich die Struktur der Einzelbeurteilungen ein wenig, wie der Auszug zu Antwort 9 zeigt. ChatGPT generiert ganze Sätze und formuliert das Feedback klarer in Stärken und Verbesserungsmöglichkeiten. Dafür verliert es an motivierenden Passagen und bleibt sachlicher.

Die Varianten 2 und 3 verfolgten das gleiche Ziel auf verschiedenen Wegen: eine Reduzierung der Datenmenge, um die Komplexität des Auftrags zu reduzieren und genauere Ergebnisse zu erzielen. Da jedes GPT-Modell über ein *Tokenlimit* verfügt, was einfach gesagt die Denkprozesse und verarbeitbare Textmenge pro Auftrag beschränkt, soll die Aufteilung der Beurteilungen auf zwei Pakete dazu beitragen, dass ChatGPT für die Beurteilung einer Antwort mehr Kapazitäten übrighat. ChatGPT generierte die Einzelbeurteilungen in den Varianten 2 und 3 in der üblichen Struktur mit leichten Verbesserungen. Begründung und Feedback wurden für alle Antworten in ganzen Sätzen, sachlich und motivierend verfasst. Eine Abnahme der Ausführlichkeit des Outputs konnte nicht in dem Ausmaß beobachtet werden, wie in den Durchläufen, wo alle 23 Antworten auf einmal bewertet wurden. In Summe führt die aufgeteilte Generierung der Beurteilungen zu einem besseren Output, aber nicht zu genaueren Beurteilungen.

Die verkürzten Promptingvarianten 4 und 5 führten zu keinen Output-Verbesserungen, belegen aber, dass kürzere Prompts zu ähnlichen Ergebnissen führen. Wie die Auszüge zu Antwort Nr. 9 oben zeigen, bleibt der Output im Wesentlichen unverändert. Ein logischer Grund dafür könnte jedoch sein, dass ChatGPT nach den vielen Testungen diesen Auftrag wiedererkennt und auch mit kürzeren Prompts ein zufriedenstellendes Ergebnis generiert. Aufgrund der Theorielage zu Prompting Techniken ist nicht davon auszugehen, dass ein knapp formulierter Prompt beim ersten Mal in dieser Form funktionieren würde. Interessanterweise verbesserte sich durch die Kürzungen jedoch die Genauigkeit der Beurteilungen. Ob das ein Zufall war oder nicht kann die einmalige Durchführung nicht belegen.

Eine klare Aufwertung des Beurteilungs-Outputs konnte die sechste Variante erreichen. Nur das Hinzufügen der Anweisung „*Verdeutliche die Verbesserung mit einem Beispiel-Satz*." in den ursprünglichen One-shot Prompt wertete das Feedback deutlich auf. Eine stichprobenartige Überprüfung zeigte, dass ein Großteil der Antworten sinnvolle Beispielsätze als Feedback erhielt, wie im Beispiel für Antwort 9 im obigen Chatauszug sichtbar. Für Antworten, die mit Niveau 1 bewertet wurden, generierte der Chatbot keinen Beispielsatz als Verbesserung, sondern einen Tipp für zukünftige S-Aufgaben. ChatGPT schrieb zum Beispiel zu Antwort Nr. 3 folgendes Feedback:

> Antwort 3: „*Ich werde Esra den ProHair-Föhn empfehlen. Warum? Begründung: Ich finde das der Preis für 24,99€ nicht zu teuer für so einen Föhn ist. Zweitens ist die Leistung bei 2000W, somit werden die Haare schneller trocken, aber auch die Energie wird bisschen gespart, als im Vergleich zu den anderen eine geringere Leistung hat und es durch ÖVE-Prüfzeichen geprüft wurde.*"
>
> **ChatGPT**
>
> „*Feedback: Sehr gut! Du hast nicht nur wichtige Kriterien verwendet, sonder auch die Optionen gegeneinander abgewogen. Tipp: Achte in Zukunft auf sprachliche Klarheit – bei längeren Sätzen kann man schnell en Überblick verlieren.*"

Nach der Studie von Hattie (2009) ist Feedback eines der wirksamsten Instrumente, um den schulischen Lernerfolg zu steigern. Variante 6 konnte zeigen, dass das Feedback von

ChatGPT weiter verfeinert werden kann. Auf diese Option wird in der Forschung der Hauptstudie sehr wahrscheinlich wieder zurückgegriffen. Den einzigen Nachteil, den diese Variante mitbringt, ist eine erhöhte Komplexität für ChatGPT in der Bearbeitung. Wenig überraschend konnte die KI die Beurteilungen nicht alle in einem Output generieren, sondern nur bis Antwort Nr. 10, hier wurde vermutlich das Tokenlimit von *GPT-4o* erreicht. Die restlichen Beurteilungen können aber einfach im nächsten Prompt mit einer kurzen Anweisung wie „*Setze fort*." abgeschlossen werden. Eine auf mehrere Outputs aufgeteilte Beurteilung stellt kein Problem dar.

Zusammenfassend zeigen manche der getesteten Varianten leichte Verbesserungen bezüglich struktureller Aspekte von ChatGPTs Beurteilungen, die potenziell in der Hauptstudie genutzt werden können. Keine der Varianten konnte jedoch eine deutliche Verbesserung in Bezug auf die Genauigkeit der Beurteilungen im Vergleich zur Testreihe aus Phase 2 erzielen.

Feedback aus dem Forschungsseminar

Ende Mai 2025 wurde der aktuelle Forschungsstand der Arbeit in einem Forschungsseminar der Physikdidaktik der Universität Wien vorgestellt. Die Ergebnisse der Pilotstudie, mit Fokus auf jene aus Phase 2, wurden diskutiert und die Seminarteilnehmer:innen sollten ihr Feedback dazu geben. Die wichtigsten Teile aus dem Gruppengespräch werden nun angeführt.

Zu Beginn des Seminars wurden die Teilnehmer:innen grob in die Forschung und die aktuelle Datenlage eingeführt. Währenddessen wurden ein paar spezifische Fragen gestellt, zum Beispiel zur Berechnung der Durchschnittsbeurteilungen oder zum Prompting. Am Ende des Vortrags folgerte ich, dass ich auf Basis der bisherigen Datenlage aktuell noch Zweifel habe, ob der Einsatz von ChatGPT für die Beurteilung von S-Kompetenz viabel ist. Die mittelmäßige Genauigkeit und Konsistenz von ChatGPTs Beurteilungen anhand des 5-stufigen Beurteilungsmodells (Gstall, 2024) überzeugten meiner Ansicht nach nicht, um den Einsatz empfehlen zu können. Diesen kritischen Blick teilten die Seminarteilnehmer:innen jedoch nicht, im Gegenteil.

In Summe zeigten sich die Physikdidaktiker:innen sehr erstaunt über die Durchschnitts-beurteilungen, die der Chatbot in der Testreihe zur Aufgabe „*Haartrockner*" generiert hat (jene zu „*Die neue Heizung*" wurden aus Zeitgründen im Seminar nicht präsentiert). Die erzielten Übereinstimmungen mit den Beurteilungen der Expertinnen wurden als beeindruckender Erfolg betrachtet. Nämlich in Anbetracht dessen, dass ChatGPT im Vergleich zu anderen KI-Modellen, wie *Gemini* oder *Copilot*, tendenziell fehleranfälliger und halluzinierender sei und dennoch so gute Ergebnisse liefert. Weiters wurde vermutet, dass eine Gruppe von Menschen in einer ähnlichen konzipierten Testreihe im Durchschnitt keine genaueren Ergebnisse als ChatGPT schaffen würde. Dieses Vorhaben wäre interessant für die Hauptstudie.

Mehrere Meinungen im Seminar machten daraufhin klar, dass nicht unbedingt zur Frage steht, ob die Beurteilung mit ChatGPT gut genug funktioniert, sondern eher, wie der gesamte Prozess (ChatGPT selbst, das Prompting oder die Kriterien des Modells) adaptiert

werden kann, um die Beurteilungsleistung noch weiter zu verbessern. Die anschließende Diskussion ergab viele Möglichkeiten für weitere Nachforschungen, die in die Hauptstudie einfließen könnten.

Als konkrete Idee für das Prompting wurde vorgeschlagen, dass man ChatGPT die Beurteilung z.B. gleich zehn Mal innerhalb eines Prompts auftragen könnte und es daraus einen Mittelwert berechnen lässt. Wobei sehr fraglich ist, ob *GPT-4o* diesen Auftrag innerhalb seines Tokenlimits bewältigen kann. Eine andere Meinung schloss sich dem mehrfachen Beurteilen lassen durch KI an und sprach von einer vorprogrammierten App, die denkbar wäre. Diese könnte ChatGPT die Beurteilungen 10-, 20- oder 30-mal durchführen und Wahrscheinlichkeiten berechnen lassen, die einer Lehrperson z.B. mitteilen „Niveau X ist mit einer Konfidenz von 85% die Beurteilung für Antwort Y". Auf Wahrscheinlichkeiten aus vielen Durchgängen ließe sich mehr vertrauen, als aus wenigen und sollte diese Konfidenz niedrig sein, kann die Lehrperson entscheiden, eine einzelne Antwort lieber selbst zu beurteilen. Zur Verbesserung des Beurteilungsmodells von Gstall (2024) bestünde auch die interessante Möglichkeit ChatGPT eine eigene Version davon erstellen zu lassen und diese zu erproben. ChatGPT ist laut eigenen Angaben dazu in der Lage Bewertungsmatrixen zu erstellen. Möglicherweise kann ein KI-Modell mit einer KI-generierten Matrix besser beurteilen.

Weiters wurde vorgeschlagen, andere Beurteilungswege in eigenen Testreihen zu untersuchen, um eventuelle Unterschiede in der Genauigkeit und Konsistenz festzustellen. Beispielsweise könnte man die Beurteilung fünf Mal ohne Prompting-Strategie generieren lassen oder auf einem anderen Gerät mit einer neuen IP-Adresse Tests durchführen. Letzteres würde Hinweise dazu liefern, ob die Arbeitsgeschichte eines Geräts Einfluss auf die Ergebnisse hat. Denkbar wäre auch eine Testreihe ohne Beurteilungsmodell durchzuführen, wo ChatGPT ohne Vorgaben eine Bewertung treffen soll. Dies wurde in der ersten Phase der Pilotstudie mit einzelnen Antworten versucht. Untersuchen ließe sich auch, wie ChatGPT im Vergleich zu Fachpersonal abschneidet, dafür müsste eine gleich konzipierte Testreihe mit Menschen durchgeführt werden.

Recht naheliegend, tauchte auch die Frage auf, ob weitere KI-Modelle getestet wurden. In der Planung dieser Masterarbeit wurde keine Untersuchung mit einem weiteren KI-Modell neben ChatGPT vorgesehen. Eine mögliche Testung mit bekannten KI's wie *Gemini* und *Copilot* oder speziell trainierten KI-Angeboten für den Schulbereich, wie Anwendungen von *MagicSchool* und *Fobizz*, wäre für einen umfassenden Vergleich interessant, aber mit viel Aufwand verbunden.

Die Teilnehmer:innen fanden den Aufbau der Pilotstudie gut gelungen und meinten, dass sie den Einsatz von ChatGPT als Beurteilungshilfe ausreichend untersuchen kann. Auch die Angaben, wie ChatGPT die generierten Beurteilungen strukturieren soll sprach den Teilnehmer:innen zu. Aus den Rückmeldungen des Seminars entstand der Eindruck, dass die aktuelle Prompting-Strategie mit den Prompts P1 und P2 in ihrer Grundform weiterverwendet werden können.

Die aus meiner Sicht zu unsicheren Ergebnisse von ChatGPT wurden mehrmals als gut genug eingestuft. Es wurde zwar beigeräumt, dass es ein Problem darstellen könnte, wenn ChatGPT etwa die Anzahl der Kriterien in einer Antwort falsch erkennt, wie es vereinzelt der Fall war. Die Tatsache, dass ChatGPT gesamt betrachtet nur teilweise Fehler und grobe Ausreißer generiert, könne aus Sicht der Seminarteilnehmer:innen verkraftet werden, wenn sich das Ausmaß, wie im der Testreihe, in Grenzen hält. Vor allem, wenn man bedenkt, dass diese Beurteilung meist nur für einzelne S-Kompetenzaufgaben gilt. Hier sei eine kleine Fehlertoleranz akzeptabel, da es nicht um wichtige Beurteilungen wie Jahresnoten geht. Abgesehen davon können Lehrpersonen auch Fehler machen meinten die Teilnehmer:innen, also müsste man die Fehlerquote der KI nicht als Verschlechterung interpretieren.

Bezüglich der abweichenden Beurteilungen kam im Seminar auch der Gedanke auf, dass die getesteten Schüler:innen-Antworten aus der Unterstufe zu Gstalls (2024) „*Haartrockner* und „*Die neue Heizung*“ möglicherweise zu kurz für ChatGPT sind. Längere Texte eignen sich vielleicht besser für eine Analyse durch die KI bzw. würden die Genauigkeit verbessern. Sollten für die Hauptstudie neue Aufgaben von Schüler:innen beantwortet werden, planmäßig die von Hackner (2025), wird versucht für diese Aufgaben längere Antworten einzufordern.

Den Teilnehmer:innen Seminars ist ebenfalls aufgefallen, dass die Unsicherheit in ChatGPTs Beurteilungen in den mittleren Niveaustufen zunimmt, was möglicherweise mit den Kriterien im Modell korreliert und Fragen für Änderungen in 5-stufigen Modell von Gstall (2024) aufwirft. Diese Unsicherheit wurde als normal eingestuft, da auch für Menschen die besten und schlechtesten Leistungen leichter zu beurteilen sind als mittelmäßige, die nun mal in den mittleren Niveaus landen.

Gegen Ende des Seminarvortrages wurden noch einige moralische und ethische Themen beim Einsatz von KI für die Beurteilung angesprochen, etwa der Umgang mit sprachlich stärkeren und schwächeren Schüler:innen bei Bewertungsaufgaben, möglicher Datenschutzprobleme beim Hochladen von Schüler:innen-Antworten oder wie es von den Lernenden aufgenommen wird, dass sie von einer KI beurteilt werden. Diese Punkte können in der abschließenden Diskussion dieser Arbeit einfließen. Wichtig zu diesen Themen war zudem für alle Anwesenden, dass die Bewertung der Lehrperson stets die Grundlinie für die Beurteilung von S-Kompetenz bleiben sollte, unabhängig davon wie gut der Beurteilungsprozess mit ChatGPT eines Tages sein könnte.

Die Diskussion im Seminar ließ viele Ideen und Ansätze für das Weitergehen der Forschung aufkommen. Die meisten der beschriebenen Vorschläge können vermutlich nicht verfolgt werden, da der Aufwand für eine Masterarbeit zu groß wäre. Sie werden vorerst nur ein Ausblick auf nachfolgende Forschungen bleiben. Abschließend trat im Seminar der Konsens auf, dass die Beurteilungsleistung von ChatGPT erstaunlich gut funktioniert und Potenzial zeigt. Die Pilotstudie konnte wichtige Erkenntnisse zum Beurteilen von S-Kompetenz mit ChatGPT und dem 5-stufigen Modell von Gstall (2024) liefern. Für die Hauptstudie wäre der Fokus auf das 3-stufige Modell (2025) aus Sicht des Seminars jedoch sinnvoller, da die Teilnehmer:innen meinen, dass die Bewertungen von ChatGPT anhand des 3-stufigen

Modells von Hackner (2025) besser funktionieren. Und auch deswegen, weil das 5-stufige Modell für die tägliche Arbeit zu fein strukturiert sei. Zusätzlich sahen viele eine große Chance in der Generierung von Feedback für Schüler:innen. Dieser Möglichkeit sollte die Hauptstudie auch weiter nachgehen, da in der Generierung von gutem Feedback sehr viel Zeitersparnis stecken könnte.

5.4. Zusammenfassende Erkenntnisse der Pilotstudie

Die Pilotstudie konnte anhand vieler explorativer Erprobungen belegen, dass *ChatGPT-4o* mit bestimmten Vorgaben Bewertungen zur S-Kompetenz generieren kann. Mit geeigneten Prompt Engineering Techniken, hier eine zweiteilige Promptvorlage mit *One-shot* Komponenten und nach Struktur des *PROPER*-Frameworks, war es möglich strukturierte Beurteilungen anhand der Richtlinien des 5-stufigen Beurteilungsmodell zur S-Kompetenz von Gstall (2024) mithilfe von ChatGPT zu erzielen. Trotz mittelmäßiger Genauigkeit und Konsistenz der Beurteilungen sowie Fehlern in ChatGPTs Outputs können die Ergebnisse aus mehreren Gründen als Erfolg interpretiert werden. Ein Hauptargument für die positive Sicht auf ChatGPTs Leistung ist, dass die Beurteilung von S-Kompetenz grundsätzlich schwierig ist und Lehrkräfte vermutlich keine besseren Ergebnisse erreichen würden. Außerdem können die Ergebnisse positiv interpretiert werden, wenn bedacht wird, dass ChatGPT im Vergleich zu anderen KI-Modellen fehleranfälliger und womöglich ungeeigneter für den Einsatz als Beurteilungshilfe ist.

Während der Pilotstudie herrschten große Zweifel daran, ob die Beurteilungen des Chatbots verlässlich genug sind. Meiner Ansicht nach waren die Bewertungen, besonders aus Phase 2, zu ungenau und inkonsistent bei mehreren Durchgängen. Teilweise sind Reaktionen von *ChatGPT-4o* nicht nachvollziehbar oder fehlerhaft. Die Tatsache, dass sich die internen Rechenprozesse nicht beobachten lassen und nicht untersucht werden kann, wie ChatGPT zu seinen Ergebnissen kommt, ließ den Einsatz als Beurteilungshilfe nach Einschätzung des Autors nicht vertrauenswürdig erscheinen.

Die Rückmeldungen des Seminars haben diese Sicht jedoch zum Teil geändert. In Anbetracht dessen, dass dieser Bewertungsauftrag eine komplexe Aufgabe für ChatGPT ist, können seine Leistungen gesamt betrachtet solide und beeindruckend gewertet werden. Außerdem scheint der Beurteilungsprozess mithilfe der zweiteiligen Promptvorlage, auch für Lehrpersonen ohne viel Vorwissen zu KI und Prompting, einfach und schnell durchführbar zu sein. Alles, was es dazu braucht, sind nach aktuellem Stand das 5-stufige Beurteilungsmodell von Gstall (2024), eine abgetippte Aufgabenstellung inklusive Antworten, die beurteilt werden sollen, die erprobten Prompts P1 und P2 und ein (kostenloser) Zugang zu ChatGPT. Die Unvermeidbarkeit, dass ein KI-Sprachmodell wie ChatGPT Fehler jeglicher Art machen kann, seien es Halluzinationen oder Schwankungen, muss wohl zum jetzigen Stand der KI-Systeme hingenommen werden. Solange die Fehlerquote niedrig bleibt, spricht diese nicht gegen den unterstützenden Einsatz ChatGPTs als Beurteilungshilfe für die Bewertung der Bewertungskompetenz.

Eine zentrale Frage zu Beginn der Forschung war des Weiteren, ob ChatGPT objektivere Beurteilungen zu den Antworten treffen kann als Lehrkräfte. Diese Frage kann nicht wirklich beantwortet werden, da nicht sicher ist, ob ChatGPT selbst objektiv bewertet. Dies lässt sich aus den Phasen der Pilotstudie schwer überprüfen, da das 5-stufige Beurteilungsmodell von Gstall (2024) bereits Lücken für Objektivität aufweist und, weil eben kein Einblick in die Rechen- und Gewichtungsprozesse von ChatGPT gegeben ist.

5.5. Folgerungen für die Hauptstudie

Fest steht nach der Pilotstudie, dass im Rahmen der ausprobierten technischen Möglichkeiten solide Bewertungen nach dem 5-stufigen Modell von Gstall (2024) mit *ChatGPT-4o* generiert werden können. Die Ergebnisse des Chatbots zeigen Limitationen, aber auch Verbesserungsmöglichkeiten und Potenzial. Genauere Beurteilungen können von ChatGPT mit dem 5-stufigen Modell wohl nicht erreicht werden. Möglicherweise eignet sich das kürzere 3-stufige Modell von Hackner (2025) mehr und ermöglicht genauere und konsistentere Bewertungen von ChatGPT. Daher scheint für die weitere Forschung ein Fokus auf das 3-stufige Modell, wie geplant sinnvoller.

6. Hauptstudie

Dieses Kapitel dokumentiert Ziel, Aufbau, Durchführung und Ergebnis der Hauptstudie, die als zweite, vertiefende Studie zur Beantwortung der Forschungsfragen beitragen soll.

6.1. Zielsetzung und Methodik

Die Pilotstudie zeigte, dass ChatGPT-Beurteilungen mit dem 5-stufigen Modell von Gstall (2024) praktikabel und reproduzierbar sind, aber der Prozess mit diesem Beurteilungsmodell in der Validität der Bewertungen an seine Grenzen stößt. Die Hauptstudie fokussiert sich nun auf die Erprobung des 3-stufigen Modells zur Beurteilung der Bewertungskompetenz von Hackner (2025) mit ChatGPT. Die Forschungsarbeit von Hackner ist zum Zeitpunkt der Hauptstudie noch nicht abgeschlossen, doch das 3-stufige Modell wurde, in der Version wie in Kap 2.4.2 gezeigt, für die Beurteilung freigegeben und in dieser Form für die Hauptstudie der vorliegenden Arbeit verwendet.

Mit den gesamten Erkenntnissen aus der Pilotstudie soll eine Testung der Beurteilungsleistung von *ChatGPT-4o* anhand des 3-stufigen Modells durchgeführt werden. Dazu werden mit neuen Aufgabenstellungen und originalen Antworten von Schüler:innen Testreihen, ähnlich wie die in Phase 2, mit ChatGPT und *zusätzlich* mit Physiklehrpersonen durchgeführt. Unmittelbar vergleichbare Bewertungsleistungen von Mensch und KI ermöglichen Rückschlüsse auf die Genauigkeit, Reproduzierbarkeit und Sicherheit von ChatGPT. So lässt sich feststellen, ob der Chatbot in diesen Faktoren schlechter, gleich gut oder besser als Menschen abschneidet. Diese Indizien sind maßgeblich für die Evaluierung von ChatGPT als Beurteilungshilfe und konnten in der Pilotstudie nicht eindeutig geklärt werden.

Weiters sollen die Testreihen der Hauptstudie doppelt, mit zwei verschiedenen Promptingvarianten, durchgeführt werden, um auszuloten, wie stark sich die Formulierung und Struktur der Prompts, auf die Ergebnisse auswirken. Dazu werden die Prompts der Pilotstudie, P1 & P2 (siehe S. 56), einmal in leicht angepasster Form (H1 & H2) und einmal deutlich veränderter Form (H3 & H4) für die Hauptstudie weiterentwickelt. Mehr zu den Hauptstudien-Prompts am Ende des nächsten Kapitels.

6.2. Aufbau, Durchführung und angepasstes Promptdesign

In der Hauptstudie wird die Oberstufen-Version des 3-stufigen Beurteilungsmodells von Hackner (2025) verwendet. Weiters wird auf neue Aufgabenstellungen und zugehörigen Antworten von Schüler:innen zurückgegriffen, die für die Evaluierung des 3-stufigen Modells erstellt wurden: Die drei Aufgabenstellungen „*Windenergie*“, „*Musikbox*“ und „*Schilddrüsenuntersuchung*“ wurden von Hackner (2025) als offene S-Kompetenzaufgaben auf Niveau der Sekundarstufe II entworfen. Für die Durchführung der Hauptstudie wurden diese drei Aufgaben von Schüler:innen der 11. Schulstufe in Wiener Schulen bearbeitet und

die Leistungen von fünf freiwilligen Physiklehrpersonen nach dem 3-stufigen Modell bewertet. Zu jeder Aufgabenstellung wurden einige Antworten ausgewählt, die eine ausgewogene Testung mit ChatGPT ermöglichen. Sie stellen eine Mischung aus eindeutig bis schwer zuordbaren Antworten nach dem 3-stufigen Modell dar. Mit diesen Materialien, dem 3-stufigen Modell, den Aufgaben „*Windenergie*“, „*Musikbox*“ und „*Schilddrüsen-untersuchung*“ mit einigen ausgewählten Antworten und der angepassten, zweiteiligen Promptvorlage, sollen neue Testreihen mit ChatGPT durchgeführt werden.

Zu jeder Aufgabenstellung wird eine Testreihe aus fünf unabhängigen Beurteilungsrunden durchgeführt, in denen ChatGPT die abgetippten Antwortsätze nach dem 3-stufigen Modell bewerten soll. Wegen des Uploadlimits von bis zu drei Dateien pro 24 Stunden in der kostenfreien Version von ChatGPT, müssen die Testreihen hintereinander, innerhalb von zehn Tagen, generiert werden.

Von Tag 1 bis Tag 5 werden die fünf Beurteilungsrunden zu „*Windenergie*“ und „*Musikbox*“ nach einheitlichem Ablauf generiert:

1. Ein neuer Chat in ChatGPT wird gestartet.
2. Das 3-stufige Beurteilungsmodell wird hochgeladen und mit Prompt H1 gesendet.
3. *ChatGPT-4o generiert eine Analyse des Modells.*
4. Die Aufgabenstellung „*Windenergie*“ inklusive dem Antwortsatz wird hochgeladen und mit Prompt H2 gesendet.
5. *ChatGPT-4o generiert Beurteilungen zu „Windenergie“.*
6. Die Aufgabenstellung „*Musikbox*“ inklusive dem Antwortsatz wird hochgeladen und mit Prompt H2 gesendet.
7. *ChatGPT-4o generiert Beurteilungen zu „Musikbox“.*
8. Die Beurteilungen zu beiden Antwortsätzen werden dokumentiert.

Von Tag 6 bis Tag 10 werden die fünf Beurteilungsrunden zu „*Schilddrüsenuntersuchung*“ ebenso einheitlich generiert:

1. Ein neuer Chat in ChatGPT wird gestartet.
2. Das 3-stufige Beurteilungsmodell wird hochgeladen und mit Prompt H1 gesendet.
3. *ChatGPT-4o generiert eine Analyse des Modells.*
4. Die Aufgabenstellung „*Schilddrüsenuntersuchung*“ inklusive dem Antwortsatz wird hochgeladen und mit Prompt H2 gesendet.
5. *ChatGPT-4o generiert Beurteilungen zu „Schilddrüsenuntersuchung“.*
6. Die Beurteilungen werden dokumentiert.

Während der Beurteilungsrunden werden Angebote und Rückfragen von ChatGPT ignoriert und es wird kein Feedback auf die Outputs gegeben. Sollte der Chatbot *Erinnerungen* anlegen, werden diese ebenfalls ignoriert und weder verwaltet noch gelöscht.

Nach Tag 10 wurden zu jeder Aufgabe fünf Beurteilungsrunden mithilfe der Prompts H1 und H2 generiert und die drei ersten Testreihen abgeschlossen. Nach dem gleichen Ablauf werden an den nachfolgenden zehn Tagen drei weitere Testreihen mit den Prompts H3 und

H4 generiert. Nach 20 Tagen wurden dementsprechend sechs KI-Testreihen mit ChatGPT durchgeführt. Das entspricht 300 Einzelbewertungen nach dem 3-stufigen Modell.

Aus den Beurteilungen der fünf Lehrpersonen werden drei menschliche Testreihen erstellt. Die Durchschnittsbeurteilungen in diesen Testreihen beruhen im Vergleich zu ChatGPT auf verschiedenen Personen und nicht aus mehreren Durchgängen.

Für jede Antwort zu den Aufgabenstellungen „*Windenergie*“, „*Musikbox*“ und „*Schilddrüsen-untersuchung*“ resultieren aus dieser Durchführung *eine* Durchschnittsbeurteilung von Physiklehrpersonen und *zwei* Durchschnittsbeurteilungen von ChatGPT. Die Durchschnittsbeurteilungen der KI können mit jener der Menschen verglichen werden und lassen erkennen, wie genau und konsistent ChatGPT im Bewerten von S-Kompetenz anhand des 3-stufigen Modells (Hackner, 2025) im Vergleich zu Physiklehrkräften ist.

In der Pilotstudie wurde die Sicherheit von ChatGPT anhand seiner Konsistenz diskutiert, sie konnte jedoch nicht direkt untersucht werden. In der Hauptstudie lässt sich die Unsicherheit der KI berechnen und mit jener der Menschen vergleichen. Dazu wird aus den fünf Beurteilungen einer Antwort die Spannweite, der am weitesten auseinanderliegenden Einstufungen bestimmt. Aus dem Vergleich der erhaltenen Werte lassen sich Aussagen über die Unsicherheit des Chatbots beim Bewerten von S-Kompetenz treffen.

Anpassung der Prompts P1 & P2 für die Hauptstudie

In der Hauptstudie wird die gleiche Prompting-Strategie angewandt, wie in der Pilotstudie. Allerdings ist nicht zu erwarten, dass die Prompts P1 & P2 (siehe S. 56), die auf das 5-stufige Modell von Gstall (2024) ausgelegt sind, problemlos für die Generierung von Beurteilungen mit dem neuen 3-stufigen Modell von Hackner (2025) funktionieren. Bevor die Hauptstudie wie geplant durchgeführt wird, müssen die Prompts P1 & P2 mit dem 3-stufigen Modell getestet und anschließend iterativ angepasst werden. In diese Anpassungen können auch Erkenntnisse, die in der Phase 3 der Pilotstudie gesammelt wurden, einfließen.

Ähnlich der Herangehensweise von Phase 1 der Pilotstudie, wurden in einzelnen Testläufen Beurteilungen mit *ChatGPT-4o* anhand des neuen 3-stufigen Modells und den Pilotstudien-Prompts P1 & P2 generiert. Die Annahme, dass die Anwendung der Prompts P1 & P2 auf das 3-stufige Modell keine modellkonformen und qualitativen Beurteilungen erzeugt, bestätigte sich schnell.

Ein großer Unterschied zwischen den Modellen von Gstall (2024) und Hackner (2025) ist die Beurteilung. Während in Gstalls Modell die inhaltlichen und sprachlichen Merkmale in einer Niveaustufe zusammenfließen, werden in Hackners Modell beide Aspekte separat mit einer Niveaustufe bewertet und das Gesamtniveau der Antwort ergibt sich dann mithilfe der Gewichtung 2:1 aus inhaltlichen Anforderungen und sprachlichen Merkmalen. Außerdem unterscheidet sich die Niveaustufenanzahl der beiden Modelle. Die Prompts P1 & P2 sind nicht auf diese differenzierte Beurteilung gemäß Hackners Modell formuliert und müssen zumindest mit dieser Information für den Chatbot erweitert werden.

Durch iteratives Vorgehen (und auch mit der Hilfe von ChatGPT) wurden die Prompts schrittweise ergänzt, umformuliert und verbessert, um auf das 3-stufige Modell abgestimmt

zu sein. Das Ergebnis sind zwei neue Prompt-Varianten: H1 & H2 sowie H3 & H4. Die erste Variante (H1 & H2) gleicht Großteils den Pilotstudien-Prompts (P1 & P2), sie wurde nur mit der Angabe auf die getrennte Beurteilung des 3-stufigen Modells erweitert. Die zweite Variante (H3 & H4) verfolgt auch die ursprüngliche Prompting-Strategie, die Prompts unterscheiden sich aber deutlich in ihrer Struktur und den Formulierungen. Diese sind noch etwas detaillierter und präziser.

Prompt-Variante H1 & H2

Prompt H1

[Beurteilungsmodell hochladen]
„Ich bin eine Lehrperson für Physik und möchte deine Unterstützung. Bitte beurteile für mich Antworten von Schüler:innen anhand eines vorgegebenen Beurteilungsmodells. Ich lade dir das Modell als File hoch. Analysiere das Modell ausführlich. Wenn dir etwas daran unklar ist, erkläre ich es dir gerne."

Prompt H2

[Aufgabenstellung und abgetippte Antworten dazu hochladen]
„Beurteile bitte die Antworten der Schüler:innen zur Aufgabe „[NAME]" im hochgeladenen File. Gehe dabei nach folgenden Schritten vor und gib alle deine Ergebnisse in einem Output an:

Schritt 1: Versetze dich in die Rolle einer Physik-Lehrperson.

Schritt 2: Analysiere die Aufgabe und alle Antworten dazu in dem PDF. Achte darauf, ob die Aufgabe einen oder mehrere Unterpunkte (z. B. a, b, ...) enthält. Wenn ja, lass alle Unterpunkte in die Beurteilung der Aufgabe einfließen / Beurteile nur die Antworten zu Unterpunkt X der Aufgabenstellung.

Schritt 3: Beurteile alle Antworten nach dem Beurteilungsmodell für S-Kompetenzaufgaben, das ich dir gegeben habe.

Schritt 4: Beurteile jede Antwort nach inhaltlichen und sprachlichen Anforderungen des Modells an. Berechne zusätzlich das Gesamtniveau pro Antwort:
- *Gewichtung: Inhalt : Sprache = 2 : 1*
- *Formel: (2 × Inhaltsniveau + Sprachniveau) ÷ 3*

Schritt 5: Formuliere die Beurteilung jeder Antwort nach diesem Beispiel:
Antwort Nr. „X"
- *Beurteilung Inhaltlich: „Niveaustufe X"*
- *Beurteilung Sprachlich: „Niveaustufe X"*
- *Gesamtniveau: „Niveaustufe X.X"*
- *Begründung für die Beurteilung: Formuliere hier kurz, warum die Antwort die entsprechende Niveaustufe erreicht hat.*
- *Feedback: (Formuliere hier ein sachliches und motivierendes Feedback. Es soll Stärken, Schwächen und Verbesserungsmöglichkeiten der Antwort beinhalten.)*

Schritt 6: Erstelle am Ende eine zusammenfassende Tabelle aller Beurteilungen mit diesen Spalten: | Antwort Nr. | Niveaustufen (Inhalt / Sprache) | Gesamtniveau | Kurze Begründung |"

Prompt-Variante H3 & H4

Prompt H3

"[Beurteilungsmodell hochladen]
Ich bin Physiklehrkraft und brauche deine Unterstützung zur Bewertung der S-Kompetenz von Schüler:innen.
Grundlage ist ein zweidimensionales Beurteilungsmodell mit getrennten Niveaus für inhaltliche Anforderungen und sprachliche Merkmale (0/–, 1/~, 2/+).
Ich lade dir das Modell als Datei hoch.
Bitte analysiere das Modell sorgfältig und gib mir Rückmeldung, falls etwas unklar ist.
Anschließend möchte ich Schüler:innenantworten von dir nach diesem Modell beurteilen lassen – möglichst konsistent, genau und modellkonform.
Ich gebe dir dann genaue Anweisungen zum Bewertungsformat."

Prompt H4

„[Aufgabenstellung und abgetippte Antworten dazu hochladen]
Beurteile die Schüler:innenantworten zur S-Kompetenzaufgabe „[Aufgabentitel]" aus dem hochgeladenen File anhand des Beurteilungsmodells, das ich dir bereits gegeben habe.

Vorgehensweise:

1. *Rolle einnehmen: Versetze dich in die Rolle einer Physiklehrkraft, die Schüler:innen fair, konstruktiv und modellkonform beurteilt.*
2. *Aufgabenverständnis: Analysiere die Aufgabenstellung und die darin enthaltenen Auswahlkriterien. Achte auf mögliche Unterpunkte (z. B. a, b, ...) der Aufgabe. Wenn es welche gibt, lass alle Unterpunkte in die Beurteilung der Aufgabe einfließen / Beurteile nur die Antworten zu Unterpunkt X der Aufgabenstellung.*
3. *Beurteilung: Beurteile jede Antwort nach dem zweidimensionalen Modell:*
   - *Inhaltliche Anforderungen (Niveaustufen 0/–, 1/~, 2/+)*
   - *Sprachliche Merkmale (Niveaustufen 0/–, 1/~, 2/+)*
4. *Berechne zusätzlich das Gesamtniveau pro Aufgabe:*
   - *Gewichtung: Inhalt : Sprache = 2 : 1*
   - *Formel: (2 × Inhaltsniveau + Sprachniveau) ÷ 3*
5. *Ausgabeformat: Gib für jede Antwort folgende strukturierte Beurteilung aus: Antwort Nr. X*
   - *Beurteilung Inhaltlich: Niveaustufe X*
   - *Beurteilung Sprachlich: Niveaustufe X*
   - *Gesamtniveau: X.X*
   - *Begründung: Kurze, klare Begründung, warum diese Niveaustufen erreicht wurden.*

- *Feedback: Sachlich und motivierend. Gehe ausführlich auf Stärken und Schwächen ein und schreibe mindestens eine konkrete Verbesserungsidee anhand eines Beispiels.*

6. *Zusammenfassung: Erstelle eine zusammenfassende Tabelle aller Beurteilungen mit den Spalten:*
   | Antwort Nr. | Niveaustufen (Inhalt / Sprache) | Gesamtniveau | Kurze Begründung |

   Achte durchgehend auf eine konsistente Struktur, klare Argumentation, fachliche Genauigkeit und eine wertschätzende Rückmeldung! Generiere die Feedbacks ausführlich und teile, wenn nötig, dafür deine Ausgabe auf zwei Outputs auf."

Mit diesen beiden Prompt-Varianten werden die Testreihen mit ChatGPT in der Hauptstudie durchgeführt. Im folgenden Kapitel werden die Ergebnisse dieser Durchführung präsentiert.

6.3. Ergebnisse

Auf den folgenden Seiten sind die Bewertungen der Physiklehrpersonen sowie die beiden Bewertungen von ChatGPT (einmal anhand der Prompts H1 & H2 und H3 und H4) pro Aufgabe tabellarisch aufgelistet. In den Tabellen **Tab. 8** – **Tab. 10** sind die drei Testreihen zur Aufgabe „*Windenergie*" dargestellt. In den Tabellen **Tab. 11** – **Tab. 13** sind die drei Testreihen zur Aufgabe „*Musikbox*" dargestellt und die Tabellen **Tab. 14** – **Tab. 16** zeigen abschließend die drei Testreihen zur Aufgabe „*Schilddrüsenuntersuchung*".
Die zusammengefassten Durchschnittsbewertungen von ChatGPT oder Lehrkräften pro Antwort und Aufgabe sind anschließend nochmal übersichtlich in den Tabellen Tab**. 17** – Tab**. 19** zusammen aufgelistet, um einen Vergleich besser zu ermöglichen.

Vorab noch einige Hinweise zu den Abkürzungen in den Ergebnistabellen 8-16: Die linke Spalte (*Ant.-Nr.*) gibt die Antwortnummer an. Jede Antwort wurde innerhalb einer Testreihe fünf Mal bewertet (*Runde 1–5* von ChatGPT oder von *Lehrkraft 1–5*). Die Bewertung gliedert sich in die Beurteilung des Inhalts (*Inh.*), der Sprache (*Spr.*) und dem resultierenden Gesamtniveau (*Ges.*). Aus den fünf Bewertungen pro Antwort wird jeweils der Mittelwert vom Inhalts-, Sprach- und Gesamtniveau berechnet. Weiters zeigen die Tabellen 8-16 die Differenzen (*Diff.*) in der ganz rechten Spalte. Diese geben die Spannweite zwischen der maximalen und minimalen Bewertungen (*Max-Min*) einer Antwort an. Diese Spannweite soll ein Maß für die Unsicherheit von ChatGPT oder den Lehrkräften sein. In Tab. Tab**. 17**: Zusammengefasste Durchschnittsbeurteilungen aus den drei Testreihen (Tab. 8-10), in denen ChatGPT und die Lehrpersonen die Antworten zur Aufgabe "Windenergie" nach dem 3-stufigen Modell bewerteten. – Tab**. 19** wurde der gleiche Farbcode, wie in der Pilotstudie, Phase 2 und 3, angelegt: Die durchschnittliche Beurteilung von *ChatGPT-4o*/*ChatGPT-5* zu einer Antwort ist grün hinterlegt, wenn sie der gleichen Durchschnittsbeurteilung der Lehrkräfte entspricht. Ist sie gelb hinterlegt, weicht die Beurteilung ein Niveau davon ab; ist sie rot hinterlegt, weicht sie um zwei Niveaus ab.
Aufgrund einer überraschenden Veröffentlichung des neuen ChatGPT-5-Modells während der Datenerhebung, wurden manche Beurteilungen der Hauptstudie mit *GPT-5* generiert.

Daher zeigt die unterste Reihe der Tabellen 8-16 zusätzlich an, mit welchen ChatGPT-Modell (*GPT-4o* oder *GPT-5*) die Beurteilungen einer Runde erzeugt wurden.

Testreihe von ChatGPT mit Variante H1 & H2 zu „*Windenergie*“																			
Ant.-Nr.	Runde 1			Runde 2			Runde 3			Runde 4			Runde 5			Mittelwert			Diff.
	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	**Spr.**	**Ges.**	**Max-Min**
1	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	0,00
2	1,00	0,00	**0,67**	1,00	0,00	**0,67**	1,00	1,00	**1,00**	1,00	0,00	**0,67**	1,00	1,00	**1,00**	1,00	0,40	**0,80**	0,33
3	0,00	0,00	**0,00**	0,00	0,00	**0,00**	0,00	0,00	**0,00**	0,00	0,00	**0,00**	0,00	0,00	**0,00**	0,00	0,00	**0,00**	0,00
4	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,00
5	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,00
6	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	1,40	1,00	**1,27**	0,67
7	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,00
8	1,00	0,00	**0,67**	1,00	0,00	**0,67**	1,00	0,00	**0,67**	1,00	0,00	**0,67**	1,00	0,00	**0,67**	1,00	0,00	**0,67**	0,00
9	1,00	1,00	**1,00**	1,00	1,00	**1,00**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	1,00	1,00	**1,00**	1,40	1,00	**1,27**	0,67
10	0,00	0,00	**0,00**	0,00	0,00	**0,00**	0,00	0,00	**0,00**	0,00	0,00	**0,00**	0,00	0,00	**0,00**	0,00	0,00	**0,00**	0,00
11	0,00	0,00	**0,00**	1,00	0,00	**0,67**	0,00	0,00	**0,00**	1,00	0,00	**0,67**	1,00	1,00	**1,00**	0,60	0,20	**0,47**	1,00
12	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	0,00	**0,67**	1,00	0,80	**0,93**	0,33
13	1,00	1,00	**1,00**	1,00	0,00	**0,67**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	0,80	**0,93**	0,33
	GPT-4o																		

***Tab. 8**: Beurteilungen von ChatGPT-4o zu „Windenergie“ mithilfe der Prompt-Variante H1 & H2. Die linke Tabellenhälfte zeigt die Bewertungen der einzelnen Runden an, die rechte Tabellenhälfte die daraus resultierenden Mittel- und Differenzwerte.*

Testreihe von ChatGPT mit Variante H3 & H4 zu „*Windenergie*“																			
Ant.-Nr.	Runde 1			Runde 2			Runde 3			Runde 4			Runde 5			Mittelwert			Diff.
	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	**Spr.**	**Ges.**	**Max-Min**
1	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	2,00	**1,33**	1,00	1,20	**1,07**	**0,33**
2	1,00	1,00	**1,00**	1,00	0,00	**0,67**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	2,00	2,00	**2,00**	1,20	1,00	**1,13**	**1,33**
3	0,00	0,00	**0,00**	0,00	0,00	**0,00**	0,00	0,00	**0,00**	0,00	0,00	**0,00**	1,00	1,00	**1,00**	0,20	0,20	**0,20**	**1,00**
4	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	**0,00**
5	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	1,00	**1,67**	2,00	1,80	**1,93**	**0,33**
6	1,00	1,00	**1,00**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	1,80	1,00	**1,53**	**0,67**
7	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	**0,00**
8	0,00	0,00	**0,00**	1,00	1,00	**1,00**	1,00	0,00	**0,67**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	0,80	0,60	**0,73**	**1,00**
9	2,00	2,00	**2,00**	2,00	2,00	**2,00**	1,00	1,00	**1,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	1,80	1,80	**1,80**	**1,00**
10	0,00	0,00	**0,00**	0,00	0,00	**0,00**	0,00	0,00	**0,00**	0,00	0,00	**0,00**	1,00	1,00	**1,00**	0,20	0,20	**0,20**	**1,00**
11	1,00	0,00	**0,67**	1,00	1,00	**1,00**	1,00	0,00	**0,67**	1,00	1,00	**1,00**	0,00	0,00	**0,00**	0,80	0,40	**0,67**	**1,00**
12	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	**0,00**
13	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	**0,00**
	GPT-4o												*GPT-5*						

Tab. 9: *Beurteilungen von ChatGPT-4o und ChatGPT-5 zu „Windenergie“ mithilfe der Prompt-Variante H3 & H4. Die linke Tabellenhälfte zeigt die Bewertungen der einzelnen Runden an, die rechte Tabellenhälfte die daraus resultierenden Mittel- und Differenzwerte.*

Testreihe der Lehrkräfte zu „*Windenergie*“																			
Ant.-Nr.	Lehrkraft 1			Lehrkraft 2			Lehrkraft 3			Lehrkraft 4			Lehrkraft 5			Mittelwerte			Diff.
	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	**Spr.**	**Ges.**	Max-Min
1	1,00	1,00	**1,00**	1,00	0,00	**0,67**	0,00	1,00	**0,33**	0,00	1,00	**0,33**	1,00	0,00	**0,67**	0,60	0,60	**0,60**	0,67
2	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,00	1,00	**0,33**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,20	1,40	**1,27**	1,67
3	2,00	1,00	**1,67**	1,00	1,00	**1,00**	0,00	1,00	**0,33**	0,00	0,00	**0,00**	0,00	1,00	**0,33**	0,60	0,80	**0,67**	1,67
4	2,00	2,00	**2,00**	2,00	2,00	**2,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	2,00	1,00	**1,67**	1,60	1,40	**1,53**	1,00
5	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	1,00	1,00	**1,00**	2,00	2,00	**2,00**	1,80	1,80	**1,80**	1,00
6	1,00	2,00	**1,33**	2,00	1,00	**1,67**	0,50	1,00	**0,67**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	1,50	1,20	**1,40**	1,00
7	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,50	1,00	**0,67**	1,00	1,00	**1,00**	1,00	2,00	**1,33**	1,30	1,60	**1,40**	1,33
8	2,00	1,00	**1,67**	1,00	1,00	**1,00**	0,00	0,50	**0,17**	1,00	0,00	**0,67**	1,00	1,00	**1,00**	1,00	0,70	**0,90**	1,50
9	2,00	1,00	**1,67**	2,00	2,00	**2,00**	0,50	0,50	**0,50**	1,00	1,00	**1,00**	1,00	2,00	**1,33**	1,30	1,30	**1,30**	1,50
10	1,00	0,00	**0,67**	1,00	1,00	**1,00**	0,50	0,00	**0,33**	1,00	0,00	**0,67**	0,00	2,00	**0,67**	0,70	0,60	**0,67**	0,67
11	1,00	1,00	**1,00**	1,00	1,00	**1,00**	0,00	0,00	**0,00**	0,00	0,00	**0,00**	0,00	1,00	**0,33**	0,40	0,60	**0,47**	1,00
12	1,00	1,00	**1,00**	1,00	2,00	**1,33**	0,00	0,00	**0,00**	2,00	2,00	**2,00**	1,00	2,00	**1,33**	1,00	1,40	**1,13**	2,00
13	1,00	1,00	**1,00**	1,00	1,00	**1,00**	0,00	0,50	**0,17**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	0,80	0,90	**0,83**	0,83

Tab. 10*: Beurteilungen der fünf Physiklehrkräfte zu „Windenergie“. Die linke Tabellenhälfte zeigt die Bewertungen der einzelnen Personen an, die rechte Tabellenhälfte die daraus resultierenden Mittel- und Differenzwerte.*

Testreihe von ChatGPT mit Variante H1 & H2 zu „*Musikbox*“																			
Ant.-Nr.	Runde 1			Runde 2			Runde 3			Runde 4			Runde 5			Mittelwerte			Diff.
	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	**Spr.**	**Ges.**	Max-Min
1	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	0,00
2	2,00	1,00	**1,67**	2,00	1,00	**1,67**	1,00	1,00	**1,00**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	1,80	1,00	**1,53**	0,67
3	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,00
4	1,00	0,00	**0,67**	1,00	0,00	**0,67**	1,00	0,00	**0,67**	1,00	0,00	**0,67**	1,00	0,00	**0,67**	1,00	0,00	**0,67**	0,00
5	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,00
6	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,00
7	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	1,00	**1,67**	2,00	1,80	**1,93**	0,33
8	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,00
9	1,00	1,00	**1,00**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	1,80	1,00	**1,53**	0,67
10	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,00
11	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,00
12	1,00	0,00	**0,67**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	0,00	**0,67**	1,00	0,00	**0,67**	1,00	0,40	**0,80**	0,33
13	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,00
	GPT-4o																		

***Tab. 11**: Beurteilungen von ChatGPT-4o „Musikbox“ mithilfe der Prompt-Variante H1 & H2. Die linke Tabellenhälfte zeigt die Bewertungen der einzelnen Runden an, die rechte Tabellenhälfte die daraus resultierenden Mittel- und Differenzwerte.*

Testreihe von ChatGPT mit Variante H3 & H4 zu „*Musikbox*“																			
Ant.-Nr.	Runde 1			Runde 2			Runde 3			Runde 4			Runde 5			Mittelwerte			Diff.
	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	**Spr.**	**Ges.**	**Max-Min**
1	1,00	0,00	**0,67**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	2,00	1,00	**1,67**	1,20	0,80	**1,07**	1,00
2	1,00	1,00	**1,00**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	1,80	1,00	**1,53**	0,67
3	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,00
4	1,00	1,00	**1,00**	1,00	0,00	**0,67**	1,00	0,00	**0,67**	1,00	0,00	**0,67**	2,00	1,00	**1,67**	1,20	0,40	**0,93**	1,00
5	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,00
6	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,00
7	2,00	1,00	**1,67**	2,00	2,00	**2,00**	2,00	1,00	**1,67**	2,00	2,00	**2,00**	2,00	1,00	**1,67**	2,00	1,40	**1,80**	0,33
8	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,00
9	2,00	1,00	**1,67**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	2,00	2,00	**2,00**	2,00	1,20	**1,73**	0,33
10	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,00
11	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,00
12	1,00	0,00	**0,67**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	0,80	**0,93**	0,33
13	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	1,00	**1,67**	2,00	1,80	**1,93**	0,33
	GPT-4o												*GPT-5*						

***Tab. 12:** Beurteilungen von ChatGPT-4o und ChatGPT-5 zu „Musikbox“ mithilfe der Prompt-Variante H3 & H4. Die linke Tabellenhälfte zeigt die Bewertungen der einzelnen Runden an, die rechte Tabellenhälfte die daraus resultierenden Mittel- und Differenzwerte.*

Testreihe der Lehrkräfte zu „*Musikbox*“																			
Ant.-Nr.	Lehrkraft 1			Lehrkraft 2			Lehrkraft 3			Lehrkraft 4			Lehrkraft 5			Mittelwerte			Diff.
	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	**Spr.**	**Ges.**	**Max-Min**
1	2,00	1,00	**1,67**	1,00	1,00	**1,00**	0,00	1,00	**0,33**	1,00	1,00	**1,00**	0,00	0,00	**0,00**	0,80	0,80	**0,80**	1,67
2	1,00	1,00	**1,00**	1,00	1,00	**1,00**	0,00	0,50	**0,17**	1,00	1,00	**1,00**	1,00	0,00	**0,67**	0,80	0,70	**0,77**	0,83
3	2,00	2,00	**2,00**	1,00	2,00	**1,33**	1,00	2,00	**1,33**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,20	1,60	**1,33**	1,00
4	2,00	1,00	**1,67**	1,00	1,00	**1,00**	1,00	0,00	**0,67**	1,00	1,00	**1,00**	1,00	0,00	**0,67**	1,20	0,60	**1,00**	1,00
5	1,00	2,00	**1,33**	2,00	1,00	**1,67**	2,00	2,00	**2,00**	1,00	0,00	**0,67**	0,00	0,00	**0,00**	1,20	1,00	**1,13**	2,00
6	2,00	2,00	**2,00**	2,00	2,00	**2,00**	1,00	0,50	**0,83**	2,00	1,00	**1,67**	1,00	1,00	**1,00**	1,60	1,30	**1,50**	1,17
7	2,00	1,00	**1,67**	1,00	1,00	**1,00**	0,50	0,50	**0,50**	2,00	1,00	**1,67**	1,00	1,00	**1,00**	1,30	0,90	**1,17**	1,17
8	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	1,00	1,00	**1,00**	2,00	2,00	**2,00**	1,80	1,80	**1,80**	1,00
9	2,00	1,00	**1,67**	2,00	1,00	**1,67**	2,00	0,50	**1,50**	2,00	2,00	**2,00**	2,00	1,00	**1,67**	2,00	1,10	**1,70**	0,50
10	2,00	2,00	**2,00**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	1,00	2,00	**1,33**	1,80	1,40	**1,67**	0,67
11	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	1,50	**1,83**	2,00	2,00	**2,00**	1,00	2,00	**1,33**	1,80	1,90	**1,83**	0,67
12	2,00	1,00	**1,67**	2,00	1,00	**1,67**	0,00	0,00	**0,00**	0,00	1,00	**0,33**	1,00	1,00	**1,00**	1,00	0,80	**0,93**	1,67
13	2,00	2,00	**2,00**	2,00	1,00	**1,67**	2,00	1,50	**1,83**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,60	1,30	**1,50**	1,00

Tab. 13: *Beurteilungen der fünf Physiklehrkräfte zu „Musikbox“. Die linke Tabellenhälfte zeigt die Bewertungen der einzelnen Personen an, die rechte Tabellenhälfte die daraus resultierenden Mittel- und Differenzwerte.*

Testreihe von ChatGPT mit Variante H1 & H2 zu *„Schilddrüsenuntersuchung“*																			
Ant.-Nr.	Runde 1			Runde 2			Runde 3			Runde 4			Runde 5			Mittelwerte			Diff.
	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Max-Min
1	1,00	1,00	**1,00**	0,00	1,00	**0,33**	0,00	0,00	**0,00**	0,00	0,00	**0,00**	0,00	0,00	**0,00**	0,20	0,40	**0,27**	1,0
2	1,00	0,00	**0,67**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	0,00	**0,67**	1,00	0,00	**0,67**	1,00	0,40	**0,80**	0,3
3	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	2,00	1,40	**1,80**	0,3
4	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	0,00	**0,67**	1,00	1,00	**1,00**	1,00	0,00	**0,67**	1,00	0,60	**0,87**	0,3
5	2,00	1,00	**1,67**	2,00	1,00	**1,67**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,40	1,00	**1,27**	0,7
6	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	1,00	**1,67**	2,00	2,00	**2,00**	2,00	1,00	**1,67**	2,00	1,60	**1,87**	0,3
7	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	0,00	**0,67**	0,00	0,00	**0,00**	1,00	0,00	**0,67**	0,80	0,40	**0,67**	1,0
8	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,0
9	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	1,00	**1,67**	2,00	2,00	**2,00**	2,00	1,80	**1,93**	0,3
10	2,00	1,00	**1,67**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	1,00	0,00	**0,67**	1,80	0,80	**1,47**	1,0
	GPT-4o																		

Tab. 14*: Beurteilungen von ChatGPT-4o zu „Schilddrüsenuntersuchung“ mithilfe der Prompt-Variante H1 & H2. Die linke Tabellenhälfte zeigt die Bewertungen der einzelnen Runden an, die rechte Tabellenhälfte die daraus resultierenden Mittel- und Differenzwerte.*

Testreihe von ChatGPT mit Variante H3 & H4 zu *„Schilddrüsenuntersuchung“*																			
Ant.-Nr.	Runde 1			Runde 2			Runde 3			Runde 4			Runde 5			Mittelwerte			Diff.
	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Inh.	Spr.	**Ges.**	Max-Min
1	1,00	2,00	**1,33**	0,00	0,00	**0,00**	1,00	0,00	**0,67**	1,00	1,00	**1,00**	1,00	2,00	**1,33**	0,80	1,00	**0,87**	1,33
2	1,00	1,00	**1,00**	0,00	0,00	**0,00**	1,00	1,00	**1,00**	0,00	1,00	**0,33**	1,00	1,00	**1,00**	0,60	0,80	**0,67**	1,00
3	2,00	1,00	**1,67**	1,00	1,00	**1,00**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	1,00	1,00	**1,00**	1,60	1,00	**1,40**	0,67
4	0,00	0,00	**0,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	0,80	0,80	**0,80**	1,00
5	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	1,00	2,00	**1,33**	1,80	1,60	**1,73**	0,67
6	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	2,00	2,00	**2,00**	0,00
7	0,00	0,00	**0,00**	0,00	0,00	**0,00**	0,00	0,00	**0,00**	0,00	2,00	**0,67**	0,00	0,00	**0,00**	0,00	0,40	**0,13**	0,67
8	2,00	1,00	**1,67**	1,00	1,00	**1,00**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	1,00	1,00	**1,00**	1,60	1,00	**1,40**	0,67
9	2,00	2,00	**2,00**	1,00	1,00	**1,00**	2,00	1,00	**1,67**	2,00	2,00	**2,00**	1,00	1,00	**1,00**	1,60	1,40	**1,53**	1,00
10	2,00	1,00	**1,67**	1,00	1,00	**1,00**	2,00	1,00	**1,67**	2,00	1,00	**1,67**	1,00	1,00	**1,00**	1,60	1,00	**1,40**	0,67
	GPT-5																		

Tab. 15: *Beurteilungen von ChatGPT-5 zu „Schilddrüsenuntersuchung“ mithilfe der Prompt-Variante H3 & H4. Die linke Tabellenhälfte zeigt die Bewertungen der einzelnen Runden an, die rechte Tabellenhälfte die daraus resultierenden Mittel- und Differenzwerte.*

Testreihe der Lehrkräfte „*Schilddrüsenuntersuchung*“																			
Ant.-Nr.	Lehrkraft 1			Lehrkraft 2			Lehrkraft 3			Lehrkraft 4			Lehrkraft 5			Mittelwerte			Diff.
	Inh.	Spr.	Ges.	Inh.	Spr.	Ges.	Inh.	Spr.	Ges.	Inh.	Spr.	Ges.	Inh.	Spr.	Ges.	Inh.	Spr.	Ges.	Max-Min
1	0,00	1,00	**0,33**	0,00	1,00	**0,33**	-	-	**2,00**	1,00	1,00	**1,00**	1,00	0,00	**0,67**	0,50	0,75	**0,87**	1,67
2	1,00	1,00	**1,00**	1,00	1,00	**1,00**	-	-	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	0,00
3	2,00	1,00	**1,67**	2,00	1,00	**1,67**	-	-	**1,00**	1,00	1,00	**1,00**	2,00	2,00	**2,00**	1,75	1,25	**1,47**	1,00
4	1,00	1,00	**1,00**	1,00	0,00	**0,67**	-	-	**1,00**	1,00	1,00	**1,00**	2,00	2,00	**2,00**	1,25	1,00	**1,13**	1,33
5	2,00	2,00	**2,00**	2,00	2,00	**2,00**	-	-	**2,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,50	1,50	**1,60**	1,00
6	2,00	2,00	**2,00**	1,00	2,00	**1,33**	-	-	**2,00**	1,00	2,00	**1,33**	2,00	2,00	**2,00**	1,50	2,00	**1,73**	0,67
7	0,00	1,00	**0,33**	1,00	1,00	**1,00**	-	-	**2,00**	1,00	1,00	**1,00**	2,00	1,00	**1,67**	1,00	1,00	**1,20**	1,67
8	2,00	1,00	**1,67**	2,00	1,00	**1,67**	-	-	**2,00**	1,00	2,00	**1,33**	2,00	2,00	**2,00**	1,75	1,50	**1,73**	0,67
9	1,00	1,00	**1,00**	1,00	1,00	**1,00**	-	-	**2,00**	1,00	1,00	**1,00**	2,00	1,00	**1,67**	1,25	1,00	**1,33**	1,00
10	2,00	1,00	**1,67**	2,00	1,00	**1,67**	-	-	**1,00**	1,00	1,00	**1,00**	1,00	1,00	**1,00**	1,50	1,00	**1,27**	0,67

***Tab. 16**: Beurteilungen der fünf Physiklehrkräfte zu „Schilddrüsenuntersuchung“. Die linke Tabellenhälfte zeigt die Bewertungen der einzelnen Personen an, die rechte Tabellenhälfte die daraus resultierenden Mittel- und Differenzwerte. In dieser Testreihe waren von Lehrkraft 3 nur die Gesamtniveaus als Daten verfügbar.*

Durchschnittsbeurteilungen zu „*Windenergie*“												
	GPT-4o (H1 & H2)				*GPT-4o/GPT-5* (H3 & H4)				Physiklehrkräfte			
Antwort	Inh.	Spr.	Ges.	Diff.	Inh.	Spr.	Ges.	Diff.	Inh.	Spr.	Ges.	Diff.
Nr. 1	1,00	1,00	**1,00**	0,00	1,00	1,20	**1,07**	0,33	0,60	0,60	**0,60**	0,67
Nr. 2	1,00	0,40	**0,80**	0,33	1,20	1,00	**1,13**	1,33	1,20	1,40	**1,27**	1,67
Nr. 3	0,00	0,00	**0,00**	0,00	0,20	0,20	**0,20**	1,00	0,60	0,80	**0,67**	1,67
Nr. 4	2,00	2,00	**2,00**	0,00	2,00	2,00	**2,00**	0,00	1,60	1,40	**1,53**	1,00
Nr. 5	2,00	2,00	**2,00**	0,00	2,00	1,80	**1,93**	0,33	1,80	1,80	**1,80**	1,00
Nr. 6	1,40	1,00	**1,27**	0,67	1,80	1,00	**1,53**	0,67	1,50	1,20	**1,40**	1,00
Nr. 7	2,00	2,00	**2,00**	0,00	2,00	2,00	**2,00**	0,00	1,30	1,60	**1,40**	1,33
Nr. 8	1,00	0,00	**0,67**	0,00	0,80	0,60	**0,73**	1,00	1,00	0,70	**0,90**	1,50
Nr. 9	1,40	1,00	**1,27**	0,67	1,80	1,80	**1,80**	1,00	1,30	1,30	**1,30**	1,50
Nr. 10	0,00	0,00	**0,00**	0,00	0,20	0,20	**0,20**	1,00	0,70	0,60	**0,67**	0,67
Nr. 11	0,60	0,20	**0,47**	1,00	0,80	0,40	**0,67**	1,00	0,40	0,60	**0,47**	1,00
Nr. 12	1,00	0,80	**0,93**	0,33	1,00	1,00	**1,00**	0,00	1,00	1,40	**1,13**	2,00
Nr. 13	1,00	0,80	**0,93**	0,33	1,00	1,00	**1,00**	0,00	0,80	0,90	**0,83**	0,83
Mittelwert	1,11	0,86	**1,03**	0,26	1,22	1,09	**1,17**	0,59	1,06	1,10	**1,07**	1,22

***Tab. 17**: Zusammengefasste Durchschnittsbeurteilungen aus den drei Testreihen (Tab. 8-10), in denen ChatGPT und die Lehrpersonen die Antworten zur Aufgabe "Windenergie" nach dem 3-stufigen Modell bewerteten.*

Durchschnittsbeurteilungen zu „*Musikbox*“												
	GPT-4o (H1 & H2)				*GPT-4o/GPT-5* (H3 & H4)				Physiklehrkräfte			
Antwort	Inh.	Spr.	Ges.	Diff.	Inh.	Spr.	Ges.	Diff.	Inh.	Spr.	Ges.	Diff.
Nr. 1	1,00	1,00	**1,00**	0,00	1,20	0,80	**1,07**	1,00	0,80	0,80	**0,80**	1,67
Nr. 2	1,80	1,00	**1,53**	0,67	1,80	1,00	**1,53**	0,67	0,80	0,70	**0,77**	0,83
Nr. 3	2,00	2,00	**2,00**	0,00	2,00	2,00	**2,00**	0,00	1,20	1,60	**1,33**	1,00
Nr. 4	1,00	0,00	**0,67**	0,00	1,20	0,40	**0,93**	1,00	1,20	0,60	**1,00**	1,00
Nr. 5	2,00	2,00	**2,00**	0,00	2,00	2,00	**2,00**	0,00	1,20	1,00	**1,13**	2,00
Nr. 6	2,00	2,00	**2,00**	0,00	2,00	2,00	**2,00**	0,00	1,60	1,30	**1,50**	1,17
Nr. 7	2,00	1,80	**1,93**	0,33	2,00	1,40	**1,80**	0,33	1,30	0,90	**1,17**	1,17
Nr. 8	2,00	2,00	**2,00**	0,00	2,00	2,00	**2,00**	0,00	1,80	1,80	**1,80**	1,00
Nr. 9	1,80	1,00	**1,53**	0,67	2,00	1,20	**1,73**	0,33	2,00	1,10	**1,70**	0,50
Nr. 10	2,00	2,00	**2,00**	0,00	2,00	2,00	**2,00**	0,00	1,80	1,40	**1,67**	0,67
Nr. 11	2,00	2,00	**2,00**	0,00	2,00	2,00	**2,00**	0,00	1,80	1,90	**1,83**	0,67
Nr. 12	1,00	0,40	**0,80**	0,33	1,00	0,80	**0,93**	0,33	1,00	0,80	**0,93**	1,67
Nr. 13	2,00	2,00	**2,00**	0,00	2,00	1,80	**1,93**	0,33	1,60	1,30	**1,50**	1,00
Mittelwert	1,74	1,48	**1,65**	0,15	1,78	1,49	**1,69**	0,31	1,39	1,17	**1,32**	1,10

***Tab. 18**: Zusammengefasste Durchschnittsbeurteilungen aus den drei Testreihen (Tab. 11-13), in denen ChatGPT und die Lehrpersonen die Antworten zur Aufgabe "Musikbox" nach dem 3-stufigen Modell bewerteten.*

Durchschnittbeurteilungen zu „*Schilddrüsenuntersuchung*“												
	GPT-4o (H1 & H2)				*GPT-5* (H3 & H4)				Physiklehrkräfte			
Antwort	Inh.	Spr.	Ges.	Diff.	Inh.	Spr.	Ges.	Diff.	Inh.	Spr.	Ges.	Diff.
Nr. 1	0,20	0,40	**0,27**	1,00	0,80	1,00	**0,87**	1,33	0,50	0,75	**0,87**	1,67
Nr. 2	1,00	0,40	**0,80**	0,33	0,60	0,80	**0,67**	1,00	1,00	1,00	**1,00**	0,00
Nr. 3	2,00	1,40	**1,80**	0,33	1,60	1,00	**1,40**	0,67	1,75	1,25	**1,47**	1,00
Nr. 4	1,00	0,60	**0,87**	0,33	0,80	0,80	**0,80**	1,00	1,25	1,00	**1,13**	1,33
Nr. 5	1,40	1,00	**1,27**	0,67	1,80	1,60	**1,73**	0,67	1,50	1,50	**1,60**	1,00
Nr. 6	2,00	1,60	**1,87**	0,33	2,00	2,00	**2,00**	0,00	1,50	2,00	**1,73**	0,67
Nr. 7	0,80	0,40	**0,67**	1,00	0,00	0,40	**0,13**	0,67	1,00	1,00	**1,20**	1,67
Nr. 8	2,00	2,00	**2,00**	0,00	1,60	1,00	**1,40**	0,67	1,75	1,50	**1,73**	0,67
Nr. 9	2,00	1,80	**1,93**	0,33	1,60	1,40	**1,53**	1,00	1,25	1,00	**1,33**	1,00
Nr. 10	1,80	0,80	**1,47**	1,00	1,60	1,00	**1,40**	0,67	1,50	1,00	**1,27**	0,67
Mittelwert	1,42	1,04	**1,29**	0,53	1,24	1,10	**1,19**	0,77	1,30	1,20	**1,33**	0,97

***Tab. 19**: Zusammengefasste Durchschnittsbeurteilungen aus den drei Testreihen (Tab. 14-16), in denen ChatGPT und die Lehrpersonen die Antworten zur Aufgabe "Musikbox" nach dem 3-stufigen Modell bewerteten.*

6.4. Auswertung der Ergebnisse

Die Hauptstudie dieser Masterarbeit konnte planmäßig durchgeführt werden und liefert wichtige Ergebnisse, die hier nun analysiert werden sollen. Eine unvorhergesehene Tatsache wirkt sich jedoch auf die Vergleichbarkeit der Daten aus: Inmitten der Testphase wurde *ChatGPT-4o* von *ChatGPT-5* als neues Basismodell abgelöst (OpenAI, 2025d). Gemeinsam mit der Betreuerin M. Korner wurde diskutiert, ob die Testphase unterbrochen werden sollte und ein Wechsel zurück zu *GPT-4o* machbar und sinnvoll wäre. Letztlich wurde beschlossen, dass die Hauptstudie ohne Unterbrechung weitergeführt und eine leichte Heterogenität der Daten in Kauf genommen wird. Vor allem aus dem Grund, da sonst eine Anknüpfung der Hauptstudie an die Pilotstudie, die nur auf dem *GPT-4o*-Modell beruht, erschwert wird.

Zuerst sollen die Beurteilungen von ChatGPT und den Physiklehrkräften verglichen werden, bevor einzelne, generierte Antworten des Chatbots qualitativ analysiert werden.

Ergebnisse zu „*Windenergie*“

In Tab. 17 sind alle Durchschnittsbeurteilungen zur Aufgabe „*Windenergie*“ aufgelistet. Verglichen mit der Beurteilung der Lehrpersonen erzielte ChatGPT Großteils identische Ergebnisse. Von den 13 Antworten ordnete ChatGPT neun (Variante H1 & H2) bzw. sieben (Variante H3 & H4) die gleiche Niveaustufe zu wie die Lehrkräfte. Die restlichen Antworten wurden von ChatGPT um ein Niveau abweichend bewertet. Für diese Ausreißer lässt sich jedoch keine Tendenz in eine Richtung, gemäß besserer bzw. schlechterer KI-Beurteilung, erkennen. Die Prompt-Variante H3 & H4 erzielte etwas weniger Übereinstimmung mit den Beurteilungen der Lehrpersonen als H1 & H2. Liegt das an der Heterogenität der Daten, da eine Beurteilungsrunde in H3 & H4 von *ChatGPT-5* generiert wurde? Vermutlich nicht, da sich die Antworten mit den abweichenden Beurteilungen mit jenen in Variante H1 & H2 decken. Dennoch sei erwähnt, dass das Modell *GPT-5* einige Antworten zu „*Windenergie*“ etwas anders beurteilte als *GPT-4o*: Vergleicht man die Bewertungen der Antworten 1, 2, 3, 5 und 10 aller Runden in Tab. 9 fällt auf, dass *ChatGPT-4o* diese Antworten immer gleich beurteilte und nur *ChatGPT-5* in Runde 5 ein anderes Gesamtniveau berechnete.

Die Mittelwerte aller sprachlichen, inhaltlichen und gesamten Durchschnittsbeurteilungen von ChatGPT (Variante H1 & H2 bzw. H3 & H4) und den Physiklehrkräften aus Tab. 17 im Vergleich zeigen, dass beide Prompt-Varianten zu sehr ähnlichen Niveaustufen führen, wie sie bei den Lehrkräften im Mittel herauskommen.
Ein klarer Unterschied zwischen menschlichen und KI-Beurteilungen lässt sich jedoch in den Differenzen hinter den durchschnittlichen Gesamtbeurteilungen erkennen. ChatGPT stufte viele Antworten zu „*Windenergie*“ mit einer Differenz von 0,00 ein – das bedeutet in allen Runden der Testreihe erhielt die Antwort das gleiche Gesamtniveau. Eine Differenz abweichend von 0,00 gibt die Spannweite zwischen dem höchsten und niedrigsten Gesamtniveau an und diese beläuft sich bei ChatGPT in Variante H1 & H2 im Mittel auf 0,26 und in Variante H3 & H4 im Mittel auf 0,59, siehe Tab. 17 unten. Hingegen beträgt die mittlere Differenz der Physiklehrkräfte 1,22. Während also die mittleren Durchschnittsbeurteilungen (inhaltlich, sprachlich und gesamt) zu „*Windenergie*“ von ChatGPT und den Lehrkräften annähernd gleich sind, zeigen sich in der Unsicherheit deutliche Unterschiede.

Ergebnisse zu „*Musikbox*“

In Tab. 18 sind die Durchschnittsbeurteilungen zur Aufgabe „*Musikbox*“ aufgelistet. Verglichen mit den Beurteilungen der Lehrpersonen erzielt ChatGPT wieder mehrheitlich übereinstimmende Gesamtbeurteilungen. Von den 13 Antworten stufte ChatGPT neun Antworten in ihrem Gesamtniveau gleich ein, wie die Lehrkräfte und nur vier weichen um ein Niveau ab. Diese vier Antworten, Nr. 2, 3, 5 und 7, wurden in beiden Prompt-Varianten H1 & H2 bzw. H3 & H4 jeweils um eine Niveaustufe höher bewertet als von den Lehrkräften. In Tab. 18 lässt sich allgemein die Tendenz erkennen, dass die Antworten zu „*Musikbox*“ von der KI etwas besser bewertet werden. Das zeigen die Mittelwerte aus den 13 Durchschnittsbeurteilungen zu Inhalt, Sprache und Gesamtniveau sehr eindrücklich.

Die Mittelwerte zu Sprache, Inhalt und Gesamtniveau von ChatGPT aus Variante H1 & H2 und H3 & H4 sind fast identisch und liegen alle etwas höher als die Mittelwerte zu Sprache, Inhalt und Gesamtniveau der Lehrkräfte. Dieser kleine Unterschied hat zur Folge, dass eine Antwort aus dem Antwortsatz zu „*Musikbox*“ von ChatGPT im Mittel mit dem Gesamtniveau 2/+ bewertet wurde und von den Lehrkräften mit 1/~ (bei Rundung der Mittelwerte in Tab. 18). Im Fall der Aufgabe „*Musikbox*“ fällt dieser Unterschied zwar nicht so sehr ins Gewicht, da die Mehrheit der Antworten einzeln betrachtet von KI und Lehrkräften das gleiche Gesamtniveau erhielten. Dennoch zeigt sich anhand dieser Aufgabe, dass bei dem kürzeren 3-stufigen Modell kleinere Unterschiede schneller zu einem Niveaustufenwechsel führen können.
Ähnlich wie bei der Aufgabe „*Windenergie*“ ist bei „*Musikbox*“ die mittlere Differenz der durchschnittlichen Gesamtbeurteilungen von ChatGPT deutlich geringer (0,15-0,36) als bei den Physiklehrkräften (1,10). Sie liegt für ChatGPT innerhalb einer halben Niveaustufe.

Ergebnisse zu „*Schilddrüsenuntersuchung*“

In Tab. 19 sind alle Durchschnittsbeurteilungen zur Aufgabe „*Schilddrüsenuntersuchung*“ dargestellt. Die KI-Beurteilungen zu dieser Aufgabe lassen sich untereinander nur schwer vergleichen, da die Testreihe mit der Prompt-Variante H1 & H2 nur mit *ChatGPT-4o* generiert wurde und die Testreihe mit Variante H3 & H4 vollständig mit *ChatGPT-5*. Es zeigen sich doch einige Unterschiede in den Bewertungen: Fünf von zehn Antworten, Nr. 1, 3, 5, 7 und 8, wurden von *GPT-4o* und *GPT-5* nicht mit der gleichen Niveaustufe beurteilt. Das könnte ein Hinweis darauf sein, dass die beiden KI-Modelle ihre Beurteilung anhand des 3-stufigen Modells etwas anders gewichten. In diesem Fall wären die Testreihen „*Windenergie*“ (H3 & H4) und „*Musikbox*“ (H3 & H4), siehe Tab. 9 und Tab. 12, durch einen kleinen Anteil ihrer Daten von *ChatGPT-5* möglicherweise verzerrt.

Verglichen mit den Beurteilungen der Lehrpersonen erzielt *ChatGPT-4o* knapp mehrheitlich übereinstimmende Gesamtbeurteilungen: Sechs der zehn Antworten stufte *GPT-4o* mit dem gleichen Gesamtniveau ein, die restlichen vier weichen um ein Niveau ab. *ChatGPT-5* erreichte mit sieben von zehn gleichen Gesamtniveaus eine etwas höhere Übereinstimmung zu den Lehrkräften, es bewertete nur drei Antworten in Summe um ein Niveau abweichend.

Die Mittelwerte zu Sprache, Inhalt und Gesamtniveau von *ChatGPT-4o* aus Variante H1 & H2 und *ChatGPT-5* aus Variante H3 & H4 sind ähnlich und gut vergleichbar mit jenen von

den Lehrpersonen. Bei „*Schilddrüsenuntersuchung*“ liegt die mittlere Differenz der KI mit über einer halben Niveaustufe (0,53 und 0,77) etwas höher als bei den anderen beiden Aufgaben. Außerdem gibt es nur zwei Fälle in den beiden KI-Testreihen zur „*Schilddrüsenuntersuchung*“, wo ChatGPT in allen fünf Runden die gleiche Beurteilung traf. Dieser Punkt könnte wieder dafürsprechen, dass die Aufgabenstellung einen Einfluss auf die Treffsicherheit der KI-Beurteilung hat. Die mittlere Differenz der Lehrkräfte zu dieser Aufgabe (0,97) liegt dafür etwas tiefer, als in „*Windenergie*“ und „*Musikbox*“.

Interpretation und Vergleich der Beurteilungen durch ChatGPT und Lehrkräfte

Welche Unterschiede und Gemeinsamkeiten lassen sich nun in der Beurteilung von ChatGPT und den Physiklehrkräfte feststellen? An dieser Stelle kann festgehalten werden, dass die Beurteilung von Antworten durch ChatGPT anhand des 3-stufigen Modells von Hackner (2025) valide Ergebnisse liefert und mittlere bis hohe Übereinstimmung zu den Beurteilungen von Physiklehrkräften zeigt. Bei den Aufgaben „*Windenergie*“ und *Schilddrüsenuntersuchung*“ konnte nicht festgestellt werden, dass ChatGPT Antworten tendenziell deutlich besser oder schlechter einstuft als die Lehrkräfte. Diese Tendenz lässt sich nur in der Aufgabe „*Musikbox*“ beobachten, wo ChatGPT die Antworten tendenziell etwas besser beurteilte.
Ein klarer Unterschied lässt sich in der Unsicherheit der Durchschnittsbeurteilungen erkennen. Die Differenz aus fünf Beurteilungen einer Antwort liegt in den KI-Testreihen durchgehend niedriger als in den Testreihen der Lehrkräfte. Weiters ist auffallend, dass die Differenz in den KI-Testreihen, wo *GPT-5* einen Teil oder alle Runden generiert hat immer etwas höher ist als in jenen Testreihen, die *GPT-4o* allein generiert hat. Woran das liegt, kann diese Studie nicht nachweisen, da diese Testreihen einerseits verschiedene Prompt-Varianten und andererseits unterschiedliche Aufgaben nutzten, deren Einfluss auf das Ergebnis nicht vollständig klar ist.

Die Analyse von abweichenden Beurteilungen zwischen ChatGPT und den Lehrkräften ist auch in der Hauptstudie wesentlich. Aufgrund des Wechsels des Beurteilungsmodells und einer veränderten Durchführung ist die Bedeutung abweichender Beurteilungen in der Hauptstudie jedoch anders zu betrachten als in der Pilotstudie. Kurz gesagt, weil das Aufkommen (stark) abweichender KI-Beurteilungen in der Hauptstudie unwahrscheinlicher bzw. weniger eindeutig ist. Warum? Das Beurteilungsmodell zur S-Kompetenz von Hackner (2025) bewertet Antworten nur auf drei möglichen Niveaustufen und nicht auf fünf, wie im Modell von Gstall (2024). Und zusätzlich ist dieses Gesamtniveau beim 3-stufigen Modell gewichtet. Dieser Modellunterschied, vor allem die Reduzierung auf drei Stufen, macht es aus Sicht des Autors unwahrscheinlicher, dass ChatGPT (vor allem stark) abweichende Bewertungen im Vergleich zu Menschen trifft. Und der zweite Grund für das weniger klare Auftreten von abweichenden KI-Beurteilung in der Hauptstudie ist die Durchführung von Testreihen bei den Physiklehrkräften, aus denen nun ebenfalls Durchschnittsbeurteilungen erhoben werden konnten. In der Pilotstudie gab es pro Antwort nur eine Beurteilung als Referenzwert, der von den beiden Expertinnen Gstall und Korner stammte. In der Hauptstudie besteht die menschliche Beurteilung nun auch aus einem Durchschnitt. Und dieser gibt nun besser Auskunft darüber, wie sicher sich Physiklehrkräfte untereinander bei der

Beurteilung einer Antwort sind. Ist die Differenz bei den Lehrkräften sehr hoch, d.h. die Sicherheit ihrer Beurteilung gering, kann im Fall der KI-Beurteilung nicht mehr eindeutig argumentiert werden, ob diese nun abweichend ist oder nicht.

Die Bedeutung abweichender KI-Bewertungen der Hauptstudie wurde damit kurz eingeordnet, nun eine kurze Analyse dazu. In keiner der Testreihen treten stark abweichende Beurteilungen von ChatGPT im Vergleich zu den Lehrkräften auf. Die Anzahl der leicht abweichenden (in Tab. Tab**. 17** – Tab**. 19** gelb markierte) Beurteilungen ändert sich pro Aufgabe, liegt allgemein jedoch klar unter 50%. Wie klar in diesen Fällen von abweichenden Bewertungen gesprochen werden kann, ist schwer zu beantworten, da die Differenz zwischen den Bewertungen der Lehrkräfte allgemein hoch ist. Für alle gelb markierten Antworten in den Tab. Tab**. 17** – Tab**. 19** liegen diese zwischen 0,67 und 2,00 Niveaustufen. Als konkretes Beispiel kann Antwort Nr. 5 zu „*Musikbox*" diskutiert werden: In beiden Testreihen (H1 & H2 und H3 & H4) beurteilte *ChatGPT-4o/-5* diese Antwort mit Niveau 2,00 bei einer Differenz von 0,00 – also eindeutig. Diese Durchschnittsbeurteilung der KI weicht absolut betrachtet um ein Niveau von der Durchschnittsbeurteilung der Lehrkräfte (1,13) ab, jedoch hat diese auch eine Differenz von 2,00 – sie ist also sehr unsicher. Darf man im Fall von Antwort Nr. 5 nun von einer abweichenden KI-Beurteilung sprechen? Beim Vergleich der Durchschnitte ja, doch unter Einbezug der im Hintergrund herrschenden Unsicherheit auf Seiten der Lehr-kräfte ist das nicht mehr so klar und dieser Fakt muss berücksichtigt werden.

Die Hauptstudie kann keine neuen Hinweise liefern, warum ChatGPT generell abweichende Beurteilungen generiert, aber teilweise bestätigt sie Vermutungen dazu, die in der Pilotstudie schon aufkamen. In der Pilotstudie verstärkte sich die Annahme, dass es oft an der Formulierung der Antwort selbst liegen kann und diese dadurch schwer einem Niveau zuzuordnen ist, egal ob mit dem 5- oder 3-stufigen Modell bzw. durch Expert:innen oder ChatGPT. Solch *kritische* Antworten scheint es womöglich auch in den Hauptstudie zu geben. Zumindest ist das Maß der Differenz ein guter Indikator dafür. Als Beispiel kann Antwort Nr. 1 zu „*Schilddrüsenuntersuchung*" diskutiert werden. Ein Blick auf die Differenzwerte von *ChatGPT-4o* (1,00), *ChatGPT-5* (1,33) und den Physiklehrkräften (1,67) in Tab. 19 zeigt, dass die einzelnen Beurteilungen zu dieser Antwort in allen Testreihen weit auseinander liegen. Solche Fälle bestätigen, dass es *kritische* Antworten gibt, die weder von ChatGPT noch Lehrkräften eindeutig mit dem Beurteilungsmodell beurteilt werden können. In der Pilotstudie trat weiters die Vermutung auf, dass auch die Aufgabenstellung einen Einfluss darauf haben könnte, wie die Beurteilung von ChatGPT auf einen Antwortsatz dazu in Summe ausfällt. Diese Vermutung wurde durch die Hauptstudie verstärkt, da auch hier Unterschiede in der Genauigkeit und Unsicherheit von ChatGPT zwischen den Aufgaben beobachtet werden kann. Beispielsweise sind die Differenzen von ChatGPT zur Aufgabe „*Musikbox*" viel niedriger im Vergleich zur Aufgabe „*Schilddrüsenuntersuchung*" und ein objektiver Vergleich der beiden Aufgabenstellungen, siehe Anhang, bestätigt (zumindest aus Sicht des Autors), dass „*Musikbox*" für Schüler:innen leichter zu beantworten bzw. für Lehrer:innen leichter anhand des 3-stufigen Modells zu beurteilen ist als „*Schilddrüsenuntersuchung*". Dementsprechend könnte argumentiert werden, dass dieser Zusammen-

hang auch für eine KI, wie ChatGPT, gültig sein kann. Diese Annahme kann jedoch nicht ausreichend belegt werden, da die Rechenprozesse von ChatGPT nicht einsehbar sind und dieser mögliche Einflussfaktor bleibt letztlich nur eine Vermutung des Autors.

Einzelbeurteilungen und Feedback von ChatGPT zu ausgewählten Aufgaben

Nun zu einer kurzen qualitativen Auswertungen der generierten Einzelbeurteilungen. Aus jeder der drei Aufgabenstellungen von Hackner (2025) wird eine Antwort ausgewählt und gegenüberstellend die generierte Beurteilung gezeigt, die *ChatGPT-4o* oder *ChatGPT-5* einmal mit Variante H1 & H2 oder H3 & H4 erzeugt hat. Die dargestellten Antworten wurden ausgewählt, da ChatGPT diese über die Testreihe hinweg sehr konstant bewertet hat. Zwei vollständige Chatauszüge einer Beurteilungsrunde von *GPT-4o* und *GPT-5* in der Hauptstudie sind im Anhang als weitere Beispiele angefügt.

Antwort 1 zu „*Windenergie*“:

a. *„Einerseits ist es gut, um erneuerbaren, guten Strom zu bekommen. Auf einen längeren Zeitraum kann es gut für die Umwelt sein. Andererseits ist die Zeit und der Aufwand viel und ob die Bevölkerung das will, hängt davon ab, wie lange der Bau dauern würde. Es muss auch viel ausgebaut werden. Außerdem wird die Umwelt sehr zerstört, was mehr Schaden als Profit bringen wird.*
b. *Die vorhandenen Informationen reichen nicht aus. Beispielsweise wäre eine Preisrange auch von Vorteil.“*

Einzelbeurteilungen:

ChatGPT-4o (Variante H1 & H2):
„Antwort Nr. 1

- ***Beurteilung Inhaltlich:*** *Niveaustufe 1/~*
- ***Beurteilung Sprachlich:*** *Niveaustufe 1/~*
- ***Gesamtniveau:*** *1.0*
- ***Begründung:*** *Zwei relevante Kriterien werden angesprochen (Umweltfreundlichkeit, Aufwand), es gibt jedoch auch Unklarheiten (z. B. „mehr Schaden als Profit“ ist unkonkret). Die Sprache ist einfach, aber es wird ein gewisser Kontrast angedeutet.*
- ***Feedback:*** *Du hast beide Seiten betrachtet, was gut ist. Versuche, deine Argumente klarer zu strukturieren und spezifischer zu machen. Begriffe wie „zerstört die Umwelt“ solltest du genauer erklären. Auch wäre ein Vergleich zu anderen Möglichkeiten hilfreich.“*

ChatGPT-4o (Variante H3 & H4):
„Antwort Nr. 1

- ***Beurteilung Inhaltlich:*** *Niveau 1 / ~*
- ***Beurteilung Sprachlich:*** *Niveau 1 / ~*
- ***Gesamtniveau:*** *(2×1 + 1) ÷ 3 =* ***1.0***
- ***Begründung:*** *Es werden mehrere relevante Aspekte genannt (Umwelt, Aufwand, Bauzeit, Zerstörung der Umwelt), jedoch bleibt unklar, welche Kriterien entscheidend für die Position sind. Kein Vergleich. Sprache etwas sprunghaft, aber kohärent.*
- ***Feedback:*** *Du hast einige wichtige Punkte angesprochen, etwa Umwelt und Aufwand. Es ist gut, dass du beide Seiten beleuchtest. Deine Argumentation könnte klarer strukturiert sein, zum Beispiel durch gezieltere Vergleiche: „Zwar ist Windkraft umweltfreundlich, jedoch könnte der Bau die Natur stark beeinträchtigen.““*

Antwort 7 zu „*Musikbox*“:

„Für das Modell „EarFun UBOOM L“, da es stabile 16h Akkulaufzeit hat, viel mehr hört man nicht am Stück und kann sie zwischendurch einfach aufladen. Es hat eine TOP

Reichweite von 30m und ist sehr wasserdicht. Die Box wiegt nur einen halben Kilo und hat eine gute Größe.“

Einzelbeurteilungen:

ChatGPT-4o (Variante H1 & H2):
„Antwort Nr. 7

- ***Beurteilung Inhaltlich:*** *Niveaustufe 2/+*
- ***Beurteilung Sprachlich:*** *Niveaustufe 2/+*
- ***Gesamtniveau:*** *2.0*
- ***Begründung:*** *Mindestens vier Kriterien; Begründung, warum gewisse Werte „ausreichen“. Klare Sprache mit Begründung, warum kein höherer Wert nötig ist.*
- ***Feedback:*** *Sehr reflektierte Antwort! Du argumentierst aus einer persönlichen Perspektive und begründest deine Wahl logisch und nachvollziehbar.“*

ChatGPT-4o (Variante H3 & H4):
„Antwort Nr. 7

- ***Beurteilung Inhaltlich:*** *Niveau 2 / +*
- ***Beurteilung Sprachlich:*** *Niveau 2 / +*
- ***Gesamtniveau:*** *(2×2 + 2) ÷ 3 =* ***2.0***
- ***Begründung:*** *Klarer Bezug zu mehreren Kriterien, explizite Wertung („TOP Reichweite“, „praktisch“), und nachvollziehbare Bewertung. Sprache differenziert und strukturiert.*
- ***Feedback:*** *Deine Argumentation ist sehr gelungen: Du erklärst verständlich, warum dir bestimmte Eigenschaften wichtig sind. Die Begriffe wie „TOP“ zeigen persönliche Einschätzung, die du gut fachlich begründest. Bravo!“*

Aus den Einzelbeurteilungen zur Antwort Nr. 1 von „*Windenergie*“ und Antwort Nr. 7 von „*Musikbox*“ lässt sich gut erkennen, dass die Struktur der Outputs gleich ist. Beide Promptvarianten erzeugten eine zweigeteilte Beurteilung von Sprache und Inhalt, dem resultierenden Gesamtniveau sowie einer Begründung für die Bewertung und ein kurzes Feedback. Diese Struktur gaben beide Varianten vor. Die Variante H3 & H4 wurde in manchen Punkten etwas ausführlicher formuliert, beispielsweise bei den Anforderungen für das Feedback, um zu prüfen, ob das einen Einfluss auf die Qualität der Beurteilung hat. Für das Modell *ChatGPT-4o* ergaben sich diesbezüglich aber kaum Unterschiede zwischen den Varianten H1 & H2 und H3 & H4. Die Begründung für die Einstufung ist nicht ausführlicher geworden und inhaltlich sehr ähnlich. Lediglich das Feedback wurde etwas ausführlicher, da die Variante H3 & H4 explizit eine konkrete Verbesserungsidee als Beispielsatz für das Feedback angibt – das ist z.B. im Feedback zu Antwort 1 zu sehen. Für *ChatGPT-4o* scheinen die beiden Prompt-Varianten wenig Unterschied zu machen. Anders sieht es bei der Antwort Nr. 10 zu „*Schilddrüsenuntersuchung*“ aus. Hier ist ein klarer Unterschied im Output der Variante H3 & H4 zu sehen, den das neuere Modell *ChatGPT-5* generiert hat:

Antwort 10 zu „*Schilddrüsenuntersuchung*“:

a. -
b. *„Lieber Onkel Max, ich würde dir die Untersuchung sehr empfehlen. Der Stoff, der benutzt wird hat eine recht geringe Strahlenbelastung. Technetium 99 ist zu sehr schwach radioaktiv und er baut sich schnell ab. Die Untersuchung liefert sehr genaue Bilder und hilft der Ärtzen sehr, deine Schilddrüse richtig zu behandeln. Liebe Grüße“*

Einzelbeurteilungen:

ChatGPT-4o (Variante H1 & H2):
*„**Antwort Nr. 10***

- ***Beurteilung Inhaltlich: 2 / +***
- ***Beurteilung Sprachlich: 1 / ~***
- ***Gesamtniveau: 1.67***
- ***Begründung:*** *Du nennst relevante Kriterien (geringe Strahlung, genaue Bildgebung, schnelle Ausscheidung), aber es fehlt ein sprachlicher Vergleich.*
- ***Feedback:*** *Deine Empfehlung ist fachlich korrekt und hilfreich. Versuche, deine Argumente sprachlich noch stärker miteinander zu verbinden, um die Wirkung deiner Argumentation zu steigern.“*

ChatGPT-5 (Variante H3 & H4):
*„**Antwort Nr. 10***

- ***Beurteilung Inhaltlich:*** *2/+*
- ***Beurteilung Sprachlich:*** *1/~*
- ***Gesamtniveau:*** *1.7*
- ***Begründung:*** *Nennung mehrerer relevanter Kriterien: geringe Strahlenbelastung, schnelles Abklingen/Abbau und diagnostische Genauigkeit — das ist inhaltlich vollständig und begründet eine Empfehlung (2/+). Sprachlich werden mehrere Argumente aufgezählt, aber es fehlt ein deutliches Abwägen mit kontrastiven Formulierungen — daher 1/~.*
- ***Feedback:***
  - ***Stärken:*** *Gute und fachlich stichhaltige Argumentation (Risiko vs. Nutzen).*
  - ***Schwächen:*** *Stilistisch noch etwas knapp, keine konkrete Zahl (Halbwertszeit), minimale Rechtschreibfehler („Ärtzen“ → „Ärzten“).*
  - ***Verbesserungsidee (Beispiel):*** *Ergänzen Sie kurz die Halbwertszeit und nutzen Sie ein kontrastierendes Bindewort: „Zwar ist es ionisierend,* ***aber*** *Technetium-99m hat eine Halbwertszeit von ca. 6 h und die Belastung ist gering; die Untersuchung liefert sehr genaue Bilder. Ich empfehle sie.““*

Die Grundstruktur der Outputs von *GPT-4o* und *GPT-5* ist auch hier gleich, jedoch generiert *GPT-5* die Begründung für die Einstufung etwas genauer und unterteilt das Feedback in die drei im Prompt H4 vorgegebenen Unterpunkte *Stärken*, *Schwächen* und *Verbesserungsidee*. Anhand der Einzelbeurteilung zu Nr. 10 zeigt sich, dass die Wahl des Prompts, wie in der Literatur vielfach nachzulesen ist, eine Auswirkung hat. In diesem Fall jedoch nur beim Vergleich zwischen den Modellen *GPT-4o* und *GPT-5* zu sehen. Möglicherweise sind für *GPT-4o* die Prompt-Varianten doch zu ähnlich formuliert, da die Outputs fast gleich ausfielen, wie in den Beispielen der drei Antworten illustriert.

Die Qualität des Feedbacks konnte im Vergleich zur Pilotstudie nur ein wenig verbessert werden, etwa durch die Konkretisierung von Verbesserungspotenzial mit einem Beispielsatz oder das gezielte Anführen von Stärken und Schwächen. Allgemein bleibt das Feedback bei *GPT-4o* dennoch wenig umfangreich, bei *GPT-5* ist dieses etwas ausführlicher. Vereinzelt enthält es Komponenten, die nach Hattie und Timperley (2007) oder Sommer (2012) zu wirkungsvollem Feedback gehören, wie in Kap. 2.3 beschrieben. Ohne die entsprechenden Angaben im Prompt generiert ChatGPT das Feedback aber logischerweise

nicht konkret nach diesen Merkmalen, geschweige denn nach einer bestimmten Feedbackmethode, wie z.B. der in Kap. 2.3 vorgestellten *Feedback-Hand* (Jonach & Wagner, 2017).

Beide KI-Modelle, *ChatGPT-4o* und *ChatGPT-5*, sind fähig dazu Beurteilungen anhand des 3-stufigen Modells von Hackner (2025) zu generieren, die im Mittel eine solide Übereinstimmung mit den Beurteilungen von Physiklehrkräften zeigen. Gemäß eines Sprachmodells sind die generierten Inhalte jedoch von Schwankungen betroffen und müssen kontrolliert werden. Durch die Hauptstudie kann weiters noch die Erkenntnis hinzugefügt werden, dass das neue ChatGPT-Modell von OpenAI, *GPT-5*, zwar zu ausführlicheren Outputs führt, aber nicht unbedingt zu genaueren und mehr übereinstimmenden Beurteilungen. Welchen Einfluss die Wahl des Prompts auf das Ergebnis genau hat, bleibt leider nicht ganz geklärt, da nicht beide Varianten (H1 & H2 sowie H3 & H4) mit beiden KI-Modellen getestet wurden. *GPT-5* generierte mit der Variante H3 & H4 einen ausführlicheren Output als *GPT-4o*.

7. Diskussion

In der vorliegenden Masterarbeit wurde eine zweigeteilte, empirische Studie zum Thema Bewertung von Bewertungskompetenz (auch *S-Kompetenz*) mithilfe künstlicher Intelligenz durchgeführt. Sie knüpft an die Forschung von Gstall (2024) und Hackner (2025) an und untersuchte die Forschungsfragen, ob und wie die Bewertung offener S-Kompetenz-aufgaben anhand der evaluierten Beurteilungsmodelle zur S-Kompetenz von Gstall und Hackner mit dem generativen KI-Sprachmodell *ChatGPT* automatisiert werden kann und welche Anzeichen in den generierten Bewertungen von ChatGPT darauf hinweisen, dass der Beurteilungsprozess praktikabel, reproduzierbar und valide ist. Die Forschung soll dazu beitragen, eine entlastende Alternative für die anspruchsvolle und aufwendige Bewertung von S-Kompetenz für Physiklehrpersonen zu finden und bestehende Limitationen von den Beurteilungsmodellen (Zeitaufwand, Handhabung, Objektivität) durch den Einsatz von KI zu minimieren.

Die Ergebnisse der aufeinander aufbauenden Pilot- und Hauptstudie haben gezeigt, dass die Beurteilung der Bewertungskompetenz von Schüler:innen durch das generative Sprach-modell *ChatGPT-4o* (sowie *ChatGPT-5*) anhand der evaluierten Beurteilungsmodelle zur S-Kompetenz (Gstall, 2024; Hackner, 2025) mit einfachen Mitteln generiert werden kann und, mit wenigen Einschränkungen, sowohl reproduzierbar als auch valide ist.
In der *Pilotstudie* konnte durch exploratives Vorgehen eine zweiteilige Promptvorlage, auf *One-shot* Komponenten (Steinmann & Piazza, 2024) basierend und nach dem *PROPER*-Framework (Gruber, 2024) strukturiert, erstellt werden, mit dem *ChatGPT-4o* strukturierte Beurteilungen nach den Vorgaben des 5-stufigen Bewertungsmodells von Gstall (2024) generiert. Die Beurteilungen konnten nach Diskussion mit der Betreuerin dieser Arbeit und einigen Expert:innen der Universität Wien als zufriedenstellend und erfolgreich betrachtet werden. Die Pilotstudie deckte das Potenzial für den Einsatz von ChatGPT im Beurteilen von S-Kompetenz auf, zeigte jedoch auch, dass die Genauigkeit, Konsistenz und Sicherheit der KI-Beurteilungen anhand Gstalls Modell an ihre Grenzen stößt, die auch mit leichten Änderungen im Prompting nicht erhöht werden können.
Die weiterführende *Hauptstudie* konzentrierte sich auf die Beurteilung neuer Aufgaben und Antworten von Schüler:innen anhand des 3-stufigen Beurteilungsmodells von Hackner (2025) für die Oberstufe. Die Ergebnisse der Hauptstudie zeigten, dass *ChatGPT-4o* und *ChatGPT-5* auch anhand des Modells von Hackner reproduzierbare und valide Bewertungen zur S-Kompetenz generieren können, deren Genauigkeit, Konsistenz und Sicherheit etwas höher sind als in der Pilotstudie. Die durchschnittliche Beurteilung von *ChatGPT-4o* und *ChatGPT-5* aus fünf Durchgängen zeigt hohe Übereinstimmung zur durchschnittlichen Beurteilung von Physiklehrkräften bei einer gleichzeitig hohen Konsistenz, also geringer Unsicherheit. In der Hauptstudie wurde die Promptvorlage aus der Pilotstudie angepasst und in zwei verschiedenen Formen getestet. Es zeigte sich, dass diese Anpassungen zu keiner signifikanten Änderungen der Beurteilung von ChatGPT führt. Lediglich die Qualität des Outputs, besonders im Vergleich zwischen *GPT-4o* und *GPT-5* erkennbar, kann damit beeinflusst werden. Dieses Ergebnis spricht dafür, dass – unter achtsamer Kontrolle auf

mögliche, KI-typische Mängel im Beurteilungsprozess – das generative KI-Sprachmodell *ChatGPT* von OpenAI (2022) auf Basis von *GPT-4o* (OpenAI, 2024) und *GPT-5* (OpenAI, 2025d) als unterstützende Hilfe zur Bewertung der S-Kompetenz verwendet werden kann.

Die Interpretation der Ergebnisse, Erkenntnisse und Auffälligkeiten dieser Masterarbeit ist herausfordernd. Die Bewertung von Bewertungskompetenz ist ein komplexes Thema, wie auch das Verstehen und Anwenden generativer Sprachmodelle. Wird ChatGPT verwendet, um die S-Kompetenz von Schüler:innen zu bewerten, gibt es viele Faktoren, die diesen Prozess und die generierten Beurteilungen in unbekannter Weise beeinflussen können.

Es ist bemerkenswert, wie nahe die Bewertungen von *ChatGPT-4o* und *ChatGPT-5* an die von Lehrkräften herankommt. Vor allem, wenn bedacht wird, wie wenig es nach den Erkenntnissen dieser Arbeit dafür braucht. Es sind nur zwei digitale Dokumente (das Modell und die Aufgabenstellung inklusive Antworten), eine Promptvorlage und eben ChatGPT. In kurzer Zeit lässt sich damit eine Beurteilung inklusive kurzem Feedback zu allen Antworten generieren. Nach den Ergebnissen der Arbeit kann eine einmalige Beurteilung des Chatbots als modellkonforme und tendenziell passende Bewertung interpretiert werden. Dennoch muss diese unbedingt überprüft werden.

Das Einlesen in die Funktionsweise generativer Sprachmodelle und das Auseinandersetzen mit effektivem Prompting sind wichtige Schritte, um KI-Modelle für bestimmte Aufgaben zu nutzen. Die Outputs werden durch gutes Prompting zufriedenstellend generiert und je nach Beurteilungsmodell und Aufgabenstellung zeigen die Bewertungen von *ChatGPT-4o* und *ChatGPT-5* mittlere bis hohe Übereinstimmung zu den Bewertungen von Physiklehrkräften. In Anbetracht der Tatsache, dass ChatGPT keine wahre, denkende Intelligenz ist, sondern ein neuronales Netzwerk, dass mit Wahrscheinlichkeitsberechnungen Lernfähigkeit zeigt und gut im Lösen komplexer Aufgaben wird, stellt sich jedoch die Frage, warum ChatGPT beim Beurteilen einer Antwort nicht immer zum gleichen Ergebnis kommt und was das für diesen speziellen Einsatz bedeutet. Der Grund dafür liegt an technischen Eigenschaften von LLMs. Aufgrund ihrer Bauweise sind die generierten Inhalte von Fehlern, Schwankungen und Ausreißern bis hin zu Halluzinationen betroffen. Das erklärt die leichten und starken Abweichungen der KI-Beurteilungen in der Pilot- und Hauptstudie. Für den Einsatz als Beurteilungshilfe heißt das, dass die Bewertungen von *ChatGPT-4o* und *ChatGPT-5* im Idealfall als Durchschnitt aus mehreren entstehen sollten (um Einzelschwankungen auszugleichen) und immer durch die Lehrperson kontrolliert werden müssen. Gleiches gilt auch für generiertes Feedback.

Damit liegt eine wichtige Limitation dieser Masterarbeit auf der Hand: Interne Rechen- und Gewichtungsprozesse von ChatGPT beeinflussen generierte Beurteilungen. Dazu gehören auch individuelle Konfigurationsmöglichkeiten (die hier nicht getestet wurden) und andere Funktionen, wie etwa *Erinnerungen* oder *Reasoning*, als Faktoren, deren Auswirkung nicht bekannt ist. Diese internen Prozesse sind intransparent und es kann nicht überprüft werden, wie genau und sorgfältig ChatGPT nach den Kriterien der Modelle von Gstall (2024) und Hackner (2025) beurteilt. Oder auch, wie genau ChatGPT die Information der Aufgabenstellung und aller Antworten wirklich entnimmt. Als Folgerung müssen die Ergebnisse der

Arbeit daher so interpretiert werden, dass die Bewertung von Bewertungskompetenz mit ChatGPT zwar funktioniert, aber nicht nachverfolgt werden kann, wie der Chatbot dazu kommt.

Die Ergebnisse legen nahe, dass auch externe Faktoren, wie die Formulierung der Prompts, die Aufgabenstellung, die Antworten der Schüler:innen sowie das zugrundeliegende Beurteilungsmodell die Beurteilung beeinflussen können. Dass unterschiedliche Prompts zu variierenden Outputs führen können, ist weitgehend bekannt und wurde auch im Rahmen dieser Arbeit mehrfach beobachtet. Wie die Hauptstudie im Fall der finalen Prompt-Varianten H1 & H2 und H3 & H4 zeigt, führen Unterschiede zwischen den Varianten zwar nicht zu abweichenden Beurteilungen. Es ist jedoch denkbar, dass die gewählten Varianten einander zu ähnlich sind und sich potenzielle Unterschiede nicht zeigten. Eine vertiefende Untersuchung der Wirkung stärker divergierender Prompts hätte hier weiterführende Erkenntnisse liefern können. Das Design der Hauptstudie ist in diesem Punkt aber zu eingeschränkt, um sichere Aussagen über die Effektstärke von verschiedenen Prompt-formulierungen auf die Beurteilung treffen zu können.

In beiden Studien dieser Arbeit viel auf, dass ChatGPT die Antworten mancher Aufgaben tendenziell besser oder schlechter als Physiklehrkräfte beurteilte oder, dass die Unsicherheit von ChatGPT im Beurteilen der Antworten zwischen verschiedenen Aufgabenstellungen schwankt. Zweiteres kann sehr gut in der Hauptstudie beim Vergleich „*Musikbox*“ und von „*Schilddrüsenuntersuchung*“ festgestellt werden. Es kann sein, dass die Komplexität der Aufgabenstellung einen Unterschied für die Beurteilung von *ChatGPT-4o* bzw. *ChatGPT-5* macht. Wenn ChatGPT für den Einsatz als Beurteilungshilfe in Frage kommt, müsste man sich folglich auch Gedanken machen, wie eine Aufgabe zur S-Kompetenz gestellt sein muss, dass ChatGPT damit gut zurechtkommt. Diese Frage könnte Untersuchungsgegenstand einer weiterführenden Arbeit sein.

Die Arbeit zeigt, dass das 3-stufige Modell von Hackner (2025) für den Gebrauch mit ChatGPT etwas geeigneter erscheint als das 5-stufige Modell von Gstall (2024), da die KI-Beurteilungen in den Testreihen der Hauptstudie mehr Übereinstimmung und Sicherheit zeigen. Gründe dafür könnten in Unterschieden der beiden Modelle zueinander liegen. Das Modell von Gstall (2024) ist länger, komplizierter und weist nach Ermessen des Autors mehr Spiel-räume auf als das Modell von Hackner (2025). Das könnten Faktoren sein, die auch für die Bewertung mit einem KI-Modell schwierig sind. Es muss allerdings aufgezeigt werden, dass die menschlichen Beurteilungen, mit denen die KI-Beurteilungen verglichen wurden in der Pilotstudie nur von zwei Personen stammten und in der Hauptstudie von fünf. Außerdem können die Beurteilungsprozesse der beiden Studien nicht direkt miteinander verglichen werden, da sich die Anzahl und die Länge der Schüler:innenantworten unterscheidet: In der Pilotstudie generierte ChatGPT innerhalb eines Auftrages Beurteilungen zu mehr als 20 kurzen gehaltenen Antworten, in der Hauptstudie waren es etwa halb so viele und die bewerteten Antworten sind länger. Die Anzahl der Antworten hat vermutlich keine starke Auswirkung auf die Beurteilungen – dieser Faktor wurde in der Pilotstudie getestet. Die Länge der Antworten könnte allerdings einen Unterschied machen und ein Grund dafür sein, warum die Genauigkeit von ChatGPT in der Pilotstudie etwas

niedriger ist als in der Hauptstudie. Möglicherweise müsste für die Beurteilung von S-Kompetenz mit ChatGPT darauf geachtet werden, dass die Antworten lang genug und das Beurteilungsmodell konkret genug für ein Sprachmodell formuliert sind. Aber auch hierzu wären vertiefte Nachforschungen notwendig.

Was bedeutet die Vermischung der beiden KI-Modelle *GPT-4o* und *GPT-5* für die Aussagekraft der Hauptstudie? Während der Datenerhebung der Hauptstudie, Anfang August 2025, wurde das Basismodell von ChatGPT zu *GPT-5* geändert (OpenAI, 2025d). Der Großteil der Hauptstudiendaten stammt von *GPT-4o*, damit ist diese immer noch aussagekräftig genug, leider nur leicht verzerrt. Das bestätigt der geringfügige Unterschied in den Beurteilungen zwischen den KI-Modellen. Die Vermischung verzerrt wie bereits erwähnt auch die Effektstärke von verschiedenen Prompt-Varianten, da *GPT-5* beispielsweise nur mit der Variante H3 & H4 getestet wurde. Als Ausgleich dazu kann die Arbeit dafür nachweisen, dass *GPT-5* ausführlichere Outputs generiert, etwa mit besserem Feedback, als *GPT-4o*.

Wie lassen sich die Ergebnisse dieser Masterarbeit nun in den aktuellen Forschungsstand einordnen oder stellen neue Erkenntnisse dar? Der Einsatz von künstlicher Intelligenz in der österreichischen Bildung wird gezielt gefördert (BMBWF, 2025a, 2025b; eEducation, 2025). Im Zentrum steht dabei die Förderung digitaler Kompetenz der Lernenden, doch auch für Lehrende soll KI einen Nutzen haben. Beispielsweise wird der Einsatz von ChatGPT für die Erstellung von Bewertungen diskutiert. Es gibt positive Ansichten (Kasneci et al., 2023; Sok & Heng, 2023) wie auch kritische Einschätzungen (Mühlhoff & Henningsen, 2024) für die Verwendung des KI-Modells zu diesem Zweck. Als wichtige Referenzquelle für diese Arbeit wurde im Theorieteil die Studie von Mühlhoff und Henningsen (2024) thematisiert. Diese kritisierte die KI-Korrekturhilfe von *Fobizz*, die auf *GPT-4*, dem Vorgängermodell von *GPT-4o* beruht, stark und stufte sie auf anhand mehrerer Mängel und Fehler, die mit technischen Eigenschaften von LLMs verbunden sind, als ungeeignet ein. Die vorliegende Arbeit kann diese Kritik teils ergänzen. Die Beurteilungen der neueren Modelle *GPT-4o* und *GPT-5* sind von Schwankungen, Inkonsistenzen und vereinzelt Halluzinationen betroffen, die den Einsatz als Beurteilungshilfe in Frage stellen. Diese lassen sich bei generativen LLMs nicht ausschalten, sondern nur durch Kontrolle erkennen und ausbessern.

In der Literatur kann bis dato keine Studie gefunden werden, die sich gezielt mit dem Bewerten von S-Kompetenz mithilfe von ChatGPT auseinandersetzt. Die Erkenntnisse dieser Arbeit sind in diesem Forschungsfeld neu und liefern einen wichtigen Beitrag zum spezifischen Einsatz von ChatGPT im Beurteilen offener Aufgabenformate zur S-Kompetenz. Im Vergleich zur Studie von Mühlhoff und Henningsen (2024), die eine vorgefertigte KI-Korrekturhilfe für Schülerarbeiten untersuchte, wurde in dieser Arbeit ein individuell entwickelter, spezifischer Beurteilungsprozess evaluiert, der direkt auf der im Internet abrufbaren Version von ChatGPT basiert, manuelles Prompting verwendet und auf Modelle zur Bewertung der S-Kompetenz zurückgreift.

Die Untersuchung von generiertem Feedback durch *ChatGPT-4o* war nicht der Fokus dieser Arbeit, dennoch können die Ergebnisse einen kleinen Beitrag zur aktuellen Forschung leisten. In der Literatur wird die Integration von generiertem Feedback durch ChatGPT für

Schüler:innen positiv gesehen (Asadi et al., 2025; Kasneci et al., 2023), aber auch kritisch betrachtet (Steiss et al., 2024; Yoon et al., 2023). Die kurzen Analysen der generierten Feedbacks zur S-Kompetenz erzeugen den Eindruck, dass *ChatGPT-4o* das Potenzial für lernförderliche Rückmeldungen hat. Dieses Potenzial wurde durch die ungeplante Testung mit dem neueren KI-Modell *ChatGPT-5* erneut bestätigt, da vor allem der Feedbackteil der Einzelbeurteilungen durch *GPT-5* noch genauer generiert wurde als von *GPT-4o*. Ob die beobachtete Feedbackqualität von ChatGPT noch weiter verbessert werden kann, etwa durch gezieltere Angaben im Prompting, oder bereits ihre Grenzen erreicht hat, kann ohne weiterführende Forschungen nicht beantwortet werden.

Auch wenn sich ChatGPT qualitativ als unterstützende Beurteilungshilfe der S-Kompetenz herausstellt, muss die mangelnde Nachhaltigkeit, die mit dieser KI aktuell einhergeht, kurz thematisiert werden.
Ein problematischer Punkt an der Betreibung von Large Language Modells ist der hohe Energieverbrauch und die damit verbundene Umweltbelastung. Die Entwicklung und Verwendung von KI-Technologie geschieht bisher oft ohne Beachtung von Nachhaltigkeitsaspekten, obwohl diese für eine verantwortungsvolle Gestaltung und Nutzung von KI-Systemen notwendig wären (Autenrieth & Schluchter, 2025). Für das Training von *GPT-3* wurde eine gewaltige Menge von 552 Tonnen CO_2-Äquivalente erzeugt (Tomlinson et al., 2024). Neben ChatGPT's täglichen CO_2-Emissionen von 23.04 kg, belastet auch die hohe Wassernutzung, die vor allem für die Stromerzeugung und Kühlung der Server gebraucht wird, die Umwelt. Das Training von *GPT-3* benötigte über 700.000 Gallonen Süßwasser, was etwa der Menge entspricht, die für die Produktion von 370 BMW-Autos notwendig ist. Für einen normalen Chat von 20-50 Antworten benötigt ChatGPT etwa 500 ml Wasser. Diese kleine Menge summiert sich über Millionen von Nutzer:innen ebenfalls zu einer großen Umweltbelastung auf (Bhaskar & Seth, 2024, S. 56–58).
Die aktuelle Belastung von ChatGPT-Modellen sieht vermutlich anders aus, da die Zahlen mit dem älteren Modell *GPT-3* zusammenhängen. Wie stark der Einfluss von ChatGPT, bzw. KI-Systemen im Allgemeinen, auf die Umwelt ist, lässt sich meist nur einschätzen. Fakt ist aber, dass diese Systeme einen erkennbaren Fußabdruck hinterlassen und negative Effekte auf die Umwelt haben. Solange die Nachhaltigkeitsdefizite, die aktuell mit ChatGPT verbunden sind, bestehen, würde der Einsatz ChatGPT's für die Beurteilung von schulischen Leistungen – hier die S-Kompetenz – auch zur steigenden Umweltbelastung beitragen. Mit dieser Folge müssen sich Lehrkräfte bewusst sein.

Die Ergebnisse dieser Masterarbeit sind für das übergeordnete Forschungsthema „*Wie bewertet man Bewertungskompetenz*?" von hoher Relevanz und ergänzen die Arbeiten von Gstall (2024) und Hackner (2025). Gstall (2024) diskutierte ausführlich, wie schwierig die Bewertung von Bewertungskompetenz führ Lehrkräfte ist. Das 5-stufige Modell von Gstall bietet eine Unterstützung, weist aber dennoch Limitationen auf (Zeitaufwand, Handhabung, Objektivität). Das 3-stufige Modell von Hackner (2025) wirkt diesen Limitationen entgegenwirken, eine Beurteilung gestaltet sich jedoch immer noch aufwendig. Kann diese Masterarbeit, die den unterstützenden Einsatz von ChatGPT als Lösung für diese Probleme untersucht, neue Erkenntnisse bringen?

Ein wesentlicher Vorteil von ChatGPT als Beurteilungshilfe liegt in der hohen Konsistenz und relativen Objektivität seiner Bewertungen. Mehrfache KI-Beurteilungen derselben Antwort ermöglichen die Berechnung eines Mittelwerts sowie einer Differenzspanne zwischen minimaler und maximaler Beurteilung. Diese Differenz fällt in der Regel gering aus und steigt nur bei schwer zu beurteilenden Antworten an. Damit bestätigt die vorliegende Arbeit die Einschätzung von de Winter (2024), dass nur wiederholte Prompts und daraus gemittelte Bewertungen eine hohe Verlässlichkeit von KI-Bewertungen bieten. Lehrkräfte können diese Differenz als Indikator für die Bewertungssicherheit von ChatGPT nutzen: Eine geringe Differenz spricht für eine verlässliche Niveaueinschätzung, während bei einer hohen Differenz eine manuelle Nachprüfung ratsam ist. *ChatGPT-4o* und *ChatGPT-5* können gewinnbringend als unterstützende Instrumente zur Beurteilung der Bewertungskompetenz eingesetzt werden, aber der Vorteil verlässlicher Beurteilungen relativiert sich durch einen höheren Zeitaufwand, da für diese mehrere Wiederholungen erforderlich sind. In der Hauptstudie waren es beispielsweise fünf Runden.

Die Arbeit hat gezeigt, dass die Beurteilungsgenerierung mit ChatGPT praktikabel, reproduzierbar und valide ist. Der nächste Schritt wäre eine weitere Forschungsarbeit, in der diese KI-gestützte Beurteilung mit Physiklehrkräften erprobt wird. Dafür könnte ein anleitender Prototyp für den Einsatz von ChatGPT als Beurteilungshilfe entworfen werden, der den erprobten Beurteilungsprozess dieser Arbeit aufgreift. In offenen Interviews könnten Lehrkräfte, die eine KI-Beurteilung testen möchten, über ihre Erfahrungen zum Umgang mit ChatGPT reflektieren, die wiederum genutzt werden könnten, um den Prototyp zu einer anleitenden Handreichung weiterzuentwickeln. Ein interessanter Aspekt wäre dahingehend auch, wie Lehrkräfte die Bewertung mithilfe von KI wahrnehmen und (moralisch) bewerten. Denkbar wäre auch eine Umfrage unter Schüler:innen, wie diese eine Bewertung oder ein Feedback durch ChatGPT finden würden. Der Einsatz von künstlicher Intelligenz eröffnet viele Wege, in die weiter-geforscht werden kann.

Zum Abschluss der Diskussion möchte ich als Autor noch meine persönlichen Erfahrungen im Umgang mit ChatGPT zusammenfassen. Bevor ich mit der Masterarbeit begonnen habe, habe ich mich kaum mit künstlicher Intelligenz und generativen Sprachmodellen beschäftigt. Diese Masterarbeit bot mir eine großartige Gelegenheit, mich tiefer in die Funktionsweise und den effektiven Einsatz dieser neuen Technologie einzuarbeiten. Ich lernte, dass vor allem gutes Prompting der Schlüssel zur effektiven Nutzung von LLMs ist und ich konnte vor allem in diesem Bereich einiges dazulernen. Der größte Lernerfolg ist für mich jedoch, dass ich nach dieser Arbeit besser einschätzen kann, was ChatGPT kann und was nicht. ChatGPT kann in vielen Bereichen eine Unterstützung bieten, aber man darf Sprachmodellen nicht blind vertrauen.
Wichtig ist, sich darüber bewusst zu bleiben, dass KI-Systeme nicht „*intelligent*“ sind. Ihre Ausgaben können stets wahr, falsch oder unsinnig sein. Zumindest aktuell ist die menschliche Intelligenz noch immer notwendig, um die Richtigkeit des Outputs einzuschätzen (Stefan, 2019, S. 25f).

8. Fazit

Diese Masterarbeit knüpft an die Arbeiten von Gstall (2024) und Hackner (2025) an und ist Teil des Forschungsthemas „*Wie bewertet man Bewertungskompetenz?*“. Die Bewertungskompetenz, auch *S-Kompetenz*, ist ein zentraler Bestandteil des naturwissenschaftlichen Unterrichts. Ihre Bewertung gestaltet sich jedoch als schwierige Aufgabe für Lehrpersonen. Gstall (2024) und Hackner (2025) entwickelten neue Modelle zur Beurteilung der S-Kompetenz, die eine transparentere, objektivere und leichter zu handhabende Bewertung ermöglichen. Ziel der vorliegenden Arbeit war es herauszufinden, ob mithilfe generativer künstlicher Intelligenz (KI) eine Bewertung der S-Kompetenz anhand der Beurteilungsmodelle von Gstall (2024) und Hackner (2025) möglich, praktikabel und reproduzierbar ist und zu validen Beurteilungen führt. Zu diesem Zweck wurde der bekannte Chatbot *ChatGPT* von OpenAI (2022) ausgewählt und getestet.

In einer zweiteiligen empirischen Studie konnte gezeigt werden, dass *ChatGPT-4o* und *ChatGPT-5* potenzielle Werkzeuge zur Bewertung von offener S-Kompetenzaufgaben sind. In der ausführlich angelegten Pilotstudie wurden Grenzen und Potenziale von *ChatGPT-4o* in diesem Kontext ausgelotet. Das Ergebnis ist ein praktikabler und reproduzierbarer Beurteilungsprozess, der auf wenigen Materialien basiert. In der Hauptstudie wurde die Validität der Beurteilungen von *ChatGPT-4o* und *ChatGPT-5* genauer untersucht. Dabei zeigte sich, dass die Bewertungen des Chatbots im Vergleich zu den Bewertungen von Physiklehrpersonen eine mittlere bis hohe Übereinstimmung aufweisen, wenn diese als Durchschnitt aus mehreren (hier fünf) Beurteilungsrunden berechnet werden.

Damit ChatGPT eine funktionierende Beurteilungen generieren kann sind viele Faktoren zu berücksichtigen. Einer der wichtigsten Faktoren ist effektives *Prompting*, also das klare und strukturierte Formulieren des Auftrags an den Chatbot. Ebenso scheint es förderlich zu sein, wenn das Bewertungsmodell klar definiert ist und keine Spielräume lässt. Die Annahme, dass die Version von Hackner (2025) eine höhere Modellklarheit für ChatGPT aufweist als die Version von Gstall (2024) und daher besser für KI-Beurteilungen geeignet ist, hat sich verstärkt. Die Ergebnisse legen zudem die Vermutung nahe, dass die Gestaltung der Aufgabenstellung und die Form der Antwort selbst einen Einfluss auf die Beurteilung durch ChatGPT haben können. Diese Effektgrößen wurden in der vorliegenden Arbeit jedoch nicht genauer untersucht. Ebenso limitieren ChatGPT-interne Rechen- und Gewichtungsprozesse, die nicht eingesehen werden können, sowie personalisierte Funktionen die Frage nach dem Zustandekommen der Beurteilungen. Aus der Arbeit kann gefolgert werden, dass diese Faktoren einen Einfluss auf die Beurteilungen haben können, doch diese aufgrund des Black-Box-Charakters eines Sprachmodells nicht ergründbar sind.

Das Ziel dieser Arbeit wurde erreicht und sie kann einen wertvollen Beitrag zur Frage „*Wie bewertet man Bewertungskompetenz?*“ leisten. Mithilfe von ChatGPT ist es möglich, Bewertungskompetenz anhand vorgegebener Beurteilungskriterien mithilfe von ChatGPT beurteilen zu lassen. Ein klarer Vorteil dieser Methode ist, dass ChatGPT tendenziell sehr konsistent und, im Sinne von mathematischen Algorithmen, objektiv beurteilt. Die Tendenz

von ChatGPT kann der Lehrperson als Richtwert für ihre eigene Beurteilung dienen. Dieser Vorteil geht allerdings mit einem hohen Zeitaufwand einher, da für eine „sichere“ KI-Beurteilung mehrere Durchgänge empfohlen werden. Dies ist ein Nachteil von ChatGPT, der mit einer zugrundeliegenden Eigenschaften aktueller KI verbunden ist. ChatGPT ist ein generatives Sprachmodell, das aufgrund seiner technischen Funktionsweise (derzeit) von nicht vermeidbaren Mängeln betroffen ist. Generierte Inhalte können Fehler in verschiedener Form aufweisen. Die Ausgaben müssen auf Richtigkeit und Sinnhaftigkeit kontrolliert werden. Ein direktes Übernehmen generierter Inhalte wäre im Kontext der Bewertung der S-Kompetenz problematisch und unverantwortlich.

Der evaluierte Beurteilungsprozess mit ChatGPT sollte für interessierte Physiklehrkräfte einfach durchführbar sein und erfordert kein tieferes Vorwissen. Laut den Erkenntnissen der Arbeit sind dafür lediglich ein (kostenloser) Zugang zu ChatGPT, eine der entwickelten Promptvorlagen (s. Seite 90 und 91) sowie zwei digitale Dokumente, die man dem Chatbot hochlädt, erforderlich: eines der beiden Bewertungsmodelle und die Aufgabenstellung mit den zu beurteilenden Antworten von Schüler:innen. Wie der erprobte Beurteilungsprozess mit ChatGPT von Lehrkräften wahrgenommen wird, wurde in dieser Arbeit nicht untersucht und müsste weiterführend erforscht werden.

Zusammenfassend lässt sich festhalten, dass sich *ChatGPT-4o* und *ChatGPT-5* unter Berücksichtigung potenzieller Fehler zur Bewertung von S-Kompetenzaufgaben eignen. Die automatisierte Bewertung kann Physiklehrkräfte unterstützen und entlasten, indem sie als Richtwert oder Vergleichsgrundlage für eine Beurteilung genutzt wird. Eine KI-gestützte Beurteilung sollte jedoch nicht als Ersatz, sondern als ergänzendes Instrument zur Beurteilung durch Lehrkräfte betrachtet werden. Eine kritische Reflexion der Grenzen von ChatGPT und künstlicher Intelligenz im Allgemeinen ist ebenso unerlässlich wie die Reflexion der damit einhergehenden pädagogischen Verantwortung.

Literaturverzeichnis

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., … Zoph, B. (2023). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., Firat, O., … Vinyals, O. (2025). *Gemini: A Family of Highly Capable Multimodal Models* (arXiv:2312.11805). arXiv. https://doi.org/10.48550/arXiv.2312.11805

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., Shafey, L. E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., … Wu, Y. (2023). *PaLM 2 Technical Report* (arXiv:2305.10403). arXiv. https://doi.org/10.48550/arXiv.2305.10403

Asadi, M., Ebadi, S., & Mohammadi, L. (2025). The impact of integrating ChatGPT with teachers' feedback on EFL writing skills. *Thinking Skills and Creativity*, *56*, 101766. https://doi.org/10.1016/j.tsc.2025.101766

Autenrieth, D., & Schluchter, J.-R. (2025). Menschliche Existenz, (Nicht)Nachhaltigkeit & Künstliche Intelligenz. AI-Safety und AI-Alignment als Reflexionsgröße der Medienpädagogik. *Medienimpulse*, *63*(1). https://doi.org/10.21243/MI-01-25-25

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., … Kaplan, J. (2022). *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback* (arXiv:2204.05862). arXiv. https://doi.org/10.48550/arXiv.2204.05862

Bernauer, M. (2024). *Einsatz von Large Language Models zur automatisierten Erstellung von Arztbriefen* [Masterarbeit]. FH Hagenberg Oberösterreich.

Bhaskar, P., & Seth, N. (2024). Environment and sustainability development: A ChatGPT perspective. In J. Singh, S. Goyal, R. Kumar Kaushal, N. Kumar, & S. Singh Sehra, *Applied Data Science and Smart Systems* (1. Aufl., S. 54–62). CRC Press. https://doi.org/10.1201/9781003471059-8

BIFIE. (2011). *Kompetenzmodell Naturwissenschaften 8. Schulstufe*. Österreichisches Kompetenzzentrum für Didaktik der Chemie (AECC Chemie). https://aeccc.univie.ac.at/fileadmin/user_upload/z_aeccc/AECC_Chemie/Fuer_Lehrer_innen/Offizielle_Dokumente/Kompetenzmodell_NAWI_Sek_I_2011_v3.pdf

BMBWF. (2023). *Handreichung: Auseinandersetzung mit Künstlicher Intelligenz im Bildungssystem (Stand 29. August 2023)*.

BMBWF. (2025a, März 29). *Künstliche Intelligenz – Chance für Österreichs Schulen*. Bundesministerium für Bildung, Wissenschaft und Forschung. https://www.bmbwf.gv.at/Themen/schule/zrp/ki.html [Zugriff am 29.03.2025]

BMBWF. (2025b, Juli 30). *Digitale Schule*. Bundesministerium für Bildung, Wissenschaft und Forschung. https://www.bmb.gv.at/Themen/schule/zrp/dibi.html [Zugriff am 30.07.2025]

Bögeholz, S., Hößle, C., Höttecke, D., & Menthe, J. (2018). Bewertungskompetenz. In D. Krüger, I. Parchmann, & H. Schecker (Hrsg.), *Theorien in der naturwissenschaftsdidaktischen Forschung* (S. 261–281). Springer-Verlag GmbH Deutschland. https://doi.org/10.1007/978-3-662-56320-5

Breit, S., Bröderbauer, S., Friedl-Lucyshyn, G., Furlan, N., Kuhn, J.-T., Laimer, G., Längauer-Hohengaßner, H., Luthe, S., Haberfellner, C., Neureiter, H., Paasch, D., Pinwinkler, M., Rieß, C., Schreiner, C., & Siller, K. (2012). *Bildungsstandards in Österreich—Überprüfung und Rückmeldung*. 64.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Budhwar, P., Chowdhury, S., Wood, G., Aguinis, H., Bamber, G. J., Beltran, J. R., Boselie, P., Lee Cooke, F., Decker, S., DeNisi, A., Dey, P. K., Guest, D., Knoblich, A. J., Malik, A., Paauwe, J., Papagiannidis, S., Patel, C., Pereira, V., Ren, S., … Varma, A. (2023). Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT. *Human Resource Management Journal*, *33*(3), 606–659. https://doi.org/10.1111/1748-8583.12524

Bundeskanzleramt. (2024, August 1). *„AI Act" der EU: Weltweit erstes staatenübergreifendes Regelwerk in Kraft*. https://www.bundeskanzleramt.gv.at/themen/europa-aktuell/2024/08/ai-act-der-eu-in-kraft.html [Zugriff am 01.08.2025]

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., … Salakhutdinov, R. (2023). PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, *24*, 1–113.

De Florio-Hansen, I. (2020). *Digitalisierung, Künstliche Intelligenz und Robotik: Eine Einführung für Schule und Unterricht*. utb GmbH. https://doi.org/10.36198/9783838554297

De Winter, J. (2024). Can ChatGPT be used to predict citation counts, readership, and social media interaction? An exploration among 2222 scientific abstracts. *Scientometrics*, *129*(4), 2469–2487. https://doi.org/10.1007/s11192-024-04939-y

De Winter, J. C. F., Dodou, D., & Stienen, A. H. A. (2023). ChatGPT in Education: Empowering Educators through Methods for Recognition and Assessment. *Informatics*, *10*(4), 87. https://doi.org/10.3390/informatics10040087

Dejanovikj, J., & Höttecke, D. (2024). Promoting Reflections on Decision-Making in Socio-Scientific Issues. *RISTAL*, *7*, 41–60. https://doi.org/10.2478/ristal-2024-0004

eEducation. (2025, März 29). *KI Initiative des BM*. eEducation. https://eeducation.at/community/ki-initiative-des-bm [Zugriff am 29.03.2025]

Eggert, S. (2008). *Bewertungskompetenz für den Biologieunterricht – Vom Modell zur empirischen Überprüfung* [Dissertation]. Georg-August-Universität Göttingen.

Eggert, S., & Bögeholz, S. (2006). Göttinger Modell der Bewertungskompetenz—Teilkompetenz „Bewerten, Entscheiden und Reflektieren" für Gestaltungsaufgaben Nachhaltiger Entwicklung. *Zeitschrift für Didaktik der Naturwissenschaften : ZfDN*, *12*, 177–197. https://doi.org/10.25656/01:31623

Eidenberger, E. (2024, November 25). *Club der Cleveren x Maximilian Böger (Live). KI-Experte & Transformationsberater* (14) [Broadcast]. https://www.nachrichten.at/nachrichten/podcasts/clubdercleveren/club-der-cleveren-x-maximilian-boeger-live;art227491,3989579 [Zugriff am 25.11.2024]

Fleischmann, A. (2022). *ChatGPT in der Hochschullehre. Wie künstliche Intelligenz uns unterstützen und herausfordern wird*. NHHL. https://www.nhhl-bibliothek.de/de/handbuch/gliederung/#/Beitragsdetailansicht/243/3700/ChatGPT-in-der-Hochschullehre---Wie-kuenstliche-Intelligenz-uns-unterstuetzen-und-herausfordern-wird

Gatt, A., & Krahmer, E. (2018). Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65–170. https://doi.org/10.1613/jair.5477

Gentsch, P. (2018). *Künstliche Intelligenz für Sales, Marketing und Service. Mit AI und Bots zu einem Algorithmic Business – Konzepte, Technologien und Best Practices*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-19147-4

Goldberg, Y. (2016). A Primer on Neural Network Models for Natural Language Processing. *Journal of Artificial Intelligence Research*, *57*, 345–420. https://doi.org/10.1613/jair.4992

Gruber, P. H. (2024). *Unterlagen zum Online-Workshop: Generative KI als Werkzeug für die Forschung. Based on: Using ChatGPT for Advanced Data Analysis 3.0*.

Gstall, D. (2024). *Wie bewertet man Bewertungskompetenz? Entwicklung und Evaluation eines Beurteilungsmodells zur S-Kompetenz für die Unterstufe* [Masterarbeit]. Universität Wien.

Hackner, C. (2025). *Wie bewertet man Bewertungskompetenz? Weiterentwicklung und Evaluation eines Beurteilungsmodells zur S-Kompetenz* ***(nicht veröffentlicht)*** [Masterarbeit]. Universität Wien.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement* (Reprinted). Routledge.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Häußler, P. (2015). Wie lässt sich der Lernerfolg messen? In E. Kircher, R. Girwidz, & P. Häußler (Hrsg.), *Physikdidaktik: Theorie und Praxis* (S. 247–294). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41745-0

Heinrichs, B., Heinrichs, J.-H., & Rüther, M. (2022). *Künstliche intelligenz*. De Gruyter.

Henderson, M., Phillips, M., Ryan, T., Boud, D., Dawson, P., Molloy, E., & Mahoney, P. (2019). Conditions that enable effective feedback. *Higher Education Research & Development*, *38*(7), 1401–1416. https://doi.org/10.1080/07294360.2019.1657807

Heston, T., & Khun, C. (2023). Prompt Engineering in Medical Education. *International Medical Education*, *2*(3), 198–205. https://doi.org/10.3390/ime2030019

Hopf, M., Schecker, H., Höttecke, D., & Wiesner, H. (2022). *Physikdidaktik kompakt*. Aulis.

Hößle, C. (2013). Ethisches Bewerten im Unterricht. *Biologie in unserer Zeit*, *43*(2), 72–74. https://doi.org/10.1002/biuz.201390034

IQS. (2025, Juli 17). *Grundlagen der Bildungsstandards*. Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen. https://www.iqs.gv.at/themen/nationale-kompetenzerhebung/grundlagen-der-nationalen-kompetenzerhebung/grundlagen-der-bildungsstandards [Zugriff am 17.07.2025]

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Chan, H. S., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Jonach, M., & Wagner, C. (2017). *Individualfeedback für Lehrkräfte. Eine Handreichung für den Einsatz von Schüler/innenfeedback in QIBB. 2. Auflage*. https://rqb.oead.at/de/evaluation-und-feedback/individualfeedback

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and

challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kipper, J. (2020). *Künstliche Intelligenz—Fluch oder Segen?* J.B. Metzler. https://doi.org/10.1007/978-3-476-05137-0

Kreutzer, R. T., & Sirrenberg, M. (2019). *Künstliche Intelligenz verstehen: Grundlagen – Use-Cases – unternehmenseigene KI-Journey*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-25561-9

Kühleitner, J. (2019). *Drei Unterrichtsbeispiele zur Erweiterung der S-Kompetenz im Physikunterricht* [Masterarbeit]. Universität Wien.

Kulmhofer-Bommer, A., & Diekmann, N. (2021). Qualitätsentwicklung im österreichischen Schulsystem – Kompetenzorientierung als Leitkonzept und dessen Implementierung. In Bundesministerium für Bildung, Wissenschaft Und Forschung (BMBWF) (Hrsg.), *Nationaler Bildungsbericht Österreich 2021* (Version 1, S. 425–469). Federal Institute for Quality Assurance of the Austrian School System (IQS). https://www.iqs.gv.at/downloads/bildungsberichterstattung/nationaler-bildungsbericht-2021

Limpert, A.-L. (2024a, Juni 14). *EU-Regulierung von KI: Das besagt der „AI Act"*. Deutsches Schulportal der Robert Bosch Stiftung. https://deutsches-schulportal.de/schule-im-umfeld/eu-regulierung-von-ki-das-besagt-der-ai-act/ [Zugriff am 14.06.2025]

Limpert, A.-L. (2024b, Juli 3). *Künstliche Intelligenz im Unterricht: Was ist erlaubt?* Deutsches Schulportal der Robert Bosch Stiftung. https://deutsches-schulportal.de/unterricht/kuenstliche-intelligenz-im-unterricht-was-ist-erlaubt/ [Zugriff am 03.07.2025]

Liu, X., Wang, J., Sun, J., Yuan, X., Dong, G., Di, P., Wang, W., & Wang, D. (2023). *Prompting Frameworks for Large Language Models: A Survey* (arXiv:2311.12785). arXiv. https://doi.org/10.48550/arXiv.2311.12785

Lo, C. K. (2023). What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences*, *13*(4), 410. https://doi.org/10.3390/educsci13040410

Lux, B. (2022). *Fake News und Verschwörungstheorien—Ein Unterrichtsbeispiel zur S-Kompetenz im Physikunterricht* [Masterarbeit]. Universität Wien.

Mader, M. (2024). *ChatGPT in der Schule. Fluch oder Segen?* [Masterarbeit]. Universität Wien.

Mühlhoff, R., & Henningsen, M. (2024). *Chatbots im Schulunterricht: Wir testen das Fobizz-Tool zur automatischen Bewertung von Hausaufgaben* (Version 5). arXiv. https://doi.org/10.48550/ARXIV.2412.06651

Oertner, M. (2024). ChatGPT als Recherchetool? Fehlertypologie, technische Ursachenanalyse und hochschuldidaktische Implikationen. *Bibliotheksdienst*, *58*(5), 259–297. https://doi.org/10.1515/bd-2024-0042

OpenAI. (2022, November 30). *Introducing ChatGPT*. OpenAI.Com. https://openai.com/index/chatgpt/ [Zugriff am 16.06.2025]

OpenAI. (2023, März 14). *GPT-4*. OpenAI.Com. https://openai.com/index/gpt-4-research/ [Zugriff am 16.06.2025]

OpenAI. (2024, Mai 13). *Hello GPT-4o*. OpenAI.Com. https://openai.com/index/hello-gpt-4o/ [Zugriff am 16.06.2025]

OpenAI. (2025a, Februar 27). *Introducing GPT-4.5*. OpenAI.Com. https://openai.com/index/introducing-gpt-4-5/ [Zugriff am 18.06.2025]

OpenAI. (2025b, April 16). *Introducing OpenAI o3 and o4-mini*. OpenAI.Com. https://openai.com/index/introducing-o3-and-o4-mini/ [Zugriff am 26.04.2025]

OpenAI. (2025c, Juni 18). *Models*. OpenAI Platform. https://platform.openai.com/docs/models [Zugriff am 18.06.2025]

OpenAI. (2025d, August 7). *Introducing GPT-5*. https://openai.com/index/introducing-gpt-5/ [Zugriff am 08.08.2025]

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Park, J., & Choo, S. (2024). Generative AI Prompt Engineering for Educators: Practical Strategies. *Journal of Special Education Technology*, 01626434241298954. https://doi.org/10.1177/01626434241298954

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, *1*(8), 9.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., Hennigan, T., Menick, J., Cassirer, A., Powell, R., Driessche, G. van den, Hendricks, L. A., Rauh, M., Huang, P.-S., … Irving, G. (2021). *Scaling Language Models: Methods, Analysis & Insights from Training Gopher* (arXiv:2112.11446). arXiv. https://doi.org/10.48550/arXiv.2112.11446

RIS. (2025a, März 29). *Bundesrecht konsolidiert: Gesamte Rechtsvorschrift für Schulorganisationsgesetz, Fassung vom 29.03.2025*. Rechtsinformationssystem des Bundes. https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10009265 [Zugriff am 29.03.2025]

RIS. (2025b, April 30). *Bundesrecht konsolidiert: Gesamte Rechtsvorschrift für Leistungsbeurteilungsverordnung, Fassung vom 30.04.2025*. Rechtsinformationssystem des Bundes. https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10009375 [Zugriff am 30.04.2025]

RIS. (2025c, Juli 23). *Bundesrecht konsolidiert: Gesamte Rechtsvorschrift für Lehrpläne – allgemeinbildende höhere Schulen, Fassung vom 23.07.2025*. Rechtsinformationssystem des Bundes. https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10008568 [Zugriff am 23.07.2025]

RIS. (2025d, Juli 23). *Bundesrecht konsolidiert: Gesamte Rechtsvorschrift für Leistungsbeurteilungsverordnung, Fassung vom 23.07.2025*. Rechtsinformationssystem des Bundes. https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10009375 [Zugriff am 23.07.2025]

RIS. (2025e, Juli 30). *Bundesrecht konsolidiert: Gesamte Rechtsvorschrift für Schulunterrichtsgesetz, Fassung vom 30.07.2025*. Rechtsinformationssystem des Bundes. https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10009600 [Zugriff am 30.07.2025]

Russell, S., & Norvig, P. (2016). *Artificial intelligence: A Modern Approach* (Third edition). Pearson Education, Limited.

Sakschewski, M. T. (2013). *Bewertungskompetenz im Physikunterricht: Entwicklung eines Messinstruments zum Themenfeld Energiegewinnung, -speicherung und -nutzung* [Dissertation, Georg-August-University Göttingen]. https://doi.org/10.53846/goediss-4614

Sander, H., & Höttecke, D. (2015). Bewertungskompetenz in der Physikdidaktik: Zwischen Rationalität und Intuition. In B. Alexandra, K. Miriam, M. Michael, S. Frank, S. Kirsten, & W. Günther (Hrsg.), *Fachlich argumentieren lernen. Didaktische Forschungen zur Argumentation in den Unterrichtsfächern* (Bd. 7, S. 167–181). Waxmann Verlag GmbH. https://www.pedocs.de/frontdoor.php?source_opus=14021

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., … Resnik, P. (2025). *The Prompt Report: A Systematic Survey of Prompt Engineering Techniques* (arXiv:2406.06608). arXiv. https://doi.org/10.48550/arXiv.2406.06608

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Sok, S., & Heng, K. (2023). ChatGPT for education and research: A review of benefits and risks. *Cambodian Journal of Educational Research*, *3*(1), 110–121.

Sommer, A. (2012). Wertschätzendes Feedback in Schulen. *Lernchancen*, *15*(86), 28–32.

Stefan, V. (2019). *Artificial Intelligence and its Impact on Young People* [Seminar Report]. Council of Europe, European Youth Centre. https://rm.coe.int/ai-report-bil-final/16809f9a88

Steinmann, N., & Piazza, A. (2024). KI-basierte Textkreation im Content Marketing: Design und Evaluation eines effektiven Prompts. *HMD Praxis der Wirtschaftsinformatik*, *61*(2), 402–417. https://doi.org/10.1365/s40702-024-01058-3

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Stucki, T., D'Onofrio, S., & Portmann, E. (2018). Chatbot – Der digitale Helfer im Unternehmen: Praxisbeispiele der Schweizerischen Post. *HMD Praxis der Wirtschaftsinformatik*, *55*(4), 725–747. https://doi.org/10.1365/s40702-018-0424-8

Tabone, W., & de Winter, J. (2023). Using ChatGPT for human–computer interaction research: A primer. *Royal Society Open Science*, *10*(9), 231053. https://doi.org/10.1098/rsos.231053

Thoppilan, R., Freitas, D. D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., … Le, Q. (2022). *LaMDA: Language Models for Dialog Applications* (arXiv:2201.08239). arXiv. https://doi.org/10.48550/arXiv.2201.08239

Tomlinson, B., Black, R. W., Patterson, D. J., & Torrance, A. W. (2024). The carbon emissions of writing and illustrating are lower for AI than for humans. *Scientific Reports*, *14*(1), 3732. https://doi.org/10.1038/s41598-024-54271-x

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models* (arXiv:2302.13971). arXiv. https://doi.org/10.48550/arXiv.2302.13971

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models* (arXiv:2307.09288). arXiv. https://doi.org/10.48550/arXiv.2307.09288

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*.

Vladova, G., & Bertheau, C. (2023). Unter dem Zeichen Künstlicher Intelligenz. Berufe, Kompetenzen und Kompetenzvermittlung der Zukunft. In C. de Witt, C. Gloerfeld, & S. E. Wrede (Hrsg.), *Künstliche Intelligenz in der Bildung* (S. 393–410). Springer Fachmedien Wiesbaden GmbH. https://doi.org/10.1007/978-3-658-40079-8

Waack, S. (2025). Hattie-Rangliste: Einflussgrößen und Effekte in Bezug auf den Lernerfolg. *VISIBLE LEARNING*. https://visible-learning.org/de/hattie-rangliste-einflussgroessen-effekte-lernerfolg/

Wang, T., Roberts, A., Hesslow, D., Scao, T. L., Chung, H. W., Beltagy, I., Launay, J., & Raffel, C. (2022). What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization? *International Conference on Machine Learning*, *PMLR*(162), 22964–22984.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent Abilities of Large Language Models* (arXiv:2206.07682). arXiv. https://doi.org/10.48550/arXiv.2206.07682

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). *ReAct: Synergizing Reasoning and Acting in Language Models* (arXiv:2210.03629). arXiv. https://doi.org/10.48550/arXiv.2210.03629

Yenduri, G., Ramalingam, M., Selvi, G. C., Supriya, Y., Srivastava, G., Maddikunta, P. K. R., Raj, G. D., Jhaveri, R. H., Prabadevi, B., Wang, W., Vasilakos, A. V., & Gadekallu, T. R. (2024). GPT (Generative Pre-Trained Transformer)—A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. *IEEE Access*, *12*, 54608–54649. https://doi.org/10.1109/ACCESS.2024.3389497

Yoon, S.-Y., Miszoglad, E., & Pierce, L. R. (2023). *Evaluation of ChatGPT feedback on ELL writers' coherence and cohesion* (arXiv:2310.06505). arXiv. https://doi.org/10.48550/arXiv.2310.06505

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2025). *A Survey of Large Language Models* (arXiv:2303.18223). arXiv. https://doi.org/10.48550/arXiv.2303.18223

Hinweis des Autors

Zur sprachlichen Verbesserung dieser Arbeit wurden die KI-Tools *DeepL* und *ChatGPT* genutzt. Die Vorschläge dieser Werkzeuge wurden in keinem Fall wortgleich übernommen, sondern stets in angepasster, eigener Form verwendet. Dieser Nutzungshinweis gilt für die gesamte Arbeit und wird an den betroffenen Textstellen nicht erneut angeführt.

Anhang

Handreichung zum 3-stufigen Beurteilungsmodell von Gstall (2024, S. 84–86)

HANDREICHUNG ZUM MODELL ZUR BEURTEILUNG DER S-KOMPETENZ

Allgemeine Beurteilungshinweise

- Die Bewertungskompetenz beschreibt die Fähigkeit von Schüler*innen, „*Standpunkte [zu] begründen und aus naturwissenschaftlicher Sicht [zu] bewerten*“. (RIS 2024)
- Dieses Modell kann zur Beurteilung von Schüler*innen-Antworten auf offene Aufgabenformate verwendet werden. Dabei muss die Aufgabenstellung jedoch mindestens zwei voneinander verschiedene Wahlmöglichkeiten bieten, die sich in mindestens zwei (bestenfalls min. 3) Kriterien voneinander unterscheiden.
- Die Wahlmöglichkeiten werden im Modell als Optionen bezeichnet. Beispiele für Optionen sind verschiedene Haartrockner.
- Die Merkmale, nach denen sich die verschiedenen Optionen unterscheiden, werden Kriterien genannt. Beispiele für Kriterien sind der Preis oder die Leistung eines Geräts.
- Bei der Beurteilung ist der Inhalt der Argumentation von zentraler Bedeutung. Auch die Satzstruktur ist relevant, der Inhalt sollte jedoch bei der Beurteilung im Vordergrund stehen.
- Fehler in der Rechtschreibung und Grammatik sind nicht in die Beurteilung miteinzubeziehen. Man sollte diese aber markieren, um die Lernenden auf ihre Fehler aufmerksam zu machen.

Auf den ersten Blick

- Das Modell ist in fünf Spalten und sechs Zeilen unterteilt
- In der ersten Spalte (*Niveau*) sind die verschiedenen Niveaustufen ersichtlich. Die Niveaus reichen von 1 bis 5 und stehen in direktem Zusammenhang mit dem österreichischen Ziffernnotensystem. Niveaustufe 2 steht beispielsweise für die Note „Gut“.
- Die zweite Spalte (*Anzahl der Kriterien*) dient als erste Beurteilungshilfe. Die Lernenden verwenden in ihrer Argumentation eine bestimmte Anzahl an Kriterien (siehe oben). Diese kann als erste Orientierung zwischen den Niveaustufen dienen. Ist die Argumentation der Lernenden jedoch besser oder schlechter als die Anforderungen dieser Stufe, kann auch eine andere Stufe gewählt werden.
- In der dritten Spalte (*Inhaltliche Anforderungen*) befinden sich inhaltliche Beschreibungen der jeweiligen Niveaus. Die Niveaus sind grundsätzlich aufeinander aufbauend, wie auch bei der Anzahl der Kriterien kann es jedoch individuell Abweichungen geben. Wenn ein Kind beispielsweise nur ein Kriterium verwendet, aber verschiedene Optionen nach diesem Kriterium miteinander vergleicht, kann eine bessere Beurteilung gegeben werden.
- Die vierte Spalte (*sprachliche Merkmale*) beinhaltet Hinweise zu Satzstrukturen, die auf dem jeweiligen Niveau zu finden sind. Wie schon bei den allgemeinen Beurteilungshinweisen vermerkt, ist jedoch primär der Inhalt für die Zuordnung zu einer Niveaustufe verantwortlich. Die sprachlichen Merkmale einer Argumentation können jedoch auch Einfluss auf die Beurteilung haben.
- Die fünfte und letzte Spalte (*Anker-Beispiele*) zeigt für jede Niveaustufe eine beispielhafte Schüler*innen-Antwort, die mit diesem Niveau zu beurteilen wäre.

Niveau	Anzahl d. Kriterien	Inhaltliche Anforderungen	sprachliche Merkmale	Anker-Beispiele
5	0	Die Begründung(en) (falls vorhanden) werden auf Basis (naturwissenschaftlichen) Alltagswissens aufgebaut und sind deswegen fachlich unangemessen *und/oder* die Entscheidung wurde intuitiv getroffen *und/oder* die gezogene Schlussfolgerung ist falsch bzw. nicht zielführend oder für die Beantwortung der Aufgabenstellung irrelevant.	Die getroffene Entscheidung (=gewählte Option) wird nicht oder nicht anhand eines Kriteriums begründet.	*„Ich wähle Option A."* *„Ich wähle Option A, weil sie am besten ist."* *(0 Kriterien)*
4	1	Niveau 5 wurde durch eine naturwissenschaftliche und begründete Entscheidung übertroffen *und* möglicherweise wurden eigene Kriterien in die Argumentation miteinbezogen, *aber* die Begründung ist möglicherweise nicht durchgehend fachlich richtig.	Die Begründung kann sprachlich erkennbar sein. Die Argumentation folgt jedoch meist einer einfachen Struktur wie *„Option A ist die beste, weil Kriterium 1."*	*„Ich wähle Option A, sie ist ziemlich günstig."* *„Ich wähle Option A, weil sie ziemlich günstig ist und mir die Farbe gefällt."* *(1 Kriterium, Farbe ist irrelevantes Kriterium)*
3	2	Niveau 4 wurde erreicht *und* Kriterien, die in der Aufgabenstellung als irrelevant vermerkt wurden (z.B. Farbe eines Objektes: *Die Farbe ist für Sara egal*), werden nicht ernsthaft in die Argumentation miteinbezogen *und* Kriterien, die in der Aufgabenstellung als wesentlich vermerkt wurden (z.B. geringer Preis: *Sara möchte beim Kauf nicht viel Geld ausgeben*), werden in die Argumentation miteinbezogen.	Die Begründung ist sprachlich klar erkennbar, zum Beispiel durch Konjunktionen wie *weil, da, denn, nämlich, … .* Die Argumentation folgt meist der Struktur: "Option A ist die beste, weil Kriterium 1 + 2."	*„Ich wähle Option A, da sie am günstigsten ist und die höchste Leistung hat."* *(2 Kriterien)*
2	≥2	Niveau 3 wurde erreicht, *aber* es kann noch Interpretationsspielraum geben, dass noch mehr als die explizit genannten Kriterien Einfluss auf die Wahl der Schüler*innen hatten.	Die Argumentation geht über folgende Struktur hinaus: „Option A ist die beste, weil Kriterium 1 + 2." und beinhaltet gegebenenfalls weitere Kriterien und/oder Adverbien wie *außerdem/auch/ zusätzlich/…*	*„Ich wähle Option A, da sie am günstigsten ist und die höchste Leistung hat. Außerdem ist sie praktisch."* *(2 Kriterien + 1 angedeutet)*
1	≥2	Niveau 2 wurde erreicht. *und* Die Argumentation für eine Option ist inhaltlich durchgehend verständlich. Es gibt keinen Interpretationsspielraum oder Zweifel. *und* Die vorliegenden Kriterien und Optionen werden genutzt, um nicht nur den eigenen Standpunkt zu rechtfertigen, sondern auch, um konträre Standpunkte zu entkräften. *und/oder* Mehrere Kriterien und/oder Optionen werden im Laufe der Argumentation berücksichtigt und begründet gegeneinander aufgewogen. Kriterien, die beim Verglichen verwendet werden, zählen auch zur Anzahl der Kriterien.	In der Argumentation ist klar erkennbar, dass Optionen sprachlich miteinander verglichen werden, zum Beispiel durch Ausdrücke wie *zwar / aber / jedoch / trotzdem / …*	*„Ich wähle Option A, da sie im Vergleich zu den anderen am günstigsten ist und auch viele Funktionen bietet. Zwar bietet Option B eine höhere Leistung, der Unterschied ist jedoch gering, weshalb auch Option A zufriedenstellende Ergebnisse liefern wird."* *(3 Kriterien)*

S-Kompetenz im Unterricht fördern

Um die S-Kompetenz im Unterricht zu fördern ist es wichtig, Schüler*innen ausreichend nach Begründungen zu ihren Aussagen zu fragen. Haben Lernende einen Standpunkt eingenommen, sollte die Lehrperson deswegen eine Begründung dafür einfordern. Dafür eignen sich die drei Aufgaben (*Lampenwahl, Haartrockner, Die neue Heizung*), die im Rahmen der Master- Arbeit „*Wie bewertet man Bewertungskompetenz? - Entwicklung und Evaluation eines Beurteilungsmodells zur S-Kompetenz für die Unterstufe*" an Schüler*innen gestellt wurden. Diese Aufgaben können auch gemeinsam im Unterricht besprochen werden, sodass die Jugendlichen verschiedene Entscheidungsstrategien kennenlernen.

Eine weitere Möglichkeit ist, dass die Lehrperson Standpunkte vorstellt, sodass die Schüler*innen diese aus naturwissenschaftlicher Sicht bewerten müssen. Eine Bewertung im Sinne der S-Kompetenz ist jedoch keine bloße Zustimmung oder Ablehnung. Auch eine Bewertung soll begründet werden.

Aufgaben zur S-Kompetenz selbst erstellen

In der Master-Arbeit „*Wie bewertet man Bewertungskompetenz? - Entwicklung und Evaluation eines Beurteilungsmodells zur S-Kompetenz für die Unterstufe*" wurden die bereits erwähnten drei Aufgaben aus Aufgaben der iKM-Testung erstellt. Da die ursprünglichen Aufgaben von iKM jedoch vielmehr die Lesekompetenz als die S-Kompetenz abfragten, wurde in der Master-Arbeit auch thematisiert, welche Anforderungen es an Aufgaben zur S-Kompetenz gibt. Lehrpersonen, die selbst Aufgaben erstellen möchten, sollten sich an folgenden Anforderungen orientieren:

- Die Aufgabenstellung muss offen formuliert werden. Fragen mit Multiple-Choice- oder true/false-Charakter führen dazu, dass es richtige Antworten gibt. Auch Aufgaben, bei denen die Antwortmöglichkeiten gereiht werden, fallen unter dieses Schema. Bei der S-Kompetenz ist es jedoch wichtig, einen Standpunkt einzunehmen und diesen möglichst ausführlich naturwissenschaftlich zu begründen.
- Für die Arbeit mit dem Modell benötigt man mindestens zwei Optionen (siehe oben) aus denen die Schüler*innen wählen können. Außerdem müssen sich diese Optionen in mehrere Kriterien (siehe oben) voneinander unterscheiden.
- Im Idealfall folgt die Aufgabenstellung den fünf Merkmalen von Socio-scientific Issues, die Sakschewski (2013) in seiner Dissertation vorgestellt hat. Zunächst muss eine Situation vorliegen, die eine Nähe zu Naturwissenschaften und Gesellschaft aufweist und diskutiert werden kann (1). Das Problem ist dabei offen, es gibt keine eindeutigen Lösungsvorschläge (2) und der Sachverhalt ist durch mehrere soziale Aspekte beeinflusst, so zum Beispiel ethische oder politische (3). Die am Ende getroffene Entscheidung ist nicht final, da sich die Daten ändern können (4). Zuletzt ist es außerdem erforderlich, die zur Verfügung stehenden Quellen kritisch zu hinterfragen (5). Aufgabenstellungen diesbezüglich sind folglich weitestgehend offen formuliert und bieten mehrere Handlungsalternativen an.

Literatur zum Nachlesen

Gstall, Denise (2024): Wie bewertet man Bewertungskompetenz? Entwicklung und Evaluation eines Beurteilungsmodells zur S-Kompetenz für die Unterstufe. Universität Wien: Master-Arbeit

RIS - Rechtsinformationssystem des Bundes. Bundesrecht konsolidiert: Gesamte Rechtsvorschrift für Lehrpläne – allgemeinbildende höhere Schulen, Fassung vom 11.08.2024. https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10008568 [Zugriff: 11.08.2024]

Sakschewski, Mark Thorsten (2013): Bewertungskompetenz im Physikunterricht: Entwicklung eines Messinstruments zum Themenfeld Energiegewinnung, -speicherung und -nutzung. Göttingen: Dissertation

Aufgabe „*Die neue Heizung*“ mit Antworten aus Gstall (2024, S. 91–98)

Familie Fischbach plant, ein neues Haus zu bauen. Die Familie überlegt, welche Brennstoffart sie für ihre neue Heizung auswählen soll.

Folgende Brennstoffe stehen zur Auswahl:

Heizmedium	Kosten pro kWh	Heizwert	Anschaffung	Ursprung
Holz	5,95 Cent	4,2 kWh/kg	muss eingekauft werden	nachwachsender Brennstoff
Pellets	5,24 Cent	5 kWh/kg	muss eingekauft werden	nachwachsender Brennstoff
Heizöl	6,26 Cent	10 kWh/l	strömt direkt ins Haus	fossiler Brennstoff
Erdgas	6,31 Cent	10,6 kWh/m^3	strömt direkt ins Haus	fossiler Brennstoff

→ Was ist der Heizwert? Der Heizwert beschreibt, wie viel Energie bei der Verbrennung eines Stoffes nutzbar ist. Ein hoher Heizwert ist somit besser.

Gib an, welche Brennstoffart Familie Fischbach wählen soll und begründe deine Entscheidung.

Antworten der Klasse:

Antwort 1:
Heizöl: weil es 10kWh/l hat was viel ist, es direkt ins Haus strömt, es einen fossilen Brennstoff Ursprung hat und weil es nicht das teuerste ist.

Antwort 2:
Heizöl, weil es sehr P praktisch ist.

Antwort 3:
Pellets, weil es der künstigste ist, es kann mehr benützt werden als Holz und ist nachwachbar.

Antwort 4:
Ich wähle die Pellets, weil der am wenigsten kostet und dazu einen guten Heizwert hat.

Antwort 5:
Ich nehme Pellets weil es günstig ist, weil es einen guten Heizwert hat und weil es ein nachwachsender Stoff ist.

Antwort 6:
Familie Fischbach sollte das Pellets wählen, es ist nicht nur Billig sondern man kann es auch öfter benutzen.

Antwort 7:

Erdgas weil es das effizienteste ist, und im Vergleich aus das günstigtste. Nachteil: schlecht für die Umwelt.

Antwort 8:
Ich würde Erdgas nehmen weil der Heizwert am Höchsten ist und es strömt direkt ins Haus also ist es praktisch.

Antwort 9:
Ich würde Pellets nehmen, da es ein nachwachsender Brennstoff ist.

Antwort 10:
Pellets weil sie gut für die Umwelt sind und ein gutes Preis leistungsverhältnis hat.

Antwort 11:
Heizöl Familie Fischbach zalt wenig muss nichts anschafen und hat nicht das Risiko einer Gasexplosion oder einem aus durch Sanktionen gegen Erdgaslieferanten bleibt, nur als Nachteil die Klimaverschmutzung.

Antwort 12:
Ich würde Erdgas nehmen da es den höchsten Heizwert hat, dafür natürlich ihre höheren Kosten aber ich finde es am effektivsten. Wenn man jedes Mal den Brennstoff einkaufen muss ist das sehr Zeitaufwendig daher ist es besser wenn es direkt ins Haus strömt.

Antwort 13:
Pellets es ist eines der billigsten Brennstoffe, aber die Familie muss bereit sein nachzukaufen.

Antwort 14:
Pellets weil, es am gunstigsten ist, nachwachsend ist. Die Anschaffung ist zwar etwas schwieriger aber trotzdem Pellets. Meine zweite erste Wahl ist Holz.

Antwort 15:
Ich würde Pellets nehmen da es billig ist und nachwachsen kann.

Antwort 16:
Ich würde Holz nehmen, weil es bilig ist und man kann Bäume wieder anpflanzen.

Antwort 17:
Ich würde Pellets nehmen weil es ein nachwachsender Brennstoff ist und besser als Holz wegen den Heizwert aber billiger als Heizöl.

Antwort 18:
Pellets weil es ein Nachwachsender brennstoff ist und ziemlich billig ist.

Antwort 19:
Pellets, da es am günstigsten ist. Einkaufen zu gehen würde mir nichts schlimmer machen. Also Pellets.

Antwort 20:
Erdgas, weil es pro kwh am günstigsten ist und nicht eingekauft werden muss.

Antwort 21:
Erdgas, weil es am meisten Leistung hat und es direkt ins Haus strömt.

Antwort 22:
Heizöl wäre am effizientestenz weil es 10kWh/l enthält mit weniger kosten (kosten sind bezahlbar) aber schlecht für die Umwelt.

Aufgabe „*Haartrockner*“ mit Antworten aus Gstall (2024, S. 99–110)

Esra benötigt einen neuen Haartrockner. Sie hat sich informiert und verschiedene Modelle gefunden. Da ihr letzter Haartrockner zu rauchen begonnen hat, möchte Esra auch darauf achten, dass der Fön Prüfzeichen besitzt.

Folgende Haartrockner stehen zur Auswahl:

Gerät	Preis	Leistung	Länge des Kabels	Prüfzeichen
Travel	9,99€	1200 W	1,8m	keine
Profi Care	11,99€	1400 W	1,8m	keine
ProHair	24,99€	2000 W	2,5m	ÖVE
SuperFön	19,99€	1800 W	2,5m	
Advanced	60,99€	2300 W	3m	

2

Zeichenerklärungen

Produkte, die das Qualitätsanforderungsniveau erfüllen, tragen dieses Zeichen.

Geprüfte Sicherheit. Einem Produkt wird bescheinigt, dass es den Anforderungen des Produktsicherheitsgesetzes entspricht.

Das ÖVE-Prüfzeichen garantiert die Geräteprüfung durch eine unabhängige, kompetente Prüfstelle.

Chinesisches Prüfzeichen: Geprüft werden elektrische Sicherheit und elektromagnetische Verträglichkeit.

→ Eine geringe Leistung bedeutet, dass der Haartrockner energiesparender ist.

→ Bei einer hohen Leistung werden die Haare schneller getrocknet.

Gib an, welchen Haartrockner Esra wählen soll und begründe deine Entscheidung.

Antworten der Klasse:

Antwort 1:
Ich würde der Esra den Advanced nehmen. Er hat gute Leistung, ist sicher und hat gute qualität. Es ist zwar ein wenig teuerer aber wird sicher lange halten mit sein super-qualität.

Antwort 2:
SuperFön, weil:

1) *Es ist nicht teuer*
2) *Es hat eine gute Leistung*
3) *Kabel ist lang*

Antwort 3:
Ich werde Esra den ProHair-Föhn empfehlen. Warum? Begründung: Ich finde das der Preis für 24,99€ nicht zu teuer für so einen Föhn ist. Zweitens ist die Leistung bei 2000W, somit werden die Haare schneller trocken, aber auch die Energie wird bisschen gespart, als im Vergleich zu den anderen eine geringere Leistung hat und es durch ÖVE-Prüfzeichen geprüft wurde.

Antwort 4:
Sie soll den Advanced kaufen weil die Leistung viel höher ist und es ein längeres Kabel hat. Advanced ist zwar teuerer als die anderen aber ist es wahrscheinlich besser.

Antwort 5:
SuperFöhn, weil

1) *Die Länge des Kabels ausreichend lang ist.*
2) *Er ist ein wenig Energiesparend ist aber auch genügend leistungsfähig.*
3) *Er ist nicht soo teuer aber er besitzt alle nötigen Anforderungen.*

Antwort 6:
Prohair weil es kostet nur 5€ mehr als der SuperFöhn, aber verschwendet mehr Energie und hat nur 0,5cm weniger als der Advanced.

Antwort 7:
Sie soll Advanced kaufen. Es ist zwar teurer, aber es hat eine Bessere Leistung als die anderen.

Antwort 8:
Sie sollte Advanced nehmen, weil der Haartrockner das Qualitätsanforderungsniveau erfüllt und weil der Föhn den Produktsicherheitsgesetzes entspricht.

Antwort 9:
Ich würde den ProHair nehmen, weil er ein langes Kabel hat und er 2000W hat.

Antwort 10:
Esra sollte den ProHair Haartrockner nehmen, weil es hat eine gute und hohe Leistung von 2000W und es hat das Prüfzeichen: ÖVE. Dieses Prüfzeichen garantiert die Geräteprüfung durch eine unabhängige, kompetente Prüfstelle.

Antwort 11:
Ich würde Esra den Advanced Föhn empfehlen, denn er wurde von zwei Prüfzeichen geprüft und weil er eine hohe Leistung hat die Esra's Haare schnell trocknet.

Antwort 12:
Ich würde den Advanced nehmen, er ist der teuerste dafür hat ist die Länge des Kabels 3m lang und es hat die meisten Prüfzeichen.

Antwort 13:
Sie soll den Advancedfön nehmen, weil das Kabel das längste ist und das Gerät eine geprüfte Sicherheit hat.

Antwort 14:
Esra sollte den Superfön um 19,99€ wählen, weil es einen Prüfzeichen hat, der Länger des Kabels gleich lang ist wie dieser von dem ProHair, obwohl es mehr kostet. Und es hat den dritt geringsten Leistung mit 1800W, so ist es auch energiesparender als der ProHair Föhner. Aber es hat trotzdem im Vergleich zu den anderen zwei Künstigeren einen hohen Leistung, und damit trocknen die Haare auch besser.

Antwort 15:

Den Superfön wäre der beste Fön weil sein Preisleistungsverhältnis finde ich am besten ist. Außerdem kostet er weniger als der Advanced oder der Prohair und benötigt weniger Leistung als die 2. Und die Länge des Kabels ist genau so lang wie der Prohair.

Antwort 16:
Der ÖVE weil er ist nicht zu teuer und es garantiert die Geräteprüfung durch eine Prüfstelle. Ich finde das der Föhn viel Leistung hat und das Kabel hat eine gute länge.

Antwort 17:
ProHair, weil die Leistung gut ist und er nicht zu teuer ist. Außerdem hat er ein Prüfzeichen.

Antwort 18:
ProHair. Dieser Haartrockner hat eine hohe Leistung und trägt das Abzeichen ÖVE das beweist das der Haartrockner eine gute Kualität besitzt. Ausserdem hat er einen bezahlbaren Preis.

Antwort 19:
Ich würde mir den „ProHair" kaufen, weil es viele vorteile bietet wie z.B. eine Kabellänge von 2,5 was so ziemlich der durchschnitt entspricht. Eine Leistung von 2000W ist es auch über dem Durchschnitt. Und vor allem vertraue ich dem Prüfzeichen der „ÖVE". Weil es unabhängig ist und eine kompetente Prüfung verspricht.

Antwort 20:
Ich würde Travel und ProfiCare nehmen weil es nicht mal ein Prüfzeichen hat. Ich würde eher den ProHair nehmen weil es erstens ein bestimmten Prüfzeichen gibt und außerdem ist es relativ billig. Die Leistung passt relativ gut und die Länge des Kabels passt eigentlich perfekt. Nicht zu lang und zu kurz.

Antwort 21:
Advanced, weil das kabel lang ist und es entspricht den anforderungen.

Antwort 22:
Ich finde sie soll den ProHair nehmen, denn sie ist nicht zu teuer. Der Kabel ist auch lang genug, als ist es perfekt!

Antwort 23:
Sie soll den ProHair wählen weil er hat genug Watt, Und ist nicht zu teuer, er hat Prüfzeichen.

Aufgabe „*Windenergie*“ mit Antworten aus Hackner (2025)

In einer kleinen Berggemeinde mit ca. 1000 Einwohnern kam dem Bürgermeister eine Idee. Er möchte gerne den Energiebedarf seiner Gemeinde mit einer neuen Windkraftanlage decken. Eine solche Anlage bietet viele Vorteile aber auch Nachteile, deshalb wird in der Sitzung des Gemeinderats über die Errichtung und den Standort der Anlage diskutiert.
(IG Windkraft, 2024) (Sturmbauer, 2013)

Für die Wahl des richtigen Standortes sollten folgende Dinge beachtet werden: Windverhältnisse (fundierte Windmessung), Infrastruktur (breite Zufahrt für LKW), genügend Abstand zu bewohnten Gebieten (1200m in NO), Nutzungsrecht der Grundstücke (kaufen, pachten) und Netzanschlussmöglichkeiten.

a. Welche Position würdest du als Bürger dieser Berggemeinde vertreten? Begründe deinen Vorschlag ausführlich. (40-60 Wörter)
b. Merke an, ob die oben genannten Informationen ausreichen, um Stellung beziehen zu können. Notiere gegebenenfalls Informationen, welche dir helfen könnten, eine fundiertere Entscheidung zu treffen.

Antworten der Klasse:

Antwort 1:

a. *Einerseits ist es gut, um erneuerbaren, guten Strom zu bekommen. Auf einen längeren Zeitraum kann es gut für die Umwelt sein. Andererseits ist die Zeit und der Aufwand viel und ob die Bevölkerung das will, hängt davon ab, wie lange der Bau dauern würde. Es muss auch viel ausgebaut werden. Außerdem wird die Umwelt sehr zerstört, was mehr Schaden als Profit bringen wird.*
b. *Die vorhandenen Informationen reichen nicht aus. Beispielsweise wäre eine Preisrange auch von Vorteil.*

Antwort 2:

a. *Eine Windkraftanlage ist zwar umweltfreundlicher, es würde aber sehr in die Natur eingreifen. Die Menschen würden immer mehr bauen wollen, was für die Pflanzen, Tiere, … nicht von Vorteil ist. Abgesehen davon sind die Kosten auch viel höher am Anfang.*
b. *Nein, man müsste auch das Stromnetz beachten. Manche Regionen haben nicht genügend für eine Windkraftanlage, was dazu führt, dass noch mehr Kosten dazu kommen.*

Antwort 3:

a. *Ja, denn sie sind sehr umweltfreundlich und verursachen keinen CO2-Ausstoß, außerdem gibt es immer Wind.*
b. *Nein, man hätte erwähnen sollen, welche Vor/Nachteile es gibt und der Bürgermeister sollte ungefähr wissen, wo genau (Adresse) er es bauen will.*

Antwort 4:

a. *Ich würde sie anbauen, da man somit Energie gewinnen kann ohne viel Umweltverschmutzung anzurichten. Außerdem macht es Sinn eine Windkraftanlage in den Bergen anzubauen, da es dort viel Wind gibt. Man schafft auch dadurch den CO2-Außstoß auf ein Minimum zu bringen.*

b. *Auch wenn die gegebenen Informationen ausführlich wirken, muss man sich fragen, ob die Bürger und Bürgerinnen dieser Region mit dem Bau einverstanden sind. Das kann man mit Hilfe einer Volksabstimmung berechnen. Außerdem muss man darauf achten, ob es auch Sinn macht für diesen Bau Geld aus dem Budget zu verwenden, also ob man nicht lieber das Geld irgendwie anders investiert.*

Antwort 5:

a. *Prinzipiell ist eine Versorgung mit Windenergie eine gute Möglichkeit, um nachhaltig Strom zu erzeugen. Außerdem könnte ein Teil des erzeugten Stroms an andere Dörfer oder Gemeinden verkauft und somit Geld verdient werden. Andererseits ist für den Bau wie oben beschreiben eine gute Infrastruktur notwendig, die ein Dorf mit 1000 Einwohnern wahrscheinlich nicht hat. Aufgrund der dadurch und durch die Grundstücksmiete entstehenden hohen Kosten wäre ich eher gegen den Bau. Eine andere Möglichkeit wäre, die Kosten mit einem anderen Dorf zu teilen und dann auch den Strom zu teilen.*
b. *Es wird nicht erwähnt, ob das Dorf die genannten Voraussetzungen erfüllt z.B. wie die Windverhältnisse sind, wie gut die Infrastruktur ist usw. Mit diesen Informationen könnte man die Kosten besser abschätzen.*

Antwort 6:

a. *Ich würde ein Grundstück wählen, dass mehr als den Mindestabstand entfernt ist, sodass es die Bewohner nicht mehr stört, als es ihnen nutzt. Zudem würde ich das Windrad in eine freie Gegend hinstellen wo viel Wind hinkommt, aber auch LKWs keine Probleme haben. Vermutlich im Tal.*
b. *Bereitschaft der Bewohner, Watt die in das Stromnetz eingespeist werden, Windgeschwindigkeit, Dezibel der Lärmbelastung auf Entfernung, Kosten, Wartungskosten*

Antwort 7:

a. *Ich wäre gegen eine Errichtung der Windkraftanlage, weil die Nachteile wie z.B. Lärmbelastung, Preis oder Aufwand deutlich den Vorteil der sauberen Energie übertreffen. Stattdessen würde ich andere Alternativen wie Solarpanele empfehlen. Da die Gemeinde in den Bergen liegt, ist die Infrastruktur nicht ausreichend und die unberührte Natur wird möglicherweise sogar beschädigt.*
b. *Nein, die Informationen reichen nicht aus. Es würde helfen zu wissen, ob es genug Platz gibt, an dem, ohne der Natur oder Bewohnern zu schaden, gebaut werden kann.*

Antwort 8:

a. *als Bürger wäre mit wichtig, dass genügend Abstand zu wohnten Gebieten herrscht. Ich wäre prinzipiell dafür, da der Bau auch Arbeitsplätze schaffen würde.*
b. *wichtige wäre noch, ob die Strompreise dadurch sinken oder steigen würden und wie viele Arbeitsplätze geschaffen werden*

Antwort 9:

a. *Ich befürworte den Bau von Windkraftwerken, da es wichtig ist die Umwelt zu schützen. Aber sie sollten so weit weg wie es geht von unserer Gemeinde sein, da sie 1. nicht schön zu betrachten sind und 2. hoch sind, 3. viel Platz brauchen (LKW-Zugang, eher inkludiert) und 4. wir die Natur in unsere Gemeinde herum behalten sollten.*

b. *Nein. Fehlend: Durchmässe Windräder, Wie ist das Gebiet? → in den Bergen Windräder bauen schwer, Kosten?,*

Antwort 10:

a. *Dafür:*
   - erneuerbare Energie ist gut
   - besseren Strom
b. *Nicht wirklich. Fehlende Infos:*
   - Sind Windkraftanlagen laut?
   - Wie groß sind sie ungefähr?
   - Gibt es schon wo anders Windkraftanlagen in der Nähe von einer Gemeinde? → Feedback der Bürger?

Antwort 11:

a. *Als Bürger würde ich die neue Windkraftanlage definitiv bauen, da es meiner Gemeinde helfen würde und ich dafür bin, jedoch nur wenn der Lärm mich nicht belasten würde.*
b. *Wie viel Energie verbraucht wird, wie teuer es ist, es reicht nicht aus einige fehlen*

Antwort 12:

a. *Ich finde diese Idee nicht sehr gut. Denn für nur 1000 Einwohner machen die Windräder zu viel Lärm und die Energie, die die Anlage erzeugt ist viel zu viel für nur 1000 Einwohner. Es würde nur ein Windrad reichen um die ganze Gemeinde zu versorgen.*
b. *Nein, die Infos reichen nicht aus um Stellung zu beziehen. Infos wie die Lautstärke und Effizienz der Windräder würde helfen.*

Antwort 13:

a. *Ich bin dagegen. Für uns 1000 Einwohner würde ein einziges Windrad reichen. Da kann man das ganze gleich sein lassen. Außerdem ist das sehr sehr viel Platz der weggenommen wird. Es schaut auch nicht schön aus. Da kann man gleich eine Fotovoltaik-anlage bauen. Ist auch mutiger (rutiger?).*
b. *Nein Ich bin ja nicht wirklich ein Teil dieser Gemeinde, außerdem ist es eine „Berggemeinde". Ich weiß nicht wie die Lage ist, ob sich die 1200m überhaupt ausgehen.*

Aufgabe „*Musikbox*“ mit Antworten aus Hackner (2025)

Du möchtest dir eine tragbare Musikbox kaufen und folgende Produkte stehen zur Auswahl:

Modell	**EarFun UBOOM L**	**Vonmählen Air Beats Mini**	**Bose SoundLink Flex**	**Mas Carney F3**	**W-KING X10**
Kunden-bewertung bei Amazon*	4.5/5 Sterne 837 Bewertungen	3.0/5 Sterne 2 Bewertungen	4.5/5 Sterne 34990 Bewertungen	4.0/5 Sterne 444 Bewertungen	4.5/5 Sterne 4734 Bewertungen
Ausgangs-leistung	28 W	25 W	20 W	3 W	70 W
Max. Akkulaufzeit	16 h	13 h	12 h	8 h	42 h
Reichweite	30 m	-	ca. 9 m	10 m	ca. 30 m
Wasser-dichtigkeit	+ + +	+ + +	+ + +	-	+ + +
Masse	650 g	254 g	590 g	218 g	3.470 g
Größe H x B x T	21 x 7,8 x 7,2 cm	8 x 9,9 x 7,5 cm	5.23 x 20.14 x 9,04 cm	9,2 x 9,2 x 4,7 cm	19,9 x 37,3 x 17 cm

(L., 2025)

Entscheide dich für ein Produkt und begründe deine Wahl mithilfe der Tabelle.

Antworten der Klasse:

Antwort 1:
Bose SoundLink Flex; weil es nicht so schwer und leicht transportabel ist. Da die Akkulaufzeit 12h hält ist es angenehmer und durch die vielen Bewertungen bestätigt sich das Produkt.

Antwort 2:
Ich nehme die Ear Fun UBOOM L weil es halbwegs energie Effizient ist, eine gute Spieldauer und eine gute Reichweite hat. Zudem ist es für mich ergonomisch und leicht.

Antwort 3:
Die Preise sind zwar nicht angegeben, man kann aber trotzdem in etwa abwägen. Ich persönlich würde den Bose Soundlink Flex wählen, weil er von den Rezensionen her sehr beliebt ist, und daher auch wahrscheinlich viel billiger als der W-King X10, welcher eine viel höhere Leistung hat. Für meinen persönlichen Nutzen reicht auch die Leistung des Bose SoundLink Flex.

Antwort 4:
EarFun UBOOM L → ziemlich leicht, wasserdicht, viele gute Bewertungen, gut, dass der Akku so lange hält

Antwort 5:
Ich würde die „Bose Soundlink Flex" kaufen, da diese die meisten Bewertungen und 4,5 Sterne hat. Die 2.-meiste hat die „W-King X10", jedoch hat die am meisten Gewicht und ist ziemlich groß, was es schwer machen kann, sie herumzutragen. Außerdem haben beide dieselbe Bewertung bei der Wasserdichte.

Antwort 6:
Bose SoundLink Flex: hat die besten und meisten Bewertungen → viele finden sie gut, Bewertung wahrscheinlich am genauesten

Akkulaufzeit nicht am höchsten, aber für mich ausreichend; sehr wasserdicht; klein genug zum Tragen, nicht zu groß und nicht zu schwer

Antwort 7:
Für das Modell „EarFun UBOOM L", da es stabile 16h Akkulaufzeit hat, viel mehr hört man nicht am Stück und kann sie zwischendurch einfach aufladen. Es hat eine TOP Reichweite von 30m und ist sehr wasserdicht. Die Box wiegt nur einen halben Kilo und hat eine gute Größe.

Antwort 8:
Ich würde mich für die „EarFun UBOOM L" Musikbox entscheiden, da sie am praktischsten ist. Sie hat nicht so viele Bewertungen wie andere, trotzdem sind sie ausreichend. Die Ausgangsleistung ist außerdem am 2.höchsten, auch wenn sie den 1.Platz nicht annähernd erreichen kann. Eigentlich wäre die „W-KING X10" Musikbox die Beste, von dem her was sie kann, aber ihr Gewicht und ihre Grö0e machen sie unpraktisch zum Transportieren. Meine gewählte Musikbox hat ausreichend Kapazitäten und ist zudem um einiges leichter und kleiner, deshalb praktisch zum Transportieren.

Antwort 9:
EarFun UBOOM L

- *Gute Bewertungen*
- *Leistung eigentlich auch gut*
- *Eine relativ gute Akkulaufzeit (für mich)*
- *30m Reichweite → das Beste von allen mit W-KING X10*
- *Wasserdicht*
- *Wiegt viel weniger als W-KING X10 → für mich besser*
- *Größe ist für mich eigentlich unwichtig, aber sieht eh gut aus*

Antwort 10:
Wenn ich mich für eines der fünf Boxen entscheiden müsste, dann für den W-KING X10. Warum? 4,5 Sterne bei 4734 Bewertungen. 70W ist einfach das Optimum an Power. Ich bin vergesslich und 42h Akkulaufzeit hört sich äußerst berauschend an. Das Gewicht, gewichtet die von mir genannten wichtigen Aspekte eher weniger.

Antwort 11:
Ich entscheide mich für die Bose SoundLink Flex, da sie die meisten und besten Bewertungen hat, außerdem ist sie im Mittelfeld was die Leistung betrifft. Die Box ist nicht zu schwer, man

kann ca. 12 Std. Musik hören und sie ist wasserdicht. Sie hat zwar die kleinste Reichweite, aber bei einer tragbaren Musikbox braucht man nicht mehr als 9 m.

Antwort 12:
Bose SoundLink Flex, weil die Akkulaufzeit OK ist, ich brauche nicht mehr als 12 h am Stück eine Musikbox. Außerdem auch gut bewertet.

Antwort 13:
Bose SoundLink Flex, weil sie auch bei 34990 Bewertungen 4,5 Sterne hat. Des Weiteren ist sie mit 590g auch nicht so schwer. Mit 20W auch in der goldenen Mitte. Akkulaufzeit von 12h ist auch okay. Das einzige was mich persönlich bisschen stören würde ist die Reichweite von 9m.

Aufgabe „*Schilddrüsenuntersuchung*“ mit Antworten aus Hackner (2025)

Bei deinem Onkel Max wird eine Schilddrüsenüberfunktion vermutet, weshalb er sich im Krankenhaus einer Untersuchung unterziehen soll. Im Rahmen der Untersuchung wird ein radioaktives Material verabreicht, weshalb dein Onkel besorgt ist. Lies dazu die bereitgestellten Informationen genau.

Infobroschüre Schilddrüsenuntersuchung

Bei der Untersuchung bekommt man eine Flüssigkeit gespritzt, welcher ein radioaktiver Stoff (Technetium 99m) beigemischt ist, der sich dann in der Schilddrüse ablagert. Technetium 99m wird als Marker bezeichnet und sendet Gammastrahlung, welche mit Röntgenstrahlung vergleichbar ist, aus. Einige Stunden nach dem Einspritzen kann mit einer besonderen Kamera die Stärke der Gammastrahlung, die vom Technetium in der Schilddrüse ausgeht, für jeden Punkt in der Schilddrüse aufgenommen werden. Daraus kann man ein Bild errechnen, das krankhafte Veränderungen im Inneren der Schilddrüse erkennbar macht. Dieses Untersuchungsverfahren ist für die Diagnose von Schilddrüsenerkrankungen anerkannt. Allerdings kann jede Belastung mit ionisierender Strahlung Zellschäden bis hin zu Krebs auslösen. Ob es eine unschädliche Untergrenze gibt, ist in der Medizin allerdings umstritten. Sechs Stunden dauert es ungefähr, bis die Hälfte des Technetiums 99m zerfallen ist (=Halbwertszeit). Schon wenige Stunden nach der Untersuchung ist das gesamte Technetium vom Körper wieder ausgeschieden. Die durch die Untersuchung entstandene Belastung liegt deutlich unterhalb der jährlichen Belastung durch die natürliche Umweltstrahlung.

Infobroschüre Herstellung und Entsorgung

Um Technetium-99m zu erhalten, muss zuerst das Mutterisotop Molybdän-99 hergestellt werden. Brennstoffplatten, welche Uran enthalten, werden bestrahlt, wodurch sich das Uran spaltet. Eines dieser Spaltprodukte ist das Molybdän-99 mit einer Halbwertszeit von 66 Stunden. Anschließend werden diese Spaltprodukte weiterverarbeitet, wobei das Molybdän-99 abgetrennt wird. Dieses wird im Anschluss an die Kliniken geliefert. Beim Zerfall von Molybdän entsteht Technetium-99m, welches dort mit speziellen Generatoren

gefiltert wird. Die Entsorgung von Technetium-99m sieht vor, dass mehrere Halbwertszeiten abgewartet werden, bevor man es über den normalen Abfall entsorgt.

(International Atomic Energy Agency, 2010-2016) (Sturmbauer, 2013) (Grotelüschen, 2017)

a. Fasse die wichtigsten Kriterien aus den Infobroschüren zusammen.
b. Schreibe deinem Onkel einen kurzen Brief (40 – 60 Wörter) mit deiner Empfehlung, beachte dabei die Informationen aus den Broschüren.

Antworten der Klasse:

Antwort 1:

a. -

b. Lieber Onkel, ich habe mich mit deiner Untersuchung näher beschäftigt und kann Dir nun meine Erkenntnisse mitteilen. Obwohl die Untersuchung ein paar Nebenwirkungen mit sich bringt, ist dies ene einmalige Sache, die Wahrscheinlichkeit, dass du jene Nebenwirkungen durch die Umwelt aufnimmst ist höher, aus diesem Grund brauchst du Dir nicht allzu große Sorgen machen. LG

Antwort 2:

a. -

b. Lieber Onkel, ich würde dir empfehlen sich Meinungen von mehreren Ärzten anzuhören und dann entscheiden. Jede Behandlung hat Nebenwirkungen somit finde ich, dass du es versuchen solltest. Die Dose des radioaktiven Stoffes ist relativ gering und sollte nicht zur Belastung führen.

Antwort 3:

a. -

b. Lieber Onkel Max, mach dir keine Sorgen wegen der Untersuchung. Die radioaktive Substanz ist sehr gering dosiert und verschwindet fast komplett nach neun Stunden. Die Methode ist bewährt und hilft Veränderungen in der Schilddrüse genau zu erkennen. Ich empfehle dir, die Untersuchung ruhig durchführen zu lassen!

Antwort 4:

a. -

b. Hallo Onkel! Ich habe gehört, dass du wahrscheinlich Probleme mit deiner Schulddrüse hast und habe dazu ein paar Infos gefunden. Dir wird wahrscheinlich eine radioaktive Flüssigkeit namens Technetium 99m gespritzt. Damit kann man Veränderungen der Schilddrüse feststellen, da sie Gammastrahlen ausstrahlt. Aber keine Sorge! Die Ärzte sagen, dass es eigentlich ungefährlich ist. Alles Liebe und gute Besserung

Antwort 5:

a. -

b. Lieber Onkel, ich habe vor kurzem eine Broschüre mit den folgenden Infos gelesen: Es ging um einen Stoff namens Technetium (99), welches als Flüssigkeit in den Körper gespritzt wird. Es wirkt als eine Art Marker und ist mit Röntgenstrahlen vergleichbar. So können gefährliche Krankheiten in der Schilddrüse erkannt werden. Dies hört sich

vielleicht gefährlich an, aber habe keine Angst! Es wird nach kurzer Zeit vom Körper ausgeschieden. Es wird hergesgtellt indem zuerst Mutterisotop Molybdän 99 hergestellt wird. Es wird der Spaltstoff weiter verarbeitet und sodurch entsteht letztendlich Technetium 99m.

Antwort 6:

a. -
b. *Lieber Onkel, ich verstehe, dass du vesorgt bist, dass die Ärzte ein radioaktives Material verabreichen wollen aber ich habe es ein bischen recherchiert. Das Material ist Technetium 99m und sendet Gammastrahlung aus. Das wird verwendet um sichergehen zu können, dass du eine Schilddrüsenerkrankung hast. Leider kann es im schlimmsten Fall zu Krebs führen, jedoch würde ich mir darum keine Sorgen machen, da nach einige Stunden das gesamte Technetium aus dem Körper ausgeschieden ist. Außerdem ist es weniger belastend als die natürliche Umweltstrahlung. Mach dir keine Sorgen, Onkel! Liebe Grüße*

Antwort 7:

a. -
b. *Lieber Onkel, ich habe gehört, dass du Sorgen wegen der Untersuchung deiner Schilddrüse hast. Doch ich habe mich damit beschäftigt und bin zum Fazit gekommen, dass du dir keine Angst machen sollst. Es stellte sich heraus, dass die Strahlung des Technetium 99m weniger belastend ist, als unsere natürliche Umweltstrahlung auf den Körper. Liebe Grüße*

Antwort 8:

a. -
b. *Hi Onkel Max, ich habe mir die Broschüre durchgelesen und ich glaube du solltest die Behandlung machen! Die Behandlung ist nicht gefährlich wenn sie von Fachleuten gemacht wird, und die Strahlung die du abbekommen wirst ist viel kleiner als die Strahlung, die du jedes Jahr von der Umwelt und so bekommst. Die Wahrscheinlichkeit, dass du Krebs bekommst oder so ist echt winzig klein, und eine Schilddrüsenüberfunktion auszuschließen oder festzulegen ist es 100% wert. Tschau Tschau und viel Glück. Das Kind einer deiner Geschwister*

Antwort 9:

a. -
b. *Lieber Onkel Max! Ich habe mitbekommen das du besorgt bis, wegen deiner bevorstehenden Untersuchung. Also, die gute Nichte, die ich bin, habe ich ein wenig recherchiert. Der radioaktive Stoff Technetium 99 wird dir einespritzt aber dessen Halbwertszeit (die Zeit in der die Hälfte des Stoffes zerfällt) beträgt nur 6 Stunden und du bist bei dieser Prozedur deutlich weniger belastet als durch die jährliche natürliche Umweltbelastung. Also Onkel, keine Sorge, es wird schon alles super laufen. Liebe Grüße, deine Nichte!*

Antwort 10:

a. -

b. *Lieber Onkel Max, ich würde dir die Untersuchung sehr empfehlen. Der Stoff, der benutzt wird hat eine recht geringe Strahlenbelastung. Technetium 99 ist zu sehr schwach radioaktiv und er baut sich schnell ab. Die Untersuchung liefert sehr genaue Bilder und hilft der Ärtzen sehr, deine Schilddrüse richtig zu behandeln. Liebe Grüße*

Chat mit *ChatGPT-4o*: Testreihe „*Windenergie*“, Runde 2, Variante H1 & H2

Prompt H1:
[Modell zur Bewertung der S-Kompetenz_Oberstufe.pdf hochgeladen]
„Ich bin eine Lehrperson für Physik und möchte deine Unterstützung. Bitte beurteile für mich Antworten von Schüler:innen anhand eines vorgegebenen Beurteilungsmodells. Ich lade dir das Modell als File hoch. Analysiere das Modell ausführlich. Wenn dir etwas daran unklar ist, erkläre ich es dir gerne.“

ChatGPT-4o:
*„Danke für das Hochladen des Dokuments. Ich habe das **„Modell zur Bewertung der S-Kompetenz_Oberstufe“** gründlich analysiert. Hier eine strukturierte Zusammenfassung und Interpretation:*

🔍 *Modellübersicht: Bewertung der S-Kompetenz in der Oberstufe*

*Das Modell unterscheidet zwei separate **Dimensionen**:*

1. ***Inhaltliche Anforderungen***
2. ***Sprachliche Merkmale***

*Jede Dimension wird **differenziert nach drei Niveaustufen**:*
0/–, 1/~, 2/+
*Zusätzlich wird für die Gesamtbewertung ein **Kriterienzähler** verwendet (wie viele relevante Kriterien verwendet wurden).*

🧠 *1. Inhaltliche Anforderungen*

Niveau	*Merkmale*	*Beispiele*
0/–	*0–1 relevante Kriterien, intuitive oder falsche Begründung, möglicherweise irrelevante Kriterien verwendet*	*„Ich wähle A, sie ist günstig.“* *„Ich wähle A, weil sie günstig ist und mir die Farbe gefällt.“*
1/~	*≥2 relevante Kriterien aus der Aufgabenstellung, aber noch Interpretationsspielraum, evtl. irrelevante Kriterien*	*„Ich wähle A, da sie am günstigsten ist und die höchste Leistung hat.“*

Niveau	*Merkmale*	*Beispiele*
2/+	*≥2 relevante Kriterien und klare, widerspruchsfreie Argumentation, Abwägung von Optionen, Vergleich, ggf. Widerlegung*	*„Ich wähle A, da sie im Vergleich zu den anderen am günstigsten ist und viele Funktionen bietet. Zwar hat B mehr Leistung, aber...“*

👉 ***Hinweis**: Kriterien, die beim Vergleichen verwendet werden, zählen auch zur Kriterienanzahl.*

🗣️ *2. Sprachliche Merkmale*

Niveau	*Merkmale*	*Beispiele*
0/–	*Einfache Struktur, wenig sprachliche Ausdifferenzierung*	*„Option A ist die beste, weil günstig.“*
1/~	*Etwas differenzierter, Verwendung einfacher Konnektoren wie „auch“, „außerdem“*	*„...ist am günstigsten und hat auch die höchste Leistung.“*
2/+	*Expliziter Vergleich durch Konnektoren wie „zwar“, „aber“, „jedoch“*	*„...zwar hat Option B eine höhere Leistung, jedoch ist A günstiger...“*

📊 *Zusätzliche Bewertung: Gesamtniveau (optional)*

*Wenn gewünscht, kann zusätzlich ein **Gesamtniveau** berechnet werden:*

$$Gesamtniveau = \frac{2 \times Inhaltsniveau + Sprachniveau}{3}$$

(z. B. 2/+ inhaltlich, 1/~ sprachlich → Gesamtniveau = (2×2 + 1)/3 = 5/3 ≈1,67)

💬 *Falls noch unklar:*

Bitte teile mir mit, ob in deinen konkreten Bewertungen...

- *...das **Gesamtniveau** zusätzlich immer berechnet werden soll.*
- *...auch **implizite Kriterien** (z. B. „praktisch“) gezählt werden dürfen oder nur explizit genannte.*
- *...du Feedback für die Schüler:innen wünschst (z. B. motivierend, begründend etc.).*
- *...die beiden Teilbereiche **getrennt ausgewiesen** werden sollen oder nur ein Gesamtniveau wichtig ist.*

Ich bin bereit, mit der Bewertung zu beginnen, sobald du mir eine Aufgabe und die Antworten gibst.“

Prompt H2

[Aufgabe Windenergie_abgetippt.pdf hochgeladen]
„Beurteile bitte die Antworten der Schüler:innen zur Aufgabe „[NAME]“ im hochgeladenen File. Gehe dabei nach folgenden Schritten vor und gib alle deine Ergebnisse in einem Output an:

Schritt 1: Versetze dich in die Rolle einer Physik-Lehrperson.

Schritt 2: Analysiere die Aufgabe und alle Antworten dazu in dem PDF. Achte darauf, ob die Aufgabe einen oder mehrere Unterpunkte (z. B. a, b, ...) enthält. Wenn ja, lass alle Unterpunkte in die Beurteilung der Aufgabe einfließen / Beurteile nur die Antworten zu Unterpunkt X der Aufgabenstellung.

Schritt 3: Beurteile alle Antworten nach dem Beurteilungsmodell für S-Kompetenzaufgaben, das ich dir gegeben habe.

Schritt 4: Beurteile jede Antwort nach inhaltlichen und sprachlichen Anforderungen des Modells an. Berechne zusätzlich das Gesamtniveau pro Antwort:

- Gewichtung: Inhalt : Sprache = 2 : 1
- Formel: (2 × Inhaltsniveau + Sprachniveau) ÷ 3

Schritt 5: Formuliere die Beurteilung jeder Antwort nach diesem Beispiel:
Antwort Nr. „X“
- Beurteilung Inhaltlich: „Niveaustufe X“
- Beurteilung Sprachlich: „Niveaustufe X“
- Gesamtniveau: „Niveaustufe X.X“
- Begründung für die Beurteilung: Formuliere hier kurz, warum die Antwort die entsprechende Niveaustufe erreicht hat.
- Feedback: (Formuliere hier ein sachliches und motivierendes Feedback. Es soll Stärken, Schwächen und Verbesserungsmöglichkeiten der Antwort beinhalten.)

Schritt 6: Erstelle am Ende eine zusammenfassende Tabelle aller Beurteilungen mit diesen Spalten: | Antwort Nr. | Niveaustufen (Inhalt / Sprache) | Gesamtniveau | Kurze Begründung |”

ChatGPT-4o:
[Gespeicherte Erinnerung aktualisiert]
„Ich habe die Aufgabe „Windenergie“ und alle 13 Schüler:innen-Antworten analysiert. Hier folgt die vollständige Beurteilung nach deinen Vorgaben:

🔍 *Analyse der Aufgabe*

Aufbau der Aufgabe:
Die Aufgabe besteht aus zwei Teilaufgaben:

- ***Teil a:*** *Stellungnahme und begründete Entscheidung aus Sicht eines Gemeindebürgers (40–60 Wörter).*
- ***Teil b:*** *Reflexion über die Informationslage + ggf. Ergänzung weiterer nötiger Informationen.*

👉 ***Für die Bewertung*** *werden* ***beide Teile (a und b)*** *einbezogen.*

✅ *Beurteilungen der Schüler:innen-Antworten*

Antwort Nr. 1
- ***Beurteilung Inhaltlich:*** *Niveaustufe 1/~*
- ***Beurteilung Sprachlich:*** *Niveaustufe 1/~*
- ***Gesamtniveau:*** *1.0*

- ***Begründung:*** *Zwei relevante Kriterien werden angesprochen (Umweltfreundlichkeit, Aufwand), es gibt jedoch auch Unklarheiten (z. B. „mehr Schaden als Profit" ist unkonkret). Die Sprache ist einfach, aber es wird ein gewisser Kontrast angedeutet.*
- ***Feedback:*** *Du hast beide Seiten betrachtet, was gut ist. Versuche, deine Argumente klarer zu strukturieren und spezifischer zu machen. Begriffe wie „zerstört die Umwelt" solltest du genauer erklären. Auch wäre ein Vergleich zu anderen Möglichkeiten hilfreich.*

Antwort Nr. 2

- ***Beurteilung Inhaltlich:*** *Niveaustufe 1/~*
- ***Beurteilung Sprachlich:*** *Niveaustufe 0/–*
- ***Gesamtniveau:*** *0.67*
- ***Begründung:*** *Es werden Umweltaspekte und Kosten angesprochen, aber die Begründung ist eher oberflächlich. Sprachlich fehlt ein klarer Aufbau mit Kontrast oder Vergleich.*
- ***Feedback:*** *Du hast einige gute Gedanken eingebracht, z. B. zur Umweltbelastung und zu Kosten. Wenn du deine Aussagen mit konkreteren Beispielen oder einem Vergleich stärkst, wird deine Argumentation überzeugender.*

Antwort Nr. 3

- ***Beurteilung Inhaltlich:*** *Niveaustufe 0/–*
- ***Beurteilung Sprachlich:*** *Niveaustufe 0/–*
- ***Gesamtniveau:*** *0.0*
- ***Begründung:*** *Die Aussage „es gibt immer Wind" ist fachlich problematisch; keine genaue Auseinandersetzung mit Kriterien; Begründung ist zu kurz und oberflächlich.*
- ***Feedback:*** *Deine Aussage zeigt Engagement für Umweltschutz, aber es fehlen konkrete Kriterien und ein realistischer Bezug zur Situation in der Gemeinde. Achte darauf, ob deine Argumente fachlich zutreffend sind.*

Antwort Nr. 4

- ***Beurteilung Inhaltlich:*** *Niveaustufe 2/+*
- ***Beurteilung Sprachlich:*** *Niveaustufe 2/+*
- ***Gesamtniveau:*** *2.0*
- ***Begründung:*** *Mehrere relevante Kriterien (CO_2-Ausstoß, Windlage, Akzeptanz in der Bevölkerung, Budget) werden gegeneinander abgewogen. Die Argumentation ist differenziert und sprachlich durch Konnektoren sehr klar.*
- ***Feedback:*** *Sehr starke, gut durchdachte Argumentation! Du hast sowohl ökologische als auch soziale und finanzielle Aspekte eingebracht. Auch der Vorschlag zur Volksabstimmung zeigt kritisches Denken. Weiter so!*

Antwort Nr. 5

- ***Beurteilung Inhaltlich:*** *Niveaustufe 2/+*
- ***Beurteilung Sprachlich:*** *Niveaustufe 2/+*
- ***Gesamtniveau:*** *2.0*
- ***Begründung:*** *Differenzierte Betrachtung: Vorteile (Stromverkauf), Nachteile (Infrastruktur, Kosten) sowie Vorschläge (Kooperation mit Nachbardorf). Sprache unterstützt Vergleiche und Alternativen.*
- ***Feedback:*** *Du hast sehr gründlich argumentiert und viele Aspekte berücksichtigt. Besonders positiv ist, dass du auch Alternativen vorschlägst. Deine Sprache macht die Argumentation nachvollziehbar und überzeugend.*

Antwort Nr. 6

- ***Beurteilung Inhaltlich:*** *Niveaustufe 1/~*
- ***Beurteilung Sprachlich:*** *Niveaustufe 1/~*
- ***Gesamtniveau:*** *1.0*
- ***Begründung:*** *Es werden sinnvolle Kriterien genannt (Abstand, Windverhältnisse, LKW-Zufahrt), aber es fehlt eine Abwägung oder Bewertung. Sprachlich etwas sprunghaft, aber mehr als Basissatzstruktur.*
- ***Feedback:*** *Du hast gute Kriterien berücksichtigt – besonders Wind und Standort. Es wäre hilfreich, wenn du deine Entscheidung noch stärker begründest und Vorteile/Nachteile gegeneinander abwägst.*

Antwort Nr. 7

- ***Beurteilung Inhaltlich:*** *Niveaustufe 2/+*
- ***Beurteilung Sprachlich:*** *Niveaustufe 2/+*
- ***Gesamtniveau:*** *2.0*
- ***Begründung:*** *Es werden mehrere relevante Kriterien angesprochen und gegeneinander abgewogen (Natur, Infrastruktur, Alternativen). Sprachlich sehr strukturiert mit Kontrasten.*
- ***Feedback:*** *Deine differenzierte Argumentation überzeugt! Besonders gelungen ist der Vergleich mit anderen Möglichkeiten wie Solarpanelen. Du setzt dich kritisch mit der Situation auseinander – das zeigt hohe Kompetenz.*

Antwort Nr. 8

- ***Beurteilung Inhaltlich:*** *Niveaustufe 1/~*
- ***Beurteilung Sprachlich:*** *Niveaustufe 0/–*
- ***Gesamtniveau:*** *0.67*
- ***Begründung:*** *Es wird ein sinnvolles Kriterium (Abstand) genannt, aber die Argumentation bleibt vage und unausgewogen. Sprachlich sehr einfach.*
- ***Feedback:*** *Du hast einen relevanten Punkt genannt, aber deine Begründung könnte konkreter und strukturierter sein. Achte auch auf vollständige Sätze und klare Vergleiche, um deine Argumente besser zu vermitteln.*

Antwort Nr. 9

- ***Beurteilung Inhaltlich:*** *Niveaustufe 1/~*
- ***Beurteilung Sprachlich:*** *Niveaustufe 1/~*
- ***Gesamtniveau:*** *1.0*
- ***Begründung:*** *Mehrere Kriterien werden genannt, aber teilweise unklar formuliert oder nicht schlüssig begründet („nicht schön zu betrachten“). Die Sprache enthält Kontraste, aber nicht sehr klar.*
- ***Feedback:*** *Du hast viele Aspekte eingebracht – das ist gut! Versuch aber, stärker zu strukturieren und deine Aussagen klar zu begründen. Was bedeutet zum Beispiel „viel Platz“ genau?*

Antwort Nr. 10

- ***Beurteilung Inhaltlich:*** *Niveaustufe 0/–*
- ***Beurteilung Sprachlich:*** *Niveaustufe 0/–*
- ***Gesamtniveau:*** *0.0*
- ***Begründung:*** *Nur stichwortartige Aussagen, keine Begründung oder Argumentation erkennbar. Sprachlich zu fragmentarisch.*

- ***Feedback:*** *Versuche, ganze Sätze zu schreiben und deine Meinung zu begründen. Auch kurze Aussagen können gut sein – wichtig ist aber, dass sie zusammenhängend und nachvollziehbar sind.*

Antwort Nr. 11

- ***Beurteilung Inhaltlich:*** *Niveaustufe 1/~*
- ***Beurteilung Sprachlich:*** *Niveaustufe 0/–*
- ***Gesamtniveau:*** *0.67*
- ***Begründung:*** *Entscheidung und Kriterium (Lärmbelastung) sind genannt, aber ohne Vergleich oder tiefere Begründung. Sprachlich einfach, ohne Verknüpfungen.*
- ***Feedback:*** *Deine Meinung ist klar erkennbar – gut! Für ein höheres Niveau brauchst du jedoch mehr inhaltliche Tiefe und sprachliche Verbindungen wie „aber" oder „dennoch".*

Antwort Nr. 12

- ***Beurteilung Inhaltlich:*** *Niveaustufe 1/~*
- ***Beurteilung Sprachlich:*** *Niveaustufe 1/~*
- ***Gesamtniveau:*** *1.0*
- ***Begründung:*** *Es wird ein Gegenargument formuliert (Lärm, Überdimensionierung), allerdings fehlen genaue Vergleiche. Sprache ist einfach, aber klar.*
- ***Feedback:*** *Du hast nachvollziehbare Bedenken geäußert – gut gemacht! Versuche, deine Aussagen noch weiter zu begründen und alternative Lösungen einzubeziehen.*

Antwort Nr. 13

- ***Beurteilung Inhaltlich:*** *Niveaustufe 1/~*
- ***Beurteilung Sprachlich:*** *Niveaustufe 0/–*
- ***Gesamtniveau:*** *0.67*
- ***Begründung:*** *Es werden Argumente genannt, aber teilweise unsachlich („schaut nicht schön aus") oder unklar („rutiger?"). Sprachlich einfach.*
- ***Feedback:*** *Deine Meinung ist deutlich – das ist wichtig! Achte darauf, sachlich zu bleiben und Begriffe zu verwenden, die verständlich und nachvollziehbar sind.*

📊 *Zusammenfassende Tabelle*

Antwort Nr.	*Niveaustufen (Inhalt / Sprache)*	*Gesamtniveau*	*Kurze Begründung*
1	*1 / 1*	*1.0*	*Zwei Kriterien, aber unklar formuliert*
2	*1 / 0*	*0.67*	*Umwelt & Kosten genannt, sprachlich schwach*
3	*0 / 0*	*0.0*	*Oberflächliche Aussagen, fachlich fraglich*
4	*2 / 2*	*2.0*	*Abgewogene Argumentation, differenziert*
5	*2 / 2*	*2.0*	*Detailliert, Alternativen berücksichtigt*
6	*1 / 1*	*1.0*	*Kriterien da, aber wenig Abwägung*
7	*2 / 2*	*2.0*	*Sehr differenziert mit Alternativen*
8	*1 / 0*	*0.67*	*Gutes Kriterium, aber vage formuliert*
9	*1 / 1*	*1.0*	*Mehrere Aspekte, teils unklar*
10	*0 / 0*	*0.0*	*Keine Argumentation erkennbar*
11	*1 / 0*	*0.67*	*Klarer Standpunkt, einfache Sprache*

Antwort Nr.	*Niveaustufen (Inhalt / Sprache)*	*Gesamtniveau*	*Kurze Begründung*
12	*1 / 1*	*1.0*	*Gegenargumente vorhanden, wenig Tiefe*
13	*1 / 0*	*0.67*	*Meinung klar, sprachlich schwach*

Wenn du möchtest, kann ich diese Ergebnisse auch als CSV oder in tabellarischer Form exportieren. Sag einfach Bescheid!“

Chat mit *ChatGPT-5*: Testreihe „*Schilddrüsenuntersuchung*“, Runde 1, Variante H3 & H4

Prompt H3:

[Modell zur Bewertung der S-Kompetenz_Oberstufe.pdf hochgeladen]
„Ich bin Physiklehrkraft und brauche deine Unterstützung zur Bewertung der S-Kompetenz von Schüler:innen.
Grundlage ist ein zweidimensionales Beurteilungsmodell mit getrennten Niveaus für inhaltliche Anforderungen und sprachliche Merkmale (0/–, 1/~, 2/+).
Ich lade dir das Modell als Datei hoch.
Bitte analysiere das Modell sorgfältig und gib mir Rückmeldung, falls etwas unklar ist.
Anschließend möchte ich Schüler:innenantworten von dir nach diesem Modell beurteilen lassen – möglichst konsistent, genau und modellkonform.
Ich gebe dir dann genaue Anweisungen zum Bewertungsformat.“

ChatGPT-5:

„Ich habe dein Bewertungsmodell gründlich analysiert.
*Es ist klar in zwei Dimensionen gegliedert – **Inhaltliche Anforderungen** und **Sprachliche Merkmale** – jeweils mit den Niveaustufen **0/–, 1/~, 2/+**.*

Hier die Kernelemente, die für eine konsistente Anwendung wichtig sind:

1. Inhaltliche Anforderungen

- ***0/–**:*
  - *0–1 relevante Kriterien, oft nur auf Alltagswissen oder irrelevante Kriterien gestützt.*
  - *Intuitive oder fachlich unpassende Begründungen, falsche oder irrelevante Schlüsse.*
- ***1/~**:*
  - *Mindestens 2 relevante Kriterien aus der Aufgabe werden einbezogen.*
  - *Evtl. Interpretationsspielraum oder zusätzliche irrelevante Kriterien.*
- ***2/+**:*
  - *Argumentation durchgehend klar und ohne Interpretationsspielraum.*
  - *Optionen und Kriterien werden gezielt verglichen oder gegeneinander abgewogen.*

2. Sprachliche Merkmale

- ***0/–**:*
  - *Sehr einfache Struktur („Option A, weil Kriterium 1“).*
- ***1/~**:*
  - *Mehrere Kriterien in einem Satz verbunden, evtl. mit Erweiterungen wie „außerdem“.*

- *2/+:*
    - *Deutlicher sprachlicher Vergleich (z. B. „zwar ... aber ...“, „jedoch“, „trotzdem“).*

Wichtige Feinheiten

- *Ein **Vergleich zwischen Optionen** zählt bei der inhaltlichen Dimension nur dann als Kriterium, wenn er auf konkrete Kriterien Bezug nimmt.*
- *Implizite Begriffe wie „praktisch“ werden nicht automatisch als fachlich relevantes Kriterium gezählt.*
- *Bei gemischten relevanten und irrelevanten Kriterien bleibt die Einstufung abhängig vom Kern der Argumentation.*

*Falls du möchtest, kann ich dir jetzt ein **präzises Bewertungsraster** ableiten, damit die späteren Schüler:innenbewertungen absolut modellkonform und konsistent bleiben.*
Möchtest du, dass ich dir so ein Raster gleich erstelle?“

Prompt H4:

[Aufgabe Schilddrüsenuntersuchung_abgetippt.pdf hochgeladen]
Beurteile die Schüler:innenantworten zur S-Kompetenzaufgabe „[Aufgabentitel]“ aus dem hochgeladenen File anhand des Beurteilungsmodells, das ich dir bereits gegeben habe.

Vorgehensweise:

1. *Rolle einnehmen: Versetze dich in die Rolle einer Physiklehrkraft, die Schüler:innen fair, konstruktiv und modellkonform beurteilt.*
2. *Aufgabenverständnis: Analysiere die Aufgabenstellung und die darin enthaltenen Auswahlkriterien. Achte auf mögliche Unterpunkte (z. B. a, b, ...) der Aufgabe. Wenn es welche gibt, lass alle Unterpunkte in die Beurteilung der Aufgabe einfließen / Beurteile nur die Antworten zu Unterpunkt X der Aufgabenstellung.*
3. *Beurteilung: Beurteile jede Antwort nach dem zweidimensionalen Modell:*
    - *Inhaltliche Anforderungen (Niveaustufen 0/–, 1/~, 2/+)*
    - *Sprachliche Merkmale (Niveaustufen 0/–, 1/~, 2/+)*
4. *Berechne zusätzlich das Gesamtniveau pro Aufgabe:*
    - *Gewichtung: Inhalt : Sprache = 2 : 1*
    - *Formel: (2 × Inhaltsniveau + Sprachniveau) ÷ 3*
5. *Ausgabeformat: Gib für jede Antwort folgende strukturierte Beurteilung aus:*
   Antwort Nr. X
    - *Beurteilung Inhaltlich: Niveaustufe X*
    - *Beurteilung Sprachlich: Niveaustufe X*
    - *Gesamtniveau: X.X*
    - *Begründung: Kurze, klare Begründung, warum diese Niveaustufen erreicht wurden.*
    - *Feedback: Sachlich und motivierend. Gehe ausführlich auf Stärken und Schwächen ein und schreibe mindestens eine konkrete Verbesserungsidee anhand eines Beispiels.*
6. *Zusammenfassung: Erstelle eine zusammenfassende Tabelle aller Beurteilungen mit den Spalten: | Antwort Nr. | Niveaustufen (Inhalt / Sprache) | Gesamtniveau | Kurze Begründung |*

Achte durchgehend auf eine konsistente Struktur, klare Argumentation, fachliche Genauigkeit und eine wertschätzende Rückmeldung! Generiere die Feedbacks ausführlich und teile, wenn nötig, dafür deine Ausgabe auf zwei Outputs auf."

ChatGPT-5:

[Nachgedacht für 1m 6s]

*„Danke — ich beurteile jetzt **nur** die Teil-b-Antworten der Aufgabe „Schilddrüsenuntersuchung" aus deiner hochgeladenen Datei anhand des zweidimensionalen Modells. Ich beziehe mich auf die Aufgaben-/Antwortdatei und auf das Bewertungsmodell.*

Hinweise zur Vorgehensweise

- *Beurteilung gilt ausschließlich für **Teil b** (der kurzen Brief an den Onkel).*
- *Bewertung nach den beiden Dimensionen **Inhaltliche Anforderungen** (0/–, 1/~, 2/+) und **Sprachliche Merkmale** (0/–, 1/~, 2/+).*
- *Gesamtniveau = (2 × Inhaltsniveau + Sprachniveau) ÷ 3 — Ergebnis gerundet auf 1 Dezimalstelle.*
- *Bei Begründungen verweise ich konkret auf das, was die Schüler:innen in ihren Briefen schreiben (paraphrasiert).*

Beurteilungen (einzeln)

Antwort Nr. 1

- ***Beurteilung Inhaltlich:** 1/~*
- ***Beurteilung Sprachlich:** 2/+*
- ***Gesamtniveau:** 1.3*
- ***Begründung:** Nennenswerte inhaltliche Punkte: Erwähnung von „Nebenwirkungen" (Risiko) und direkter Vergleich mit der Umweltstrahlung („Wahrscheinlichkeit ... ist höher"). Es fehlen jedoch konkrete Fakten (Halbwertszeit, Exkretion, diagnostischer Nutzen), die ein eindeutiges, modellkonformes Argument ergeben würden. Sprachlich wird ein Gegensatz formuliert („Obwohl ... deshalb ..."), das zeigt, dass Risiken genannt und relativiert werden — deswegen 2/+.*
- ***Feedback (stärkend, konkret):***
  - ***Stärken:** Die Schülerin/der Schüler zeigt, dass er/sie die zentrale Sorge (Nebenwirkungen) wahrnimmt und relativiert sie durch einen Vergleich zur natürlichen Strahlung — das ist eine geeignete Argumentationsstrategie.*
  - ***Schwächen:** Es fehlen präzise Fakten (z. B. Halbwertszeit 6 h, schnelle Ausscheidung) und der Hinweis auf den diagnostischen Nutzen der Untersuchung. Die Formulierung ist etwas knapp; außerdem Tippfehler („ene") sollten korrigiert werden.*
  - ***Konkrete Verbesserungsidee (Beispiel):** Ergänze eine Zahl und ein kurzes Abwägen: „Zwar kann ionisierende Strahlung Zellschäden verursachen, **aber** die verabreichte Menge ist sehr gering (Halbwertszeit von Technetium-99m: ca. 6 h) und wird nach wenigen Stunden ausgeschieden; zugleich liefert die Untersuchung wichtige Bilder zur Diagnose. Deshalb brauchst du dir wenig Sorgen zu machen."*
    (Diese Version zeigt explizit das Abwägen — klarer, fachlicher und modellkonform.)

Antwort Nr. 2

- ***Beurteilung Inhaltlich:** 1/~*
- ***Beurteilung Sprachlich:** 1/~*

- ***Gesamtniveau:*** *1.0*
- ***Begründung:*** *Die Antwort nennt zwei in der Aufgabe relevante Punkte: Vorhandensein von Nebenwirkungen (Risiko) und dass die Dosis „relativ gering" ist (Entlastung). Die Empfehlung, mehrere Ärzte zu hören, ist sinnvoll, aber nicht auf die Informationsbroschüre bezogen. Sprachlich listet die Schülerin/der Schüler mehrere Argumente, verknüpft sie aber nicht als Abwägung (kein klarer Gegensatzausdruck).*
- ***Feedback:***
    - ***Stärken:*** *Erwähnung der geringen Dosis ist relevant und korrekt gewichtet.*
    - ***Schwächen:*** *Fehlende konkrete Fakten (Halbwertszeit, Exkretion, Vergleich zur natürlichen Strahlung) und fehlender Hinweis auf den diagnostischen Nutzen. Die Empfehlung „von mehreren Ärzten" ist eher ein Ratschlag zur Entscheidungsfindung als eine begründete Bewertung der Untersuchung.*
    - ***Verbesserungsidee (Beispiel):*** *Ergänzen und verbinden:*
      „Die Dosis ist sehr gering und das Technetium wird nach einigen Stunden ausgeschieden; die Untersuchung liefert jedoch verlässliche Bilder zur Diagnose. Du kannst zusätzlich ärztliche Zweitmeinungen einholen, aber die Broschüre spricht dafür, die Untersuchung durchzuführen."

Antwort Nr. 3

- ***Beurteilung Inhaltlich:*** *2/+*
- ***Beurteilung Sprachlich:*** *1/~*
- ***Gesamtniveau:*** *1.7*
- ***Begründung:*** *Sachlich starke Punkte: niedrige Dosierung, Hinweis auf Ausscheidung (hier mit „neun Stunden" genannt) und klarer Verweis auf den bewährten diagnostischen Nutzen. Mindestens zwei konkrete, relevante Kriterien sind benannt und bilden eine stimmige Empfehlung — daher 2/+. Sprachlich sind mehrere Argumente vorhanden, aber es fehlt ein explizites Abwägen mit Konnektoren wie „jedoch / zwar ... aber", deshalb 1/+.*
- ***Feedback:***
    - ***Stärken:*** *Gute Kombination aus Risikoabschätzung (geringe Dosis/Ausscheidung) und Betonung des Nutzens (bewährtes Verfahren). Das ist genau die Art von Argumentation, die das Modell belohnt.*
    - ***Schwächen:*** *Kleine Ungenauigkeit („neun Stunden" — in der Broschüre steht die Halbwertszeit ≈ 6 Stunden und „Schon wenige Stunden nach der Untersuchung ist das gesamte Technetium vom Körper wieder ausgeschieden"). Präzision ist wichtig.*
    - ***Verbesserungsidee (Beispiel):*** *Korrigiere die Zahl und füge einen kurzen Gegensatz: „Zwar handelt es sich um ionisierende Strahlung,* ***doch*** *die verabreichte Menge ist gering (Halbwertszeit ≈ 6 h) und wird rasch ausgeschieden; die Methode ist etabliert und hilft, krankhafte Veränderungen zu erkennen. Ich empfehle deshalb die Untersuchung."*

Antwort Nr. 4

- ***Beurteilung Inhaltlich:*** *0/–*
- ***Beurteilung Sprachlich:*** *0/–*
- ***Gesamtniveau:*** *0.0*
- ***Begründung:*** *Die Antwort beschreibt kurz, was passiert (Technetium 99m, Gammastrahlung) und gibt eine beruhigende Schlussbemerkung („Aber keine Sorge!"), nennt aber keine weiteren sachlich relevanten Kriterien (keine Halbwertszeit/Exkretion, kein Vergleich zur Umwelt, kein klarer Hinweis auf den diagnostischen Nutzen als Argument). Damit fehlen 2+ Kriterien → 0/–. Sprachlich ist die Argumentation sehr einfach.*

- ***Feedback:***
    - ***Stärken:*** *Einfache, verständliche Beschreibung des Verfahrens.*
    - ***Schwächen:*** *Zu oberflächlich für eine begründete Empfehlung; es fehlen die für Entscheidung relevanten Informationen.*
    - ***Verbesserungsidee (Beispiel):*** *Füge zwei Fakten und eine kurze Begründung hinzu: „Dir wird Technetium-99m gespritzt; zwar sendet es Gammastrahlung,* ***aber*** *die Dosis ist gering (Halbwertszeit ca. 6 h) und die Belastung liegt deutlich unter der jährlichen natürlichen Strahlung. Die Untersuchung hilft, Veränderungen in der Schilddrüse zu erkennen — daher würde ich sie empfehlen.“*
      (So werden Beschreibung + Abwägung + Empfehlung verbunden.)

Antwort Nr. 5

- ***Beurteilung Inhaltlich:*** *2/+*
- ***Beurteilung Sprachlich:*** *2/+*
- ***Gesamtniveau:*** *2.0*
- ***Begründung:*** *Umfassende Zusammenstellung: Technetium als Marker, Vergleich mit Röntgenstrahlung, diagnostischer Nutzen, Hinweis auf schnelle Ausscheidung und Erklärung zur Herstellung. Die Antwort nutzt eine kurze Gegenüberstellung („hört sich gefährlich an, aber …“), zeigt also Abwägung und mehrere relevante Kriterien — daher 2/+ in beiden Dimensionen.*
- ***Feedback:***
    - ***Stärken:*** *Sehr informativ und modellkonform: mehrere relevante Kriterien, klarer Satz mit Abwägung.*
    - ***Schwächen:*** *Etwas viel Herstellungs-/Prozessdetail für einen kurzen Beraterbrief (kann die Lesbarkeit stören).*
    - ***Verbesserungsidee (Beispiel):*** *Kürze die Herstellungsdetails zugunsten eines fokussierteren Arguments:*
      „Technetium-99m wird als Markersubstanz eingesetzt und liefert präzise Bilder; obwohl es radioaktiv ist, wird es schnell ausgeschieden und die Belastung ist gering. Deshalb: Untersuchung empfehlen.“
      (So bleibt die inhaltliche Qualität erhalten, Stil ist treffender für einen 40–60-Worte-Brief.)

Antwort Nr. 6

- ***Beurteilung Inhaltlich:*** *2/+*
- ***Beurteilung Sprachlich:*** *2/+*
- ***Gesamtniveau:*** *2.0*
- ***Begründung:*** *Mehrere relevante Punkte: Benennung des Isotops + Gammastrahlung, diagnostischer Zweck, offenes Eingeständnis des Krebsrisikos, relativierende Fakten (Ausscheidung nach einigen Stunden, geringere Belastung als natürliche Strahlung). Die Schülerin/der Schüler nennt Risiko und widerlegt/relativiert es — deutliches Abwägen, daher 2/+ in beiden Dimensionen.*
- ***Feedback:***
    - ***Stärken:*** *Sehr ausgewogene Argumentation: Risiko wird ehrlich genannt, dann fachlich entkräftet (Ausscheidung, Vergleich zur Umwelt). Modellkonform und für einen Familienbrief gut geeignet.*
    - ***Schwächen:*** *Kleinere Tippfehler/Stylistik („vesorgt“, „ein bischen“).*
    - ***Verbesserungsidee (Beispiel):*** *Zahl ergänzen und leicht formaler Ton:*
      „Zwar besteht theoretisch ein geringes Krebsrisiko, ***doch*** *die verabreichte Menge ist*

sehr klein (Halbwertszeit ≈ 6 h) und die Substanz wird rasch ausgeschieden; die Belastung liegt unter der jährlichen Umweltstrahlung. Darum: Untersuchung empfehlen."

Antwort Nr. 7

- ***Beurteilung Inhaltlich:*** *0/–*
- ***Beurteilung Sprachlich:*** *0/–*
- ***Gesamtniveau:*** *0.0*
- ***Begründung:*** *Kurz und knapp: Fazit „du brauchst keine Angst zu haben" mit einem einzigen Beleg (Strahlung von Tc-99m weniger belastend als natürliche Umwelt). Da nur ein (einfacher) relevant benannter Punkt vorhanden ist und kein Abwägen oder Zusatzinfo, fällt die inhaltliche Einstufung auf 0/–. Sprachlich ist die Aussage ebenfalls sehr einfach.*
- ***Feedback:***
  - ***Stärken:*** *Zielgerichtete Beruhigung — empathisch.*
  - ***Schwächen:*** *Zu wenig inhaltliche Substanz zur Begründung einer Empfehlung. Für das Modell sind mindestens zwei nachvollziehbare Kriterien nötig.*
  - ***Verbesserungsidee (Beispiel):*** *Mindestens ein weiteres Faktum ergänzen (z. B. Exkretion oder diagnostischer Nutzen):*
    „Die Belastung ist geringer als die natürliche jährliche Strahlung ***und*** *das Technetium wird nach wenigen Stunden ausgeschieden. Außerdem liefert die Untersuchung wichtige Bilder zur Diagnose — daher brauchst du dir kaum Sorgen zu machen."*

Antwort Nr. 8

- ***Beurteilung Inhaltlich:*** *2/+*
- ***Beurteilung Sprachlich:*** *1/~*
- ***Gesamtniveau:*** *1.7*
- ***Begründung:*** *Enthält mehrere relevante Punkte: geringe Strahlenmenge, Vergleich zur Umwelt, sehr geringe Wahrscheinlichkeit, an Krebs zu erkranken, und klarer Bewertungsauftrag („solltest die Behandlung machen"). Das ist inhaltlich stark. Sprachlich ist der Ton jedoch sehr umgangssprachlich („echt winzig klein", „100% wert") und weniger formal/abgewogen — deshalb 1/~.*
- ***Feedback:***
  - ***Stärken:*** *Klare, überzeugende Empfehlung mit mehreren relevanten Argumenten.*
  - ***Schwächen:*** *Zu salopper Ton für einen Brief an einen Onkel und etwas überdramatisierte Formulierungen („100% wert", „echt winzig klein"). Für Schulbeurteilung zählt auch angemessener Stil.*
  - ***Verbesserungsidee (Beispiel):*** *Formalisieren und Zahlen ergänzen:*
    „Die Strahlenbelastung ist deutlich kleiner als die jährliche Umweltstrahlung; das Risiko ist sehr gering. Die Untersuchung hilft, eine Überfunktion sicher festzustellen — ich würde sie deshalb empfehlen."

Antwort Nr. 9

- ***Beurteilung Inhaltlich:*** *2/+*
- ***Beurteilung Sprachlich:*** *2/+*
- ***Gesamtniveau:*** *2.0*
- ***Begründung:*** *Sehr gutes Modellbeispiel: nennt das Isotop, die Halbwertszeit (~6 h) und den Vergleich zur jährlichen natürlichen Belastung — mehrere klare, fachliche Kriterien und eine deutliche Relativierung des Risikos. Sprachlich ist ein klares Abwägen erkennbar (Inforation → Relativierung → beruhigende Empfehlung).*

- *Feedback:*
  - ***Stärken:*** *Präzise Angabe (Halbwertszeit) und sinnvoller Vergleich — sehr gutes, modellkonformes Argument. Ton ist freundlich und angemessen.*
  - ***Schwächen:*** *Leichte Tippfehler („besorgt bis“ → „besorgt bist“).*
  - ***Verbesserungsidee (Beispiel):*** *Kleinere Korrekturen und ggf. kurze Ergänzung des Nutzens:*
    „...die Halbwertszeit beträgt etwa 6 h und die Belastung ist geringer als die natürliche Jahresstrahlung. Die Untersuchung liefert genaue Bilder, deshalb würde ich sie empfehlen.“

Antwort Nr. 10

- ***Beurteilung Inhaltlich:*** *2/+*
- ***Beurteilung Sprachlich:*** *1/~*
- ***Gesamtniveau:*** *1.7*
- ***Begründung:*** *Nennung mehrerer relevanter Kriterien: geringe Strahlenbelastung, schnelles Abklingen/Abbau und diagnostische Genauigkeit — das ist inhaltlich vollständig und begründet eine Empfehlung (2/+). Sprachlich werden mehrere Argumente aufgezählt, aber es fehlt ein deutliches Abwägen mit kontrastiven Formulierungen — daher 1/~.*
- ***Feedback:***
  - ***Stärken:*** *Gute und fachlich stichhaltige Argumentation (Risiko vs. Nutzen).*
  - ***Schwächen:*** *Stilistisch noch etwas knapp, keine konkrete Zahl (Halbwertszeit), minimale Rechtschreibfehler („Ärtzen“ → „Ärzten“).*
  - ***Verbesserungsidee (Beispiel):*** *Ergänzen Sie kurz die Halbwertszeit und nutzen Sie ein kontrastierendes Bindewort:*
    „Zwar ist es ionisierend, ***aber*** *Technetium-99m hat eine Halbwertszeit von ca. 6 h und die Belastung ist gering; die Untersuchung liefert sehr genaue Bilder. Ich empfehle sie.“*

Zusammenfassende Tabelle aller Beurteilungen

Antwort Nr.	*Niveaustufen (Inhalt / Sprache)*	*Gesamtniveau*	*Kurze Begründung*
1	*1/~ / 2/+*	*1.3*	*Nennung von Risiko + Vergleich zur Umwelt; fehlende Zahlen und diagnostischer Nutzen.*
2	*1/~ / 1/~*	*1.0*	*Nennung „geringe Dosis“ + Nebenwirkungen; wenig Fakten und kein starkes Abwägen.*
3	*2/+ / 1/~*	*1.7*	*Gute Kombination aus niedriger Dosis/Ausscheidung und diagnostischem Nutzen; kleine Zahlungenauigkeit.*
4	*0/– / 0/–*	*0.0*	*Beschreibung vorhanden, aber zu wenig kriteriale Begründung für Empfehlung.*
5	*2/+ / 2/+*	*2.0*	*Umfassend, mit Abwägung und mehreren relevanten Kriterien (inkl. Ausscheidung).*
6	*2/+ / 2/+*	*2.0*	*Sehr ausgewogen: Risiko benannt und fachlich entkräftet (Ausscheidung, Vergleich).*
7	*0/– / 0/–*	*0.0*	*Nur eine kurze Beruhigung ohne ausreichende fachliche Begründung.*

Antwort Nr.	*Niveaustufen (Inhalt / Sprache)*	*Gesamtniveau*	*Kurze Begründung*
8	*2/+ / 1/~*	*1.7*	*Mehrere relevante Argumente, aber zu umgangssprachlicher Ton; fehlt formales Abwägen.*
9	*2/+ / 2/+*	*2.0*	*Sehr gutes Modellbeispiel: Halbwertszeit genannt, Vergleich zur Umwelt, klare Empfehlung.*
10	*2/+ / 1/~*	*1.7*	*Mehrere relevante Kriterien und Empfehlung; sprachlich noch ausbaufähig (Kontrastierungen).*

*Wenn du möchtest, kann ich jetzt **(A)** die Briefe derjenigen Antworten, die unter 2.0 liegen, in eine überarbeitete, modellkonforme Fassung (40–60 Wörter) umschreiben — pro Brief eine Musterformulierung — oder **(B)** dir ein kurzes Bewertungsraster / Checkliste für die schnelle, wiederholbare Bewertung erstellen (z. B. „Checkliste: nenne mindestens 2 Kriterien, benenne Halbwertszeit/exkretion oder diagnostischen Nutzen, formuliere ein kurzes Abwägen mit ‚zwar... aber...'"), damit du die verbleibenden Schüler:innen selbst effizient und konsistent bewerten kannst. Welche Option hättest du gern?"*