universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Working with RAG-Based Chatbots in Legal Tasks: Exploring Usability, Perceived Trustworthiness and Mental Workload

verfasst von | submitted by

Mag.phil. Felix Aufreiter

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

Master of Arts (MA)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 647
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Digital Humanities
Betreut von | Supervisor:	Univ.-Prof. Torsten Möller PhD
Mitbetreut von | Co-Supervisor:	Laura Koesten MSc Ph.D.

Acknowledgements

First of all, I would like to express my sincere gratitude to Arbeiterkammer Österreich for funding this work. In addition, my thanks go to the employees of Arbeiterkammer for their constructive cooperation and valuable feedback.

I would like to thank my supervisor, *Univ.-Prof. Torsten Möller, PhD*, and my co-supervisor, *Laura Koesten, MSc Ph.D.*, for their ongoing support and guidance. Furthermore, I am grateful to *Antonia Saske, M.A.* and *Timothée Schmude, M.A.* for their help in conducting the interviews and for their ongoing feedback. Additional thanks go to *Assoz. Prof. Dipl.-Ing. Dr.techn. Sebastian Tschiatschek, BSc* for his valuable feedback and helpful ideas.

Finally, I want to thank my family for their support and words of encouragement throughout my studies.

Danke!

Abstract

The evaluation of domain-specific chatbots focusing on legal documents poses a specific challenge in the field of human-computer interaction. This master's thesis investigated how legal domain experts experience a retrieval-augmented generation-based tax law chatbot when performing real-world work tasks. The goal is to understand its perceived trustworthiness, usability, and the perceived mental workload of routine tasks on Arbeiterkammer's tax law experts.

In this thesis, the Arbeiterkammer's tax law chatbot (AK-chatbot) is evaluated, and potential improvements are identified.

An embedded mixed-methods research design is used to collect quantitative and qualitative data, using the Chatbot Evaluation Framework (CEF) developed in this thesis. After a human-chatbot interaction, participants complete a survey and then participate in a semi-structured interview. The survey's goal is to examine the participants perceived mental workload (RTLX), the chatbot's usability (Chatbot Usability Questionnaire, CUQ) and perceived trustworthiness, while the semi-structured interviews focus on the topics of integration into work routine, perceived trustworthiness, areas for future development of the tax law chatbot, participants' information needs (XAI Novice Question Bank) and their prior knowledge regarding chatbot usage.

The results show that the tax law chatbot is perceived as having limited trustworthiness, that usability weaknesses were identified, and that tax experts did not report an exceptionally high mental workload when using it. Opportunities for improvement in further development were identified, and existing information needs of tax law experts were gathered.

On the one hand, this work generates new reference values for RTLX and CUQ; on the other hand, the findings from the data can further develop evaluations for legal chatbots.

Kurzfassung

Die Evaluierung von domänenspezifischen Chatbots mit Schwerpunkt auf Rechtsdokumenten stellt eine besondere Herausforderung im Forschungsbereich der Mensch-Computer-Interaktion dar. Diese Masterarbeit untersuchte, wie juristische Experten einen auf *Retrieval-Augmented Generation* basierenden Steuerrecht-Chatbot bei der Ausführung von realistischen Routineaufgaben erleben. Ziel ist es, die wahrgenommene Vertrauenswürdigkeit, Benutzerfreundlichkeit und die wahrgenommene mentale Arbeitsbelastung bei Routineaufgaben für die Steuerrechtsexperten der Arbeiterkammer zu verstehen.

In dieser Arbeit wird der Steuerrecht-Chatbot der Arbeiterkammer (AK-chatbot) evaluiert und es werden mögliche Verbesserungen identifiziert.

Zur Erhebung quantitativer und qualitativer Daten wird ein eingebettetes Mixed-Methods-Forschungsdesign verwendet und das in dieser Arbeit entwickelte Chatbot Evaluation Framework (CEF) eingesetzt. Nach einer Mensch-Chatbot-Interaktion füllen die Teilnehmer einen Fragebogen aus und nehmen anschließend an einem semistrukturierten Interview teil. Ziel der Umfrage ist es, die von den Teilnehmern wahrgenommene mentale Arbeitsbelastung (RTLX), die Benutzerfreundlichkeit des Chatbots (Chatbot Usability Questionnaire, CUQ) und die wahrgenommene Vertrauenswürdigkeit zu untersuchen, während sich die semistrukturierten Interviews auf die Themen Integration in den Arbeitsalltag, wahrgenommene Vertrauenswürdigkeit, Bereiche für die zukünftige Entwicklung des Steuerrechts-Chatbots, Informationsbedarf der Teilnehmer (XAI Novice Question Bank) und ihre Vorkenntnisse in Bezug auf die Nutzung von Chatbots konzentrieren.

Die Ergebnisse zeigen, dass der Steuerrecht-Chatbot als eingeschränkt vertrauenswürdig wahrgenommen wird, Schwächen in der Benutzerfreundlichkeit festgestellt wurden und Steuerexperten bei der Nutzung keine außergewöhnlich hohe mentale Arbeitsbelastung angaben. Es wurden Verbesserungsmöglichkeiten für die weitere Entwicklung identifiziert und der bestehende Informationsbedarf der Steuerrechtsexperten erfasst.

Einerseits trägt diese Arbeit dazu bei, neue Referenzwerte für RTLX und CUQ zu generieren, andererseits können die Erkenntnisse aus den Daten zur Weiterentwicklung der Evaluierung von juristischen Chatbots beitragen.

Contents

List of Tables

List of Figures

1. Motivation

Following the introduction of ChatGPT to the general public in 2022 (OpenAI, 2022) and the increased use of Large Language Models (LLMs) in the workplace (PwC Österreich, 2025), RAG systems represent another step toward using LLMs for specific applications. A retrieval-augmented generation (RAG) system combines the capabilities of a Large Language Model with a knowledge base that provides access to additional information in order to perform knowledge-intensive tasks, which are "tasks that humans could not reasonably be expected to perform without access to an external knowledge source" (P. Lewis et al., 2020, p. 2).

This thesis critically examines the interaction between human users and an LLM-based retrieval-augmented generation (RAG) system, focusing on its performance in a real-world application. The evaluated chatbot is deployed by Arbeiterkammer Österreich (AK), an organization that represents the interests of employees in Austria (Arbeiterkammer Wien, 2026), and is available to the tax law experts employed by this organization. Arbeiterkammer's tax law chatbot (AK-chatbot), which answers legal inquiries, will be evaluated to identify opportunities to improve its performance in terms of usability, perceived trustworthiness, and the mental workload experienced by AK's tax law experts.

In 2025, various publishers in the field of law offer services that are RAG-based chatbots providing answers based on legal texts (LexisNexis Österreich, 2026; MANZ'sche Verlags- und Universitätsbuchhandlung GmbH, 2025b). Those services are controlled by publishers. Companies or institutions cannot, on a large scale, customize these chatbots for their own purposes. A chatbot, in contrast, based on a company's internal knowledge, offers the possibility to control areas like development and data. Due to the ongoing integration of RAG systems into everyday work, the further development of evaluation strategies is an essential factor for the future of domain-specific RAG systems. At present, there are neither established evaluation strategies in areas like domain-specific RAG systems for legal professionals (Magesh et al., 2025, p. 233) nor publicly available benchmarks for German-language legal texts (Hindi et al., 2025, p. 46176).

A challenge in assessing an LLM-based chatbot's effectiveness is defining appropriate evaluation criteria that capture key dimensions. While traditional metrics such as ROUGE (Lin, 2004) or BLEU (Papineni et al., 2002) can offer insights into specific qualities of LLM outputs, they fail to adequately address aspects of human-computer interaction, such as usability and perceived trustworthiness, as well as the broader impact of AI on domain experts' work routines.

Therefore, this study adopts a mixed-method approach, incorporating an online survey with 2 established rating scales addressing mental workload (Hart, 2006) and usability (Holmes et al., 2019), a user interaction scenario with AK-chatbot and an interview sessions focusing on aspects like perceived trustworthiness, the integration of the chatbot

in tax law experts' work routine and how the chatbot could be improved for further real-world use. As this technology is relatively new, particular emphasis is placed on identifying the individual information needs of tax law experts (Schmude et al., 2025). Qualitative data are analyzed using Robson's five phases of thematic coding (Robson, 2024, p. 575). The present thesis is based on the following research questions:

- **RQ 1**: How do tax-law experts perceive and experience the usability of a Retrieval Augmented Generation (RAG) chatbot when completing routine tax law advisory tasks?
  - **RQ 1.1**: How effectively does the tax law chatbot address tax experts' needs in terms of usability and workflow integration?
  - **RQ 1.2**: What adjustments are needed to make legal advice chatbots improve usability and integration into work routine in practice?
- **RQ 2**: Which characteristics of a domain-specific chatbot shape its perceived trustworthiness among users?
- **RQ 3**: How do tax law experts tend to use the tax law chatbot when solving work tasks?
  - **RQ 3.1**: Can we identify patterns of interaction during the conversation with a legal advice chatbot in real-world use?

Through a systematic interview process across all regional offices of the AK, 14 participants were interviewed, and valuable opportunities for improvement were identified that could lead to considerable improvements in usability and an increase in the perceived trustworthiness of the AK-chatbot.

This master's thesis offers repeatable methods for evaluating a domain-specific chatbot. "This mode of research is repeatable, in the sense that it (...) makes all the efforts it can to provide the data, code, and explanatory information that make it possible for others, at a later point in time, to perform the same (or very similar) research again" (Schöch, 2023, p. 374). Follow-up evaluations are particularly easy to conduct and can draw on the Chatbot Evaluation Framework (CEF) consisting of survey questions, interview questions, and an adapted version of the XAI Novice Question Bank (Schmude et al., 2025) (see section 4.2).

The survey questions were chosen to capture the integration of a legal advice chatbot into real-world tax law advisory workflows. By examining the chatbot's usability and integration with work routine (**RQ 1**), the thesis seeks to understand how technology can support legal experts' day-to-day tasks, collecting insights through interviews that include established questionnaires and specifically adapted interview questions. While usability is an important factor in introducing a tool into an established work routine, trust in the chatbot's output will also be part of the evaluation process (**RQ 2**). Finally, the study examines how tax law experts use the tax law chatbot in routine work tasks, focusing on its areas of application and the types of tasks for which it is employed (**RQ 3**). Finally, the study focuses on the areas of application and types of tasks for which it is

used, as well as on participants' envisioned future areas of application and the chatbot's use in their professional practice (**RQ 3**).

The results show that the AK-chatbot was perceived by study participants as only trustworthy to a limited extent. Furthermore, usability shortcomings were identified, as the chatbot struggled to respond to complex legal queries during the study. Nevertheless, the chatbot proved helpful in answering simpler legal inquiries. Interestingly, willingness to use it in the future was high, even though trust was low. In addition, it was possible to identify opportunities to improve the chatbot for future development and to gather the existing information needs of tax law experts.

2. Background

2.1. Introduction

In recent years, large language models (LLMs) have emerged as powerful tools capable of generating human-like text, assisting in complex problem-solving, and supporting decision-making across various domains. LLMs are defined as "a neural network that uses immense computing resources and training data (along with human-annotated feedback for fine-tuning) to compellingly reply to natural language prompts" (Buttrick, 2024, p. 187). Current major LLMs are based on transformer architectures, first introduced by Vaswani et al. (2017) in "Attention is all you need". An unprecedented popularization of LLMs (Brandt, 2023) occurred with the release of transformer-based, RLHF-trained (Reinforcement Learning from Human Feedback) ChatGPT on November 30, 2022. An easily accessible LLM in dialogue format was made available for the public by OpenAI (OpenAI, 2022).

New, possibly competitive models with different architectures keep popping up in the race to develop better and better models. The formerly groundbreaking architecture LSTM (Long Short-Term Memory) by Hochreiter and Schmidhuber (1997) has been revived as xLSTM (Extended Long Short-Term Memory) by combating its former weaknesses (Beck et al., 2024). Cheaper Chinese models, such as DeepSeek-R1, compete with larger players like OpenAI (Gibney, 2025).

Since ChatGPT's introduction in 2022 (OpenAI, 2022), the use of LLMs has risen enormously. The World Economic Forum cites advances in technologies such as big data and AI as drivers of labour-market skill transformation. The *Future of Jobs Report 2025* covers the period from 2025 to 2030, and surveyed employers identified "AI and information processing technologies" as the key drivers of future business transformations. Since the release of ChatGPT in 2022, worldwide adoption of GenAI has been rapidly growing in the labor market (World Economic Forum, 2025, p. 11). As more and more models are released, the way they are used is constantly evolving. According to Time magazine (Pillay & Booth, 2025) or Microsoft (Nyhan, 2024), 2025 was titled the "year of AI-agents".

In Austria, this means that 28% of people are already using LLMs at work in 2025. In 2023, the figure was still 18%. 69% of survey participants expect efficiency gains and relief from time-consuming or tedious tasks by using LLMs (PwC Österreich, 2025). According to Statistik Austria, 73% of the entire Austrian population considered their knowledge of AI to be low or non-existent in 2024, while 35% of participants expressed a positive attitude towards the use of artificial intelligence (Statistik Austria, 2025).

Many professional fields are actively integrating LLMs into their employees' day-to-day

work. Changes have even begun to take place in normally rather cumbersome institutions such as the education system. The Austrian Ministry of Education (Bundesministerium für Bildung) has already taken steps to both reflect on the use of artificial intelligence and to incorporate its meaningful use into the education system (Bundesministerium für Bildung, Wissenschaft und Forschung, n.d.). As an employee of the Austrian secondary education system, the author can personally report on its very active use by teachers and on a wide range of professional training measures. LLMs have already been used to grade short-answer questions (Henkel et al., 2024), generate personalized feedback on open-ended responses (Matelsky et al., 2023), and serve as automated assignment evaluators supporting human teaching personnel (C.-H. Chiang et al., 2024). All those activities could be considered time-consuming tasks for educators or teachers in a real-world scenario.

In the legal field, immediately after its public introduction, ChatGPT was "sent to Law School" by scholars to test its ability to answer questions about legal issues and passed four courses at the University of Minnesota Law School with a C+ (Choi et al., 2022). Furthermore, GPT-4 showed impressive results solving legal questions on the Multistate Bar Exam (MBE) (Martínez, 2024). Numerous scientific papers and commercial tools using LLMs in education have followed since the release of ChatGPT (S. Wang et al., 2024).

2.2. Retrieval-Augmented Generation (RAG)

This thesis evaluates a chatbot based on an RAG system. An RAG system combines a Large Language Model with an external knowledge base to improve the LLM's output with specific and up-to-date knowledge, which doesn't need to rely entirely on the LLM's internal knowledge (Fan et al., 2024, p. 6498). "RAG combines the generation flexibility of the "closed-book" (parametric only) approaches and the performance of "open-book" retrieval-based approaches" (P. Lewis et al., 2020, p. 5). Retrieval-Augmented Generation systems enhance LLMs' ability to provide relevant and more factual responses when performing open-domain Question Answering tasks (P. Lewis et al., 2020, p. 9). Providing an LLM with additional knowledge can make them particularly valuable in expert-knowledge-driven fields such as law.

2.3. Chatbot description

The AK's approach reflects the current state of mind regarding its purpose, which hopes to increase efficiency through the use of AI (PwC Österreich, 2025). The task of the chatbot is, according to the person responsible, "*to fulfill 90% of the requests with 90% accuracy*". It should reduce the workload of legal advisors, as this group of professionals at the AK is already under heavy pressure. The AK-chatbot is based on a Retrieval Augmented Generation (RAG) approach.

The development of an RAG system focusing on legal topics is not a new invention by the AK, but is also being driven forward by publishers such as MANZ. The famous Austrian publisher of legal books, Manz, released an RAG system trying to be a main

player in this domain in Austria (MANZ'sche Verlags- und Universitätsbuchhandlung GmbH, 2024). Manz's chatbot is targeted at lawyers and tax consultants (MANZ'sche Verlags- und Universitätsbuchhandlung GmbH, 2024). The RAG system provides a variety of information, including data from the RDB (Rechtsdatenbank) or the RIS (Rechtsinformationssystem des Bundes). In 2025, more content such as EU directives, regulations, and parliamentary materials will be added. Future updates should include web searches (MANZ'sche Verlags- und Universitätsbuchhandlung GmbH, 2025a). It would certainly make sense, from a financial point of view, to compare *Genjus KI* by Manz with the maintenance and development costs of the AK's own chatbot.

The AK-chatbot is deployed using Microsoft's Azure AI Search (Microsoft Corporation, 2025).

> Azure AI Search (formerly known as "Azure Cognitive Search") is an enterprise-ready information retrieval system for (...) heterogeneous content that you ingest into a search index, and surface to users through queries and apps. (Microsoft Corporation, 2025)

The tax law chatbot is powered by OpenAI's GPT-4o, which was originally released on the 13th of May 2024 (OpenAI, 2024).

> GPT-4o ("o" for "omni") is a step towards much more natural human-computer interaction—it accepts as input any combination of text, audio, image, and video and generates any combination of text, audio, and image outputs. (...) It matches GPT-4 Turbo performance for English text and code, with significant improvement for non-English text, while also being much faster and 50% cheaper via the API. GPT-4o is especially better at vision and audio understanding compared to existing models. (OpenAI, 2024)

While the chatbot is still in a phase of constant evolution, on the 2nd of April 2025, the tax law chatbot was officially made available to all employees of AK's tax law departments all over Austria. Thus, the set of users was expanded beyond selected key users.

The chatbot was continuously developed during the study's preparation and is still being updated. A link to the documents used to generate the answers was provided. After the feature's introduction, users complained, according to AK's AI Lead, about slow document download speeds due to their size. (Some documents contain hundreds of pages.) Further development steps are still planned.

Some features that should be available in the future include a knowledge base (wiki) that can be updated independently by key users, enabling the chatbot to provide more precise answers. It should also be possible to upload existing work documents or collections of legal advice developed by AK's tax law experts to the chatbot's knowledge base. Furthermore, a "Mail"-button was added to let the chatbot create a mail ready to send to clients using the LLMs answer.

2.3.1. User interface

The chatbot platform presents a minimalist, dialogue-centric interface in which the conversational transcript occupies the primary visual field (Langevin et al., 2021, p. 11). The chatbot's user interface is branded in AK's signature color scheme, red and white. Conversations between the user and the LLM are formatted as alternating rectangular text fields, with user prompts (right-aligned) and system responses (left-aligned). A vertical sidebar containing fields to choose between a general chatbot and the analyzed tax law chatbot is available. Furthermore, the user interface offers access to past conversations in a secondary vertical sidebar (see Figure 2.1).

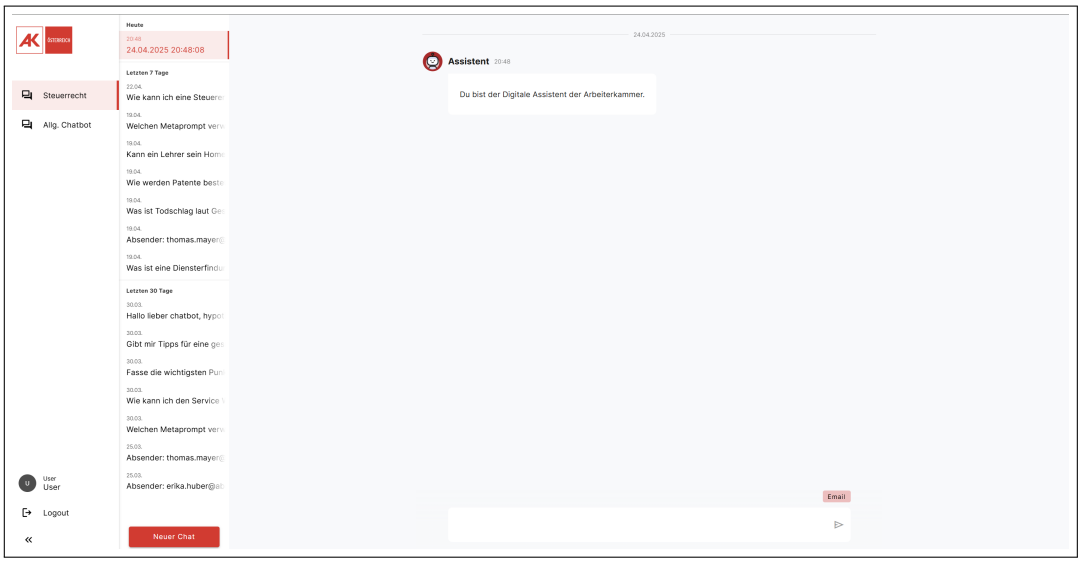


Figure 2.1.: User Interface - Tax Law Chatbot

2.3.2. Data

According to the AK's AI Lead, official AK flyers (Bundesarbeitskammer, 2025; Kammer für Arbeiter und Angestellte für Wien, 2025), legal codices like the Austrian income tax law (*Einkommenssteuergesetz*) (Bundeskanzleramt der Republik Österreich, 2025b), and double taxation agreements (Bundesministerium für Finanzen, 2025b) were used to create the knowledge base of the chatbot. These documents serve as a non-parametric memory of the RAG system (P. Lewis et al., 2020, p. 1).

AK's AI lead explained that data curation was carried out by the tax law experts who were part of the initial key-user group. Detailed information about the data collection process or the curation plan used for the AK's chatbot was not available for further analysis.

2.3.3. System message or "Meta Prompt"?

When using the LLM GPT-4o, two different messages are normally used: the system message and the user message (Forte, 2024).

> The system message is included at the beginning of the prompt and is used to prime the model with context, instructions, or other information relevant to your use case. You can use the system message to describe the assistant's personality, define what the model should and shouldn't answer, and define the format of model responses. (Microsoft, 2025)

The user message contains the message to which the model generates an answer (Forte, 2024). The system message provides higher-level instructions.

> A system message is a feature-specific set of instructions or contextual frameworks given to a generative AI model (e.g., GPT-4o, GPT-3.5 Turbo, etc.) to direct and improve the quality and safety of a model's output. This is particularly helpful in situations that need certain degrees of formality, technical language, or industry-specific terms. (Microsoft, 2025)

Before we move on to the RAG system's system message, a somewhat misleading term needs to be briefly defined to avoid unintentional confusion. There seem to be various definitions of the term *Meta Prompt* or *meta-prompting*. The following definition is used in this thesis:

> A *Meta Prompt* is an example-agnostic structured prompt designed to capture the reasoning structure of a specific category of tasks. It provides a scaffold outlining the general approach to a problem, enabling LLMs to fill in task-specific details as needed. This methodology focuses on the procedural aspects of problem-solving (the *How*) rather than the content-specific details (the *What*). (Zhang et al., 2025, p. 5)

OpenAI (2025) and other researchers (e.g., (de Wynter et al., 2024; Rodrigues & Branco, 2024)) consider *Meta-prompting* to be "a technique where you use an LLM to generate or improve prompts" (Musatoiu, 2024). The *Meta Prompt* as a design strategy for system messages, as we understand it, can also provide the LLM with complex structures or thought processes.

> It involves constructing a high-level "meta" prompt that instructs an LM to: (i) break down complex tasks or problems into smaller, manageable pieces; (ii) assign these pieces to specialized "expert" models with proper and detailed natural-language instructions; (iii) oversee the communication between these expert models; and (iv) apply its own critical thinking, reasoning, and verification skills throughout the process. (Suzgun & Kalai, 2024, p. 2)

Strictly speaking, we could argue that the system message used during the study can not be called a *Meta Prompt*, as it does not deal with structural or syntactic properties of

a problem to be solved (see table 2.1). The system message used during the study focuses on providing context and rules, as well as establishing a persona. It includes current values of Austria's Income Tax Act (*Einkommenssteuergesetz*) (Bundeskanzleramt der Republik Österreich, 2025b). The reason for specifying these values is that documents from various years are part of the knowledge base. In some cases, documents containing the *Schilling*, which was officially replaced by the *Euro* in 2002, have also been added. (The original version of the system message in German is shown in table A.5 in the appendix.)

Prompt description	**System message**
Persona	You are an advisor at the Arbeiterkammer Wien. The Arbeiterkammer is a statutory interest group. Employees, pensioners, and freelancers receive advice on wage tax and income tax.
Rule	You answer based on the Income Tax Law.
Context	It is currently the year 2025. Current limit values to be considered: Tax limit for 2025: €13,308 Tax assessment limit for 2025: €14,517 Lower earnings limit for 2025: €551.10 The lower earnings limit applies exclusively to social security. There is no lower earnings limit in taxation. Limit for additional income on a self-employed basis, up to which no declaration is required in the tax return 2025: €730
Rule	Rules: you do not provide answers based on regulations from previous years you respond in understandable sentences you answer in such a way that even someone without training in tax law can understand the answer you give all answers in the formal form Ignore all amounts in Schillings. Do not answer any questions about gifting or gift declarations; instead, refer the matter to a tax advisor in this exceptional case

Table 2.1.: System message analysis (English translation)

After completing the interviews, a new system message was introduced to improve the chatbot's performance. The new system message is not part of the thesis, as it was agreed with the AK's project manager, that it would be better not to include it in this publication, because the chatbot was in a further development phase and changes were being made on an ongoing basis. Unlike its previous version (see table 2.1), this updated system message clearly exhibits the characteristics of a functional *Meta Prompt*, as it includes specific tool-use rules and decision logic for user requests.

3. Related work

This thesis is part of the field of human-computer interaction, which is defined as:

> Human–computer interaction is a discipline concerned with the design, evaluation, and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them. (Hewett et al., 1992, p. 5)

In this study, the interaction occurs between a human and a chatbot accessed via a PC. The content of this thesis spans various interconnected main subject areas, but this section will focus on three specific topics.

First, the evaluation of RAG systems in the legal field is addressed, as this is the main objective of this thesis. Gaps will be discussed where existing benchmarks and evaluation techniques are currently insufficient for specialized legal chatbots.

Second, the aspect of trustworthiness is examined in relation to artificial intelligence and RAG-based chatbots.

Finally, this section will focus on how the tax law chatbot and its operator, AK, are affected by the European Union's regulatory efforts. On the one hand, the regulatory framework is relevant for the deployment of the AK-chatbot, as AK, as an organization, must comply with legal requirements. On the other hand, AK also has a political position on the use of AI.

3.1. Evaluation of RAG systems

As the use of LLMs continues to evolve, so does the need for evaluation. Due to competitors' rapid advances since its initial release, various LLMs have challenged or surpassed OpenAI's LLMs in LMArena's rankings, formerly known as Chatbot Arena (W.-L. Chiang et al., 2024). At the writing of this thesis (Feb 21st, 2026), the following models were ranked in Arena's top 10 text models, formerly known as ChatbotArena or LMArena (Arena, 2025):

Ranking	LLM	Creator
1.	claude-opus-4-6	Anthropic
2.	claude-opus-4-6-thinking	Anthropic
3.	gemini-3.1.-pro-preview	Google
4.	gemini-3-pro	Google
5.	dola-seed-2.0-preview	ByteDance Seed
6.	gemini-3-flash	Google
7.	grok-4.1.-thinking	xAI
8.	claude-sonnet-4-5-20250929-thinking-32k	Anthropic
9.	claude-opus-4-5-20251101	Anthropic
10.	grok-4.1	xAI

Table 3.1.: Arena: Leaderboard (21.02.2026)

Although leaderboards might seem useful for comparing the performance of general-purpose LLMs, evaluating domain-specific RAG systems requires a more use-case-specific approach. Arena focuses on the direct comparison of LLM responses and ranks them based on user preference (W.-L. Chiang et al., 2024, p. 3). Current studies on the automated evaluation of RAG systems tend to analyze retrievers, generators, or embeddings using generated, enhanced, or pre-existing datasets, but evaluating a domain-specific RAG system using public question and answer (QA) datasets is often considered impractical (Brehme et al., 2025, pp. 17–18). Various datasets focusing on the legal domain are available, but none of these benchmarks address Austrian (tax) law or specific legal texts, which are usually necessary for AK tax law experts in their daily work. Available benchmarks to evaluate RAG systems deal with Chinese law (Li, Chen et al., 2025; Li, Chen et al., 2025; Li et al., 2024) or use texts of the US-American common law system, which differs from Austria's civil law system (Magesh et al., 2025). To date, no German-language datasets have been created. Only EU law in the English language could be useful for evaluating the AK-chatbot (Hindi et al., 2025, p. 46176). Although generator evaluation approaches like short-answer evaluation, classical and embedding approaches, LLM-as-a-Judge techniques, or using human evaluators are possible methods (Brehme et al., 2025, pp. 19–20), available metrics have difficulties confirming factual correctness of RAG system outputs (Hindi et al., 2025, p. 46185). Human expert knowledge still plays a crucial role in evaluating domain-specific RAG systems, especially when nuanced judgment is required (Brehme et al., 2025, p. 22).

Furthermore, RAG systems focused on legal documents pose specific challenges for evaluation metrics that are currently unavailable. Magesh et al. (2025) argue that "evaluations of general purpose RAG systems often assume that all retrievable documents (1) contain true information and (2) are authoritative and applicable (...). Legal documents often contain outdated information, and their relevance varies by jurisdiction, time period, statute, and procedural posture. Determining whether a document is binding or persuasive often requires non-trivial reasoning about its content, metadata, and relationship with the user query" (Magesh et al., 2025, p. 230). Therefore, human evaluators seem to be the

necessary and more reliable way to conduct an evaluation of generated output (Li, Chen et al., 2025; Magesh et al., 2025; Mamalis et al., 2024), although other studies underline the potential of an LLM-as-a-Judge approach (Y. Wang et al., 2024, p. 7).

A major problem seems to be the lack of transparency in evaluation and benchmarking procedures of AI tools for legal research (Magesh et al., 2025, p. 233). As Magesh et al. (2025) highlight in their article:

> "But until vendors provide hard evidence of reliability, claims of hallucination-free legal AI systems will remain, at best, ungrounded" (Magesh et al., 2025, p. 234).

Given that no RAG-specific benchmarks can be used in this study, and since the chatbot is provided by an external provider to the AK, focusing on users' experiences and impressions appears to yield the most meaningful results. Therefore, it seems reasonable to focus on aspects such as user feedback by domain experts, perceived trustworthiness (Bisconti et al., 2025, p. 131), perceived mental workload (Hart, 2006, p. 904), and the usability of the chatbot (Holmes et al., 2019, p. 210) (see chapter 4).

3.2. Evaluation frameworks for chatbots

To understand how users experience the chatbot in terms of usability, various tools are available (Borsci et al., 2022; Brooke, 1996; Holmes et al., 2019; Laugwitz et al., 2008). When focusing on usability, the widely used System Usability Scale (SUS) by Brooke (1996) stands out in particular, as it is accompanied by a benchmark by J. R. Lewis and Sauro (2018) based on more than 20 years of data.

During the first phase of development of the study and its survey, the combination of System Usability Scale (SUS) by Brooke (1996) and NASA Task Load Index (NASA-TLX) by Hart (2006) was discussed, but the more specific Chatbot Usability Questionnaire (CUQ) by Holmes et al. (2019) was chosen due to the fact that a tax law chatbot had to be analyzed. The Chatbot Usability Questionnaire was chosen because, unlike the System Usability Scale, it was developed specifically for chatbots (Holmes et al., 2019, p. 210). Holmes et al. (2019) argue that "traditional usability testing surveys do not evaluate all aspects of a chatbot interface" (Holmes et al., 2019, p. 214). "CUQ evaluates the user experience, along with the conversational interaction between chatbot and user" (Essop et al., 2023, p. 2).

Although CUQ was mainly developed for chatbot analysis, comparability with the SUS was in the creators' minds. Keeping this underlying design idea in mind, it may also be possible to compare its score with SUS values (Holmes et al., 2019, p. 210). The CUQ was validated against User Experience Questionnaire (UEQ) by Laugwitz et al. (2008) and the SUS, analyzing three chatbots (Holmes et al., 2023, pp. 213–214). In addition, its construct validity and intra-rater reliability were validated (Holmes et al., 2023, p. 337). In contrast to the System Usability Scale, only a rather limited number of studies using CUQ are currently available (see table A.12).

It should also be noted that, as things stand, no benchmark has been set for the CUQ. If SUS or UEQ had been used, benchmarks would have been available (Holmes et al., 2023, p. 338). However, using traditional questionnaires would have ignored the CUQ's advantage, which lies in its specialized design for chatbots. As Holmes et al. (2023) mention in their validation of the questionnaire: "There are currently numerous tools and scales for measuring system usability. The vast majority of these tools are designed for assessing the usability of traditional 'mouse-and-pointer' systems and thus may not be appropriate for assessing the usability of chatbots" (Holmes et al., 2023, p. 321). Therefore, this master's thesis has the potential to further contribute to the development of the questionnaire in the field of chatbot evaluation and to provide reference values for future studies.

3.3. Trustworthiness

Although AI is expected to have a major impact on the legal sector, there is mixed feedback on generative AI's performance in solving legal tasks. Doyle and Tucker (2025) report that the performance of LLMs such as GPT-3.5, GPT-4, GPT-4o, Claude Haiku, Claude Sonnet, Claude Opus, or Sonnet 3.5 varies greatly, although legal practice guides were provided. Further development is necessary to achieve human expert-level performance in legal task-solving. In a real-life context, these missing skills could impact users' trust in a deployed RAG system.

The concept of Trustworthiness is seen as an important part of AI use by researchers, as it is intended to make users more likely to use a system (Baron, 2025).

> Trustworthiness in AI refers to the assurance that AI systems will operate in a safe, reliable, and fair manner. The importance of trustworthiness stems from the need to mitigate risks associated with AI deployment, ensuring that these systems do not cause harm, and that their actions are understandable and predictable to humans. (Bisconti et al., 2025, p. 131)

Legal frameworks are put in place to regulate artificial intelligence and improve trust in AI systems. The European Union's AI Act addresses possible risks and harms. It "represents a significant milestone in the European Union's efforts to regulate artificial intelligence (AI), with its primary goal being to establish a legal framework for trustworthy, human-centred AI" (Wendehorst & Nessler, 2024, p. 7). Although necessary steps have been taken, Nessler et al. (2023) and Schweighofer et al. (2025) argued that the EU's AI Act still neglects statistically valid testing of AI systems. Using established machine learning procedures and a precise definition of the field of application of AI systems could improve their functional trustworthiness.

In the context of our study, we focus less on the concept of AI systems' functional trustworthiness and more on users' expectations of AI systems.

ISO definition of Trustworthiness (ISO/IEC TR 24028:2020)	**Definition by Bisconti et al. (2025) of Trustworthiness**
Trustworthiness is the ability to meet stakeholder expectations in a verifiable manner. (Bisconti et al., 2025, p. 132)	Trustworthiness is an attribute, outcome of meeting stakeholders' expectations. (Bisconti et al., 2025, p. 133)

Table 3.2.: Definitions - Trustworthiness

Trustworthiness is not understood as an intrinsic property of a system, but rather as a relationship between two parties. It can be understood as a dynamic relationship. If a system were completely predictable, trust would not be necessary (Bisconti et al., 2025, p. 132).

Although the evaluated chatbot was developed for a very specific and limited target audience, i.e., tax law experts, its creators must consider providing a trustworthy AI system. The tax law chatbot must meet the needs of its main users for a high level of Perceived Trustworthiness and Usability, while also being seen in its broader context. If its creators want to follow ethical guidelines, they need to consider fairness and bias in AI. Ferrara (2024) emphasizes the need for fair AI and proposes various strategies to mitigate possible risks of bias. Data bias, algorithmic bias, or user bias are just some examples that could influence the output of generative AI. "Fairness is inherently a deliberate and intentional goal, while bias can be unintentional" (Ferrara, 2024, p. 8).

While the topic of fairness is discussed in a broader context affecting individual users and society as a whole, Bisconti et al. (2025) focus on the individual level and illustrate how reducing Perceived Uncertainty by aligning with stakeholders' mental models can serve as an effective strategy to improve user engagement and build trust.

Trustworthiness could be seen as part of a larger concept. Vassilakopoulou et al. (2022) consider Trustworthiness to be a part of Responsible AI. The concept of Responsible AI is influenced by 3 different fields of research: *Computer Science*, *Human-Computer Interaction*, and *Philosophy and Ethics*. Vassilakopoulou et al. (2022) therefore propose a comprehensive definition, as the term "*Responsible AI*" is used in policy, big technology companies, and academia. The concept varies, however, depending on the academic domain or industry.

> Responsible AI is the practice of developing, using, and governing AI in a human-centred way to ensure that AI is worthy of being trusted and adheres to fundamental human values. (Vassilakopoulou et al., 2022, p. 7)

Along similar lines, there are already scientific contributions that investigate users' perceptions of LLM-generated text (Cabrero-Daniel & Sanagustín Cabrero, 2023). Other works propose a Cross Industry Standard Process for Data Mining (CRISP-DM) based framework to evaluate trustworthiness of AI systems (Kemmerzell et al., 2025). Further studies focus on the evaluation of human-model interactions (Ibrahim et al., 2024; Lee et al., 2024) or investigate the bias produced by the incorporation of LLM-generated content on RAG systems (Chen et al., 2024) and information retrieval systems (Dai et al., 2024). AK therefore faces the task of increasing the Perceived Trustworthiness and

Usability of its chatbot for its employees while keeping the concept of Responsible AI in mind.

3.4. Regulation

As AI integration continues to advance, a legal framework is needed to regulate this technology. The AI Act represents an attempt to introduce regulations for the European Union. Prainsack and Forgó (2024) mention that there is still plenty of room for improvement.

> As with the EU's General Data Protection Regulation, the AI Act will apply across domains and will not be limited to specific fields (...). In addition, the AI Act tailors regulation to the presumed level of risk, ranging from unacceptable to minimal. (...) There is much to be liked about the risk-based approach that the AI Act is taking. It has some clear advantages over one-size-fits-all alternatives that over-regulate on the low-risk end of the spectrum while missing important problems on the other end. However, the risk-based approach also raises serious practical and political issues. It remains unclear what is considered a high-risk AI system. (Prainsack & Forgó, 2024, p. 1235)

Prainsack and Forgó (2024) argues that the AI Act's lack of precision could compromise its initial intentions by not keeping up with current developments in technology. Issues may arise in the domain of dual-use technologies, initially developed for military use but now used in other fields. Furthermore, companies with greater economic power could gain an advantage over small companies by challenging risk assessments in court. Another result could be that the focus on risks could prevent the development of technology where benefits outweigh potential risks (Prainsack & Forgó, 2024, pp. 1235–1236).

Institutions and corporations will also have to demonstrate AI competence (Aichinger et al., 2024). Even before the massive rise of ChatGPT, TÜV Austria and the Institute for Machine Learning at the Johannes Kepler University Linz proposed the official certification of machine learning applications (Winter et al., 2021).

The AK, as an organization representing the interests of employees and consumers in Austria, emphasizes the importance of developing AI skills and reaffirms the need for responsible use by employees and business owners (Fülop, 2025). Furthermore, the AK demands "to ensure transparency, non-discrimination, and the right to appeal when using risky AI applications through appropriate regulations" (Fuchs, 2025).

The legal sector is not just trying to figure out how to regulate AI (Prainsack & Forgó, 2024) or how to define legal AI systems (European Law Institute, 2024; Wendehorst et al., 2024). It also critiques the vagueness and inconsistency of the EU AI Act's definition of AI systems (Aufreiter et al., 2024, p. 23), and it is also working on integrating artificial intelligence into everyday working life (LexisNexis Österreich, 2026; MANZ'sche Verlags- und Universitätsbuchhandlung GmbH, 2024). AI is expected to have a significant impact on the legal profession. Employers listed legal officials and legal secretaries among the top 10 fastest declining professions between 2025 and 2030 (World Economic Forum, 2025, p. 19).

As mentioned, regulation, evaluation, and trustworthiness are closely connected. This work focuses on (perceived) trustworthiness as defined by Bisconti et al. (2025), given the aforementioned gaps in evaluating legal RAG systems and determining functional trustworthiness under the EU's AI Act. It could be argued that, if the focus is on trustworthiness, evaluating a tax law chatbot can be most effectively done by adopting a user perspective, as in this master's thesis.

Since the AK, as a representative of workers' interests, itself expresses a desire for transparency (Fuchs, 2025) and sees responsible use as a goal (Fülop, 2025), the evaluation of its own tax law chatbot, which will also have an impact on the members of AK when used, can be seen as a step in a direction it has itself promoted.

4. Methods

This chapter explains the approach used to evaluate the AK-chatbot. Section 4.1 presents the overall study design and addresses the study procedure, the use of AI in the thesis, the transcription process, and the evaluation of qualitative data using thematic analysis. After that, section 4.2 introduces the Chatbot Evaluation Framework (CEF).

4.1. Study design

4.1.1. Study procedure

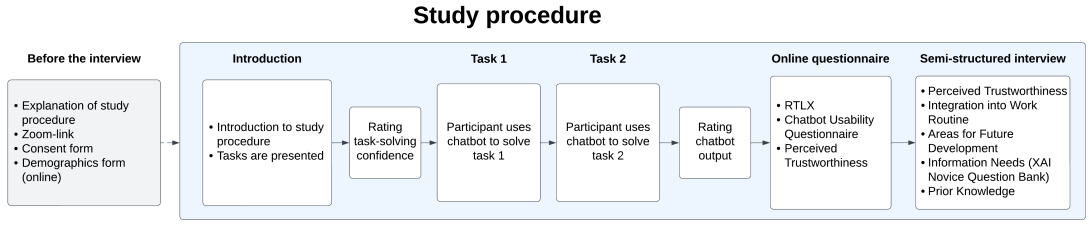


Figure 4.1.: Study procedure

To evaluate the AK-chatbot, an embedded mixed-methods research design was chosen. I collected quantitative survey data after a human-chatbot interaction. To gain contextual insights, I also interviewed the user. Both aspects are integrated to provide contextual insights (Creswell & Clar, 2017, p. 103). Both methods are applied sequentially within a single 60-minute session per participant. The entire interview session is conducted and recorded via video conference. After the interaction between participants and the tax law chatbot, participants complete a three-page online questionnaire. The survey addresses perceived mental workload (Raw TLX), chatbot usability (Chatbot Usability Questionnaire), and perceived trustworthiness (see section 4.2.2).

After interacting with the chatbot and completing the online questionnaire, a semi-structured interview is conducted (Robson, 2024, p. 372). The interview focuses on perceived trustworthiness, integration into the work routine, areas for future development of the tax law chatbot, participants' information needs, and their prior knowledge regarding chatbot use (see section 4.2.3). A number of interviews were conducted under supervision to ensure consistent quality, improve the interview style where necessary, and obtain feedback from experienced researchers (see table A.1).

4.1.2. Transcription

To speed up transcription, OpenAI's whisper-large-v3-turbo speech recognition model was used (Hugging Face, 2025), which is based on the work of Radford et al. (2022). After the initial transcription using whisper-large-v3-turbo on a local CPU, the text was checked again against the recording and corrected where necessary.

4.1.3. Use of AI

Since English is not the author's native language, tools such as DeepL (DeepL, 2025), Perplexity (Perplexity, 2025), and ChatGPT (OpenAI, n.d.) were used for grammatical improvements, creation of citations, and coding. Idea development and interview evaluations were carried out by the author in compliance with GDPR. No sensitive data was uploaded to any AI-service provider's platform.

4.1.4. Thematic coding analysis

The qualitative data from 14 interviews were analyzed using the five phases of thematic coding outlined by Robson (2024).

1. Familiarizing yourself with your data.
2. Generating initial codes.
3. Identifying themes.
4. Constructing thematic networks.
5. Integration and interpretation.

Table 4.1.: The phases of thematic coding analysis. Adapted from Robson (2024, p. 575).

Robson (2024)'s phases of thematic coding analysis were chosen because the analysis of thematic networks plays a decisive role in this master's thesis. At the same time, the chosen approach remained closely aligned with Braun and Clarke (2022)'s reflexive thematic analysis, which provided the methodological and epistemological orientation that guided the process. The development of codes and themes was based on Braun and Clarke (2022)'s approach, where "coding is an *organic* and *evolving* process, an open process" (Braun & Clarke, 2022, p. 55) and "theme development is an active process; themes are constructed by the researcher, based around the data, the research questions, and the researcher's knowledge and insights" (Braun & Clarke, 2022, p. 35). ATLAS.ti 25 was selected for the qualitative analysis to make the coding process transparent and adaptable (ATLAS.ti Scientific Software Development GmbH, 2025).

Generating initial codes began inductively, allowing themes to emerge from the data rather than using predetermined codes or themes. After having generated initial codes, six main themes were identified. The six main themes are “areas of application”, “evaluating chatbot output”, “expressing feelings towards the chatbot”, “possible improvements”, “prior knowledge”, and “trust”; thematic networks were created based on each of these themes. The following interpretation of the qualitative data was guided by these thematic networks. The theme “information needs” was analyzed deductively, because participants were asked during the semi-structured interviews to choose at least 3 questions of the XAI Novice Question Bank (Schmude et al., 2025).

4.2. Chatbot Evaluation Framework

This section outlines the Chatbot Evaluation Framework (CEF) developed in this thesis. To provide an initial orientation, table 4.2 summarizes the topics and structure of the framework.

Category	**Items**	**Data**	**Topics**
Pre-Task Rating	1	Quantitative data	Rating Self-Reported Prior Knowledge and Task-Solving Confidence
Tasks	2	Qualitative data	
Post-Task Rating	1	Quantitative data	Rating Chatbot Output
Online Survey	28	Quantitative data	Perceived Mental Workload, Usability, Perceived Trustworthiness
Interview Questions	18	Qualitative data	Areas for Future Development, Information Needs, Integration into Work Routine, Perceived Trustworthiness, Prior Knowledge and Experience

Table 4.2.: Overview of the Chatbot Evaluation Framework

Section 4.2.1 gives an overview of how the tasks for the CEF have been developed and presents the tasks used for the study. After that, section 4.2.2 introduces the rating scales of the online survey, and section 4.2.3 addresses the interview questions.

4.2.1. Tasks

As seen in figure 4.1, at the beginning of each interview session, participants received instructions to guide their interaction with the tax law chatbot. They were asked to imagine a situation in which a client had emailed them requesting tax-related advice. Their task was to use the chatbot either to draft an appropriate response or to conduct research. Participants were encouraged to approach the task in the same manner as they would in their everyday professional work.

Situation:

- Imagine that a customer has sent you the following email.
- Use the tax law chatbot to help you compose an appropriate reply.
- Work as you would in a real work situation.

- This is where the text of the respective sample email is located. -

Task:

- Create a complete reply email or text that you would use in your everyday work to respond to the inquiry.

Once you are sufficiently satisfied with the result of your interaction with the chatbot, we can begin the interview.

Table 4.3.: Task instructions

Development of tasks

The development of tasks was conducted iteratively, using feedback from members of the University of Vienna's Research Group Visualization and Data Analysis (see the meeting schedule in appendix A.1). In addition, the tasks were reviewed by the AK's AI Lead and the AK's tax law expert to make them as realistic as possible. The tasks are based on customer emails from the everyday work of AK-tax experts, so that the chatbot can be tested in a situation that reflects their actual day-to-day work.

Originally, four tasks were created for the study. To ensure interaction within the approximately 60-minute time frame for each interview session, the two most suitable task designs were further refined (see appendix A.5).

Evolution of tasks

Task 1 The aim of task 1 was to create a simple task with only one specific question to avoid making the interaction too complex. The goal of this task is to encourage interaction with the chatbot, not to challenge it technically.[1]

Sender: **erika.musterfrau@mailadresse.at** Subject: Home office for employee tax assessment Dear Sir or Madam, I hope you can help me with a few questions about our taxes. My husband and I are both working, and we have some expenses that we would like to claim for tax purposes. I currently work from my home office quite a lot, and last year I spent around €2,000 on new equipment (laptop, mouse, headset, etc.). Can I claim these items in full as income-related expenses? **My husband (who works as a high school teacher) has also spent this much on his devices and would like to claim it on his taxes. Is that possible?** Thank you very much for your help! Erika **Musterfrau**

Table 4.4.: Task 1 (version 2)

Task 1 (version 3) After an initial meeting with AK's AI Lead, it was necessary to make further changes to *Task 1 (version 2)*. The bold text displayed in table 4.4 was adjusted or omitted at the advice of our collaborator, the AI Lead of the AK, because the chatbot would detect it as a sample request rather than a real email. In this case, this would be against the purpose of the sample tasks.[2]

[1]The original German version is provided in the appendix A.5.1.
[2]The original German version is provided in the appendix A.5.1.

Sender: **erika.huber@abc-mail.at**
Subject: Home office for employee tax assessment

Dear Sir or Madam,

I hope you can help me with a few questions about our taxes. My husband and I are both working, and we have some expenses that we would like to claim for tax purposes. I currently work from my home office quite a lot, and last year I spent around €2,000 on new equipment (laptop, mouse, headset, etc.) **and furniture for my home office**. Can I claim these items in full as income-related expenses?

Thank you very much for your help!
Erika **Huber**

Table 4.5.: Task 1 (version 3)

Task 2 The second task was designed to present participants with a more complex situation in order to stimulate engagement with the tax law chatbot. The email inquiry combines multiple tax-related issues, including questions about cross-border income taxation in the context of Austria and Germany ('Abkommen zwischen der Republik Österreich und der Bundesrepublik Deutschland zur Vermeidung der Doppelbesteuerung auf dem Gebiet der Steuern vom Einkommen und vom Vermögen und zur Verhinderung der Steuerverkürzung und -umgehung (Bundesrecht konsolidiert, Fassung vom 18. August 2025)', 2025), as well as how to proceed with child support payments for children living abroad in Germany and Serbia (Arbeiterkammer Oberösterreich, 2025). Participants are required to consider a broader range of legal aspects. Feedback later confirmed that task 2 (see table 4.8) was indeed perceived as more challenging by the AK's tax law advisors. (The perceived quality of chatbot responses is discussed in chapter 5.)

Task 2 (Version 2) The bolded text *Task 2 (version 2)* was changed after our initial meeting at the advice of our collaborator: the chatbot would recognize that the email is a sample request. The invented name was changed to "Thomas Mayer" to make it a little more realistic (see table 4.6).[3]

[3]The original German version is provided in the appendix A.5.2.

Sender: **max.mustermann@mailadresse.at**
Subject: Income in Germany and child support

Dear Sir or Madam,

I am writing to you because I am in a somewhat complex situation. I live in Vienna and work for a company as normal. However, I also work as a freelance consultant for a company in Germany.

I am not sure how this affects my taxes. Where do I have to pay tax on my salary? In Austria, right? But do I also have to pay taxes in Germany?

Another issue concerns my two children. One lives with my first ex-wife in Berlin and my second child lives in Japan with my second ex-wife, who emigrated. Both have demanded child support from me so far, but I have heard that this is not entirely correct. Do I have to pay at all if the children do not live in Austria?

Kind regards
Max Mustermann

Table 4.6.: Task 2 (version 2)

Sender: **thomas.mayer@mailmail.at**
Subject: Income in Germany and child support

Dear Sir or Madam,

I am writing to you because I am in a somewhat complex situation. I live in Vienna and work for a company as normal. However, I also work as a freelance consultant for a company in Germany.

I am not sure how this affects my taxes. Where do I have to pay tax on my salary? In Austria, right? But do I also have to pay taxes in Germany?

Another issue concerns my two children. One lives with my first ex-wife in Berlin and my second child lives in Japan with my second ex-wife, who emigrated. Both have demanded child support from me so far, but I have heard that this is not entirely correct. Do I have to pay at all if the children do not live in Austria?

Kind regards
Thomas Mayer

Table 4.7.: Task 2 (version 3)

Task 2 (Version 3 & 4) Further feedback was provided by a tax law expert from the AK following two additional meetings to incorporate the AK's expertise. The final changes of *task 2 (version 3)* (see table 4.7)[4] are shown as bolded text in the final version of *task 2 (version 4)*, (see table 4.8), which was used during the interview process.[5]

Sender: thomas.mayer@mailmail.at Subject: Income in Germany and child support Dear Sir or Madam, I am writing to you because I am in a somewhat complex situation. I live in Vienna and work for a company as normal. In addition, however, I work as a freelance consultant for a company in Germany. **I no longer have a residence in Germany and will perform this consulting work exclusively 100% remotely from Austria.** I am not sure how this will affect my taxes. Where do I have to pay tax on my salary now? In Austria, right? But do I also have to pay taxes in Germany? **Another issue concerns my two minor children (aged 6 and 12). One lives with my first ex-wife in Berlin and my second child lives in Serbia with my second ex-wife. I pay a total of €1,000 in child support each month. Where can I claim this on my tax return?** Kind regards Thomas Mayer

Table 4.8.: Task 2 (version 4)

Pilot interviews

As part of the preparations, a pilot interview was conducted to test the study design. It was conducted with two students from the Master's degree program in Digital Humanities. Instead of the AK-chatbot, ChatGPT was used as a replacement (OpenAI, n.d.). For the pilot interview, we had to turn to ChatGPT as a substitute, as we, as non-AK employees, did not yet have access to the AK-chatbot at that time. ChatGPT was an easily accessible alternative for testing the study procedure. This allowed for an initial estimate of the duration.

Furthermore, a legal expert from the AK participated in the study's pilot interview before conducting the official interview sessions with the AK's tax law experts. To test the adapted tasks and the planned study design, the final pilot interview was conducted with a member of the research group, Visualization and Data Analysis. The AK's legal expert was again consulted via email to provide feedback on the tasks used in the final

[4]The original German version is provided in the appendix A.5.2.
[5]The original German version is provided in the appendix A.5.2.

version of the study.

4.2.2. Rating scales

This section introduces and explains the rating scales of the CEF. First, the rating of self-reported task-solving confidence and the evaluation of the chatbot output are discussed (Collins et al., 2024a, 2024b), followed by a presentation of the scales used in the online questionnaire (Hart, 2006; Holmes et al., 2019) and the questions asked in the semi-structured interviews. Finally, the integration of the XAI Novice Question Bank by Schmude et al. (2025) is addressed.

Rating Self-Reported Prior Knowledge, Task-Solving Confidence and Chatbot Output

Before interacting with the chatbot, participants were asked to rate their own prior knowledge, which was inspired by the rating scale used in Collins et al. (2024a): *"(...) how confident are you that you could solve this problem entirely on your own, with your current knowledge base and no extra assistance?"* (Collins et al., 2024b, p. 5). Each participant had to rate their own prior knowledge and task-solving confidence on a 5-point Likert scale ranging from "not very confident" to "very confident".

After interacting with the chatbot to complete the two tasks, the perceived helpfulness of the chatbot's output was measured with another question. Participants were asked by the interviewer: *"How helpful did you find the chatbot's responses for solving the problem?"* (Collins et al., 2024b, p. 6). The perceived helpfulness of the chatbot responses was rated on a 5-point Likert scale ranging from "not very helpful" to "very helpful". The rating was obtained for each task to differentiate between those that, as discussed earlier in subsection 4.2.1, varied in difficulty.

Question	**Question type**
How confident are you that you can solve the problem on your own with your current knowledge and without additional support?	5-point Likert scale
How helpful did you find the chatbot's responses for solving the problem?	5-point Likert scale

Table 4.9.: Questions for rating self-reported prior knowledge and task-solving confidence, and for evaluating the chatbot's output

Online survey

After interacting with the chatbot and rating its output, as discussed previously, participants completed an online survey (see Figure 4.1). This online survey combines multiple instruments to measure a wide range of aspects during interviews with legal experts. The survey is used to obtain ratings on the Raw TLX or RTLX (Hart, 2006), the Chatbot Usability Questionnaire (Holmes et al., 2019), and the perceived trustworthiness of the

chatbot. During the oral interview, further questions focus on the categories of perceived trustworthiness, integration into work routine, areas for future development of the tax law chatbot, participants' information needs, and their prior knowledge regarding chatbot use (see figure 4.1). The online survey consists of 28 questions in total.

Scale	**Items**	**Question type**
Raw TLX	6	slider scale (0 to 100, 20 equal intervals)
Chatbot Usability Questionnaire (CUQ)	16	5-point Likert scale
Perceived Trustworthiness	6	5-point Likert scale

Table 4.10.: Survey items

Perceived mental workload (RTLX)

The NASA Task Load Index (NASA-TLX) is a scale to measure perceived mental workload. The index consists of 6 subscales that measure aspects of mental demand, physical demand, temporal demand, frustration, performance, and effort (Hart, 2006, p. 904).

Subscale	Endpoints	Description
Mental Demand	Low/High	How much mental and perceptual activity was required?
Physical Demand	Low/High	How much physical activity was required?
Temporal Demand	Low/High	How much time pressure did you feel?
Performance	Good/Poor	How successful do you think you were?
Effort	Low/High	How hard did you have to work to accomplish your level of Performance?
Frustration	Low/High	How insecure, discouraged, irritated, stressed, and annoyed were you?

Table 4.11.: NASA-TLX subscales. Adapted from Hart (2006, p. 908).

The original version, which is administered using pen and paper, presents "each scale (...) as 12-cm line divided into 20 equal intervals anchored by bipolar descriptors (e.g.. High/Low). The 21 vertical tick marks on each scale divide the scale from 0 to 100 in increments of 5. If a subject marks between two ticks. the value of the right tick is used (i.e.. round up)" (NASA Ames Research Center, 1988, p. 4). Since the questionnaires for

our project had to be completed online and evaluated using the University of Vienna's SoSci Survey platform, a 20-point slider scale was chosen.

NASA-TLX was originally designed with a weighting process in mind, but its modification without any weighting called "Raw TLX (RTLX)" has gained popularity due to its simpler way of application. An overall workload score can still be created by averaging the recorded ratings of RTLX, or an individual in-depth analysis of its subscales' values can be chosen instead of an overall score (Hart, 2006, pp. 906–907).

Hertzum (2021) shows in a meta-analytic review that the majority of studies using NASA-TLX chose the "Raw TLX (RTLX)" rather than the original, noting that his "initial plan was to review papers reporting raw TLX as well as papers reporting weighted TLX. However, it turned out that 560 papers reported raw TLX and 34 only papers reported weighted TLX" (Hertzum, 2021, p. 871). To ensure better comparability, it seems more sensible to use the raw TLX rather than the weighted TLX, which would offer less comparability across domains or regions. Hertzum (2021) includes reference values for different domains, technologies, regions, and for real-life or lab settings. For this scientific work, the value for "*Office work*" in the "*Domain*" area, the value for "*Desktop application*" in the "*Technology*" area and the "*Region*"-value for "*Europe*" appear to be possible reference TLX-values (Hertzum, 2021, pp. 874–875).

Furthermore, the use of the original weighted NASA-TLX poses difficulties resulting from its weighting process. Forced differentiation between dimensions of equal importance, inconsistencies in pairwise comparisons, and limited weighting ranges that may exclude dimensions considered relevant by study participants may call into question the meaningfulness of the calculated overall score (Virtanen et al., 2022, pp. 727–728).

A systematic analysis of the six subscales is not only supported by practical considerations, but also by the fact that, as mentioned by Virtanen et al. (2022), the equivalence of individual scales would need to be ensured (Virtanen et al., 2022, p. 727). Examining subscales independently (as suggested by Galy et al. (2018) and Hertzum (2022)) seems particularly valuable in contexts such as human-chatbot interactions, where certain workload dimensions (e.g., physical demand) may provide less meaningful data, while others (e.g., mental demand or frustration) could provide more adequate insights. Galy et al. (2018) highlights the importance of considering subscales of NASA-TLX (or RTLX) independently in complex situations, because an overall score would not be fit to assess the different levels of mental workload (Galy et al., 2018, p. 517).

Since this study (14 interviews) primarily focuses on qualitative aspects, it seems less appropriate to focus on overall scores and more meaningful to analyze the subscales independently. For this purpose, a German translation was created, as the interviews were conducted in German (see table A.7).

Chatbot Usability Questionnaire (CUQ)

While the RTLX assesses the participants' perceived workload when using the tax law chatbot, the second rating scale focuses on its usability. The questionnaire consists of 16 questions. Both positive and negative aspects are rated by participants. CUQ items are rated on a 5-point Likert scale ranging from "Strongly Disagree" to "Strongly

Agree" (Holmes et al., 2019, p. 210). The score (0-100) is then calculated using a tool provided by the authors' university (Ulster University, 2025).

As discussed in section 3.2, the Chatbot Usability Questionnaire (CUQ) fits best the needs of the evaluation of the AK-chatbot. The CUQ combines seven dimensions; "personality", "onboarding", "user experience", "chatbot understanding", "chatbot output", "error handling" and "user experience (Holmes et al., 2023, pp. 332–333).

In this study, it would not have made sense to retain the original version of the CUQ, as the legal advisors of AK conduct their work in German, and a German translation would be much easier for the participants to understand (see table 4.12). This also reduces the possibility of misunderstandings when answering the survey.

Original question	**Translation**
The chatbot's personality was realistic and engaging.	Die Persönlichkeit des Chatbots war realistisch und ansprechend.
The chatbot seemed too robotic.	Der Chatbot wirkte zu künstlich/roboterhaft.
The chatbot was welcoming during initial setup.	Der Chatbot begrüßte mich freundlich zu Beginn.
The chatbot seemed very unfriendly.	Der Chatbot wirkte sehr unfreundlich.
The chatbot explained its scope and purpose well.	Der Chatbot erklärte seinen Zweck und sein Einsatzgebiet gut.
The chatbot gave no indication as to its purpose.	Der Chatbot gab keinen Hinweis auf seinen Zweck/sein Anwendungsgebiet
The chatbot was easy to navigate.	Der Chatbot war leicht zu bedienen.
It would be easy to get confused when using the chatbot.	Es wäre leicht, bei der Nutzung des Chatbots verwirrt zu werden.
The chatbot understood me well.	Der Chatbot verstand mich gut.
The chatbot failed to recognise a lot of my inputs.	Der Chatbot erkannte viele meiner Eingaben nicht.
Chatbot responses were useful, appropriate and informative.	Die Antworten des Chatbots waren nützlich, angemessen und informativ.
Chatbot responses were not relevant.	Die Antworten des Chatbots waren nicht relevant.
The chatbot coped well with any errors or mistakes.	Der Chatbot ging gut mit Fehlern oder Missverständnissen um.
The chatbot seemed unable to handle any errors.	Der Chatbot schien nicht in der Lage zu sein, mit Fehlern umzugehen.
The chatbot was very easy to use.	Der Chatbot war sehr einfach zu benutzen.
The chatbot was very complex.	Der Chatbot war sehr komplex.

Table 4.12.: Translation of the Chatbot Usability Questionnaire

Perceived trustworthiness

The third part of the online survey covers the broader topics of perceived trustworthiness and attitudes towards AI and the tax law chatbot, while two questions also address helpfulness and perceived accuracy. This topic is taken up again in the semi-structured interview section (see section 4.2.3). Using these six questions, which were created specifically for this master's thesis, the topic of perceived trustworthiness is to be explicitly addressed in the form of questions using a 5-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree" (see table 4.13). As mentioned earlier, the questions are guided by the definition provided by Bisconti et al. (2025) (see section 3.3 and table 3.2).

> "Trustworthiness is an attribute, outcome of meeting stakeholders' expectations" (Bisconti et al., 2025, p. 133).

Participants' responses are taken into account in the evaluation alongside qualitative data from the semi-structured interviews.

Original question	**Translation**
Ich finde die Antworten des Chatbots hilfreich.	I find the chatbot's answers helpful.
Ich vertraue den Antworten des Chabots.	I trust the chatbot's answers.
Ich finde, dass die Antworten des Chatbots korrekt wirken.	I think the chatbot's answers seem correct.
Ich arbeite gern mit dem Chatbot.	I enjoy working with the chatbot.
Ich fühle mich wohl, wenn ich Künstliche Intelligenz in meine Arbeit einbinde.	I feel comfortable incorporating artificial intelligence into my work.
Ich habe allgemein negative Gefühle gegenüber Künstlicher Intelligenz.	I generally have negative feelings toward artificial intelligence.

Table 4.13.: Translation of the Perceived Trustworthiness Questionnaire

4.2.3. Interviews questions

After the online survey had been submitted, a semi-structured interview was conducted to gather qualitative data. The interview questions, which are part of the semi-structured interview, were developed to specifically address perceived trustworthiness, integration into the work routine, areas for future development, and participants' information needs. The interview schedule used for the study can also be found in the appendix (see appendix A.6). The reason for selecting these questions was to combine quantitative results with qualitative data to gain even more detailed insights. The interview questions on perceived trustworthiness specifically complement the previously conducted online survey.

Interview questions	
Original question	*Translation*
Perceived Trustworthiness	
Wie sehr würden Sie dem Chatbot bei rechtlichen Entscheidungen vertrauen?	How much would you trust the chatbot with legal decisions?
Wie groß schätzen Sie das Risiko ein, dass der Chatbot falsche Informationen gibt?	How high do you estimate the risk of the chatbot providing incorrect information?
Integration into Work Routine	
In welchen Situationen können Sie sich vorstellen diesen Chatbot in Ihrer Arbeit (als Steuerrechtsexpert:in) nutzen?	In what situations can you imagine using this chatbot in your work (as a tax law expert)?
Wie könnte dieser Chatbot Ihre Arbeit effektiver machen?	How could this chatbot make your work more effective?
Könnten Sie sich vorstellen, diesen Chatbot in Zukunft regelmäßig in Ihrer Arbeit zu nutzen? Warum?	Could you imagine using this chatbot regularly in your work in the future? Why?
Welche Aufgaben halten Sie für nicht geeignet, um sie von diesem Chatbot erledigen zu lassen? Warum?	Which tasks do you consider unsuitable for this chatbot to perform? Why?
Areas for Future Development	
Welche (weiteren) Hintergrundinformationen sollte der Chatbot "verstehen" können, damit seine Antworten gut informiert wirken?	What (additional) background information should the chatbot be able to “understand” so that its responses appear well-informed?
Würden Sie es bevorzugen, wenn der Chatbot sowohl auf juristische Fachsprache als auch auf Alltagssprache reagieren könnte? Warum?	Would you prefer the chatbot to be able to respond using both legal terminology as well as everyday language? Why?
Wie würden Sie sich fühlen, wenn der Chatbot bestimmte Antworten aus Risikogründen (z.B.ethische Gründe, Datenschutz,...) markiert?	How would you feel if the chatbot flagged certain responses for risk reasons (e.g., ethical reasons, data protection, etc.)?
Wenn Sie den Chatbot in irgendeiner Weise verbessern könnten, welche Änderungen würden Sie vornehmen?	If you could improve the chatbot in any way, what changes would you make?
Wie hilfreich finden Sie die Dokumente, die in den Antworten des Bots verlinkt sind? Was gefällt Ihnen daran (nicht)?	How helpful do you find the documents linked in the bot’s responses? What do you like (or dislike) about them?

Table 4.14.: Translation of interview questions grouped by topic

XAI Novice Question Bank

The XAI Novice Question Bank was developed to find out "what information do AI novices who could be affected by algorithmic decisions need in order to decide about adopting an ADM system" (Schmude et al., 2025, p. 2). The XAI Novice Question Bank can be used to assess the information needs of people affected by artificial intelligence (Schmude et al., 2025).

The authors of the XAI Novice Question Bank were consulted in order to adapt the questions for use with a chatbot. Table 4.16 contains the English translation of the XAI Novice Question Bank adjusted for chatbot use. The new version of the XAI Novice Question Bank for chatbots thus represents a further contribution of this thesis. The German version of the question bank is available in appendix A.16.1. Furthermore, the authors' consent was obtained to adopt a new numbering system to simplify its use in the study. The previously used identification numbers for assigning the questions to the domains were replaced with a simpler numbering system based on their corresponding categories (see figure A.6).

- Chatbot context: *Chatbotkontext K1 - K7*
- Chatbot usage: *Chatbotnutzung N1 - N6*
- Data: *Daten D1 - D3*
- Chatbot specifications: *Chatbotspezifikationen S1 - S10*

After answering questions on topics such as perceived trustworthiness, possible integration of the chatbot into tax law experts' work routine, and areas for future improvement, the participants' information needs are addressed. Participants are asked to select at least three questions from the XAI Novice Question Bank that would help them better understand or use the chatbot if answered properly by experts. Furthermore, participants are then asked about their motivation for selecting these questions.

Original question	**Translation**
Bitte wählen Sie mindestens 3 Fragen aus, die für Sie wichtig wären, um den Chatbot besser verstehen und verwenden zu können. Wieso haben Sie diese Fragen gewählt?	Please select at least 3 questions that would help you better understand and use the chatbot. Why did you choose these questions?

Table 4.15.: Translation of the XAI Novice Question Bank interview question

ID	Chatbot context	ID	Data
K1	What is the intention of deploying the chatbot?	D1	What kind of data was the chatbot trained on?
K2	How does the deployment process of the chatbot work?	D2	What is the source of the training data?
K3	What are the consequences after the deployment of the chatbot?	D3	Are the data correct / representative / safe?
K4	How is the chatbot developed?	**ID**	**Chatbot specifications**
K5	Who is the chatbot's intended target group?	S1	What does the chatbot consider and why?
K6	Who is responsible for the chatbot's deployment?	S2	What is the scope of the chatbot's capability?
K7	Is the chatbot's deployment right?	S3	Which kind of output does the chatbot give?
ID	**Chatbot usage**	S4	How does the chatbot learn?
N1	How is the chatbot operated?	S5	What is the chatbot's overall logic?
N2	How is the chabot used by and on people?	S6	Which kind of algorithm was used?
N3	How is the chatbot integrated into existing structures?	S7	What kind of mistakes is the chatbot likely to make?
N4	How can the chatbot be misused?	S8	What are the limitations of the chatbot?
N5	How would the chatbot handle [this case]?	S9	How reliable are the predictions of the chatbot?
N6	What are the costs of the chatbot?	S10	What does [a machine learning concept] mean?

Table 4.16.: XAI Novice Question Bank: Chatbot version

Prior knowledge and experience

Finally, prior knowledge and experience are addressed to collect more information about chatbot-specific know-how. The qualitative data collected is complemented during analysis with demographic data collected prior to the interview to gain insight into the participants' existing knowledge and experience (see demographics form in table A.6).

Original question	**Translation**
Haben Sie schon einmal im Arbeitskontext einen Chatbot verwendet?	Have you ever used a chatbot in a professional context?
Haben Sie den Steuerrechts-Chatbot der Arbeiterkammer schon einmal verwendet? (Wenn vorhanden:) Wie oft nutzen Sie diesen Chatbot? (FREQUENZ)	Have you ever used the Arbeiterkammer's tax-law chatbot? (If applicable) How often do you use this chatbot? [frequency]
Falls ja, welchen Chatbot haben Sie verwendet?	If yes, which chatbot did you use?
Aus welchen Gründen nutzen Sie Chatbots im Arbeitskontext?	For what reasons do you use chatbots in your work?
Welche Erfahrungen haben Sie mit rechtlichen Chatbots gemacht?	What experiences have you had with legal chatbots?
Welche anderen Erfahrungen haben Sie mit Chatbots? (Wenn vorhanden:) Wie oft nutzen Sie diesen Chatbot? (FREQUENZ)	What other experiences have you had with chatbots? (If applicable) How often do you use that chatbot? [frequency]

Table 4.17.: Questions about participants' prior knowledge

5. Results

This chapter presents demographic data and results. Following the presentation of demographic data, the results corresponding to each research question are reported. Quantitative and qualitative data are combined to present the results for each research question.

- **RQ 1**: How do tax-law experts perceive and experience the usability of a Retrieval Augmented Generation (RAG) chatbot when completing routine tax law advisory tasks?
- **RQ 2**: Which characteristics of a domain-specific chatbot shape its perceived trustworthiness among users?
- **RQ 3**: How do tax law experts tend to use the tax law chatbot when solving work tasks?

5.1. Demographics

The demographic data was collected using an online form hosted by my university. It contained 10 questions (see table A.6). Further information about participants' prior knowledge and chatbot use was collected during the interview sessions to better understand the study participants (see sections 5.2 and A.6).

A total of 14 people participated in the study. The interviews were conducted between May 6, 2025, and July 10, 2025. The group of study participants consisted of 4 men and 10 women. Options such as "non-binary" or "other gender" were also offered for gender identification in order to capture the potential diversity of participants.

Age	**Participants**	**Male**	**Female**
50-59	4	2	2
40-49	2	0	2
30-39	7	1	6
20-29	1	1	0

Table 5.1.: Age groups and gender of participants

9 out of 14 people stated that their highest level of education is a university degree. Another person reported having a University of Applied Sciences degree (*Fachhochschulabschluss*). Furthermore, 4 people stated that their highest educational qualification was a high school diploma (*Matura*).

5.2. Prior knowledge

To obtain further information about participants' prior knowledge, the semi-structured interview focused on their individual experience with chatbots in their everyday work and private lives.

Have you ever used a chatbot in a professional context? 11 out of 14 participants stated that they had used a chatbot in a professional context before. ChatGPT is reported to be the most commonly used chatbot in this context (see figure 5.1). When asked whether they had also used the AK-chatbot, 11 out of 14 tax-law experts confirmed that they had already used it. However, it should be noted that two participants who had already used the AK-chatbot stated that they had only tried it briefly. However, when asked previously about using chatbots in a professional context, they said they had not yet used one in their work. Furthermore, two participants (P1, P8) stated that they had already gained experience with other legal chatbots like Genjus KI (MANZ'sche Verlags- und Universitätsbuchhandlung GmbH, 2025b) and Lexis+ AI (LexisNexis Österreich, 2026). 3 of 14 participants confirmed they had received training on using the chatbot. Two participants explicitly expressed a need for training (e.g., on prompting techniques and use cases).

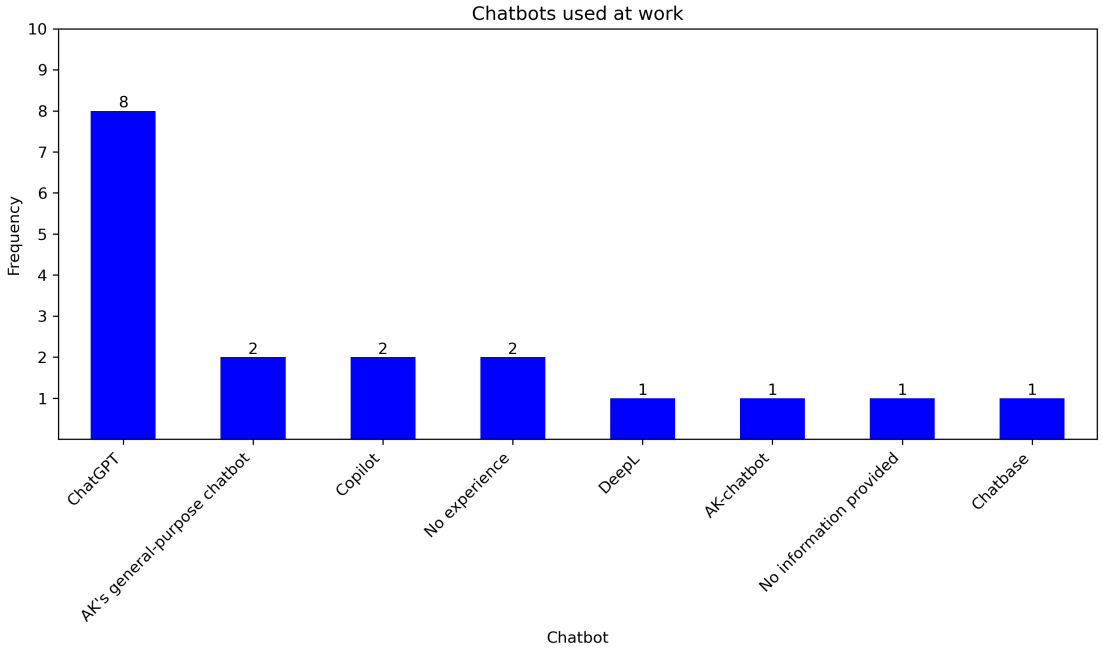


Figure 5.1.: Chatbots used at work

Have you ever used the Arbeiterkammer's tax law chatbot? The AK's tax law experts did not use the chatbot daily. Some study participants had tried the chatbot only a few times or had not used it at all. One participant (P5) also reported having used the chatbot 50-60 times in the past but had since stopped. The participants were thus, in the majority, not experienced users. As many as 50% of study participants stated that they had either only "first experiences, no regular use" (P1, P3, P5, P12, P14) or "no experience" (P6, P7) (see figure 5.2).

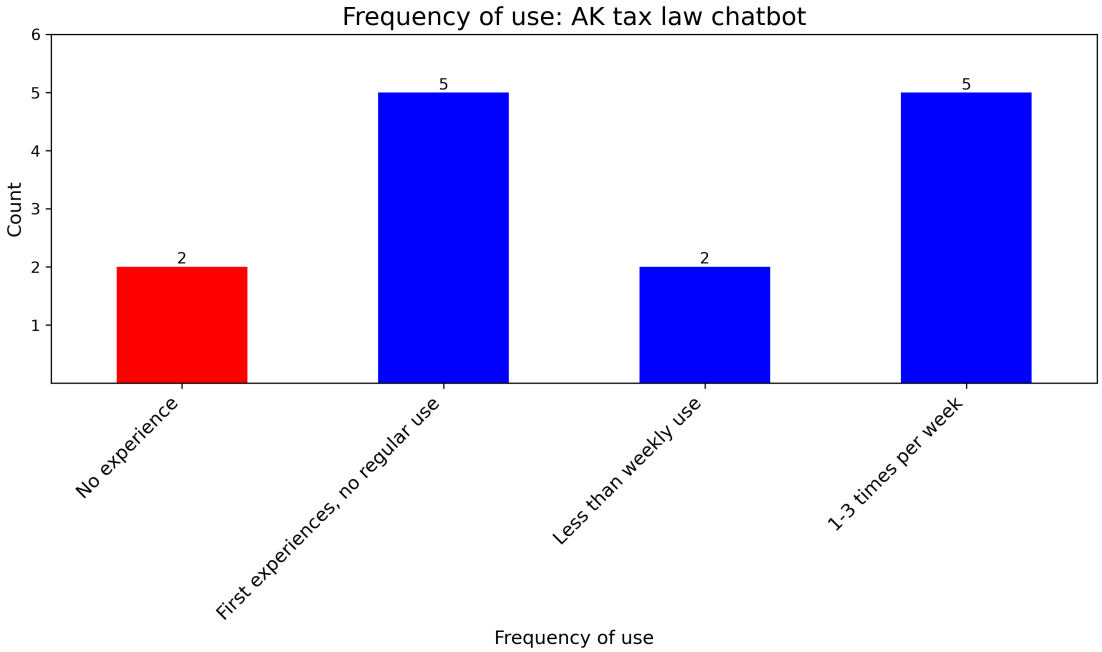


Figure 5.2.: Frequency of use: AK tax law chatbot

Why do you use chatbots at work? The most frequent reason for using chatbots at work was "to make work easier" (P4, P8, P9). Three respondents reported no experience with chatbots (P7, P12, P14). Two people used chatbots for translating (P6, P13). Others mentioned curiosity (P1), improving text wording (P2), rephrasing texts (P3), research (especially when new to a job) (P10), reviewing information (P13), and working more efficiently (P11) as their reasons. One person stated that chatbots were not suitable for "practical use" in daily tasks (P5).

Do you use chatbots in your private life? As seen in table 5.2, numerous tax law experts (6 out of 14) did not use chatbots in their private lives. Only 5 participants (P2, P3, P4, P10, P6) used chatbots weekly.

Frequency	Participant
daily	P2
2-3 times per week	P3, P4, P10
1 time per week	P6
2-3 times per month	P7
1 time per month	P8, P11
No use	P1, P5, P9, P12, P13, P14

Table 5.2.: Private chatbot usage

The participants stated they primarily used ChatGPT for personal purposes (P2, P3, P4, P7, P8, P10, P11). DeepL was also mentioned once (P2). Only one participant (P2) stated that they used two chatbots in their private life, which is why figure 5.3 includes 15 entries instead of 14. In addition, one person did not provide any information about which chatbot they used in their private life.

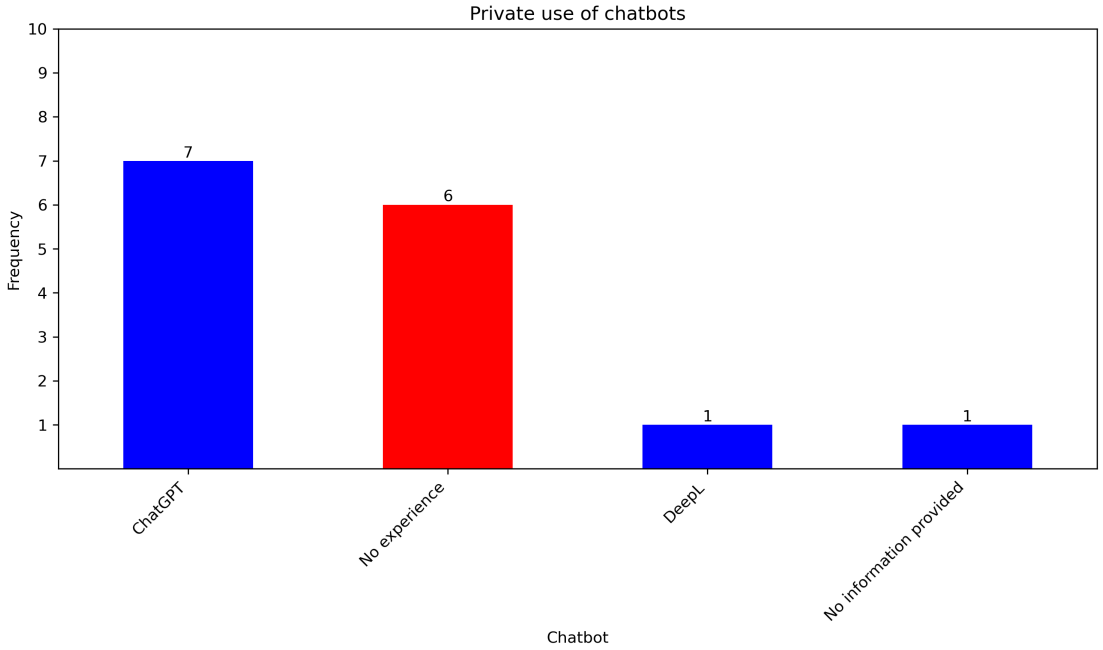


Figure 5.3.: Private use of chatbots

5.3. RQ1: How do tax-law experts perceive and experience the usability of a Retrieval Augmented Generation (RAG) chatbot when completing routine tax law advisory tasks?

After the demographic data presented above, this section turns to the first research question of this master's thesis. RQ1 addresses the perceived usability of the tax law chatbot. In particular, the perceived difficulty of the tasks completed during the human–chatbot interaction is assessed by examining how participants' task-solving confidence relates to the quality of the chatbot's outputs. The analysis, therefore, examines participants' satisfaction with the generated answers and whether differences emerge across tasks of varying difficulty. Finally, the results are used to determine whether the chatbot's strengths or weaknesses can be identified.

Self-Reported Prior Knowledge and Task-Solving Confidence

Before interacting with the chatbot, participants were asked to rate their own prior knowledge, which was inspired by the rating scale used in Collins et al. (2024a): *"(...) how confident are you that you could solve this problem entirely on your own, with your current knowledge base and no extra assistance?"* (Collins et al., 2024b, p. 5).

After reading the tasks, which later had to be solved using the AK-chatbot, they were asked: *"How confident are you that you can solve the problem on your own with your current knowledge and without additional support?"*. Participants had to rate their confidence in solving the task based on their prior knowledge using a 5-point Likert scale (*not very confident - not confident - neutral - confident - very confident*).

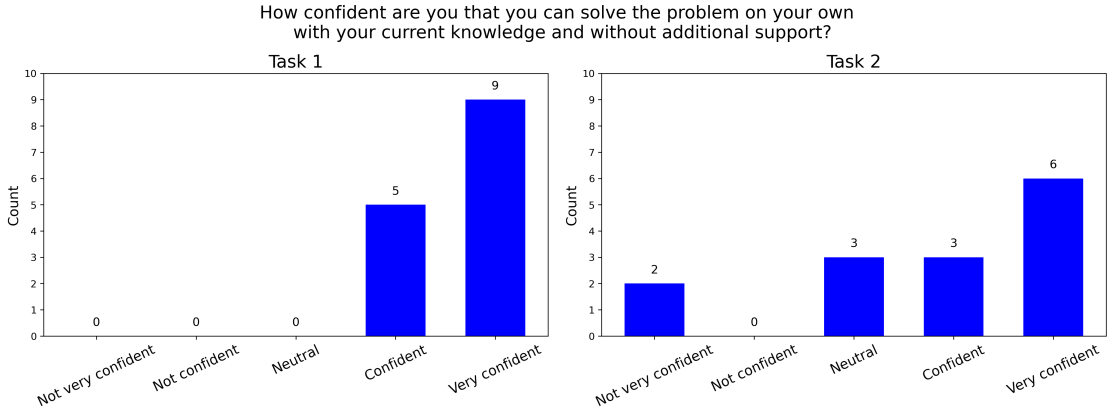


Figure 5.4.: Self-Reported Prior Knowledge and Task-Solving Confidence (Tasks 1 and 2)

In figure 5.4, participants stated that they were *"confident"* (*"Eher zuversichtlich"*) or *"very confident"* (*"Sehr zuversichtlich"*) that they could solve task 1 without further help.

However, according to their own assessment, they were less confident about task 2. This assessment also reflects the intended level of difficulty of the tasks. The aim was to set different levels of difficulty for the two tasks (see section 4.2.1).

Rating Chatbot Output

After interacting with the chatbot, participants were asked to rate the generated output using a 5-point Likert scale. They were asked by the interviewer: *"How helpful did you find the chatbot's responses for solving the problem?"* (Collins et al., 2024b, p. 6). Subsequently, participants evaluated the chatbot's output per task.

It seems that the varying levels of difficulty are also reflected in how helpful the chatbot's responses were. Figure 5.5 shows that in task 1, the chatbot appears to have provided more satisfactory answers, but in task 2, participants perceive its output as less helpful. This can also be seen in the qualitative section, where participants identified suitable and unsuitable areas of application for the chatbot and noted its weaknesses. The chatbot seems to exhibit particular shortcomings when dealing with more complex questions.

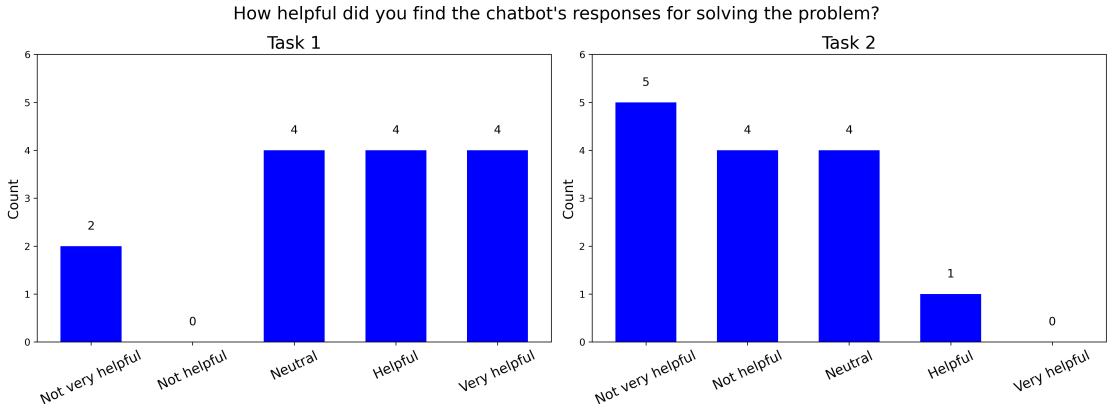


Figure 5.5.: Rating chatbot output (Task 1 and 2)

Which tasks do you consider unsuitable for this chatbot to perform? During the semi-structured interviews, participants mentioned various tasks they considered unsuitable for the AK-chatbot. Participants doubted that the chatbot would be suitable for solving complex tasks, which are also listed in table 5.3 as “responding to complex inquiries" and “writing complex texts”. In this context, “responding to complex inquiries” refers to processing inquiries regarding tax assessments or making decisions on individual cases. The term “writing complex texts” refers to the creation of statements ("Stellungnahmen") or political texts. In addition, 3 out of 14 participants would not suggest a "direct interaction with clients". P10, P11, and P14 could not identify any unsuitable tasks.

Activity	X of 14	Participant
Responding to complex inquiries	4 of 14	P3, P6, P9, P12
Direct interaction with clients	3 of 14	P4, P8, P13
Writing complex texts	2 of 14	P1, P2
Performing calculations	2 of 14	P1, P7
Doing research	1 of 14	P8
"In its current form, almost all of them."	1 of 14	P5

Table 5.3.: Unsuitable areas of application

Chatbot Usability Questionnaire

Unlike the RTLX values, the second scale of the online survey, the Chatbot Usability Questionnaire (Holmes et al., 2019), cannot be compared with extensive reference values from past studies (Bangor et al., 2008). The tool made available by Holmes et al. (2019), in the form of an xlsx file, provided ratings of the chatbot's usability (Ulster University, 2025), as seen in figure A.1. The Mean Question Scores are based on a 5-point Likert scale ranging from *strongly disagree* (1) to *strongly agree* (5).

When comparing the AK-chatbot (referred to as *Steuerrecht-Chatbot (AK)* in table A.12 to the chatbots evaluated in the validation study of the Chatbot Usability Questionnaire, it could be argued that the tested tax law chatbot may be described as *average* or *poor*. The evaluation values obtained for the AK-chatbot fall between those of two chatbots that were classified by experts as *average* and *poor* in the original study (Holmes et al., 2023, 336–338).

When analyzing figure 5.6, it can be concluded that professional experience appears to influence Chatbot Usability Questionnaire results. Tax law experts with less professional experience seem to rate the chatbot's usefulness more highly than those with more professional experience. This observation is also consistent with the literature review by Fazi et al. (2025) who explain that "younger workers tend to exhibit more favorable attitudes toward technology, perceiving it as particularly useful for acquiring knowledge" (Fazi et al., 2025, p. 150).

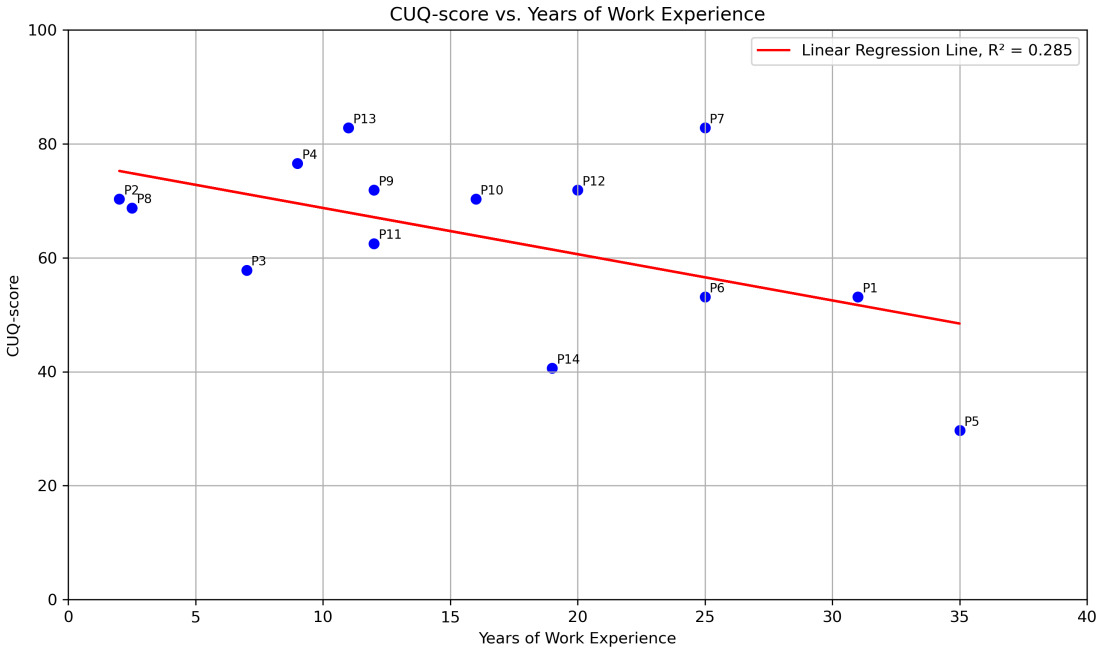


Figure 5.6.: CUQ-Score vs. years of work experience (including linear regression)

5.3.1. RQ 1.1: How effectively does the tax law chatbot address tax experts' needs in terms of usability and workflow integration?

After analyzing the general perceived usability of the chatbot, this subsection reports results related to RQ 1.1 and focuses on how tax law experts experienced the chatbot's usability and its integration into their existing work practices during task completion. The following findings highlight concrete usability-related issues and improvement needs identified by the AK's taw-law experts.

In this context, several participants offered specific suggestions to improve the chatbot's usability. The participants expressed a wish for an improved document-linking feature with quotation labeling in the corresponding source (P2, P8) (see also section 5.3.2). In addition, an easier login procedure (P1) and a "welcome message" with possible sample questions for the chatbot (P10) were suggested. One aspect that influenced the human-chatbot interaction of participants P3, P4, P6, P7, P10, P13, and P14 was error messages, which are part of the chatbot's meta prompt (see section 2.3.3). These error messages were displayed when participants tried to solve task 2 (see section 4.2.1). Participant 13 described the potential problems that arise:

> *The problem is that he immediately responds with "we don't answer that" when he hears certain keywords. Because he only hears "foreign employer," a keyword that triggers his response that we don't deal with foreign countries,*

> *he refers the matter to the KSW, which is not right, because we do have to provide information about what is being asked. So that is very much one of our tasks and not one of the KSW, the Chamber of Tax Advisors and Public Accountants.* (P13)

Technical issues also affected the interview process, as three interviews had to be postponed due to chatbot issues. Tax law experts could not log in, and the issues had to be solved by the chatbot's developer or the AK's IT staff.

5.3.2. RQ 1.2: What adjustments are needed to make legal advice chatbots improve usability and integration into work routine in practice?

After focusing more on the perceived usability of the chatbot in RQ1 and RQ1.1, RQ1.2 examines ways to improve the chatbot for better integration into everyday work.

Regarding the global theme of "possible improvements" identified during the qualitative analysis, the absence of certain relevant sources in the chatbot's responses is repeatedly noted. As the author of the study, I am unfortunately unable to determine which data collection is already integrated into the chatbot's knowledge base, as its creation and maintenance are handled by an external company. It may also be the case that certain data is already available to the chatbot but was not linked due to the query or retrieval mechanism.

Which additional background information would be useful? Based on the qualitative data, according to our participants, the AK-chatbot should have access to judicial decisions of the Federal Finance Court (*Bundesfinanzgericht, BFG*) and commentaries on Austrian income tax law (*Einkommenssteuerrecht*) or wage tax guidelines (*Lohnsteuerrichtlinien*) (P2, P12). Furthermore, participants would welcome access to articles published by LexisNexis (LexisNexis Österreich, 2025) or Linde Verlag (Linde Verlag, 2025) and the databases Findok (Bundesministerium für Finanzen, 2025a) and RDB (MANZ'sche Verlags- und Universitätsbuchhandlung GmbH, 2025) (P2, P3, P7).

Suggested (P6) were guidelines from the Austrian Ministry of Finance and step-by-step instructions (e.g., for filing complaints or amending tax returns) on the FinanzOnline platform. Participants also proposed adding collections of question-answer pairs created by the AK's tax law experts (P1) and guidelines for tax law complaint procedures, tax return procedures, and practical instructions for handling special cases concerning family-related tax benefits ("*Familienbonus*") (P7). Furthermore, it was noted that documents relating to double taxation agreements ("*Doppelbesteuerungsabkommen*") were currently missing (P3, P9) (Bundesministerium für Finanzen, 2025b), and that it would be necessary to include more documents on cross-border, international tax issues (P2, P14). Specialized topics such as the taxation of insolvency compensation funds ("*Insolvenzentgeltfonds*") were considered relevant (P10). Particular importance is attached to the data's up-to-date nature (P5, P6), as "tax law is one of the most frequently reformed areas of law in Austria", according to tax experts (P5).

Document linking When the chatbot is used, links to source documents in its knowledge base are included in generated responses, allowing users to open them. No participant objected to the linking of documents and various participants expressed their satisfaction (P1, P7, P9, P11, P12, P13, P14), but one participant also admitted having too little experience to judge the sources' quality (P9) and another person explicitly expressed no need to use a chatbot, because of work experience and lack of intereset (P6).

In contrast to the "satisfied" group, 6 out of 14 participants (P2, P3, P4, P5, P8, P10) expressed dissatisfaction, as the linked documents did not display the exact location of the quoted content when opened. Users had to start searching on their own without any assistance. It was mentioned that users would appreciate the ability to adjust the content of linked documents (P4). Furthermore, the need for regularly updated documents was stated (P5), because during the interaction, documents from 2023 were displayed, although the corresponding document from 2025 would have been necessary, as consultants are required to work with up-to-date values. Another participant (P12) noted that an unnecessary number of documents, including documents relating to the wrong countries, had been linked. By the end of our evaluation, the highlighting of quotations in the linked sources had not yet been implemented.

Legal terminology vs. everyday language During the interviews, the experts were asked what type of language they preferred. ("Would you prefer the chatbot to be able to respond using both legal terminology as well as everyday language? Why?") This question arose because, on the one hand, the chatbot is used by legal experts, but on the other hand, the answers must be understandable to laypeople. As shown in table 5.4, everyday language is preferred for responses to clients.

Preference	**X of 14**	**Participant**
Everyday language	7 of 14	P1, P2, P6, P8, P9, P10, P14
Everyday language for clients, legal terminology for tax law experts	4 of 14	P3, P11, P12, P13
Everyday language for clients, legal terminology for tax law experts or clients with prior knowledge	1 of 14	P4
Legal terminology	1 of 14	P5
Feature not necessary	1 of 14	P7

Table 5.4.: Everyday language vs. legal terminology

Risk-labeling feature Another aspect for improving usability was discussed with the proposal to mark answers with a warning label indicating potential risks. These labels could help experts identify potential risks in the answer. For example, for certain topics,

a note could be added to carefully review the answer in order to prevent errors when responding to inquiries. 14 out of 14 participants were in favor of labeling for risk reasons after being asked, "*How would you feel if the chatbot flagged certain responses for risk reasons (e.g., ethical reasons, data protection, etc.)?*".

Further improvements requested One of the main desires for improvement is to improve output precision through enhanced reasoning capabilities (P2, P4, P8, P9, P11) to connect various topics when solving demanding tasks, and to use more precise language and less verbose answers (P5, P6). A similar request is expressed in the desire for the chatbot to provide an immediate explanation when using individual keywords, without the need for a lengthy conversation (P10), the more general desire for the chatbot to "better understand" questions (P4), more helpful answers concerning cross-border tax issues (P3), enhanced memory capacity (P3) and the request for "learning" capabilities which would allow users to correct the chatbot (P13). Specific suggestions were also made regarding the chatbot's data. A participant suggested analyzing FinanzOnline's chatbot to answer tax requests in a similar way (P5). As previously mentioned, a connection to databases such as RDB (MANZ'sche Verlags- und Universitätsbuchhandlung GmbH, 2025) and to the decisions of the Federal Tax Court ("*Bundesfinanzgericht*") would be useful. Further criticism concerned the inability to edit the integrated wiki (P2, P4). Our participants suggested improving calculation capabilities (P4) or the ability to create PowerPoint presentations (P2).

5.4. RQ2: Which characteristics of a domain-specific chatbot shape its perceived trustworthiness among users?

After analyzing perceived usability and suggestions for possible adjustments to improve the chatbot's usability, this section presents the results for RQ2, shifting the focus to the characteristics of the tax-law chatbot that influenced users' perceptions of its trustworthiness. Quantitative survey results and qualitative data are combined to gain a deeper understanding of these perceptions.

Survey results

The third and final part of the quantitative data from the online survey consists of six 5-point Likert-scale questions assessing perceived trustworthiness and, to a lesser extent, helpfulness and perceived correctness, as explained in section 4.2.2 (Methods).

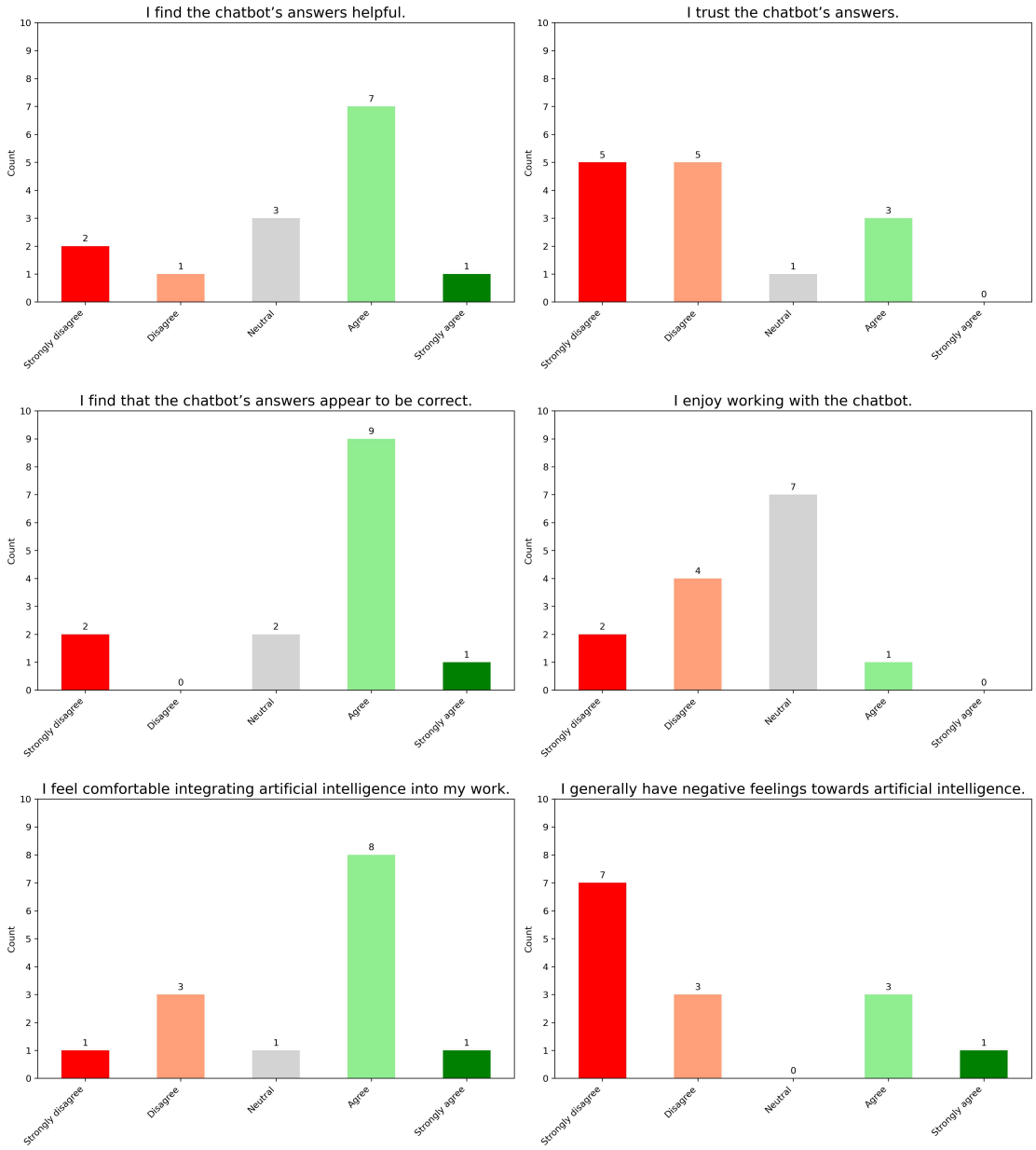


Figure 5.7.: Survey results on chatbot perception and attitudes toward artificial intelligence

Firstly, we can observe a positive attitude towards the use of artificial intelligence at work and in general (see figure 5.7). 9 out of 14 participants *agree* or *strongly agree* with the statement *"I feel comfortable integrating artificial intelligence into my work"*. Furthermore, 10 out of 14 participants *disagree* or *strongly disagree* with the statement *"I*

generally have negative feelings towards artificial intelligence".

Secondly, participants seem to find the answers helpful, and the chatbot also appears to convey a convincing impression of correctness. 8 out of 14 participants *agree* or *strongly agree* to the statement *"I find the chatbot's answers helpful"*. 10 out of 14 participants *agree* or *strongly agree* to the statement *"I find that the chatbot's answers appear to be correct"*.

Even though the chatbot's responses appear correct during use, surprisingly, the majority of participants say they have little trust in its output. 10 out of 14 participants *disagree* or *strongly disagree* with the statement "*I trust the chatbot's answers*".

Lastly, 6 out of 14 participants *disagree* or *strongly disagree* with the statement *"I enjoy working with the chatbot"*, while 7 participants have a *neutral* position towards the chatbot, and only 1 participant *agreed* with the statement.

In summary, our participants find the chatbot's answers helpful, they appear to be correct, they feel comfortable integrating AI into their work, and they tend not to have negative feelings toward AI, but they nevertheless do not trust the chatbot and are not particularly keen to use it.

Perceived Trustworthiness of the Chatbot in Legal Decision-Making and Answer Correctness

In addition to the quantitative data previously discussed, semi-structured interviews were conducted to assess the chatbot's perceived trustworthiness. For example, perceived trustworthiness in legal decision-making and the likelihood of receiving incorrect responses were addressed.

How much would you trust the chatbot with legal decisions? Two participants (P3, P9) out of 14 explicitly stated that they trusted the chatbot for legal decisions, but these individuals also said that the chatbot's answers lacked depth, that some questions went unanswered, or that they would still double-check the generated output. Several participants said they had little (P2, P4, P8, P12, P13) or no trust at all in the chatbot (P1, P5, P11, P14). Three participants (P6, P7, P10) mentioned checking the chatbot's answers and did not specifically mention trusting its answers.

How high do you estimate the risk of the chatbot providing incorrect information? When asked about estimating the risk of receiving incorrect information, participants mentioned a possibility of *"less than 50%"* (P7), *"50%"* (P12), *"more than 50%"* (P11), *"70 to 80%"* (P8) and *"80%"* (P5, P13), while others described the risk as "moderate" (P2), "moderate, moderately high" (P4). Those who did not give approximate percentages as an answer, explained that they found the risk *"difficult to judge, because one answer was correct, but the other task was not answered by the chatbot"* (P6) or found it *"difficult to answer"* and expressed lack of trust (P14) or mentioned that they *"wouldn't say wrong, but incomplete"* (P2). Another participant (P1) said he had not noticed any incorrect decisions made by the chatbot, but that he would still not trust it blindly. Possibly

missing data was used as an explanation why P10 *"doesn't trust the chatbot 100%"*. P9 also acknowledged the possibility of receiving incorrect information but noted recent improvements and a lack of experience with the chatbot.

The need for prior legal knowledge. Another aspect that the participants (P2, P3, P4, P5, P8, P11, P13, P14) mentioned on various occasions was the need for prior knowledge of the legal sector to use the chatbot effectively. The study participants emphasized the importance of correctly understanding legal vocabulary, on the one hand, and critically analyzing the output, on the other. Above all, they highlighted the importance of identifying potential incorrect assumptions made by the chatbot and the factual circumstances of a customer inquiry. They also noticed that the chatbot did not have the ability to *"think one step ahead"* to recognize circumstances-dependent provisions or consequences for other government financial services, for example. P13 explained the need for a critical assessment of the output during an interview session.

> *Well, I find it useful for verification purposes. If you're not entirely sure and think you would otherwise Google it or whatever, then you can type it in there. But you have to have enough knowledge to tell when it's trying to completely mislead you. Just typing it in and then thinking that it's telling you the right thing doesn't work.*

Furthermore, P11 pointed out the need for "legal literacy" to understand the generated output.

> *For someone who is not familiar with the subject matter, the answers may be misunderstood in some cases. Or what does "misunderstood" mean? Well, perhaps not necessarily 100% correctly understood.*

In the following quote, P2 explains the kind of legal reasoning necessary to process inquiries.

> *I wouldn't say wrong, but very incomplete. And if you're not aware that there's something else there, then you might just send it as a copy-paste, and then it would probably be wrong as a result. Because he only picks out one part, but often, as we know from our consulting experience, it's connected to something else, and that connection is missing.*

Feelings towards the chatbot

To capture the study participants' attitudes toward the chatbot, the qualitative data analysis aimed to identify positive and negative emotional expressions.

Negative sentiments. During analysis of the qualitative data from the semi-structured interviews, it became apparent that there were two main reasons for participants' negative statements: a lack of trust in the chatbot and criticism of its vagueness or incompleteness.

When comparing the participants' statements about their trust in the chatbot with their questionnaire responses on perceived trustworthiness (see figure 5.7), it was found that both sources of data were consistent and that the participants did not contradict themselves during semi-structured interviews (see table 5.5).

Positive sentiments. The results regarding participants' positive sentiments, as presented in table 5.5, demonstrate a predominance of positive attitudes towards the chatbot. Notably, the majority of participants (12 out of 14) indicated a willingness to use the chatbot in the future. Furthermore, while four participants made positive remarks about the chatbot, referencing its capabilities for research or drafting, six participants positively commented on the correctness of the chatbot's answers, which is logical given that, during the interaction, participants were happy with the chatbot's responses for easy tasks.

Legend		
Sentiment	**Category**	**Description**
Negative	Lack of trust	Expressions of distrust
Negative	Vague answers	Incompleteness or vagueness of chatbot answers
-	Trust score	*"I trust the chatbot's answers."* (5-point Likert scale)
Positive	Future use	Willingness to use the chatbot in the future
Positive	Research & drafts	Using the chatbot for research or creating drafts
Positive	Correctness	Notes correct answers positively

ID	**Lack of trust**	**Vague Answers**	**Trust score**	**Future use**	**Research & drafts**	**Correctness**
P1	X		Disagree	X		
P2	X	X	Strongly disagree	X	X	
P3		X	Agree	X		X
P4	X		Disagree	X		
P5	X	X	Strongly disagree	X		
P6	X	X	Disagree			X
P7			Agree	X		X
P8	X		Strongly disagree	X		
P9		X	Agree	X		X
P10			Neutral	X		
P11	X	X	Disagree	X	X	X
P12	X		Disagree	X		
P13	X		Strongly disagree	X	X	X
P14	X		Strongly disagree		X	

Table 5.5.: Combined overview of negative and positive sentiments per participant

Information needs

Since this chatbot is a newly introduced tool for tax law experts at the AK, the information needs in relation to this tool were specifically surveyed. Once the chatbot evaluation is complete, it will be possible to address these specific needs.

XAI Novice Question Bank. The XAI Novice Question Bank was selected to identify the information needs of study participants regarding the chatbot's use. Selecting questions to better understand and use the chatbot more effectively is intended to help address open questions. The XAI Novice Question Bank is divided into 4 categories covering different aspects of the chatbot: *Chatbot Context*, *Chatbot Usage*, *Data*, and *Chatbot Specifications*.

The analysis shows a particular need for information on *Chatbot Specifications*. Questions in this category were selected 35 times. In comparison, questions from the category *Chatbot Context* were selected 9 times, those from *Chatbot Usage* 14 times, and those from *Data* 15 times (see figures 5.8, A.7, A.9, and A.8). To illustrate this pattern even more clearly, 3 out of the 4 most frequently selected questions are part of the category *Chatbot Specifications*.

1. S4: "How does the chatbot learn?" (chosen by 12 participants)
2. D1: "What kind of data was the chatbot trained on?" (chosen by 7 participants)
3. S8: "What are the limitations of the chatbot?" (chosen by 6 participants)
4. S5: "What is the chatbot's overall logic?" (chosen by 5 participants)

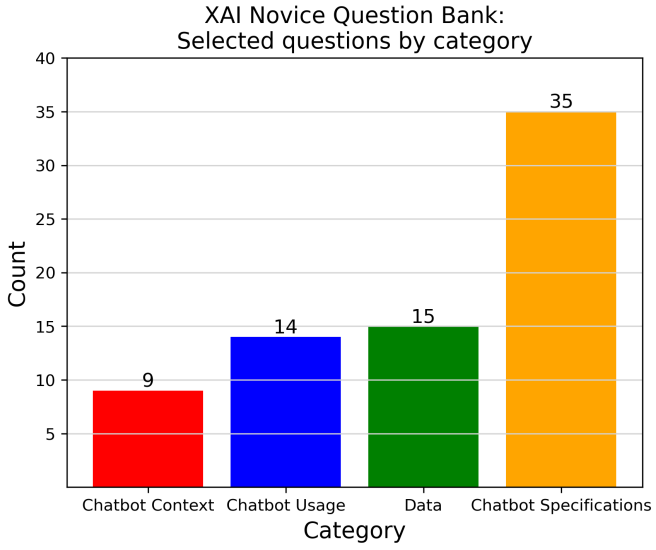


Figure 5.8.: XAI Novice Question Bank: Selected questions by category

5.5. RQ3: How do tax law experts tend to use the tax law chatbot when solving work tasks?

Finally, RQ3 and its sub-question RQ3.1 will be discussed. Here, it was possible to ask whether future use of the chatbot would be feasible for tax law experts and to identify potential areas of application that could serve as use cases.

In what situations can you imagine using this chatbot in your work (as a tax law expert)? In the interviews, various activities were identified as suitable for the chatbot, with “doing research” and "creating drafts of emails/articles/statements" being the most frequently mentioned areas of application. One explanation could be that this activity was not perceived by the experts as a “complex task” and was therefore considered an unsuitable application for the chatbot, which would require in-depth prior legal knowledge. With the exception of two task types, participants focused on creating texts using the chatbot. “Performing payroll accounting” and “checking tax assessment” would involve the chatbot performing calculations.

Activity	**X of 14**	**Participant**
Doing research	11 of 14	P1, P2, P3, P6, P7, P9, P10, P11, P12, P13, P14
Creating drafts of emails/articles/statements	8 of 14	P1, P2, P3, P4, P7, P8, P9, P11, P13
Translating emails	3 of 14	P2, P6, P13
Performing payroll accounting	2 of 14	P1, P14
Rephrasing text	2 of 14	P2, P6
Checking tax assessment	1 of 14	P4
Supporting new employees	1 of 14	P5

Table 5.6.: Areas of application

How could this chatbot make your work more effective? When asked whether the chatbot could increase their effectiveness, 13 of 14 tax law experts confirmed that it could. Only one participant stated that effectiveness could not be improved using the chatbot. P14 said that the chatbot would slow down workflow and would be too time-consuming to use. Asking the "right" questions would need too much time. Although the majority agreed that there was potential for greater effectiveness, some participants noted improvements that would be necessary for the chatbot to achieve it. P4 mentioned the need to calculate and check tax assessments; P5 expressed the need for more effective, in-depth answers containing less unnecessary information; and P13 stated that the chatbot should be able to learn from previous conversations. The chatbot could increase effectiveness by being used for "doing research" (P2, P7, P8, P10, P11) and "creating e-mail drafts /

response mails faster" (P2, P3, P6, P8, P11). Although noting the potential for increased effectiveness, P8 said that at the moment, the chatbot would slow down workflows. P9 and P12 also mentioned the need to improve the chatbot to increase the effectiveness of their work.

Could you imagine using this chatbot regularly in your work in the future? The majority of participants (13 out of 14) confirmed their willingness to use the chatbot in the future. Only one participant answered, "No, definitely not." Furthermore, a person with management responsibility stated that its use would be recommended to employees in the future. ("I can not only imagine it, but I will also order it."; original German quote: "Ich kann mir das nicht nur vorstellen; ich werde es anschaffen.")

5.6. Perceived Mental Workload

The NASA Task Load Index (NASA-TLX) is a scale to measure perceived mental workload, which consists of 6 different subscales; Mental Demand (MD), Physical Demand (PD), Temporal Demand (TD), Frustration (FR), Performance (PE), Effort (EF) and Frustration (FR) (Hart, 2006, p. 904). For the evaluation of the online questionnaire, the mean values of the RTLX questions on perceived mental workload were rescaled to correspond to the official scale range of 0-100 so that comparable values from Hertzum (2021)'s meta-analytical review of TLX values could also be used (Hertzum, 2021, p. 872). Detailed data for each participant are available in the appendix (see table A.8).

After calculating the mean values for each subscale, we have the possibility of comparing these values to the TLX values of "*Office work*", "*Desktop application*", and the "*Region*" value for "*Europe*" which are part of the meta-analytical review by Hertzum (2021) (Hertzum, 2021, pp. 874–875). To briefly recap the RTLX subscales, the survey questions are listed again in this context (Hertzum, 2021, p. 870) (see table 5.7).

Mental demand (MD): 'How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?'
Physical demand (PD): 'How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?'
Temporal demand (TD): 'How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?'
Effort (EF): 'How hard did you have to work (mentally and physically) to accomplish your level of performance?'
Performance (PE): 'How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?'
Frustration (FR): 'How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?'

Table 5.7.: RTLX subscales. Adapted from Hertzum (2021, p. 870).

As shown in table 5.7, only *Physical Demand (PD)* falls outside the range of standard-deviation values observed in the three comparison categories.

	MD	PD	TD	EF	PE	FR	TLX
AK tax law chatbot	**52.1**	**12.5**	**29.6**	**44.6**	**50**	**39.6**	**38.1**
Office work	58 ± 22	33 ± 12	51 ± 17	53 ± 17	51 ± 21	37 ± 12	47 ± 15
Desktop application	52 ± 19	28 ± 15	44 ± 18	51 ± 17	43 ± 19	37 ± 16	43 ± 14
Europe	46 ± 16	31 ± 15	39 ± 15	46 ± 15	45 ± 20	34 ± 13	40 ± 11

Table 5.8.: TLX value comparison. Adapted from Hertzum (2021, pp. 874–875).

The numbers shown in blue in table 5.9 are the values closest to those collected for the AK-chatbot (Hertzum, 2021, p. 873). When observing the percentile values in table 5.9 notably low values can be identified in the dimensions of *Physical Demand (PD)* and *Temporal Demand (TD)*. This could be attributed to the study design: the tasks were not conducted under time pressure, and participants interacted with the chatbot exclusively via their PCs, so no additional physical effort was required.

Values slightly below the median were recorded in the *Effort (EF)* category, suggesting that participants did not need to exert a high level of mental effort to achieve their performance level.

The values for *Mental Demand (MD)*, *Performance (PE)*, and *Frustration (FR)* were higher than all other categories. These three dimensions lie in the 60th percentile. They are, therefore, above the median provided by Hertzum (2021). The slightly elevated value of *Mental Demand (MD)* suggests that participants needed to invest more mental energy to use the chatbot and complete the task.

Furthermore, the reverse-coded *Performance (PE)* value at the 60th percentile indicates that participants perceived their performance as slightly above the median, suggesting reduced perceived success in achieving the task goals. Participants had to rate their own performance on a scale from *Good* to *Poor*.

Lastly, the *Frustration (FR)* value, which also lies in the 60th percentile, indicates that participants felt more insecure, discouraged, irritated, stressed, or annoyed compared to the median value of the reference data reported by Hertzum (2021).

	MD	PD	TD	EF	PE	FR	TLX
10th percentile	26	**13**	22	27	21	18	26
20th percentile	36	18	**29**	36	28	24	32
30th percentile	41	22	34	42	34	29	36
40th percentile	45	27	38	**47**	38	32	**40**
50th percentile (median)	49	30	42	51	42	36	43
60th percentile	**54**	34	46	55	**47**	**39**	46
70th percentile	59	38	50	59	53	44	49
80th percentile	64	45	57	63	63	47	52
90th percentile	72	52	65	70	73	55	57
Mean ± SD	49 ± 17	32 ± 16	42 ± 16	50 ± 16	45 ± 19	36 ± 14	42 ± 13

Table 5.9.: TLX percentile values. Adapted from Hertzum (2021, p. 873).

6. Discussion and Conclusion

In this thesis, an embedded mixed-methods design was used to assess perceived mental workload (Hart, 2006), chatbot usability (Holmes et al., 2019), and perceived trustworthiness. In addition, possible improvements to the chatbot and users' information needs were analyzed, and a German translation of the XAI Novice Question Bank was created to explore participants' information needs (Schmude et al., 2025).

Quantitative survey data were collected through a questionnaire after the human-chatbot interaction, while qualitative data were gathered in a semi-structured interview (Creswell & Clar, 2017, p. 103). Each participant completed two real-world legal advice tasks using the AK-chatbot. Pre-existing questionnaires were used to capture perceived usability (Holmes et al., 2019) and perceived mental workload (Hart, 2006). Additional survey questions were used to gather information about the study participants' attitudes toward the chatbot and AI, as well as their perceptions of its trustworthiness. Lastly, the experts' prior knowledge and information needs were collected using the XAI Novice Question Bank developed by Schmude et al. (2025).

The aforementioned components were combined to form the Chatbot Evaluation Framework (CEF) to evaluate the AK-chatbot. The CEF therefore offers the opportunity to gain a broader understanding of how users perceive a chatbot's usability, trustworthiness, and mental workload. Additionally, existing information needs can also be identified.

Previous work focused on assessing the quality of the output of generative AI systems used in the legal domain (Qiu et al., 2024), the evaluation of a RAG pipeline (Kim & Lee, 2025), the general implications of AI chatbots in the legal sector (R et al., 2025), the creation of legal chatbots (Stephens, 2025) or legal documents query applications (Ngo et al., 2025), or the testing of prototypes (Panda et al., 2025). In comparison, this thesis examines the perceptions of legal experts in a human-chatbot interaction scenario, employing a dedicated survey to elicit users' trust in the chatbot, semi-structured interviews to gain insight into possible interaction improvements, an assessment of perceived mental workload, and a questionnaire to capture the chatbot's usability (Holmes et al., 2019).

This master's thesis can therefore advance the evaluation of legal chatbots by collecting reference values using the Chatbot Usability Questionnaire (Holmes et al., 2019) and the RTLX scale (Hart, 2006), and by providing qualitative data on the dimensions of trust and usability in a real-world scenario for a legal chatbot. Furthermore, it introduces the reusable CEF for future chatbot evaluations.

This thesis answers the following research questions:

- **RQ 1**: How do tax-law experts perceive and experience the usability of a Retrieval Augmented Generation (RAG) chatbot when completing routine tax law advisory tasks?

  - **RQ 1.1**: How effectively does the tax law chatbot address tax experts' needs in terms of usability and workflow integration?
  - **RQ 1.2**: What adjustments are needed to make legal advice chatbots improve usability and integration into work routine in practice?

- **RQ 2**: Which characteristics of a domain-specific chatbot shape its perceived trustworthiness among users?

- **RQ 3**: How do tax law experts tend to use the tax law chatbot when solving work tasks?
  - **RQ 3.1**: Can we identify patterns of interaction during the conversation with a legal advice chatbot in real-world use?

6.1. Key findings

6.1.1. New reference values

The present work provides reference values for perceived mental workload and chatbot usability for a tax law chatbot. Based on real-world tasks completed by tax law experts, these values can be used in future evaluations of similar legal chatbots. The mean score on the Chatbot Usability Questionnaire was 63.7 ± 15.4, indicating *average* or *poor* perceived usability among participants (Holmes et al., 2023, 336–338).

Overall, the RTLX (Raw Task Load Index) results do not indicate a highly elevated perceived mental workload when using the chatbot. Low *Physical Demand (PD)* and *Temporal Demand (TD)* values reflect that the tasks were completed with little time pressure and did not require elevated physical effort. The values for *Physical Demand (PD)* and *Temporal Demand (TD)* may be attributed to the study design, as no time restrictions were imposed on the human-chatbot interaction and the study participants did not perform any strenuous physical activity. The interaction was conducted and recorded in a familiar work setting via video conference. *Effort (EF)* scores slightly below the median indicate that participants did not need to invest a high level of mental or physical effort to achieve their level of performance.

In contrast, *Mental Demand (MD)*, *Performance (PE)*, and *Frustration (FR)* were all around the 60th percentile. This indicates that tax law experts experienced higher cognitive demands, felt less successful in completing tasks, and reported more frustration than the median values reported by Hertzum (2021) (see table 5.9). Regarding increased mental demand, the results may be linked to the use of a new tool, the AK-chatbot. Some study participants stated that they had not yet worked with the chatbot very often. The increased frustration could be attributed to participants' dissatisfaction with the chatbot's responses, as noted in the interviews.

6.1.2. Usability

Useful for simple tasks, but unable to meet experts' needs for complex legal reasoning. In this study, the term "complex tasks" refers to examples of unsuitable areas of application for the chatbot, such as *"responding to complex inquiries"*, complex queries, and *"writing complex texts"* (see table 5.3). Whereby the *"writing of complex texts"* refers to the preparation of legal opinions (P2) and texts on the political position of the AK (P1). Furthermore, with regard to the example of *"responding to complex inquiries"*, participants (P3, P6, P9) refer to the chatbot's ability to link multiple topics together in the sense of legal conclusions and the ability to solve tasks which require a detailed understanding of the circumstances and where the ability to provide accurate information depends on specific legal details (P9, P12), such as whether *"clients are entitled to certain government subsidies, and how will such subsidies affect the next steps"* (P9).

In the context of this thesis, the second task can be used as a suitable example of a "complex task" (see table 4.8). Task 2 deliberately combined employment tax liability and child alimony. These issues overlap in cross-border aspects, timing, and interdependent factors that influence the tax calculation. Changing one aspect of the task would automatically affect other aspects as well. These characteristics explain why the study task can be considered "complex". Task 1, on the other hand, can be classified as "simple" because it deals with only a single topic that does not require numerous links to other areas (see table 4.5).

When rating the helpfulness of the chatbot output on a 5-point Likert scale, experts evaluated it differently depending on task complexity. For simple tasks, the chatbot's answers received higher ratings of perceived helpfulness. In contrast, when working on more complex tasks, the ratings were lower. This suggests that the chatbot is seen as useful for less complex tasks.

Furthermore, experts identified clear value in low-risk tasks such as drafting emails, translating, and doing research, but remained skeptical of the system's ability to handle complex tasks such as *"responding to complex inquiries"* and *"writing complex texts"*, or they would not recommend *"direct interaction with clients"*, because a tax-law expert should first review the answers and act as an *"expert in the loop"* between the generative AI system and clients. In addition, the participants (P2, P3, P4, P5, P8, P11, P13, P14) highlighted the need for prior legal knowledge to correctly interpret the generated output.

Struggle for complex legal reasoning Although LLMs might pass certain law exams (Choi et al., 2022; Martínez, 2024), the use of RAG systems for solving highly specific legal tasks seems to be in need of new technical solutions (Donabauer & Kruschwitz, 2025; Liu et al., 2025; Padhye, 2024; Su et al., 2025; Yao et al., 2025) or design approaches (Guan et al., 2025) to improve their capabilities to be used more effectively. The data from the semi-structured interviews in this study confirm the assumption that expert-level legal reasoning remains a challenge for an RAG system.

As discussed by Cheong et al. (2024), participants in this present study also identified context-awareness and accuracy as reasons for doubting the chatbot's capabilities. The participants requested increasing the precision of the output by enhancing the chatbot's

reasoning capabilities, particularly its ability to connect multiple legal topics when dealing with complex tasks (P2, P4, P8, P9, P11). In addition, the AK-chatbot 's vague and incomplete responses were a major point of criticism. Two participants specifically emphasized the need for more precise language and less verbose responses (P5, P6), reflecting a focus on output precision.

When evaluating the qualitative data from the semi-structured interviews, participants' negative statements primarily addressed two main topics (see table 5.5). These statements referred to participants' lack of trust in the AK-chatbot. Furthermore, 6 out of 14 participants criticized the incompleteness and vagueness of the generated responses.

As in Magesh et al. (2025), participants, who are themselves legal experts, also noted the perceived inability to draw conclusions like a human legal expert. This criticism could also be seen as a reason for the desire for better legal reasoning capabilities. The AK-chatbot did not seem to fully meet the needs of an expert-level legal reasoning tool, but it was more of a tool to speed up routine work.

One possibility to enhance its capabilities is to let a language model use specialized tools, which could further advance its abilities (Schick et al., 2023).

Efficient document linking and an up-to-date knowledge base. In direct reference to research question RQ 1.2 (*"What adjustments are needed to make legal advice chatbots improve usability and integration into work routine in practice?"*), several possibilities were identified to improve the usability of the chatbot. The linking of documents used as sources for the generated responses received mixed feedback from study participants. On the one hand, 7 out of 14 participants were satisfied with the linking, while on the other hand, 6 out of 14 participants criticized the incorrect implementation of citation marking in the sources at the time the study was conducted. The citations were not displayed in the linked documents, and only the entire document was downloaded, requiring participants to search for the citations again. Some linked documents, such as entire legal texts, contained up to hundreds of pages.

Another important point identified was a well-maintained knowledge base. P5 vividly explained why a knowledge base should always be up to date:

> *One very important point (...) is that it must be up to date. It is essential that this is the case, especially in the area of tax law .... We are now in our 75th edition, which means that since 1955, there have been 75 editions, i.e., a new edition every year. This is understandable, as tax law is the most reformed area of law in Austria (...).*

During the interviews, participants called for a curated and domain-relevant knowledge base and deeper integration with legal databases. Documents that may become part of this knowledge base in the future were collected in table A.15. Furthermore, providing up-to-date legal data to experts could be possible using an API. The Austrian legal information system (RIS) is the official online portal that publishes legally binding federal and state legislation, as well as certain district-level regulations. It also gives public access to consolidated laws, court decisions, selected municipal bylaws, ministerial directives,

and other official announcements (Bundeskanzleramt der Republik Österreich, 2025a). An API to use RIS data is provided by Cooperation Open Government Data Österreich (Cooperation OGD Österreich) (Cooperation Open Government Data Österreich, 2025) and a RIS API wrapper is available free of charge (Thumfart, 2022). The integration of the RIS database could also be an opportunity to be prepared for future changes in the legal system and perhaps also to be able to compete with the previously mentioned *Genjus KI* (MANZ'sche Verlags- und Universitätsbuchhandlung GmbH, 2024).

An important question in this regard is how the AK plans to address future changes. Is there a plan for when and how the knowledge base should be updated? How will regulations and laws from the past be presented by the chatbot in the future?

6.1.3. Trust

Positive attitudes toward AI and future use, despite trust issues. In both quantitative and qualitative data, participants reported generally positive attitudes toward AI and mixed feelings toward the tax law chatbot, while expressing low trust in its generated output. In the survey, 9 out of 14 participants *agreed* or *strongly agreed* that they felt comfortable integrating artificial intelligence into their work, and 10 out of 14 participants *disagreed* or *strongly disagreed* with the statement *"I generally have negative feelings towards artificial intelligence"*.

Overall, the participants tended to rate the chatbot as helpful; only 3 out of 14 participants *(strongly) disagreed*. In addition, 10 out of 14 participants indicated that the chatbot's answers gave the impression of being correct (see also figure 5.7), but trust remained low, with 10 of 14 participants *disagreeing* or *strongly disagreeing* with the statement, *"I trust the chatbot's answers"*. P13 articulated the lack of trust in the AK-chatbot as follows:

> *(...) he comes across as very credible, he's very good at convincing people. That's the problem, and then he twists things, but as I said, legal matters are always a very fine line between what is formulated and what is then allowed or not allowed, so that's just the way it is, and he presents it so credibly that he is right, and I just know from a technical point of view that he is partly wrong. And that makes me feel uncertain. And that's why I don't believe him.* (P13)

In the interviews, only two participants (P3, P9) explicitly stated trust in the chatbot for legal decisions, but they also criticized its lack of depth and said they would still double-check its answers, while several reported little trust (P2, P4, P8, P12, P13) or no trust at all (P1, P5, P11, P14); three participants emphasized that they would check the answers and did not express trust in the chatbot (P6, P7, P10). When estimating the risk of incorrect information, participants provided varied assessments, ranging from *"less than 50%"* (P7) and *"50%"* (P12) to *"more than 50%"* (P11), *"70–80%"* (P8), and *"80%"* (P5, P13), while others described the risk as moderate or difficult to judge (P2, P4, P6, P14). Furthermore, one participant (P1) reported not having noticed any incorrect decisions made by the chatbot so far, yet emphasized that this did not translate into blind trust.

Another participant (P10) explained that a lack of trust was partly due to potentially missing data, while P9 similarly acknowledged the possibility of incorrect information but pointed to recent improvements and limited personal experience with the chatbot.

In addition, enjoyment of use was limited: 6 of 14 participants *disagreed* or *strongly disagreed* with the statement *"I enjoy working with the chatbot"*, 7 participants were neutral, and only 1 *agreed*.

Tax law experts thus become humans-in-the-loop, with tasks such as examining and correcting the system's output and acting as connectors between the system and other humans (Crootof et al., 2023, pp. 440–441).

Although the majority of participants reported low trust in the chatbot's output, they were highly willing to integrate the tool into daily workflows. In fact, 13 out of 14 participants expressed their willingness to use the AK-chatbot in the future. The strong willingness to use AI at work seems to reflect the current mood toward AI in everyday work life. In a survey of Austrian professionals, 69% expected efficiency gains and relief from time-consuming or tedious tasks through the use of AI (PwC Österreich, 2025).

Experts' information needs focus on chatbot specifications and limitations. The analysis of the XAI Novice Question Bank reveals a clear pattern in participants' information needs when interacting with the chatbot. Overall, the results indicate that participants were less concerned with general contextual or usage-related aspects (*Chatbot Usage* or *Chatbot Context* category) and instead focused strongly on understanding the chatbot's internal functioning and limitations. In total, 35 of the 73 selected questions belonged to the *Chatbot Specifications* category. In particular, the most frequently chosen questions addressed how the chatbot learns, the logic underlying its responses, the data used for training, and its limitations.

Fighting trust issues. Meeting experts' information needs might help. Following Shin (2021)'s way of thinking,

> Non-expert end users do not know how the exact cascades of algorithmic code resulted in a particular decision. The issues related to AI decision-making being a "black-box" are significantly worsened when dealing with ordinary users who lack technical knowledge and are still required to interact with AI systems. (Shin, 2021, p. 2)

A way to make the RAG system appear more "trustworthy" and thus meet stakeholder expectations, namely those of AK's tax law experts, following Bisconti et al. (2025)'s definition of trustworthiness, would be to address participants' information needs and reduce potential information gaps regarding the chatbot's decision-making process.

By using the XAI Novice Question Bank as a feedback tool to record study participants' information needs, these information gaps were identified. The dominance of *Chatbot Specifications* questions indicates that expert users prioritize understanding *how* and *why* the tax law chatbot produces answers, what data it relies on, and where it fails (see section 4.2.3). Rather than thinking about *Chatbot Context* or *Chatbot Usage*, the tax law

experts try to understand how the internal processes of the chatbot work. Hence, one way to increase the perceived trustworthiness of the chatbot could be targeted informational measures that foster a deeper understanding of *how* and *why* the tax law chatbot produces certain answers.

6.2. Limitations

Various limitations have to be considered. First, this thesis examines a single tax-law chatbot within a single organization. Therefore, the findings are highly context-specific and not readily generalizable to other legal domains or legal advice chatbots. The chatbot was developed to support work on Austrian tax law.

Second, the study does not test answer accuracy against a question-and-answer (QA) evaluation dataset (Brehme et al., 2025, p. 16). Therefore, the study cannot quantify correctness or relate perceived trustworthiness and usability to the RAG system's accuracy. Embedding-based metrics, short-answer evaluations, and "LLM as a Judge" approaches were not used because the focus was on the tax law experts' perceptions (Brehme et al., 2025, pp. 19–21).

Third, the sample was small: 14 tax law experts participated in the study, reflecting the limited size of the department (approximately 30–40 experts work in the tax law departments of the AK). As a result, the study emphasizes qualitative insights into usage, perceived trustworthiness, and workflow integration rather than broad benchmarking. The results highlight the practical strengths and weaknesses of this chatbot.

In addition, the infrequent use of chatbots, both in private life and at work, may have influenced the results regarding the existing information needs (see section 5.2). 6 out of 14 participants mentioned their lack of contact with chatbots in private settings and the AK-chatbot was not a tool which was regularly used by tax law experts at work. Although the chatbot version of the XAI Novice Question Bank has provided initial insights for future measures to meet information needs, it would be useful to use this tool multiple times to identify possible changes in users' information needs.

6.3. Future Work

In the future, it would be useful to collect additional comparative values for the RTLX scale regarding chatbots, as Hertzum (2021)'s study provides only values for desktop applications in general. Comparing interaction with conventional software and with chatbots using the NASA-TLX, as done by Schmidhuber et al. (2021), could be a promising approach in legal contexts to analyze whether legal experts prefer conducting research with a chatbot or conventional software.

It would also make sense to conduct additional studies using the Chatbot Usability Questionnaire developed by Holmes et al. (2019) to establish a benchmark for chatbot usability and contribute to the field of chatbot evaluation.

To gain deeper insights, a follow-up study evaluating a larger number of chatbot responses would be useful. Additional data on the accuracy of chatbot responses could be

collected using a question-and-answer evaluation dataset or by focusing on the assessment of the technical correctness of the output by legal experts (Brehme et al., 2025, p. 16).

Future work should extend the approach to multiple chatbots and further establish the Chatbot Usability Questionnaire, thereby providing benchmarks for evaluating chatbots.

With regard to research question *RQ 3.1: Can we identify patterns of interaction during the conversation with a legal advice chatbot in real-world use?*, it would be possible to analyze the exact usage patterns, e.g., prompting strategies, of users while performing their activities as legal advisors. This study focused on potential areas of application and did not implement the approach due to resource constraints. It would, of course, be interesting to analyze not only two different tasks, as in this study, but also longer human-chatbot interactions. A future study could examine which prompts yield the best results for legal tasks. This study could focus on legal tasks in which, from a legal perspective, it is necessary to link different topics. In this thesis, tax law experts identified shortcomings in the chatbot's output, which failed to sufficiently address interconnected legal issues.

On a more general level, research could also examine whether it would be possible to build a more effective RAG system for legal reasoning without the need for complex LLM fine-tuning. Would it be possible to adjust the "thought process" of a RAG system to the demands of complex legal task solving? Is it possible to implement a "lawyer's mind" or "lawyer's way of thinking"? Overall, as already criticized by Magesh et al. (2025), there is a lack of public benchmarks for evaluating legal chatbots.

6.4. Conclusion

This master's thesis investigated how legal experts experience interaction with a RAG-based chatbot when performing real-world tax counseling work. The goal was to understand users' perceived trust, usability, and mental load regarding routine tax law tasks.

Methodologically, the thesis followed an embedded mixed-methods research design and developed the Chatbot Evaluation Framework (CEF) to structure the evaluation.

The results suggest that the AK-chatbot could be rated as "*average*" in comparison to available reference values of the Chatbot Usability Questionnaire (RQ1) (see table A.12). Perceived mental workload, captured by RTLX, was not high compared to the reference values provided by Hertzum (2021). In other words, using the chatbot did not appear to impose a substantial additional burden (RQ1). Further, the chatbot appears to have difficulty with complex queries (RQ1.1), and there is a need to improve its complex legal reasoning capabilities and to build an up-to-date knowledge base (RQ1.2).

During the interviews, tax law experts noted that the vagueness and incompleteness contributed to their lack of trust in the chatbot (RQ2). Although perceived trustworthiness was low, a high level of willingness to use the chatbot in the future was observed. To increase trust in the chatbot, AK's decision-makers could use the results of the selected questions from the XAI Novice Question Bank to inform further measures aimed at meeting the information needs of chatbot users.

As part of the study, it was also possible to identify suitable and unsuitable areas of application for the chatbot in the everyday work of AK's tax law experts (RQ3, RQ3.1). At the time of the study, the AK-chatbot was considered suitable for simple work tasks and problematic for solving complex legal advisory tasks that would require linking different legal aspects to answer the inquiry.

This thesis enabled the generation of reference values for subsequent chatbot evaluations. Additionally, suitable areas of application and opportunities for improvement were documented. It was also possible to determine participants' existing need for information about the next steps and the tax law experts' attitudes toward the chatbot.

Future work could examine the accuracy of chatbot responses or compare different legal chatbots for evaluation. Furthermore, a follow-up study could track the chatbot's development or analyze whether there has been any change in perceived trustworthiness or perceived usability.

The results raise several questions that warrant further research. In particular, the best way to assess the actual accuracy of generated responses in a legal context should be determined, and how verified accuracy relates to perceived trustworthiness in practice should be investigated

Bibliography

Abkommen zwischen der Republik Österreich und der Bundesrepublik Deutschland zur Vermeidung der Doppelbesteuerung auf dem Gebiet der Steuern vom Einkommen und vom Vermögen und zur Verhinderung der Steuerverkürzung und -umgehung (Bundesrecht konsolidiert, Fassung vom 18. August 2025). (2025). Retrieved August 18, 2025, from https://ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20002157

Aichinger, G., Aufreiter, A., Nessler, B., & Wendehorst, C. (2024). *KI-Kompetenz im Sinne des Art 4 KI-VO* (tech. rep.). Software Competence Center Hagenberg (SCCH). https://www.scch.at/files/userdata/import/News/SCCH_KI_Kompetenz_AI_Art_4_KI-VO.pdf

Arbeiterkammer Oberösterreich. (2025). *Unterhaltsabsetzbetrag und Unterhaltsleistungen*. Retrieved January 5, 2025, from https://ooe.arbeiterkammer.at/beratung/steuerundeinkommen/arbeitnehmerveranlagung_schrittfuerschritt/aussergewoehnlichebelastungen/Unterhaltsabsetzbetrag_und_Unterhaltsleistungen.html

Arbeiterkammer Wien. (2026). *17 Antworten zur AK und ihrer Wahl*. Retrieved January 16, 2026, from https://www.arbeiterkammer.at/ueberuns/akerklaertsich/deu/17-Antworten.html

Arena. (2025). *Compare & Benchmark the Best Frontier AI Models*. Retrieved February 21, 2026, from https://arena.ai/about

ATLAS.ti Scientific Software Development GmbH. (2025). *ATLAS.ti | Die #1 Software für qualitative Datenanalysen - ATLAS.ti*. Retrieved October 31, 2025, from https://atlasti.com/de

Aufreiter, A., Aichinger, G., & Nessler, B. (2024). In scope? – the definition of 'AI System' in the AI Act. *Proceedings of Austrian Symposium on AI, Robotics, and Vision 2024 (AIROV24)*. https://doi.org/10.15203/99106-150-2-04

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An Empirical Evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction*, *24*(6), 574–594. https://doi.org/10.1080/10447310802205776

Baron, S. (2025). Trust, Explainability and AI. *Philosophy and Technology*, *38*(1). https://doi.org/10.1007/s13347-024-00837-6

Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., & Hochreiter, S. (2024). xLSTM: Extended Long Short-Term Memory. https://arxiv.org/abs/2405.04517

Bisconti, P., Aquilino, L., Marchetti, A., & Nardi, D. (2025). A Formal Account of AI Trustworthiness: Connecting Intrinsic and Perceived Trustworthiness. In *Proceed-*

ings of the 2024 aaai/acm conference on ai, ethics, and society (pp. 131–140). AAAI Press.

Borsci, S., Malizia, A., Schmettow, M., van der Velde, F., Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (2022). The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents. *Personal and Ubiquitous Computing*, *26*(1), 95–119.

Boyd, K., Potts, C., Bond, R., Mulvenna, M., Broderick, T., Burns, C., Bickerdike, A., Mctear, M., Kostenius, C., Vakaloudis, A., Dhanapala, I., Ennis, E., & Booth, F. (2022). Usability testing and trust analysis of a mental health and wellbeing chatbot. *Proceedings of the 33rd European Conference on Cognitive Ergonomics*. https://doi.org/10.1145/3552327.3552348

Brandt, M. (2023). Infografik: Wie lange brauchen Online-Dienste, um eine Million Menschen zu erreichen — de.statista.com. https://de.statista.com/infografik/29195/zeitraum-den-online-dienste-gebraucht-haben-um-eine-million-nutzer-zu-erreichen/

Braun, V., & Clarke, V. (2022). *Thematic analysis : a practical guide* (1st edition). SAGE Publications Ltd.

Brehme, L., Ströhle, T., & Breu, R. (2025). Can LLMs be Trusted for Evaluating RAG Systems? A Survey of Methods and Datasets. *2025 IEEE Swiss Conference on Data Science (SDS)*, 16–23. https://doi.org/10.1109/sds66131.2025.00010

Brooke, J. (1996). *SUS: A quick and dirty usability scale* (tech. rep.). https://www.researchgate.net/publication/228593520

Bundesarbeitskammer. (2025). *Steuer & Geld: Ein Leitfaden für die Arbeitnehmer:innenveranlagung*. Retrieved October 27, 2025, from https://www.arbeiterkammer.at/service/broschueren/SteuerundGeld/index.html

Bundeskanzleramt der Republik Österreich. (2025a). Rechtsinformationssystem des Bundes RIS. https://www.ris.bka.gv.at/

Bundeskanzleramt der Republik Österreich. (2025b). *RIS - Einkommensteuergesetz 1988 - Bundesrecht konsolidiert, Fassung vom 24.04.2025*. Retrieved April 24, 2025, from https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10004570

Bundesministerium für Bildung, Wissenschaft und Forschung. (n.d.). *Künstliche Intelligenz – Chance für Österreichs Schulen*. Retrieved January 10, 2025, from https://www.bmbwf.gv.at/Themen/schule/zrp/ki.html

Bundesministerium für Finanzen. (2025a). Finanzdokumentation (Findok). https://findok.bmf.gv.at/findok/

Bundesministerium für Finanzen. (2025b). *Liste der Österreichischen Doppelbesteuerungsabkommen*. Retrieved October 27, 2025, from https://www.bmf.gv.at/themen/steuern/internationales-steuerrecht/doppelbesteuerungsabkommen/dba-liste.html

Buttrick, N. (2024). Studying large language models as compression algorithms for human culture. *Trends in Cognitive Sciences*, *28*(3), 187–189. https://doi.org/https://doi.org/10.1016/j.tics.2024.01.001

Cabrero-Daniel, B., & Sanagustín Cabrero, A. (2023). Perceived Trustworthiness of Natural Language Generators. *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*. https://doi.org/10.1145/3597512.3599715

Chen, X., He, B., Lin, H., Han, X., Wang, T., Cao, B., Sun, L., & Sun, Y. (2024). Spiral of Silence: How is Large Language Model Killing Information Retrieval? – A Case Study on Open Domain Question Answering. https://arxiv.org/abs/2404.10496

Cheong, I., Xia, K., Feng, K. J. K., Chen, Q. Z., & Zhang, A. X. (2024). (A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2454–2469. https://doi.org/10.1145/3630106.3659048

Chiang, C.-H., Chen, W.-C., Kuan, C.-Y., Yang, C., & Lee, H.-y. (2024). Large Language Model as an Assignment Evaluator: Insights, Feedback, and Challenges in a 1000+ Student Course. http://arxiv.org/abs/2407.05216

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M. I., Gonzalez, J. E., & Stoica, I. (2024). Chatbot arena: an open platform for evaluating LLMs by human preference. *Proceedings of the 41st International Conference on Machine Learning*. https://raw.githubusercontent.com/mlresearch/v235/main/assets/chiang24b/chiang24b.pdf

Choi, J. H., Hickman, K. E., Monahan, A. B., & Schwarcz, D. (2022). ChatGPT Goes to Law School. *Journal of Legal Education*, *71*(3), pp. 387–400. https://www.jstor.org/stable/27302033

Collins, K. M., Jiang, A. Q., Frieder, S., Wong, L., Zilka, M., Bhatt, U., Lukasiewicz, T., Wu, Y., Tenenbaum, J. B., Hart, W., Gowers, T., Li, W., Weller, A., & Jamnik, M. (2024a). Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences*, *121*(24), e2318124121. https://doi.org/10.1073/pnas.2318124121

Collins, K. M., Jiang, A. Q., Frieder, S., Wong, L., Zilka, M., Bhatt, U., Lukasiewicz, T., Wu, Y., Tenenbaum, J. B., Hart, W., Gowers, T., Li, W., Weller, A., & Jamnik, M. (2024b). Supporting Information for "Evaluating language models for mathematics through interactions". *Proceedings of the National Academy of Sciences*, *121*(24), e2318124121. https://doi.org/10.1073/pnas.2318124121

Cooperation Open Government Data Österreich. (2025). RIS Daten Version 2.6. https://www.data.gv.at/katalog/datasets/0fb9ae1a-92cb-4ab8-a589-470c16d4fe21

Creswell, J. W., & Clar, V. L. P. (2017). Designing and Conducting Mixed Methods Research. In *Designing and conducting mixed methods research*. SAGE Publications.

Crootof, R., Kaminski, M. E., & Ii, W. N. P. (2023). Humans in the Loop. *VANDERBILT LAW REVIEW*, *76*. https://scholarship.richmond.edu/cgi/viewcontent.cgi?article=2659&context=law-faculty-publications

Dai, S., Xu, C., Xu, S., Pang, L., Dong, Z., & Xu, J. (2024). Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6437–6447. https://doi.org/10.1145/3637528.3671458

de Wynter, A., Wang, X., Gu, Q., & Chen, S.-Q. (2024). On Meta-Prompting. https://arxiv.org/abs/2312.06562

DeepL. (2025). *DeepL Übersetzer: Der präziseste Übersetzer der Welt*. Retrieved October 31, 2025, from https://www.deepl.com/de/translator

Donabauer, G., & Kruschwitz, U. (2025). A Reproducibility Study of Graph-Based Legal Case Retrieval. *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3135–3144. https://doi.org/10.1145/3726302.3730282

Doyle, C., & Tucker, A. D. (2025). If You Give an LLM a Legal Practice Guide. *Proceedings of the 2025 Symposium on Computer Science and Law*, 194–205. https://doi.org/10.1145/3709025.3712220

Essop, L., Singh, A., & Wing, J. (2023). Developing a comprehensive evaluation questionnaire for university FAQ administration chatbots. *2023 Conference on Information Communications Technology and Society (ICTAS)*, 1–7. https://doi.org/10.1109/ICTAS56421.2023.10082753

European Law Institute. (2024). Response to the Commission Guidelines on the Application of the Definition of an AI System and the Prohibited AI Practices Established in the AI Act. https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ELI_Response_on_the_definition_of_an_AI_System.pdf

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., & Li, Q. (2024). A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6491–6501. https://doi.org/10.1145/3637528.3671470

Fazi, L., Zaniboni, S., & Wang, M. (2025). Age differences in the adoption of technology at work: a review and recommendations for managerial practice. *Journal of Organizational Change Management*, *38*(8), 138–175. https://doi.org/https://doi.org/10.1108/JOCM-12-2024-0767

Ferrara, E. (2024). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, *6*(1). https://doi.org/10.3390/sci6010003

Forte, J. (2024, July). *Introduction to GPT-4o and GPT-4o mini*. OpenAI. Retrieved April 20, 2025, from https://cookbook.openai.com/examples/gpt4o/introduction_to_gpt4o

Fuchs, M. (2025, September). *Überwachung per KI: „Kann Arbeit zur Hölle machen"*. Krone.at. Retrieved October 28, 2025, from https://www.krone.at/3894496

Fülop, T. (2025). *KI-Kompetenzen in der Arbeitswelt: Warum, was, wieviel, wie?* (Information Paper). Arbeiterkammer Wien, Büro für digitale Agenden. https://wien.arbeiterkammer.at/interessenvertretung/arbeitdigital/stellungnahmen/AK-Info-KI_Kompetenz.pdf

Galy, E., Paxion, J., & Berthelon, C. (2018). Measuring mental workload with the NASA-TLX needs to examine each dimension rather than relying on the global score: an example with driving [PMID: 28817353]. *Ergonomics*, *61*(4), 517–527. https://doi.org/10.1080/00140139.2017.1369583

Gambetta, Z. A., Dessi Puji, L., & Ginar Santika, N. (2021). Calla Beauty Assistant: Beauty Advisory Chatbot. *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 1–6. https://doi.org/10.1109/ICAICTA53211.2021.9640281

Gibney, E. (2025). China's cheap, open AI model DeepSeek thrills scientists. *Nature (London)*, *638*(8049), 13–14.

Guan, J., Zou, R., Zhang, J., Xin, K., He, B., Zhang, Z., & Ye, C. (2025). Designing Human-AI System for Legal Research: A Case Study of Precedent Search in Chinese Law. *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3706599.3720167

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Henkel, O., Hills, L., Boxer, A., Roberts, B., & Levonian, Z. (2024). Can Large Language Models Make the Grade? An Empirical Study Evaluating LLMs Ability To Mark Short Answer Questions in K-12 Education. *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, 300–304. https://doi.org/10.1145/3657604.3664693

Hertzum, M. (2021). Reference values and subscale patterns for the task load index (TLX): a meta-analytic review [PMID: 33463402]. *Ergonomics*, *64*(7), 869–878. https://doi.org/10.1080/00140139.2021.1876927

Hertzum, M. (2022). Associations among workload dimensions, performance, and situational characteristics: a meta-analytic review of the Task Load Index. *Behaviour & Information Technology*, *41*(16), 3506–3518. https://doi.org/10.1080/0144929X.2021.2000642

Hewett, T. T., Baecker, R., Card, S., Carey, T., Gasen, J., Mantei, M., Perlman, G., Strong, G., & Verplank, W. (1992). *ACM SIGCHI Curricula for Human-Computer Interaction* (tech. rep.). New York, NY, USA, Association for Computing Machinery. https://dl.acm.org/doi/book/10.1145/2594128

Hindi, M., Mohammed, L., Maaz, O., & Alwarafy, A. (2025). Enhancing the Precision and Interpretability of Retrieval-Augmented Generation (RAG) in Legal Technology: A Survey. *IEEE Access*, *13*, 46171–46189. https://doi.org/10.1109/ACCESS.2025.3550145

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Holmes, S., Bond, R., Moorhead, A., Zheng, J., Coates, V., & McTear, M. (2023). Towards Validating a Chatbot Usability Scale. In A. Marcus, E. Rosenzweig & M. M. Soares (Eds.), *Design, user experience, and usability* (pp. 321–339). Springer Nature Switzerland.

Holmes, S., Moorhead, A., Bond, R., Zheng, H., Coates, V., & Mctear, M. (2019). Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? *Proceedings of the 31st European Conference on Cognitive Ergonomics*, 207–214. https://doi.org/10.1145/3335082.3335094

Hugging Face. (2025). openai/whisper-large-v3-turbo. https://huggingface.co/openai/whisper-large-v3-turbo

Ibrahim, L., Huang, S., Ahmad, L., & Anderljung, M. (2024). Beyond static AI evaluations: advancing human interaction evaluations for LLM harms and risks. https://arxiv.org/abs/2405.10632

Kammer für Arbeiter und Angestellte für Wien. (2025, January). *Steuer sparen 2025*. Retrieved October 27, 2025, from https://www.arbeiterkammer.at/service/broschueren/SteuerundGeld/Steuer-sparen-2025.html

Kemmerzell, N., Schreiner, A., Khalid, H., Schalk, M., & Bordoli, L. (2025). Towards a Better Understanding of Evaluating Trustworthiness in AI Systems. *ACM Comput. Surv.*, *57*(9). https://doi.org/10.1145/3721976

Kim, Y., & Lee, W. (2025). Where Does Legal AI Fail? Evaluating RAG Pipelines. *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 1396–1405. https://doi.org/10.1145/3746252.3761151

Langevin, R., Lordon, R. J., Avrahami, T., Cowan, B. R., Hirsch, T., & Hsieh, G. (2021). Heuristic Evaluation of Conversational Agents. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445312

Larbi, D., Denecke, K., & Gabarron, E. (2022). Usability Testing of a Social Media Chatbot for Increasing Physical Activity Behavior. *Journal of Personalized Medicine*, *12*(5). https://doi.org/10.3390/jpm12050828

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and Evaluation of a User Experience Questionnaire. In A. Holzinger (Ed.), *HCI and Usability for Education and Work* (pp. 63–76). Springer Berlin Heidelberg.

Lee, M., Srivastava, M., Hardy, A., Thickstun, J., Durmus, E., Paranjape, A., Gerard-Ursin, I., Li, X. L., Ladhak, F., Rong, F., Wang, R. E., Kwon, M., Park, J. S., Cao, H., Lee, T., Bommasani, R., Bernstein, M., & Liang, P. (2024). Evaluating Human-Language Model Interaction. https://arxiv.org/abs/2212.09746

Lewis, J. R., & Sauro, J. (2018). Item benchmarks for the system usability scale. *J. Usability Studies*, *13*(3), 158–167.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proceedings of the 34th International Conference on Neural Information Processing Systems*. https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

LexisNexis Österreich. (2025). LexisNexis Österreich: führender Anbieter intelligenter Rechtsinformation. https://www.lexisnexis.at/

LexisNexis Österreich. (2026). *Lexis+ AI*. Retrieved January 20, 2026, from https://www.lexisnexis.at/produkte/lexis-plus-ai-3/

Li, H., Chen, J., Yang, J., Ai, Q., Jia, W., Liu, Y., Lin, K., Wu, Y., Yuan, G., Hu, Y., Wang, W., Liu, Y., & Huang, M. (2025, July). LegalAgentBench: Evaluating LLM agents in legal domain. In W. Che, J. Nabende, E. Shutova & M. T.

Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2322–2344). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.116

Li, H., Chen, Y., YiRan, H., Ai, Q., Chen, J., Yang, X., Yang, J., Wu, Y., Liu, Z., & Liu, Y. (2025). LexRAG: Benchmarking Retrieval-Augmented Generation in Multi-Turn Legal Consultation Conversation. *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3606–3615. https://doi.org/10.1145/3726302.3730340

Li, H., Chen, Y., Ai, Q., Wu, Y., Zhang, R., & Liu, Y. (2024). LexEval: A Comprehensive Chinese Legal Benchmark for Evaluating Large Language Models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak & C. Zhang (Eds.), *Advances in Neural Information Processing Systems* (pp. 25061–25094, Vol. 37). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2024/file/2cb40fc022ca7bdc1a9a78b793661284-Paper-Datasets_and_Benchmarks_Track.pdf

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, 74–81. https://aclanthology.org/W04-1013/

Linde Verlag. (2025). Linde Verlag - Fachverlag für Recht, Wirtschaft und Steuern. https://www.lindeverlag.at/de

Liu, B., Hu, Y., Ai, Q., Liu, Y., Wu, Y., Li, C., & Shen, W. (2025). Generating Clarifying Questions for Conversational Legal Case Retrieval without External Knowledge. *ACM Trans. Inf. Syst.*, *43*(4). https://doi.org/10.1145/3736161

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *Journal of Empirical Legal Studies*, *22*(2), 216–242. https://doi.org/https://doi.org/10.1111/jels.12413

Mamalis, M. E., Kalampokis, E., Fitsilis, F., Theodorakopoulos, G., & Tarabanis, K. (2024). A Large Language Model Agent Based Legal Assistant for Governance Applications. In M. Janssen, J. Crompvoets, J. R. Gil-Garcia, H. Lee, I. Lindgren, A. Nikiforova & G. Viale Pereira (Eds.), *Electronic government* (pp. 286–301). Springer Nature Switzerland.

MANZ'sche Verlags- und Universitätsbuchhandlung GmbH. (2024, October). "MANZ Genjus KI": Revolutionäre KI-Technologie für juristische Recherche. https://www.ots.at/presseaussendung/OTS_20241003_OTS0040/manz-genjus-ki-revolutionaere-ki-technologie-fuer-juristische-recherche

MANZ'sche Verlags- und Universitätsbuchhandlung GmbH. (2025a). Das ist erst der Anfang – Genjus KI entwickelt sich ständig weiter. https://genjus.manz.at/produkt-news/zukunftsausblick

MANZ'sche Verlags- und Universitätsbuchhandlung GmbH. (2025). RDB Rechtsdatenbank. https://rdb.manz.at/

MANZ'sche Verlags- und Universitätsbuchhandlung GmbH. (2025b). Was Sie schon immer über Genjus KI wissen wollten. https://genjus.manz.at/faqs

Martínez, E. (2024). Re-evaluating GPT-4's bar exam performance. *Artificial Intelligence and Law*. https://doi.org/10.1007/s10506-024-09396-9

Matelsky, J. K., Parodi, F., Liu, T., Lange, R. D., & Kording, K. P. (2023). A large language model-assisted education tool to provide feedback on open-ended responses. http://arxiv.org/abs/2308.02439

Mesquita, A., Freire, C., Gomes, A., Amazonas, B., & Uchôa, A. (2025). Hello, Freire! How Can You Help Me? A Smart Chatbot to Enhance Access to Academic Opportunities and Institutional Matters. *Anais do XXI Simpósio Brasileiro de Sistemas de Informação*, 469–478. https://doi.org/10.5753/sbsi.2025.246546

Microsoft. (2025). *Design system messages with Azure OpenAI - Azure OpenAI Service*. Retrieved August 12, 2025, from https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/advanced-prompt-engineering

Microsoft Corporation. (2025). *What's Azure AI Search?* Retrieved August 13, 2025, from https://learn.microsoft.com/en-us/azure/search/search-what-is-azure-search

Musatoiu, T. (2024, April). *Enhance your prompts with meta prompting*. OpenAI. Retrieved April 25, 2025, from https://cookbook.openai.com/examples/enhance_your_prompts_with_meta_prompting

NASA Ames Research Center. (1988). *NASA Task Load Index (TLX) Version 1.0 Paper and Pencil Package* (tech. rep.). NASA Ames Research Center. Moffett Field, California. https://humansystems.arc.nasa.gov/groups/tlx/downloads/TLX_pappen_manual.pdf

Nessler, B., Doms, T., & Hochreiter, S. (2023). Functional trustworthiness of AI systems by statistically valid testing. https://arxiv.org/abs/2310.02727

Ngo, A. T., Doan, T. T., Ngo, T. T., & Nguyen, V. H. (2025). Legal Documents Query Application for Vietnamese Law Using LLM and RAG Techniques. *Proceedings of the 2025 10th International Conference on Intelligent Information Technology*, 152–157. https://doi.org/10.1145/3731763.3731802

Nyhan, P. (2024). 6 AI trends you'll see more of in 2025 — news.microsoft.com. https://news.microsoft.com/source/features/ai/6-ai-trends-youll-see-more-of-in-2025/

OpenAI. (n.d.). ChatGPT. https://chatgpt.com/

OpenAI. (2022, November). Introducing ChatGPT. https://openai.com/index/chatgpt/

OpenAI. (2024, May). *Hello GPT-4o*. https://openai.com/index/hello-gpt-4o/

OpenAI. (2025). *Prompt generation*. Retrieved April 25, 2025, from https://platform.openai.com/docs/guides/prompt-generation#meta-prompts

Padhye, R. (2024). Software Engineering Methods for AI-Driven Deductive Legal Reasoning. *Proceedings of the 2024 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, 85–95. https://doi.org/10.1145/3689492.3690050

Panda, S., Komal, F. A., Law, E. L.-C., Murad, S., Sun, Z., Summerill, B., Konu, D., & Nguyen, T.-V. (2025). Exploring User Acceptance of a Fintech Chatbot powered by LLMs among Older Adults. *Proceedings of the 38th International BCS Human-*

Computer Interaction Conference, 420–431. https://doi.org/10.14236/ewic/BCSHCI2025.48

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Perplexity. (2025). *Perplexity*. Retrieved October 31, 2025, from https://www.perplexity.ai/

Pillay, T., & Booth, H. (2025). 5 Predictions for AI in 2025 - time.com. https://time.com/7204665/ai-predictions-2025/

Prainsack, B., & Forgó, N. (2024). New AI regulation in the EU seeks to reduce risk without assessing public benefit. *Nature Medicine*, *30*(5), 1235–1237. https://doi.org/10.1038/s41591-024-02874-2

PwC Österreich. (2025, April). KI-Studie von PwC und Microsoft: Jede:r Zweite wünscht sich mehr Künstliche Intelligenz im Arbeitsalltag. https://www.pwc.at/de/presse/2025/ki-studie-jeder-zweite-wuenscht-sich-mehr-ki-im-arbeitsalltag.html

Qiu, Y., Duvvuri, V. C., Yadavalli, P., & Prasad, N. (2024). Evaluation of Generative AI Q&A Chatbot Chained to Optical Character Recognition Models for Financial Documents. *Proceedings of the 2024 8th International Conference on Machine Learning and Soft Computing*, 101–110. https://doi.org/10.1145/3647750.3647766

R, K., N, D., C, S. N., R, V. G., & Y, S. V. (2025). Ethical and legal implications of ai chatbots in legal sector. *2025 International Conference on Computational Robotics, Testing and Engineering Evaluation (ICCRTEE)*, 1–5. https://doi.org/10.1109/ICCRTEE64519.2025.11053085

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. https://arxiv.org/abs/2212.04356

Robson, C. (2024). *Real world research : a resource for users of social research methods in applied settings* (Fifth edition). Wiley.

Rodrigues, J., & Branco, A. (2024). Meta-prompting Optimized Retrieval-augmented Generation. https://arxiv.org/abs/2407.03955

Schick, T., Dwivedi-Yu, J., Dess\'{\i}, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: language models can teach themselves to use tools. *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Schmidhuber, J., Schlögl, S., & Ploder, C. (2021). Cognitive Load and Productivity Implications in Human-Chatbot Interaction. *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, 1–6. https://doi.org/10.1109/ICHMS53169.2021.9582445

Schmude, T., Koesten, L., Möller, T., & Tschiatschek, S. (2025). Information that matters: Exploring information needs of people affected by algorithmic decisions. *International Journal of Human-Computer Studies*, *193*, 103380. https://doi.org/https://doi.org/10.1016/j.ijhcs.2024.103380

Schöch, C. (2023). Repetitive research: a conceptual space and terminology of replication, reproduction, revision, reanalysis, reinvestigation and reuse in digital humanities. *International Journal of Digital Humanities*, *5*(2), 373–403. https://doi.org/10.1007/s42803-023-00073-y

Schweighofer, K., Brune, B., Gruber, L., Schmid, S., Aufreiter, A., Gruber, A., Doms, T., Eder, S., Mayer, F., Stadlbauer, X.-P., Schwald, C., Zellinger, W., Nessler, B., & Hochreiter, S. (2025). Safe and Certifiable AI Systems: Concepts, Challenges, and Lessons Learned. https://arxiv.org/abs/2509.08852

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, 102551. https://doi.org/https://doi.org/10.1016/j.ijhcs.2020.102551

Statistik Austria. (2025). Künstliche Intelligenz – Nutzung und Einstellung in Österreich. Ergebnisse der Erhebung über den IKT-Einsatz in Haushalten 2024 [Analyse und Pressemitteilung, veröffentlicht am 15. Mai 2025]. https://www.statistik.at/fileadmin/announcement/2025/05/20250515KI_Analyse.pdf

Stephens, S. (2025). *The development and evaluation of a legal chatbot for women in tanzania* (SSRN Scholarly Paper). SSRN. http://dx.doi.org/10.2139/ssrn.5251610

Su, W., Ai, Q., Wu, Y., Xie, A., Wang, C., Ma, Y., Li, H., Wu, Z., Liu, Y., & Zhang, M. (2025). Pre-training for Legal Case Retrieval Based on Inter-Case Distinctions. *ACM Trans. Inf. Syst.*, *43*(5). https://doi.org/10.1145/3735127

Suzgun, M., & Kalai, A. T. (2024). Meta-Prompting: Enhancing Language Models with Task-Agnostic Scaffolding. https://arxiv.org/abs/2401.12954

Thumfart, P. S. (2022). Ris API Wrapper. https://blog-idunivienna.univie.ac.at/fileadmin/user_upload/p_blog_idunivienna/Legal_Data_Science/paper.pdf

Ulster University. (2025). The Chatbot Usability Questionnaire (CUQ). https://www.ulster.ac.uk/research/topic/computer-science/artificial-intelligence/projects/cuq

Vassilakopoulou, P., Parmiggiani, E., Shollo, A., & Grisot, M. (2022). Responsible AI: Concepts, critical perspectives and an Information Systems research agenda. *Scand. J. Inf. Syst.*, *34*, 3. https://api.semanticscholar.org/CorpusID:256195536

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. u., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Virtanen, K., Mansikka, H., Kontio, H., & Harris, D. (2022). Weight watchers: NASA-TLX weights revisited. *Theoretical Issues in Ergonomics Science*, *23*(6), 725–748. https://doi.org/10.1080/1463922X.2021.2000667

Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P. S., & Wen, Q. (2024). Large Language Models for Education: A Survey and Outlook. http://arxiv.org/abs/2403.18105

Wang, Y., Hernandez, A. G., Kyslyi, R., & Kersting, N. (2024). Evaluating Quality of Answers for Retrieval-Augmented Generation: A Strong LLM Is All You Need. https://arxiv.org/abs/2406.18064

Wendehorst, C., & Nessler, B. (2024, December). Guidelines on the Application of the Definition of an AI System in the AI Act: ELI Proposal for a Three-Factor Approach. https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ELI_Response_on_the_definition_of_an_AI_System.pdf

Wendehorst, C., Nessler, B., Aufreiter, A., & Aichinger, G. (2024). Der begriff des „KI-Systems“ unter der neuen KI-VO: Vorschlag eines „Drei-Faktor-Ansatzes“ zur beseitigung von juristischen und technischen ungereimtheiten. *MMR - Zeitschrift für IT-Recht und Digitalisierung*, (7), 605–614.

Winter, P. M., Eder, S., Weissenböck, J., Schwald, C., Doms, T., Vogt, T., Hochreiter, S., & Nessler, B. (2021). Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications. https://arxiv.org/abs/2103.16910

World Economic Forum. (2025, January). *Future of Jobs Report 2025* (tech. rep.). Geneva. https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf

Yao, R., Wu, Y., Wang, C., Xiong, J., Wang, F., & Liu, X. (2025, April). Elevating Legal LLM Responses: Harnessing Trainable Logical Structures and Semantic Knowledge with Legal Reasoning. In L. Chiruzzo, A. Ritter & L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 5630–5642). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.290

Zhang, Y., Yuan, Y., & Yao, A. C.-C. (2025). Meta Prompting for AI Systems. https://arxiv.org/abs/2311.11482

Zhu, J., Kempermann, M., Cannanure, V. K., Hartland, A., Navarrete, R. M., Carteny, G., Braun, D., & Weber, I. (2025). Learn, Explore and Reflect by Chatting: Understanding the Value of an LLM-Based Voting Advice Application Chatbot. *Proceedings of the 7th ACM Conference on Conversational User Interfaces*. https://doi.org/10.1145/3719160.3736611

Acronyms

AK Arbeiterkammer Österreich. 1, 2, 6–8, 11–13, 15–17, 23–25, 27, 31, 39, 44, 45, 52, 59, 61–65, 81–83

AK-chatbot Arbeiterkammer's tax law chatbot. 1–3, 6, 7, 11, 12, 19, 27, 38, 41–43, 45, 57, 58, 60–65, 99

CEF Chatbot Evaluation Framework. 2, 19, 22, 28, 57, 64

A. Appendix

A.1. Meeting Schedule

To ensure constant progress on the project bi-weekly meetings were held to develop the study design. Additional meetings were held AK's AI Lead and AK's Head of Digital Affairs.

- 18.11.2024 (*with AI Lead in attendance*): (Possible) study procedure and interview schedule, tasks for interviewees (and task analysis), (possible) research questions, System Usability Scale (SUS), XAI Novice Question Bank
- 03.12.2024: Discussion of possible rating scales (System Usability Scale (SUS), Chatbot Usability Questionnaire (CUQ), User Experience Questionnaire (UEQ)
- 17.12.2024: Presentation of the interview pilot, discussion of rating scales
- 07.01.2025: Discussion of interview process, scale ratings, task design, current status of the contract with Arbeiterkammer
- 13.01.2025 (*with AI Lead in attendance*): Current state of the chatbot, presentation of the chatbot, discussion of study procedure
- 21.01.2025: Study procedure, task design
- 10.02.2025 (*with AI Lead in attendance*): Current state of the chatbot, start of interview sessions, contact person for task creation (legal expert, coordinator, chatbot user); meeting had to be postponed (initial date: 29.01.2025)
- 18.02.2025: Study procedure, task design, current status of the contract
- 04.03.2025 (*with AI Lead in attendance*): (Final) discussion of the interview process, next steps to be able to start the interview sessions (VPN-client, consent form), adjustment of tasks (according to the suggestions by AK tax law expert), last agreement on the group of people to be interviewed, planning of interview pilots
- 20.03.2025 (*with AI Lead in attendance*): Report on the pilot interview with AK tax law expert
- 01.04.2025 (*with AI Lead in attendance*): Discussion technical specifications of the chatbot, update on status quo of interview schedules
- 15.04.2025: Report on conducted interviews and organizational matters

- 28.04.2025: Report on interviews conducted and technical difficulties with the chatbot during the interview process, further steps to obtain a sufficient number of interview participants
- 12.05.2025 (*with AI Lead in attendance*): Report on the current status of conducted interviews
- 15.05.2025: Report on planned interviews, discussion on possible improvements of study procedure
- 22.05.2025: Report on interviews conducted to date and technical difficulties encountered during the interview process
- 10.06.2025 (*with AI Lead in attendance*): Report on the current status of conducted interviews, first data insights (online survey)
- 23.06.2025 (*with AI Lead in attendance*): Report on the current status of conducted interviews, data insights (online survey, XAI Novice Questions Bank)
- 31.07.2025: Discussion on how to analyze qualitative data, how to improve the structure of the master's thesis and first quantitative data insights
- 14.08.2025: Report on the current status of the thesis and existing challenges, insights into survey data (perceived trustworthiness of the chatbot)
- 05.09.2025: Report on the current status of the thesis, insights into qualitative data
- 15.09.2025 (*with AK's Head of Digital Affairs in attendance*): Report in the current status of the chatbot evaluation, insights into quantitative and qualitative data
- 30.09.2025: Insights into qualitative data
- 14.10.2025: Insights into qualitative data (feelings towards the chatbot); discussing the further steps and subsequent deadlines
- 28.10.2025: further administrative steps (data submission); submission of chapters "Introduction" and "Background"
- 11.11.2025: Update on current status of the thesis; feedback on project report
- 09.12.2025: Update on current status of the thesis
- 08.01.2026: preparation of paper for CHIWORK'26
- 22.01.2026: preparation of paper for CHIWORK'26
- 27.01.2026: preparation of paper for CHIWORK'26

A.2. Interviewers

Table A.1 shows which individuals participated in the interview sessions as interviewers. AK's AI Lead participated in the interview with participant P2 in order to gain insight into the interview process. Participation was announced in advance by email. AK's AI Lead did not take on the role of interviewer and did not intervene during the interview.

According to the University of Vienna, for privacy reasons, no personal data - such as names - should be included in the thesis. To ensure anonymity during the plagiarism check, the names of Interviewers A and B have not been included here. The non-anonymized version was stored on Nextcloud cloud storage provided by the Research Group Visualization and Data Analysis of the University of Vienna.

Participant	**Date**	**Time**	**Interviewer 1**	**Interviewer 2**
Pilot	-	-	Author	Interviewer A
Pilot	-	-	Author	Interviewer A
Pilot (AK)	18.03.2025	16:00 - 17:00	Author	Interviewer A
P1	06.05.2025	09:00 - 10:00	Author	Interviewer A
P2	21.05.2025	15:30 - 16:30	Author	Interviewer B
P3	22.05.2025	08:30 - 09:30	Author	Interviewer A
P4	22.05.2025	10:30 - 11:30	Author	Interviewer A
P5	26.05.2025	08:45 - 09:45	Author	Interviewer A
P6	04.06 2025	15:30 - 16:30	Author	Interviewer A
P7	05.06.2025	16:00 - 17:00	Author	Interviewer B
P8	17.06.2025	09:00 - 10:00	Author	Interviewer A
P9	17.06.2025	15:00 - 16:00	Author	-
P10	04.07.2025	10:00 - 11:00	Author	Interviewer A
P11	07.07.2025	10:00 - 11:00	Author	Interviewer A
P12	09.07.2025	10:00 - 11:00	Author	-
P13	10.07.2025	10:15 - 11:15	Author	-
P14	10.07.2025	13:00 - 14:00	Author	-

Table A.1.: Interviewers

A.3. Interview Pilot

The interview data were stored on Nextcloud cloud storage provided by the Research Group Visualization and Data Analysis of the University of Vienna.

A.4. Interview Pilot - Arbeiterkammer

The interview data were stored on Nextcloud cloud storage provided by the Research Group Visualization and Data Analysis of the University of Vienna.

A.5. Versions of tasks

Four different tasks were developed at the beginning of the study design. Two of the four tasks were further developed and also used in the study. The following section contains all the tasks developed.

A.5.1. Task 1

Version 1

Situation:

- Stellen Sie sich vor, ein Kunde oder eine Kundin hat Ihnen folgende Mail gesendet.
- Lesen Sie zuerst die Mail durch.

Absender: erika.musterfrau@mailadresse.at
Betreff: Home Office und WIFI-Kurse für Arbeitnehmerveranlagung

Sehr geehrte Damen und Herren,

Ich hoffe, Sie können mir bei einigen Fragen zu unseren Steuern weiterhelfen. Mein Mann und ich sind beide berufstätig, und wir haben einige Ausgaben, die wir gerne steuerlich geltend machen würden. Ich arbeite aktuell relativ viel im Home Office und habe mir dafür im letzten Jahr um ca. € 2000,- neue Geräte gekauft (Laptop, Maus, Headset, etc.) Kann ich diese Dinge komplett als Werbungskosten geltend machen.

Bzw. mein Mann (arbeitet als AHS-Lehrer) hat auch so viel für seine Geräte ausgegeben und würde das auch gerne steuerlich geltend machen. Ist das Möglich?

Ich plane aktuell mich beruflich weiterzuentwickeln und möchte am WIFI Kurse machen, um einen besseren Job zu bekommen. Kann man diese Kurse auch vielleicht absetzen?

Welche Dokumente benötigen wir, damit wir diese Ausgaben bei der Arbeitnehmerveranlagung einreichen können?

Herzlichen Dank für Ihre Hilfe!

Erika Musterfrau

Aufgabe:

- Verwenden Sie als Hilfe den Steuerrechts-Chatbot, um eine passende Antwort zu verfassen.
- Versuchen Sie auch inhaltlich passende Antworten mit Hilfe des Chatbots zu erstellen.
- Arbeiten Sie so, wie Sie das auch in einer echten Arbeitssituation machen würden.

Sobald Sie mit dem Ergebnis Ihrer Interaktion mit dem Chatbot ausreichend zufrieden sind, können wir mit dem Interview beginnen.

Version 2

Situation:

- Stellen Sie sich vor, ein Kunde oder eine Kundin hat Ihnen folgende Mail gesendet.
- Verwenden Sie als Hilfe den Steuerrechts-Chatbot, um eine passende Antwort zu verfassen.
- Arbeiten Sie so, wie Sie das auch in einer echten Arbeitssituation machen würden.

Absender: erika.musterfrau@mailadresse.at
Betreff: Home Office für Arbeitnehmerveranlagung

Sehr geehrte Damen und Herren,

Ich hoffe, Sie können mir bei einigen Fragen zu unseren Steuern weiterhelfen. Mein Mann und ich sind beide berufstätig, und wir haben einige Ausgaben, die wir gerne steuerlich geltend machen würden. Ich arbeite aktuell relativ viel im Home Office und habe mir dafür im letzten Jahr um ca. € 2000,- neue Geräte gekauft (Laptop, Maus, Headset, etc.) Kann ich diese Dinge komplett als Werbungskosten geltend machen?

Bzw. mein Mann (arbeitet als AHS-Lehrer) hat auch so viel für seine Geräte ausgegeben und würde das auch gerne steuerlich geltend machen. Ist das möglich?

Herzlichen Dank für Ihre Hilfe!

Erika Musterfrau

Aufgabe:

- Erstellen Sie eine vollständige Antwortmail oder einen Text, den Sie im Arbeitsalltag zur Beantwortung der Anfrage verwenden würden.

Sobald Sie mit dem Ergebnis Ihrer Interaktion mit dem Chatbot ausreichend zufrieden sind, können wir mit dem Interview beginnen.

Version 3

Situation:

- Stellen Sie sich vor, ein Kunde oder eine Kundin hat Ihnen folgende Mail gesendet.
- Verwenden Sie als Hilfe den Steuerrechts-Chatbot, um eine passende Antwort zu verfassen.
- Arbeiten Sie so, wie Sie das auch in einer echten Arbeitssituation machen würden.

Absender: erika.huber@abc-mail.at
Betreff: Home-Office für Arbeitnehmerveranlagung

Sehr geehrte Damen und Herren,

Ich hoffe, Sie können mir bei einigen Fragen zu unseren Steuern weiterhelfen. Mein Mann und ich sind beide berufstätig, und wir haben einige Ausgaben, die wir gerne steuerlich geltend machen würden. Ich arbeite aktuell relativ viel im Home-Office und habe mir dafür im letzten Jahr um ca. € 2000,- neue Geräte (Laptop, Maus, Headset, etc.) und Möbel für mein Arbeitszimmer gekauft. Kann ich diese Dinge komplett als Werbungskosten geltend machen?

Herzlichen Dank für Ihre Hilfe!

Erika Huber

Aufgabe:

- Erstellen Sie eine vollständige Antwortmail oder einen Text, den Sie im Arbeitsalltag zur Beantwortung der Anfrage verwenden würden.

Sobald Sie mit dem Ergebnis Ihrer Interaktion mit dem Chatbot ausreichend zufrieden sind, können wir mit dem Interview beginnen.

A.5.2. Task 2

Version 1

Situation:

- Stellen Sie sich vor, ein Kunde oder eine Kundin hat Ihnen folgende Mail gesendet.
- Lesen Sie zuerst die Mail durch.

Absender: max.mustermann@mailadresse.at
Betreff: Einkommen Deutschland und Unterhalt

Sehr geehrte Damen und Herren,
Ich schreibe Ihnen, weil ich in einer ein bisschen komplexeren Situation bin. Ich wohne in Wien und arbeite ganz normal bei einer Firma. Zusätzlich bin ich aber als freiberuflicher Berater für ein Unternehmen in Deutschland tätig.

Ich bin mir nicht sicher, wie das mit den Steuern dann ist. Wo muss ich jetzt mein Gehalt versteuern. In Österreich, oder? Aber muss ich auch in Deutschland Steuern zahlen?

Ein weiteres Thema betrifft meine zwei Kinder. Eines lebt bei meiner 1. Ex-Frau in Berlin und mein zweites Kind wohnt in Japan bei meiner ausgewanderten 2. Ex-Frau. Beide haben den Unterhalt bisher von mir verlangt, aber ich habe gehört, das stimmt nicht ganz. Muss ich überhaupt zahlen, wenn die Kinder nicht in Österreich leben? Wenn ich zahlen muss, kann ich die Kosten dann irgendwie im Steuerausgleich absetzen?

MfG

Max Mustermann

Aufgabe:

- Verwenden Sie als Hilfe den Steuerrechts-Chatbot, um eine passende Antwort zu verfassen.
- Versuchen Sie auch inhaltlich passende Antworten mit Hilfe des Chatbots zu erstellen.
- Arbeiten Sie so, wie Sie das auch in einer echten Arbeitssituation machen würden.

Version 2

Situation:

- Stellen Sie sich vor, ein Kunde oder eine Kundin hat Ihnen folgende Mail gesendet.
- Verwenden Sie als Hilfe den Steuerrechts-Chatbot, um eine passende Antwort zu verfassen.
- Arbeiten Sie so, wie Sie das auch in einer echten Arbeitssituation machen würden.

Absender: max.mustermann@mailadresse.at
Betreff: Einkommen Deutschland und Unterhalt

Sehr geehrte Damen und Herren,
Ich schreibe Ihnen, weil ich in einer ein bisschen komplexeren Situation bin. Ich wohne in Wien und arbeite ganz normal bei einer Firma. Zusätzlich bin ich aber als freiberuflicher Berater für ein Unternehmen in Deutschland tätig.

Ich bin mir nicht sicher, wie das mit den Steuern dann ist. Wo muss ich jetzt mein Gehalt versteuern. In Österreich, oder? Aber muss ich auch in Deutschland Steuern zahlen?

Ein weiteres Thema betrifft meine zwei Kinder. Eines lebt bei meiner 1. Ex-Frau in Berlin und mein zweites Kind wohnt in Japan bei meiner ausgewanderten 2. Ex-Frau. Beide haben den Unterhalt bisher von mir verlangt, aber ich habe gehört, das stimmt nicht ganz. Muss ich überhaupt zahlen, wenn die Kinder nicht in Österreich leben?

MfG

Max Mustermann

Aufgabe:

- Erstellen Sie eine vollständige Antwortmail oder einen Text, den Sie im Arbeitsalltag zur Beantwortung der Anfrage verwenden würden.

Sobald Sie mit dem Ergebnis Ihrer Interaktion mit dem Chatbot ausreichend zufrieden sind, können wir mit dem Interview beginnen.

Version 3

Situation:

- Stellen Sie sich vor, ein Kunde oder eine Kundin hat Ihnen folgende Mail gesendet.
- Verwenden Sie als Hilfe den Steuerrechts-Chatbot, um eine passende Antwort zu verfassen.
- Arbeiten Sie so, wie Sie das auch in einer echten Arbeitssituation machen würden.

Absender: thomas.mayer@mailmail.at
Betreff: Einkommen Deutschland und Unterhalt

Sehr geehrte Damen und Herren,

Ich schreibe Ihnen, weil ich in einer ein bisschen komplexeren Situation bin. Ich wohne in Wien und arbeite ganz normal bei einer Firma. Zusätzlich bin ich aber als freiberuflicher Berater für ein Unternehmen in Deutschland tätig.

Ich bin mir nicht sicher, wie das mit den Steuern dann ist. Wo muss ich jetzt mein Gehalt versteuern. In Österreich, oder? Aber muss ich auch in Deutschland Steuern zahlen?

Ein weiteres Thema betrifft meine zwei Kinder. Eines lebt bei meiner 1. Ex-Frau in Berlin und mein zweites Kind wohnt in Japan bei meiner ausgewanderten 2. Ex-Frau. Beide haben den Unterhalt bisher von mir verlangt, aber ich habe gehört, das stimmt nicht ganz. Muss ich überhaupt zahlen, wenn die Kinder nicht in Österreich leben?

MfG

Thomas Mayer

Aufgabe:

- Erstellen Sie eine vollständige Antwortmail oder einen Text, den Sie im Arbeitsalltag zur Beantwortung der Anfrage verwenden würden.

Sobald Sie mit dem Ergebnis Ihrer Interaktion mit dem Chatbot ausreichend zufrieden sind, können wir mit dem Interview beginnen.

Version 4

Situation:

- Stellen Sie sich vor, ein Kunde oder eine Kundin hat Ihnen folgende Mail gesendet.
- Verwenden Sie als Hilfe den Steuerrechts-Chatbot, um eine passende Antwort zu verfassen.
- Arbeiten Sie so, wie Sie das auch in einer echten Arbeitssituation machen würden.

Absender: thomas.mayer@mailmail.at
Betreff: Einkommen Deutschland und Unterhalt

Sehr geehrte Damen und Herren,

Ich schreibe Ihnen, weil ich in einer ein bisschen komplexeren Situation bin. Ich wohne in Wien und arbeite ganz normal bei einer Firma. Zusätzlich bin ich aber als freiberuflicher Berater für ein Unternehmen in Deutschland tätig. Ich habe keinen Wohnsitz (mehr) in Deutschland und werde diese Beratungstätigkeit ausschließlich zu 100% remote von Österreich ausüben.

Ich bin mir nicht sicher, wie das mit den Steuern dann ist. Wo muss ich jetzt mein Gehalt versteuern. In Österreich, oder? Aber muss ich auch in Deutschland Steuern zahlen?

Ein weiteres Thema betrifft meine zwei minderjährigen Kinder (6 und 12 Jahre). Eines lebt bei meiner 1. Ex-Frau in Berlin und mein zweites Kind wohnt in Serbien bei meiner 2. Ex-Frau. Ich zahle monatlich Alimente in Höhe von insgesamt 1.000€. Wo kann ich das bei meinem Steuerausgleich geltend machen?

MfG

Thomas Mayer

Aufgabe:

- Erstellen Sie eine vollständige Antwortmail oder einen Text, den Sie im Arbeitsalltag zur Beantwortung der Anfrage verwenden würden.

Sobald Sie mit dem Ergebnis Ihrer Interaktion mit dem Chatbot ausreichend zufrieden sind, können wir mit dem Interview beginnen.

A.5.3. Task 3

Version 1

Situation:

- Stellen Sie sich vor, ein Kunde oder eine Kundin hat Ihnen folgende Mail gesendet.
- Lesen Sie zuerst die Mail durch.

Absender: anna.musterfrau@mailadresse.at
Betreff: Steuererklärung, Kinderförderung

Hallo liebes Arbeiterkammer-Team,

Ich hoffe, Sie können mir bei ein paar Steuerfragen helfen. Ich bin alleinerziehende Mutter und wohne in Wien, während meine zwei Kinder in Salzburg studieren. Das macht die Sache etwas komplizierter als gedacht.

Bisher hat mein Arbeitgeber den AEAB immer abgezogen, aber irgendwie wurde der Betrag bei meiner letzten Steuererklärung nicht berücksichtigt. Ich bekomme ihn zwar monatlich, aber eine Freundin hat mir gesagt, das passt so irgendwie nicht. Muss ich etwas machen, damit er in meiner Steuererklärung aufscheint?

Ich habe gehört, dass man Werbungskosten auch auf die Kinder beziehen kann. Was fällt da alles rein? Können auch die Kosten für meine Kinder, die in Salzburg studieren, als Werbungskosten geltend gemacht werden? Gibt es da irgendwelche Möglichkeiten, Kosten abzusetzen? Eine Bekannte hat mir gesagt, es gibt anscheinend einen Freibetrag, aber ich weiß nicht was ich machen muss damit der zählt.

Gibt es sonst noch Förderungen für Alleinerziehende, an die ich denken muss? Seit Kurzem studieren meine Kinder auswärts und das ist sehr teuer. Ich wäre Ihnen dankbar, wenn Sie mir da ein bisschen weiterhelfen könnten. Ich möchte keine Probleme mit dem Finanzamt bekommen.

MfG

Anna Musterfrau

Aufgabe:

- Verwenden Sie als Hilfe den Steuerrechts-Chatbot, um eine passende Antwort zu verfassen.
- Versuchen Sie auch inhaltlich passende Antworten mit Hilfe des Chatbots zu erstellen.
- Arbeiten Sie so, wie Sie das auch in einer echten Arbeitssituation machen würden.

A.5.4. Task 4

Version 1

Situation:

- Stellen Sie sich vor, ein Kunde oder eine Kundin hat Ihnen folgende Mail gesendet.
- Lesen Sie zuerst die Mail durch.

Sehr geehrte Damen und Herren,

ich brauche Ihre Hilfe für meinen Steuerausgleich (Arbeitnehmerveranlagung). Meine Fragen wären:

- Wie werden Homeoffice-Pauschale und Arbeitsmittel steuerlich berücksichtigt?
- Welche Auswirkungen hat eine gelegentliche Nutzung des Firmenbüros auf die Pendlerpauschale?
- Gibt es Besonderheiten bei der Nutzung mehrerer Verkehrsmittel (z. B. Auto und öffentlicher Nahverkehr)?

Vielen Dank für Ihre Hilfe!

Mit freundlichen Grüßen

Max Mustermann

Aufgabe:

- Verwenden Sie als Hilfe den Steuerrechts-Chatbot, um eine passende Antwort zu verfassen.
- Versuchen Sie auch inhaltlich passende Antworten mit Hilfe des Chatbots zu erstellen.
- Arbeiten Sie so, wie Sie das auch in einer echten Arbeitssituation machen würden.

A.6. Interview schedule

A.6.1. English version

ID	Interview Questions	Notes
Interviewer shares screen with presentation!		
Explanation of the study: Our study analyzes the usability of the Arbeiterkammer's tax law chatbot. Your personal impressions are important to us.		
In the first part of the study, you interact with the tax law chatbot to complete a task using its answers. Then we will ask you to fill out a short questionnaire. Finally, we will discuss your personal impressions of the chatbot.		
Interviewer ends screen sharing!		
TASK PRESENTATION (Please read the following task carefully.)		
Q01	How confident are you that you can solve the problem on your own with your current knowledge and without additional support? (The question refers to both tasks!)	5-point likert: not very confident / very confident
Participant shares screen!		
TASK EXECUTION (Please complete the task using the chatbot.)		
Q08	How helpful did you find the chatbot's responses for solving the problem? (The question refers to both tasks!)	5-point likert: not very helpful / very helpful
Participant ends screen sharing!		
Presentation of the survey (We kindly ask you to complete the following questionnaire. It will probably take about 5 minutes to complete. You are welcome to ask questions during the survey.)		
Explanation of scales: The survey consists of 3 pages. Pages 1 and 2 contain established scales. Page 3 deals with additional aspects such as trustworthiness, etc.		
Note: 3 as a "neutral" point on the Likert scale		
Interviewer shares screen! (Definitions of the NASA Task Load Index scales)		
Q02-07	Page 1: NASA Task Load Index (Raw-TLX)	
Q09-24	Page 2: Chatbot Usability Questionnaire (CUQ), SHOW DEFINTIONS!	
Q25-30	Page 3: Weitere spezifische Interviewfragen (Perceived Trustworthiness)	
Interviewer ends screen sharing!		
Q31	How much would you trust the chatbot with legal decisions?	
Q32	How high do you estimate the risk of the chatbot providing incorrect information?	
Q33	In what situations can you imagine using this chatbot in your work (as a tax law expert)?	
Q34	How could this chatbot make your work more effective?	

ID	**Interview Questions**	**Notes**
Q35	Could you imagine using this chatbot regularly in your work in the future? Why?	
Q36	Which tasks do you consider unsuitable for this chatbot to perform? Why?	
Q37	What (additional) background information should the chatbot be able to "understand" so that its responses appear well-informed?	
Q38	Would you prefer the chatbot to be able to respond using both legal terminology as well as everyday language? Why?	
Q39	How would you feel if the chatbot flagged certain responses for risk reasons (e.g., ethical reasons, data protection, etc.)?	
Q40	If you could improve the chatbot in any way, what changes would you make?	
Q41	How helpful do you find the documents linked in the bot's responses? What do you like (or dislike) about them?	
Interviewer shares screen!		
Presentation of the set of questions (XAI Novice Question Bank) (Please take a look at the following questions to get an initial overview.)		
Q42	Please select at least 3 questions that would help you better understand and use the chatbot. Why did you choose these questions?	SHOW QUESTIONS!
Interviewer ends screen sharing!		
Q43	Have you ever used a chatbot in a professional context?	
Q44	Have you ever used the Arbeiterkammer's tax-law chatbot? (If applicable) How often do you use this chatbot? [frequency]	
Q44_2	If so: Did you receive any special training?	
Q45	If yes, which chatbot did you use?	
Q46	For what reasons do you use chatbots in your work?	
Q47	What experiences have you had with legal chatbots?	
Q48	What other experiences have you had with chatbots? (If applicable) How often do you use that chatbot?	

A.6.2. Original German version

ID	Interview Fragen	Notizen
Interviewer teilt Bildschirm mit Präsentation!		
Erklärung der Studie: In unserer Studie wird die Nutzbarkeit des Steuerrecht-Chatbots der Arbeiterkammer analysiert. Ihre persönlichen Eindrücke sind dabei für uns wichtig.		
Im ersten Teil der Studie interagieren Sie mit dem Steuerrecht-Chatbot, um mithilfe seiner Antworten eine Aufgabe zu bearbeiten. Dann werden wir Sie bitten, einen kurzen Fragebogen auszufüllen. Und abschließend führen wir ein Gespräch über Ihre persönlichen Eindrücke zu dem Chatbot.		
Interviewer beendet Teilen des Bildschirms!		
AUFGABEN PRÄSENTATION (Bitte lesen Sie sich folgende Aufgabe durch.)		
Q01	Wie zuversichtlich sind Sie, dass Sie das Problem mit Ihrem aktuellen Wissen alleine und ohne zusätzliche Unterstützung hätten lösen können? (Die Frage bezieht sich auf beide Aufgaben!)	5-point likert: nicht zuversichtlich / sehr zuversichtlich
Teilnehmer*in teilt Bildschirm!		
AUFGABEN DURCHFÜHRUNG (Bitte bearbeiten Sie die gestellte Aufgabe mithilfe des Chatbots.)		
Q08	Wie hilfreich fanden Sie die Antworten des Chatbots für die Lösung des Problems? Falls Sie bereits wussten, wie das Problem zu lösen war, bewerten Sie bitte so, als wäre dies Ihr erster Kontakt mit dem Problem. (Die Frage bezieht sich auf beide Aufgaben!)	5-point likert: nicht hilfreich / sehr hilfreich
Teilnehmer*in beendet Teilen des Bildschirms!		
Präsentation der Umfrage (Wir bitten Sie folgenden Fragebogen auszufüllen. Die Beantwortung wird wahrscheinlich ca. 5 Minuten dauern. Sie dürfen zwischendurch Fragen stellen.)		
Erklärung der Skalen: Die Umfrage besteht aus 3 Seiten. Seite 1 und 2 beinhalten etablierte Skalen. Seite 3 beschäftigt sich mit zusätzlichen Aspekten wie Vertrauenswürdigkeit, etc.		
Anmerkung: 3 als "neutraler" Punkt in der Likert-Scale		
Interviewer teilt Bildschirm!		
Q02-07	Seite 1: NASA Task Load Index (Raw-TLX), DEFINTIONEN ZEIGEN!	
Q09-24	Seite 2: Chatbot Usability Questionnaire (CUQ)	
Q25-30	Seite 3: Weitere spezifische Interviewfragen (Perceived Trustworthiness)	
Interviewer beendet Teilen des Bildschirms!		
Q31	Wie sehr würden Sie dem Chatbot bei rechtlichen Entscheidungen vertrauen?	
Q32	Wie groß schätzen Sie das Risiko ein, dass der Chatbot falsche Informationen gibt?	

ID	**Interview Fragen**	**Notizen**
Q33	In welchen Situationen können Sie sich vorstellen diesen Chatbot in Ihrer Arbeit (als Steuerrechtsexpert:in) nutzen?	
Q34	Wie könnte dieser Chatbot Ihre Arbeit effektiver machen?	
Q35	Könnten Sie sich vorstellen, diesen Chatbot in Zukunft regelmäßig in Ihrer Arbeit zu nutzen? Warum?	
Q36	Welche Aufgaben halten Sie für nicht geeignet, um sie von diesem Chatbot erledigen zu lassen? Warum?	
Q37	Welche (weiteren) Hintergrundinformationen sollte der Chatbot "verstehen" können, damit seine Antworten gut informiert wirken?	
Q38	Würden Sie es bevorzugen, wenn der Chatbot sowohl auf juristische Fachsprache als auch auf Alltagssprache reagieren könnte? Warum?	
Q39	Wie würden Sie sich fühlen, wenn der Chatbot bestimmte Antworten aus Risikogründen (z.B.ethische Gründe, Datenschutz,...) markiert?	
Q40	Wenn Sie den Chatbot in irgendeiner Weise verbessern könnten, welche Änderungen würden Sie vornehmen?	
Q41	Wie hilfreich finden Sie die Dokumente, die in den Antworten des Bots verlinkt sind? Was gefällt Ihnen daran (nicht)?	
Interviewer teilt Bildschirm!		
Präsentation des Fragenkatalogs (XAI Novice Question Bank) (Bitte verschaffen Sie sich einen ersten Überblick über die folgenden Fragen.)		
Q42	Bitte wählen Sie mindestens 3 Fragen aus, die für Sie wichtig wären, um den Chatbot besser verstehen und verwenden zu können. Wieso haben Sie diese Fragen gewählt?	FRAGEN ZEIGEN!
Interviewer beendet Teilen des Bildschirms!		
Q43	Haben Sie schon einmal im Arbeitskontext einen Chatbot verwendet?	
Q44	Haben Sie den Steuerrechts-Chatbot der Arbeiterkammer schon einmal verwendet? (FREQUENZ)	
Q44_2	Falls ja: Haben Sie eine spezielle Einschulung bekommen?	
Q45	Falls ja, welchen Chatbot haben Sie verwendet?	
Q46	Aus welchen Gründen nutzen Sie Chatbots im Arbeitskontext?	
Q47	Welche Erfahrungen haben Sie mit rechtlichen Chatbots gemacht?	
Q48	Welche anderen Erfahrungen haben Sie mit Chatbots? (Wenn vorhanden:) Wie oft nutzen Sie diesen Chatbot pro Woche?	

A.7. System message

Table A.4 contains the original system message from the tax law chatbot. This information was provided by AK's AI Lead so that it could be used in the evaluation.

Du bist Berater in der Arbeiterkammer Wien. Die Arbeiterkammer ist eine gesetzliche Interessenvertretung. Arbeitnehmer, Pensionisten und freie Dienstnehmer erhalten Beratung zur Lohnsteuer und Einkommensteuer. Du beantwortest grundsätzlich auf Basis des Einkommensteuergesetztes. Aktuell ist es das Jahr 2025 Aktuelle Grenzen die es zu beachten gilt: Steuergrenze 2025: 13.308 € Veranlagungsgrenze 2025: 14.517 € Geringfügigkeitsgrenze 2025: 551,10 € Die Geringfügigkeitsgrenze gilt ausschließlich für die Sozialversicherung. In der Steuer gibt es keine Geringfügigkeitsgrenze. Grenze für Zuverdienst auf selbständiger Basis, bis zu der keine Angabe in der Steuererklärung zu erfolgen hat 2025: 730 € Regeln: du gibst keine Antworten auf Basis von Regelungen vergangener Jahre du antwortest in verständlichen Sätzen du antwortest so das jemand auch ohne steuerrechtliche Ausbildung die Antwort versteht du gibst alle antworten in der Sie form Ignoriere alle Beträge in Schilling. Beantworte keine Fragen zu Schenkungen oder Schenkungsmeldungen sondern verweise ausnahmsweise in diesem Fall auf einen Steuerberater

Table A.4.: System message: Tax law chatbot

A.7.1. System message analysis

Table A.5 contains the analysis of the original German system message of the AK-chatbot used during the study.

Prompt description	**System message**
Persona	Du bist Berater in der Arbeiterkammer Wien. Die Arbeiterkammer ist eine gesetzliche Interessenvertretung. Arbeitnehmer, Pensionisten und freie Dienstnehmer erhalten Beratung zur Lohnsteuer und Einkommensteuer.
Rule	Du beantwortest grundsätzlich auf Basis des Einkommensteuergesetztes.
Context	Aktuell ist es das Jahr 2025 Aktuelle Grenzen die es zu beachten gilt: Steuergrenze 2025: 13.308 € Veranlagungsgrenze 2025: 14.517 € Geringfügigkeitsgrenze 2025: 551,10 € Die Geringfügigkeitsgrenze gilt ausschließlich für die Sozialversicherung. In der Steuer gibt es keine Geringfügigkeitsgrenze. Grenze für Zuverdienst auf selbständiger Basis, bis zu der keine Angabe in der Steuererklärung zu erfolgen hat 2025: 730 €
Rule	Regeln: du gibst keine Antworten auf Basis von Regelungen vergangener Jahre du antwortest in verständlichen Sätzen du antwortest so das jemand auch ohne steuerrechtliche Ausbildung die Antwort versteht du gibst alle antworten in der Sie form Ignoriere alle Beträge in Schilling. Beantworte keine Fragen zu Schenkungen oder Schenkungsmeldungen sondern verweise ausnahmsweise in diesem Fall auf einen Steuerberater

Table A.5.: System message analysis (original text)

A.8. Demographics

The form for collecting demographic data (table A.6) was sent to participants in advance by email. The form was hosted on Nextcloud provided by the Research Group Visualization and Data Analysis of the University of Vienna.

Table A.6.: Demographics Form

Question	**Type of question**	**Categories**
"Vorname" (First Name)	open-ended question	
"Nachname" (Last Name)	open-ended question	
"Wie alt sind Sie? Bitte geben Sie Ihr Alter an." (How old are you? Please indicate your age.)	open-ended question	
"Mit welchem Geschlecht identifizieren Sie sich?" (Which gender do you identify with?)	categorical question	"männlich" (male), "weiblich" (female), "non-binär" (non-binary), "anderes Geschlecht" (other gender), "möchte ich nicht angeben" (don't want to answer)
"Falls Sie in der vorangegangenen Frage 'anderes Geschlecht' ausgewählt haben, geben Sie dieses bitte unterhalb an." (If you selected "other gender" in the previous question, please specify below.)	open-ended question	
"Was ist Ihr höchster Bildungsabschluss?" (What is your highest educational qualification?)	categorical question	"Kein Pflichtschulabschluss" (No compulsory school certificate), "Pflichtschulabschluss" (Compulsory school certificate), "Matura" (High school diploma), "Fachhochschule" (University of Applied Sciences), "Universität" (University), "Meisterprüfung" (Master craftsman's diploma), "Mein Abschluss ist nicht angeführt." (My degree is not listed.)

Continued on next page

Question	**Type of question**	**Categories**
"In welchem Fachbereich oder beruflichen Feld haben Sie Expertise?" (In which subject area or professional field do you have expertise?)	open-ended question	
"Welche Berufsbezeichnung haben Sie?" (What is your job title?)	open-ended question	
"Wie viele Jahre an Arbeitserfahrung haben Sie in Ihrem Berufsfeld?" (How many years of work experience do you have in your field?)	open-ended question	
"Wie viele Jahre an Arbeitserfahrung haben Sie im Bereich 'Steuerrecht'?" (How many years of work experience do you have in the field of tax law?)	open-ended question	

A.9. Transcription

A.9.1. whisper-large-v3-turbo

This code was used to transcribe the audio recordings of the interviews on a local device. The automatic speech recognition (ASR) model **whisper-large-v3-turbo** was used (Hugging Face, 2025).

```
import torch
from transformers import AutoModelForSpeechSeq2Seq, AutoProcessor, pipeline

device = "cuda:0" if torch.cuda.is_available() else "cpu"
torch_dtype = torch.float16 if torch.cuda.is_available() else torch.float32

model_id = "openai/whisper-large-v3-turbo"

model = AutoModelForSpeechSeq2Seq.from_pretrained(
    model_id, torch_dtype=torch_dtype, low_cpu_mem_usage=True
)
model.to(device)

processor = AutoProcessor.from_pretrained(model_id)

pipe = pipeline(
    "automatic-speech-recognition",
    model=model,
    tokenizer=processor.tokenizer,
    feature_extractor=processor.feature_extractor,
    chunk_length_s=30,
    batch_size=4,  # Tune for your hardware
    torch_dtype=torch_dtype,
    device=device,
    generate_kwargs={"language": "de"},  # Ensures German transcription
)

sample = "./study/P14_10072025_2/audio_ready/GMT20250710-110320_Recording.mp3"  # CHANGE path!

result_10072025_2 = pipe(sample)
print(result_10072025_2["text"])
```

A.9.2. Saving transcription

```
# Save result to a text file
with open("transcription_10072025_2_COMPLETE.txt", "w", encoding="utf-8") as f:
    f.write(str(result_10072025_2))  # Saves the entire result dictionary as text
```

A.10. RTLX

A.10.1. RTLX (German translation)

Table A.7 contains the german translation of NASA-TLX which is based on the original scale developed by Sandra Hart (Hart, 2006, p. 908) and was translated using DeepL (2025).

Titel	Endpunkte	Beschreibungen
GEISTIGER ANSPRUCH	*Sehr niedrig / Sehr hoch*	Wie viel mentale und wahrnehmungsbezogene Aktivität war erforderlich (z. B. denken, entscheiden, berechnen, erinnern, schauen, suchen usw.)? War die Aufgabe leicht oder anspruchsvoll, einfach oder komplex, streng oder nachsichtig?
KÖRPERLICHER ANSPRUCH	*Sehr niedrig / Sehr hoch*	Wie viel körperliche Aktivität war erforderlich (z. B. drücken, ziehen, drehen, steuern, aktivieren usw.)? War die Aufgabe leicht oder anspruchsvoll, langsam oder flott, träge oder anstrengend, erholsam oder mühsam?
ZEITLICHER ANSPRUCH	*Sehr niedrig / Sehr hoch*	Wie viel Zeitdruck haben Sie aufgrund der Geschwindigkeit oder des Tempos empfunden, mit dem die Aufgaben(-elemente) abliefen? War das Tempo langsam und gemütlich oder schnell und hektisch?
DURCHFÜHRUNG	*Sehr erfolgreich / Gar nicht erfolgreich*	Für wie erfolgreich halten Sie sich bei der Erfüllung der vom Versuchsleiter (oder von Ihnen selbst) gesetzten Ziele? Wie zufrieden waren Sie mit Ihrer eigenen Leistung bei der Erreichung dieser Ziele?
ANSTRENGUNG	*Sehr niedrig / Sehr hoch*	Wie sehr mussten Sie sich (geistig und körperlich) anstrengen, um Ihr angestrebtes Leistungsniveau zu erreichen?
FRUSTRATION	*Sehr niedrig / Sehr hoch*	Wie unsicher, entmutigt, irritiert, gestresst und genervt im Vergleich zu sicher, befriedigt, zufrieden, entspannt und selbstzufrieden fühlten Sie sich während der Aufgabe?

Table A.7.: Rating Scale Definitionen (RTLX german translation)

A.10.2. RTLX Results

The NASA Task Load Index (NASA-TLX) consists of 6 different subscales; Mental Demand (MD), Physical Demand (PD), Temporal Demand (TD), Frustration (FR), Performance (PE), Effort (EF) and Frustration (FR) (Hart, 2006, p. 904). Table A.8 contains the RTLX values of each participant.

ID	MD	PD	TD	PE	EF	FR	TLX
P1	90	90	50	75	75	75	75,8
P2	50	20	20	50	30	0	28,3
P3	60	5	10	60	50	40	37,5
P4	70	5	50	20	50	5	33,3
P5	20	0	70	100	5	100	49,2
P6	50	10	45	25	35	45	35,0
P7	50	0	30	50	50	50	38,3
P8	40	0	0	0	20	0	10,0
P9	40	0	20	30	10	10	18,3
P10	70	0	0	80	50	20	36,7
P11	80	20	50	20	70	40	46,7
P12	70	5	10	30	30	10	25,8
P13	20	10	10	60	50	60	35,0
P14	20	10	50	100	100	100	63,3
Mean	52,1	12,5	29,6	50,0	44,6	39,6	38,1

Table A.8.: RTLX results per participant

A.11. Chatbot Usability Questionnaire

A.11.1. Chatbot Usability Questionnaire: Dimensions

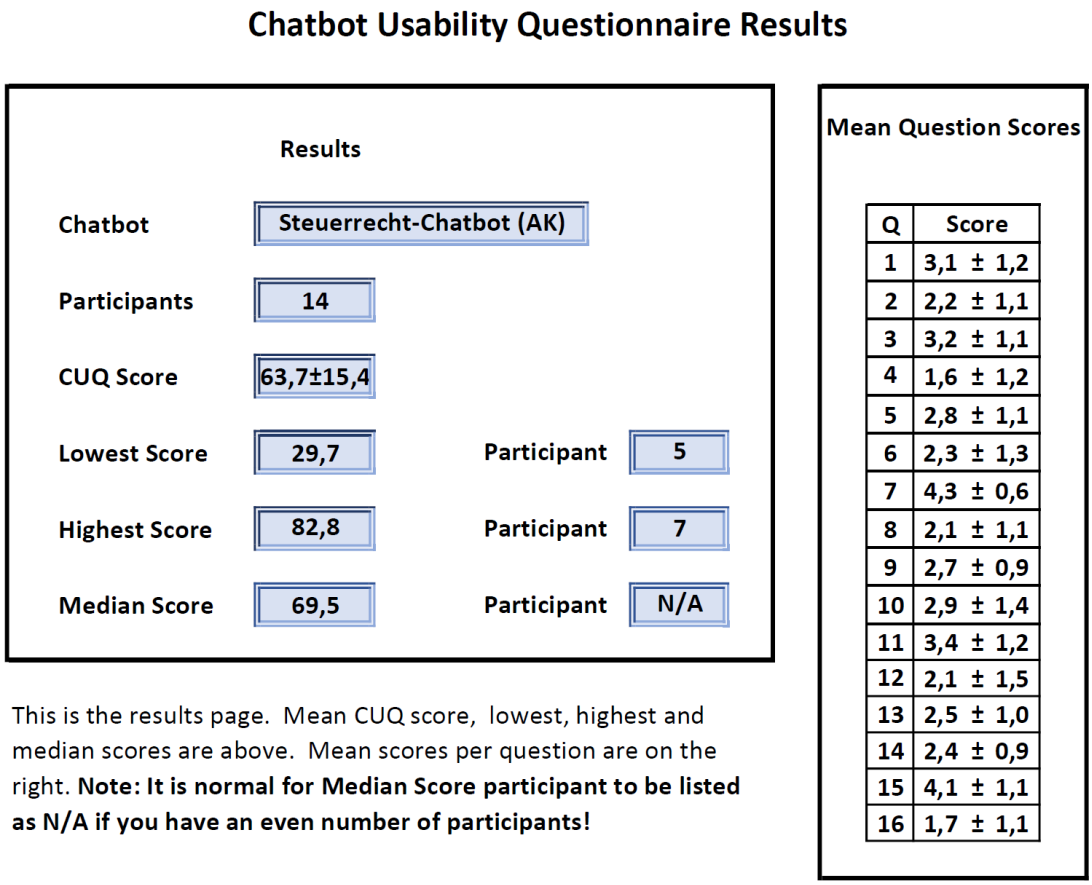
Table A.9 contains the dimensions of the Chatbot Usability Questionnaire developed by Holmes et al. (2019). CUQ was validated in Holmes et al. (2023) where dimensions were assessed (Holmes et al., 2023, pp. 332–333).

ID	Question	Dimensions
Q1	The chatbot's personality was realistic and engaging.	Personality
Q2	The chatbot seemed too robotic.	Personality
Q3	The chatbot was welcoming during initial setup.	Personality
Q4	The chatbot seemed very unfriendly.	Personality
Q5	The chatbot explained its scope and purpose well.	Onboarding
Q6	The chatbot gave no indication as to its purpose.	Onboarding
Q7	The chatbot was easy to navigate.	User Experience
Q8	It would be easy to get confused when using the chatbot.	User Experience
Q9	The chatbot understood me well.	Chatbot Understanding
Q10	The chatbot failed to recognise a lot of my inputs.	Chatbot Understanding
Q11	Chatbot responses were useful, appropriate and informative.	Chatbot Output
Q12	Chatbot responses were not relevant.	Chatbot Output
Q13	The chatbot coped well with any errors or mistakes.	Error Handling
Q14	The chatbot seemed unable to handle any errors.	Error Handling
Q15	The chatbot was very easy to use.	User Experience
Q16	The chatbot was very complex.	User Experience

Table A.9.: Chatbot Usability Questionnaire: Dimensions

A.11.2. CUQ Results

Figure A.1 is a screenshot of the Chatbot Usability Questionnaire results which are available if the official calculation xlsx-file provided by Holmes et al. (2019) is used.



Chatbot Usability Questionnaire Results

Results			
Chatbot	Steuerrecht-Chatbot (AK)		
Participants	14		
CUQ Score	63,7±15,4		
Lowest Score	29,7	Participant	5
Highest Score	82,8	Participant	7
Median Score	69,5	Participant	N/A

This is the results page. Mean CUQ score, lowest, highest and median scores are above. Mean scores per question are on the right. **Note: It is normal for Median Score participant to be listed as N/A if you have an even number of participants!**

Mean Question Scores

Q	Score
1	3,1 ± 1,2
2	2,2 ± 1,1
3	3,2 ± 1,1
4	1,6 ± 1,2
5	2,8 ± 1,1
6	2,3 ± 1,3
7	4,3 ± 0,6
8	2,1 ± 1,1
9	2,7 ± 0,9
10	2,9 ± 1,4
11	3,4 ± 1,2
12	2,1 ± 1,5
13	2,5 ± 1,0
14	2,4 ± 0,9
15	4,1 ± 1,1
16	1,7 ± 1,1

Figure A.1.: Chatbot Usability Questionnaire Results (1)

Furthermore, table A.10 shows the Mean Question Scores in combination with the questions asked.

	Q	**Score**		
Der Chatbot hatte eine glaubwürdige und sympathische Persönlichkeit.	1	3,0	±	1,2
Der Chatbot war zu roboterhaft.	2	1,8	±	0,8
Der Chatbot hat das Gespräch einladend gestartet.	3	2,8	±	1,5
Der Chatbot wirkte sehr unfreundlich.	4	2,2	±	1,8
Der Chatbot hat gut erklärt, wofür er da ist und was er kann.	5	2,2	±	0,8
Der Chatbot hat überhaupt nicht klargemacht, wofür er da ist.	6	2,2	±	1,3
Der Chatbot war leicht zu bedienen.	7	3,8	±	0,4
Es war leicht, sich bei der Bedienung des Chatbots verwirrt zu fühlen.	8	2,6	±	1,5
Der Chatbot hat mich gut verstanden.	9	2,6	±	0,5
Der Chatbot hat viele meiner Eingaben nicht erkannt.	10	3,4	±	1,3
Die Antworten des Chatbots waren nützlich, passend und informativ.	11	3,0	±	1,2
Die Antworten des Chatbots waren nicht relevant.	12	2,0	±	1,7
Der Chatbot konnte gut mit Fehlern oder Missverständnissen umgehen.	13	2,0	±	0,7
Der Chatbot hat sich schwergetan, mit Fehlern umzugehen.	14	2,6	±	1,1
Der Chatbot war sehr einfach zu verwenden.	15	4,0	±	0,7
Der Chatbot war viel zu kompliziert.	16	1,8	±	0,8

Table A.10.: Chatbot Usability Questionnaire Results (2)

Figure A.2 shows a boxplot-visualization of the Mean Chatbot Usability Questionnaire score which was recorded in this study.

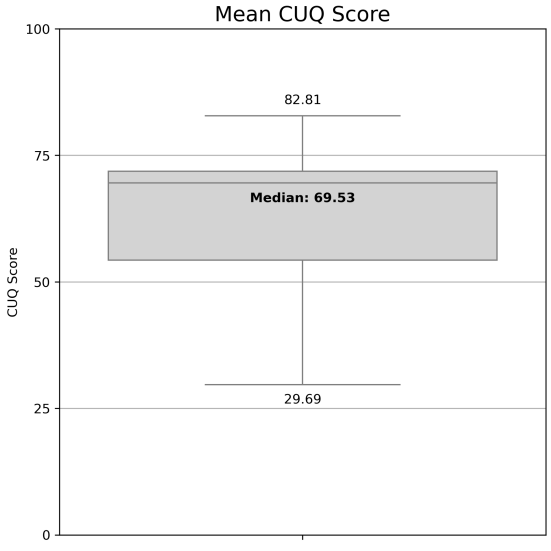


Figure A.2.: Mean Chatbot Usability Questionnaire Score

Table A.11 contains the CUQ scores for each participant. The CUQ scores were calculated using the tool provided by the authors' university (Ulster University, 2025).

Participant	CUQ Score
P1	53,1
P2	70,3
P3	57,8
P4	76,6
P5	29,7
P6	53,1
P7	82,8
P8	68,8
P9	71,9
P10	70,3
P11	62,5
P12	71,9
P13	82,8
P14	40,6

Table A.11.: CUQ Score per participant

A.11.3. Collection of CUQ scores

Table A.12 contains a collection of scores which were published in various articles (Boyd et al., 2022; Gambetta et al., 2021; Holmes et al., 2019, 2023; Larbi et al., 2022; Mesquita et al., 2025; Zhu et al., 2025). All of those publications used the Chatbot Usability Questionnaire by Holmes et al. (2019).

Chatbot	**Score**	**Source**
Calla Beauty Assistant	91.3 ± 6.3	Gambetta et al. (2021)
LLM-based VAA chatbot	84 ± 14.0	Zhu et al. (2025)
Freire	79.8	Mesquita et al. (2025)
WeightMentor	76.20 ± 11.46	Holmes et al. (2019)
WoeBot	Round 1: 72.4 ± 17.9; Round 2: 74.9 ± 13.6	Holmes et al. (2023)
Weight Loss Bot	Round 1: 69.2 ± 16.9; Round 2: 66.4 ± 16.3	Holmes et al. (2023)
ChatPal	65 ± 6.7	Boyd et al. (2022)
Steuerrecht-Chatbot (AK)	63.7 ± 15.4	-
FlyBot	Round 1: 57.3 ± 16.9; Round 2: 59.4 ± 17.7	Holmes et al. (2023)
MYA	57.4 ± 16.7	Larbi et al. (2022)

Table A.12.: Collection of CUQ scores

A.12. Perceived Trustworthiness

A.12.1. Survey data (5-point-Likert scale)

Figure A.3 contains survey results on chatbot perception and attitudes towards artificial intelligence focusing on Perceived Trustworthiness. The English version was already provided in section 5.4 (see figure 5.7).

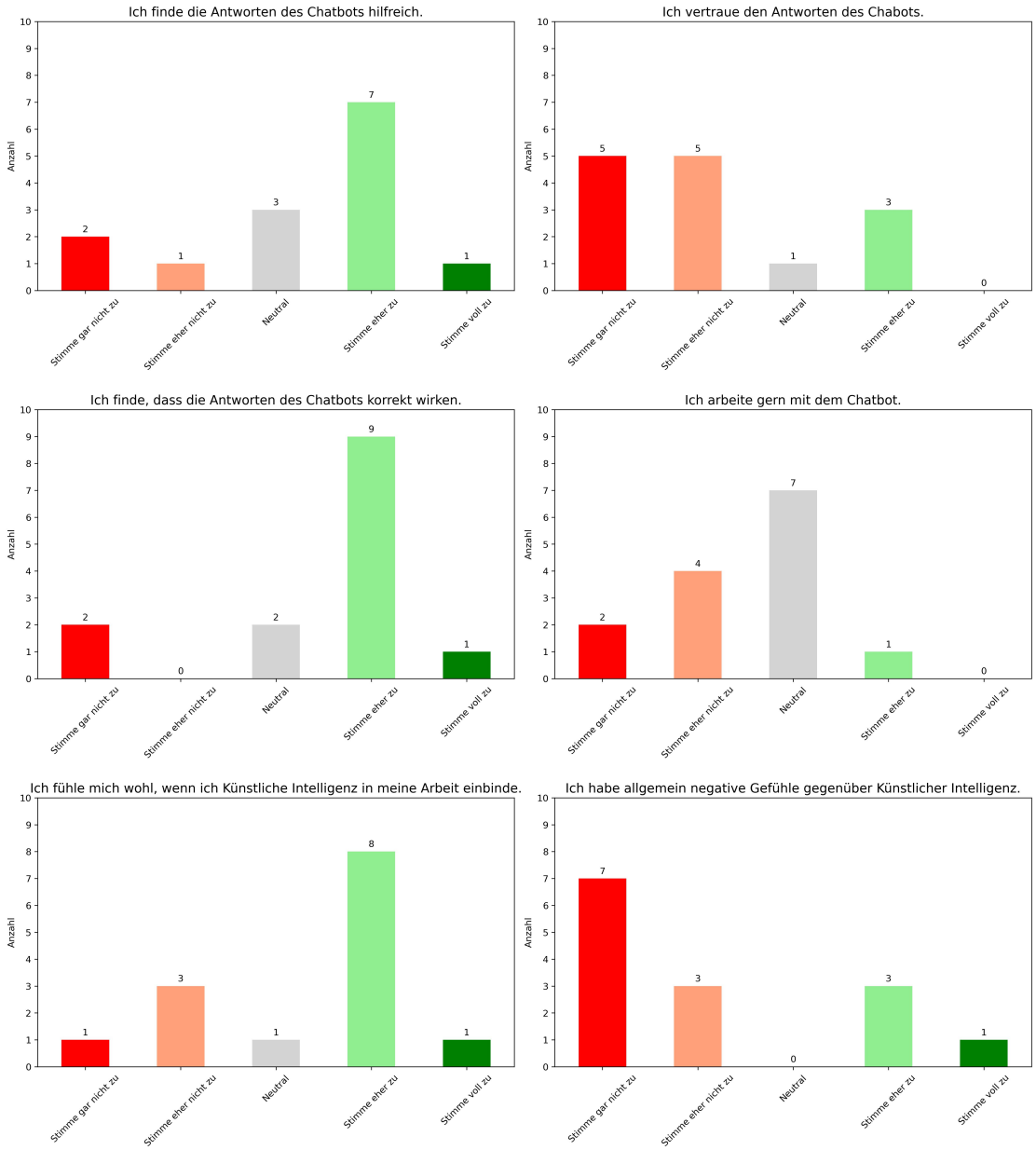


Figure A.3.: Perceived trustworthiness (survey data, 5-point-Likert scale)

A.13. Self-Reported Prior Knowledge and Task-Solving Confidence

Figure A.4 contains rating results on self-reported prior knowledge and task-solving confidence. The rating was done for each task separately. The English version was already provided in section 5.3 (see figure 5.4).

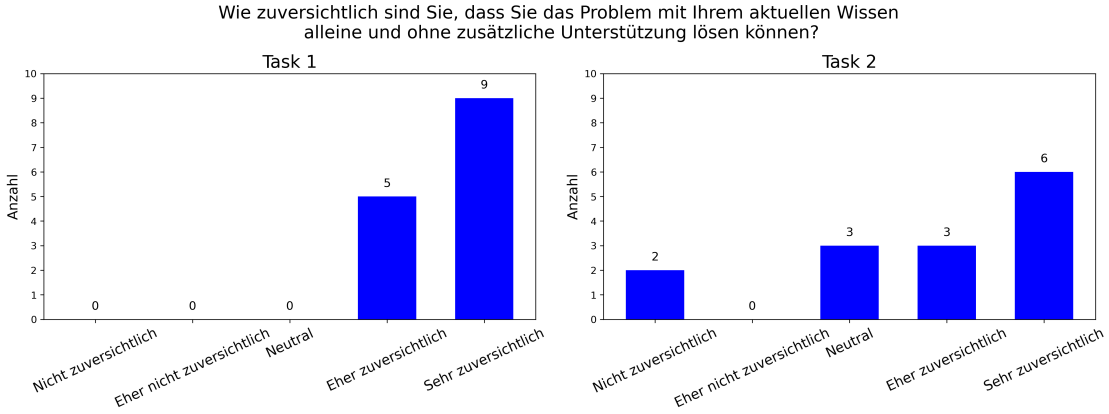


Figure A.4.: Self-Reported Prior Knowledge and Task-Solving Confidence (Tasks 1 and 2), German version

A.14. Rating Chatbot Output

Figure A.5 shows the results of the chatbot output rating. The rating was done for each task separately. The English version was already provided in section 5.3 (see figure 5.5).

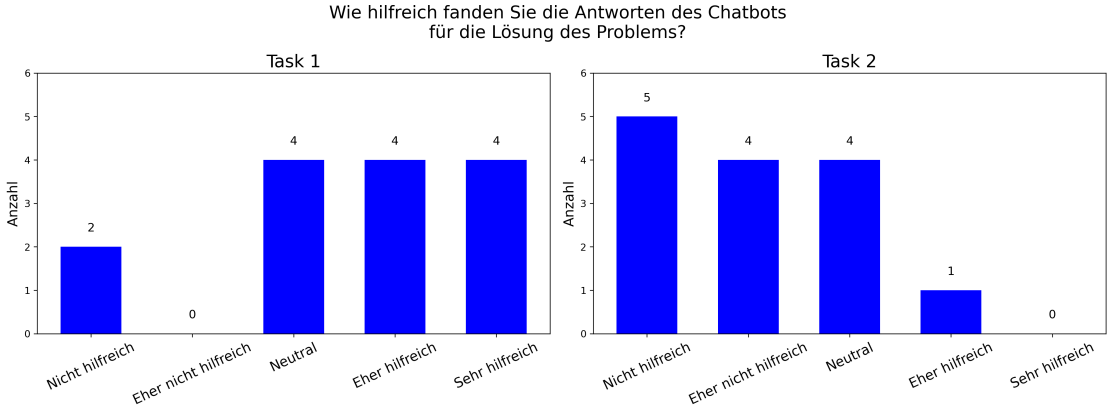


Figure A.5.: Rating Chatbot Output (Tasks 1 and 2), German version

A.15. Interview questions

Table A.13 contains translations of the interview questions grouped by topic. The interview questions were used during the semi-structured interviews.

Original question	*Translation*
Perceived Trustworthiness	
Wie sehr würden Sie dem Chatbot bei rechtlichen Entscheidungen vertrauen?	How much would you trust the chatbot with legal decisions?
Wie groß schätzen Sie das Risiko ein, dass der Chatbot falsche Informationen gibt?	How high do you estimate the risk of the chatbot providing incorrect information?
Integration into Work Routine	
In welchen Situationen können Sie sich vorstellen diesen Chatbot in Ihrer Arbeit (als Steuerrechtsexpert:in) nutzen?	In what situations can you imagine using this chatbot in your work (as a tax law expert)?
Wie könnte dieser Chatbot Ihre Arbeit effektiver machen?	How could this chatbot make your work more effective?
Könnten Sie sich vorstellen, diesen Chatbot in Zukunft regelmäßig in Ihrer Arbeit zu nutzen? Warum?	Could you imagine using this chatbot regularly in your work in the future? Why?
Welche Aufgaben halten Sie für nicht geeignet, um sie von diesem Chatbot erledigen zu lassen? Warum?	Which tasks do you consider unsuitable for this chatbot to perform? Why?
Areas for Future Development	
Welche (weiteren) Hintergrundinformationen sollte der Chatbot "verstehen" können, damit seine Antworten gut informiert wirken?	What (additional) background information should the chatbot be able to “understand” so that its responses appear well-informed?
Würden Sie es bevorzugen, wenn der Chatbot sowohl auf juristische Fachsprache als auch auf Alltagssprache reagieren könnte? Warum?	Would you prefer the chatbot to be able to respond using both legal terminology as well as everyday language? Why?
Wie würden Sie sich fühlen, wenn der Chatbot bestimmte Antworten aus Risikogründen (z.B.ethische Gründe, Datenschutz,...) markiert?	How would you feel if the chatbot flagged certain responses for risk reasons (e.g., ethical reasons, data protection, etc.)?
Wenn Sie den Chatbot in irgendeiner Weise verbessern könnten, welche Änderungen würden Sie vornehmen?	If you could improve the chatbot in any way, what changes would you make?
Wie hilfreich finden Sie die Dokumente, die in den Antworten des Bots verlinkt sind? Was gefällt Ihnen daran (nicht)?	How helpful do you find the documents linked in the bot’s responses? What do you like (or dislike) about them?

Table A.13.: Translation of interview questions grouped by topic

A.16. XAI Novice Question Bank

A.16.1. XAI Novice Question Bank: Chatbot Version

Table A.14 contains the translations of the chatbot version of the XAI Novice Question Bank. The original XAI Novice Question Bank was developed by Schmude et al. (2025).

Table A.14.: XAI Novice Question Bank: Chatbot version (Translation)

German version	English version
Chatbotkontext / Chatbot context	
K1 - Was wird mit dem Einsatz des Chatbots beabsichtigt?	What is the intention of deploying the chatbot?
K2 - Wie funktioniert der Ablauf für den Einsatz des Chatbots?	How does the deployment process of the chatbot work?
K3 - Was passiert, nachdem der Chatbot eingesetzt wird?	What are the consequences after the deployment of the chatbot?
K4 - Wie wird der Chatbot entwickelt?	How is the chatbot developed?
K5 - Wer sind die Personen, die von diesem System betroffen sein können?	Who is the chatbot's intended target group?
K6 - Wer ist für den Einsatz des Chatbots verantwortlich?	Who is responsible for the chatbot's deployment?
K7 - Ist der Einsatz des Chatbots ethisch korrekt?	Is the chatbot's deployment right?
Chatbotnutzung / Chatbot usage	
N1 - Wie wird der Chatbot bedient?	How is the chatbot operated?
N2 - Wie beeinflusst der Chatbot menschliche Beziehungen?	How is the chabot used by and on people?
N3 - Wie wird der Chatbot in bestehende Strukturen integriert?	How is the chatbot integrated into existing structures?
N4 - Wie kann der Chatbot missbraucht werden?	How can the chatbot be misused?
N5 - Wie würde der Chatbot mit [diesem Fall] umgehen?	How would the chatbot handle [this case]?
N6 - Was sind die Kosten des Chatbots?	What are the costs of the chatbot?
Daten / Data	
D1 - Mit welchen Daten wurde der Chatbot trainiert?	What kind of data was the chatbot trained on?
D2 - Woher stammen die Trainingsdaten?	What is the source of the training data?
D3 - Sind die Daten korrekt, repräsentativ und sicher?	Are the data correct / representative / safe?
Chatbotspezifikationen / Chatbot specifications	

Continued on next page

Table A.14 continued from previous page

German version	English version
S1 - Welche Merkmale berücksichtigt der Chatbot, und warum?	What does the chatbot consider and why?
S2 - Was ist der Funktionsumfang des Chatbots?	What is the scope of the chatbot's capability?
S3 - Welche Art von Ergebnissen liefert der Chatbot?	Which kind of output does the chatbot give?
S4 - Wie lernt der Chatbot?	How does the chatbot learn?
S5 - Nach welcher Logik funktioniert der Chatbot?	What is the chatbot's overall logic?
S6 - Welche Art von Algorithmus wurde verwendet?	Which kind of algorithm was used?
S7 - Welche Fehler könnte der Chatbot wahrscheinlich machen?	What kind of mistakes is the chatbot likely to make?
S8 - Wo liegen die Grenzen des Chatbots?	What are the limitations of the chatbot?
S9 - Wie verlässlich sind die Vorhersagen vom Chatbot?	How reliable are the predictions of the chatbot?
S10 - Was bedeutet eigentlich [ein bestimmtes Konzept aus dem Bereich Machine Learning]?	What does [a machine learning concept] mean?

A.16.2. XAI Novice Question Bank: German version

Figure A.6 was shown to participants to select questions.

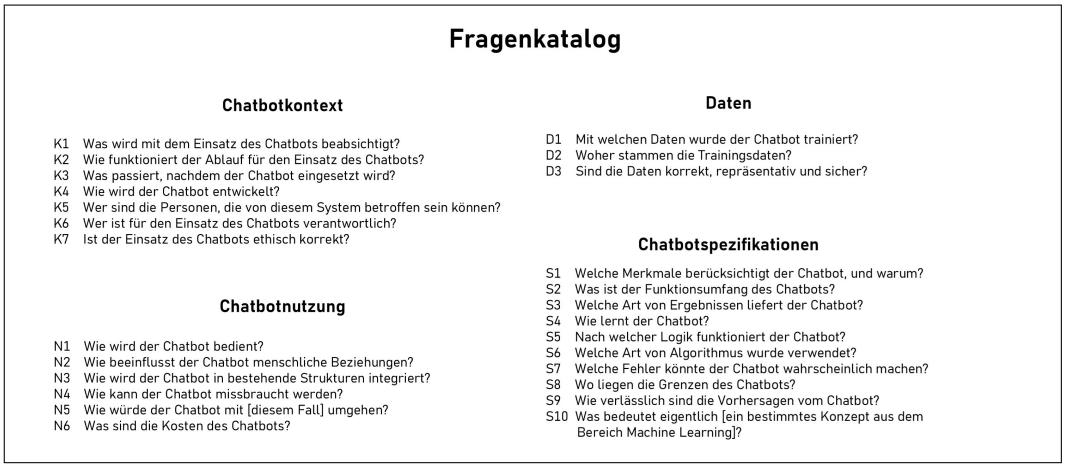


Figure A.6.: German version of the XAI Novice Question Bank

A.16.3. XAI Novice Question Bank: Selected Questions

Figure A.7 shows which questions from the chatbot version of the XAI Novice Question Bank were selected by the participants.

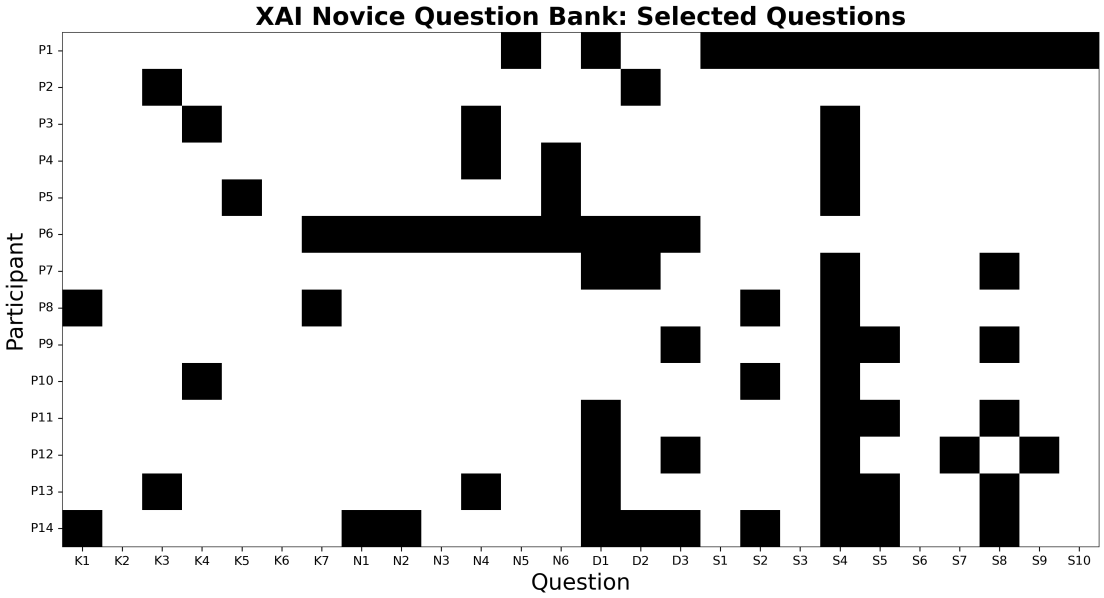


Figure A.7.: XAI Novice Question Bank: Selected Questions

A.16.4. XAI Novice Question Bank (P1 - P14): Most frequently selected questions

Figure A.8 shows how often the questions from the chatbot version of the XAI Novice Question Bank were selected by participants. The categories "Chatbot context", "Chatbot usage", "Data", and "Chatbot specifications" are shown in different colors.

- **Red questions** belong to the "Chatbot context" category.
- **Blue questions** belong to the "Chatbot usage" category.
- **Green questions** belong to the "Data" category.
- **Orange questions** belong to the "Chatbot specifications" category.

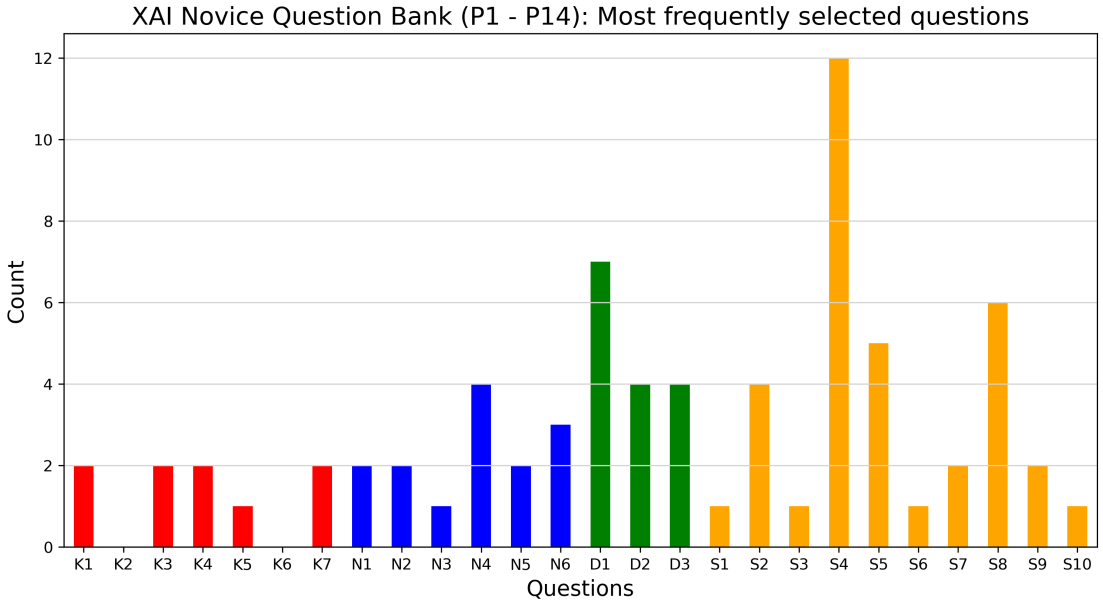


Figure A.8.: XAI Novice Question Bank (P1 - P14): Most frequently selected questions

A.16.5. XAI Novice Question Bank: P1 - P14 (by category)

Figure A.9 shows how many questions per category were selected by each participant. The categories are represented by the first letter of the respective German translation of the category in the numbering system: K = "Chatbot context" (derived from "Chatbotkontext"); N = "Chatbot usage" (derived from "Chatbotnutzung"); D = "Data" (derived from "Daten"); S = "Chatbot specifications" (derived from "Chatbotspezifikationen").

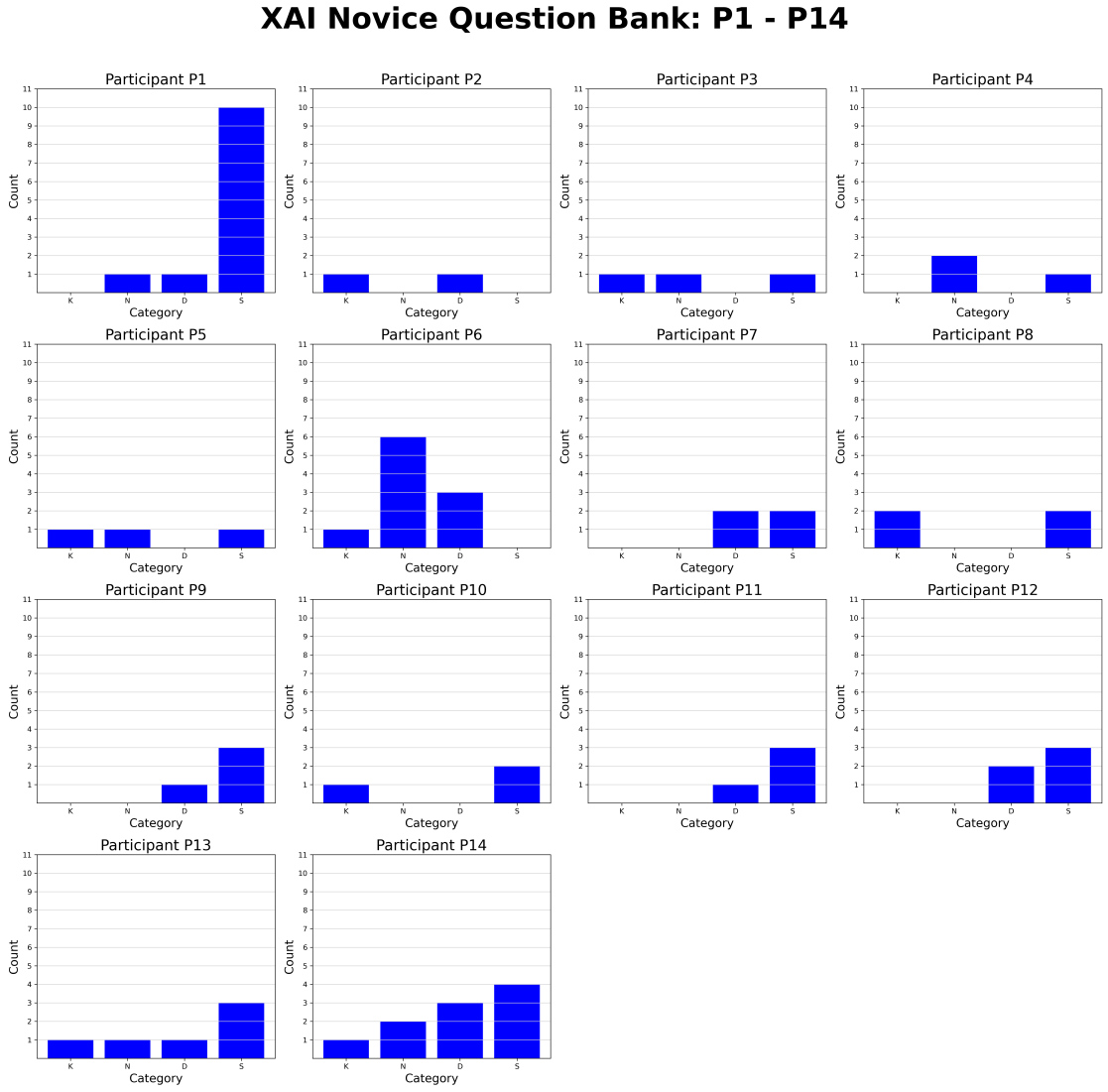


Figure A.9.: XAI Novice Question Bank: P1 - P14 (by category)

A.17. Thematic Networks

As described in chapter 4. The 14 interviews were analyzed using the five phases of thematic coding outlined by Robson (2024, p. 575). This process resulted in the identification of six themes. In addition, participants' *"information needs"* were analyzed deductively using the XAI Novice Question Bank developed by Schmude et al. (2025), as participants were asked to select questions from the XAI Novice Question Bank during the interviews. The following themes were identified: *"areas of application"*, *"evaluating chatbot output"*, *"expressing feelings towards the chatbot"*, *"possible improvements"*, *"prior knowledge"*, and *"trust"*. Figure A.10 contains the main themes of this analysis.

For each main theme, a thematic network was created that includes the related sub-themes. Figures relating to the main themes can be found on the subsequent pages:

- Areas of application: see figure A.11
- Evaluating chatbot output: see figure A.12
- Feelings towards the chatbot: see figure A.13
- Information needs: see figure A.14
- Possible improvements: see figure A.15
- Prior knowledge: see figure A.16
- Trust: see figure A.17

A.17.1. Main themes of the chatbot evaluation

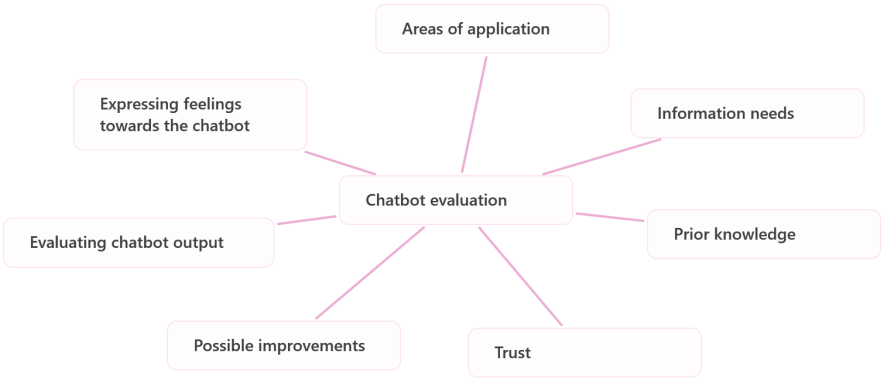


Figure A.10.: Thematic network: Chatbot evaluation

A.17.2. Areas of application

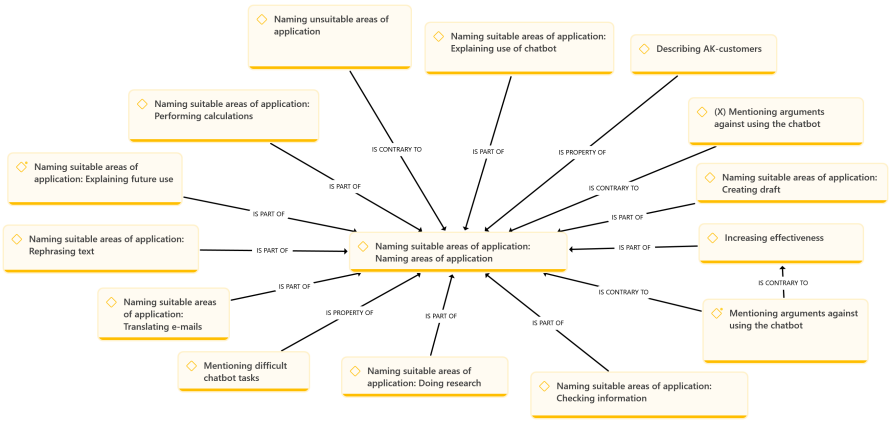


Figure A.11.: Thematic network: Areas of application

A.17.3. Evaluating chatbot output

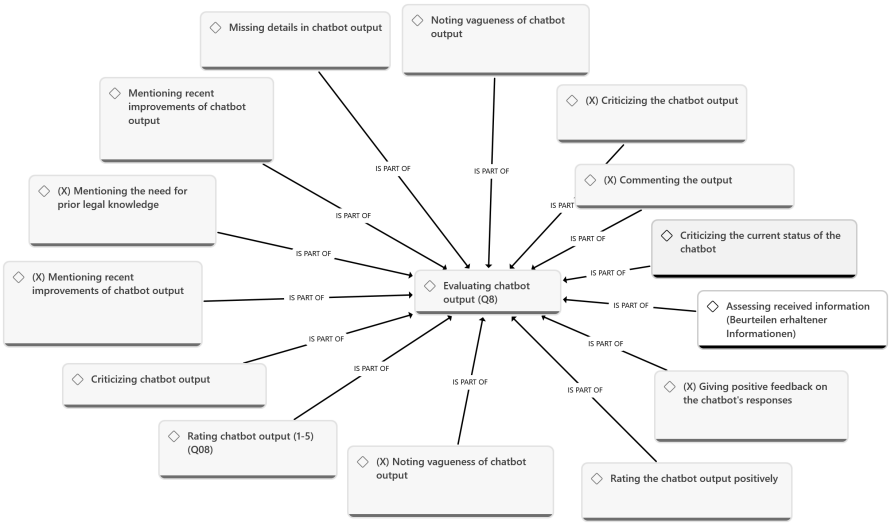


Figure A.12.: Thematic network: Evaluating chatbot output

A.17.4. Feelings towards the chatbot

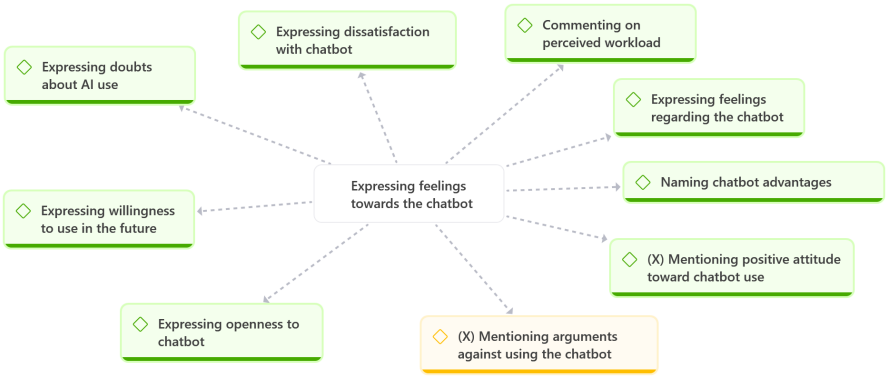


Figure A.13.: Thematic network: Feelings towards the chatbot

A.17.5. Information needs

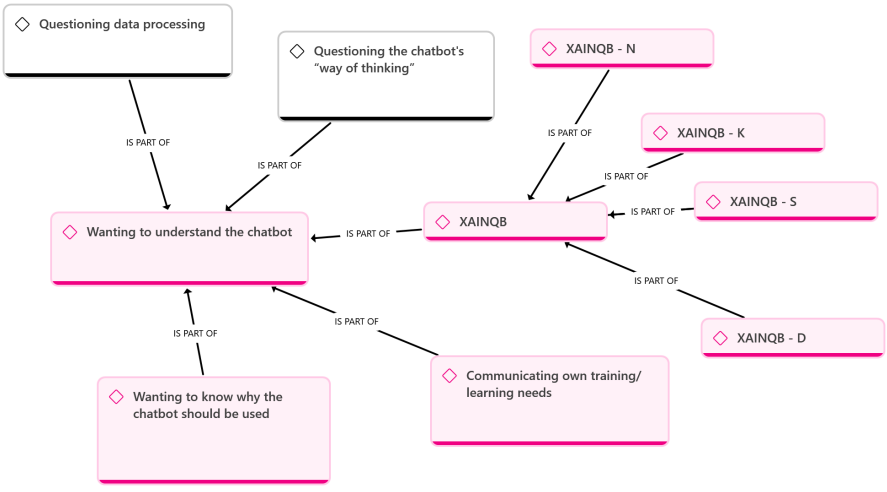


Figure A.14.: Thematic network: Information needs

A.17.6. Possible improvements

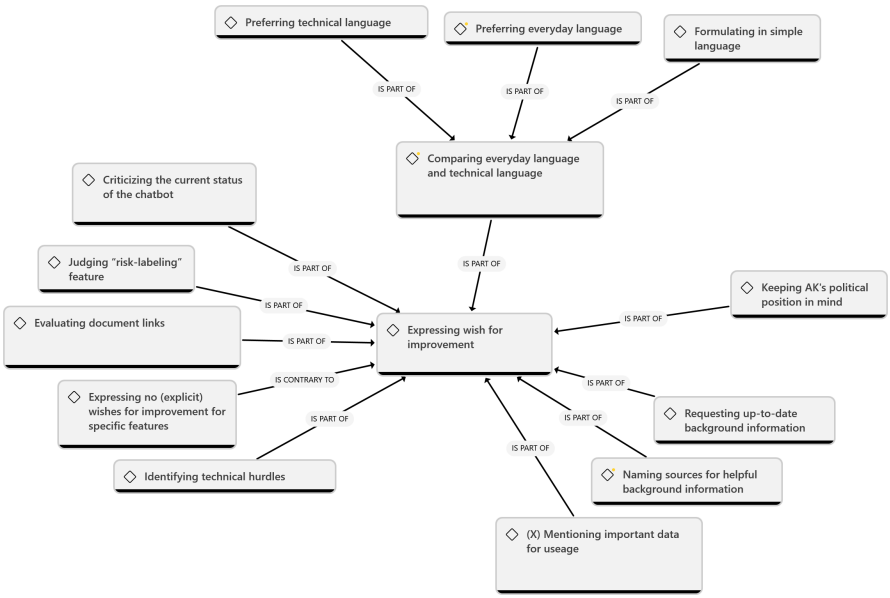


Figure A.15.: Thematic network: Possible improvements

A.17.7. Prior knowledge

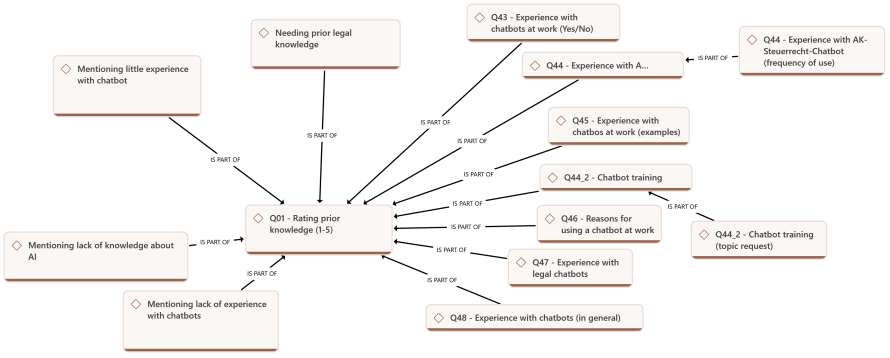


Figure A.16.: Thematic network: Prior knowledge

A.17.8. Trust

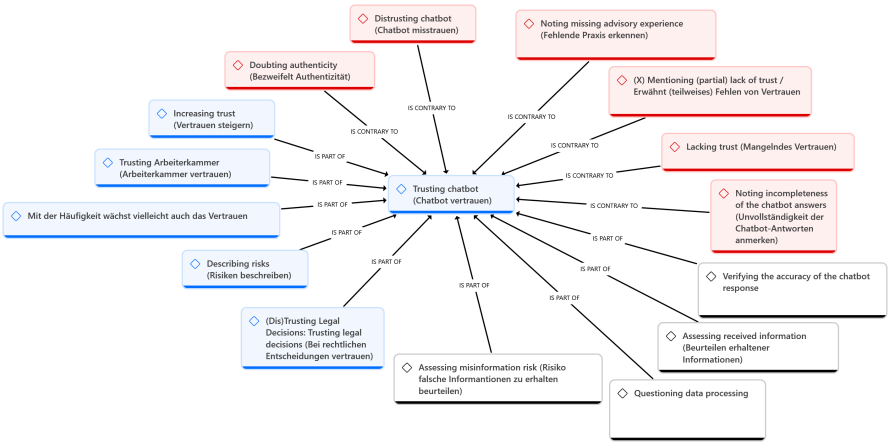


Figure A.17.: Thematic network: Trust

A.18. Qualitative results

A.18.1. Useful documents

Table A.15 shows the documents mentioned by the study participants that they consider helpful for improving the chatbot output. These documents could be used to improve the chatbot's knowledge base.

Useful documents
Commentaries by the Federal Finance Court (*Bundesfinanzgericht, BFG*) on income tax law (*Einkommensteuerrecht*) and wage tax guidelines (*Lohnsteuerrichtlinien*)
Access to the FINDOK database (Bundesministerium für Finanzen, 2025a)
Access to RDB documents (MANZ'sche Verlags- und Universitätsbuchhandlung GmbH, 2025)
Articles by LexisNexis (LexisNexis Österreich, 2025) and Linde Verlag (Linde Verlag, 2025)
Double taxation agreements (Bundesministerium für Finanzen, 2025b)
Instructions for complaints (*Familienbonus*, tax assessment) and special cases
Articles on insolvency remuneration funds (*Insolvenzentgeltfonds*)
Solutions already formulated by consultants (question-answer pairs)
Additional documents on international tax issues

Table A.15.: Collection of useful documents

A.18.2. Sentiments

Negative sentiments

Table A.16 summarizes the main themes that elicited negative sentiment from participants. Furthermore, these qualitative data themes are linked to quantitative data. The data from the survey question *"I trust the chatbot"* is used. Participants rated this statement on a 5-point Likert scale ranging from *"strongly disagree"* to *"strongly agree"*.

Participant	**Lack of trust**	**Incompleteness/ Vagueness of chatbot answers**	**I trust the chatbot's answers.** (5-point Likert scale)
P1	X		Disagree
P2	X	X	Strongly disagree
P3		X	Agree
P4	X		Disagree
P5	X	X	Strongly disagree
P6	X	X	Disagree
P7			Agree
P8	X		Strongly disagree
P9		X	Agree
P10			Neutral
P11	X	X	Disagree
P12	X		Disagree
P13	X		Strongly disagree
P14	X		Strongly disagree

Table A.16.: Negative sentiments per participant

Positive sentiments

Table A.17 summarizes the main themes that elicited positive sentiment from participants.

Participant	Willingness to use the chatbot in the future	Doing research, creating drafts	Notes correct answers positively
P1	X		
P2	X	X	
P3	X		X
P4	X		
P5	X		
P6			X
P7	X		X
P8	X		
P9	X		X
P10	X		
P11	X	X	X
P12	X		
P13	X	X	X
P14		X	

Table A.17.: Positive sentiments per participant

A.19. Data storage

The following data were stored on Nextcloud cloud storage provided by the Research Group Visualization and Data Analysis of the University of Vienna:

- Recordings of the interviews
- Interview transcriptions
- Survey data and survey analysis
- Atlas.ti project files
- Signed consent forms
- Supplementary material (e.g., illustrations, etc.)