universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Evaluating the Validity of LLM-Based Psychological Assessments

A Comparative Study Using the Attributional Style Questionnaire

verfasst von | submitted by

Jakob Armin Friedrich Möller BSc

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

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 840
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Psychologie
Betreut von | Supervisor:	Univ.-Prof. Mag. Dr. Claus Lamm

Abstract

Recent advancements in artificial intelligence, specifically Large Language Models (LLMs) like OpenAI's GPT-4o, offer innovative opportunities to overcome limitations of traditional psychological assessments. This thesis evaluates the validity of an LLM-enhanced version of the Attributional Style Questionnaire (ASQ), comparing dynamically personalized and AI-generated vignettes against traditional static scenarios. Employing a within-subject design with 112 participants, this study investigates comparability with traditional methods, test-retest reliability, emotional engagement, and user experience. Attribution patterns could be partly replicated but differed between conditions. Emotional engagement notably improved with personalized negative scenarios, enhancing participant satisfaction and perceived usability without increasing cognitive load. The findings suggest the potential of integrating structured and standardized AI-driven personalization in psychological assessments for improved ecological validity, as well as user experience. Avenues for future research to leverage generative AI without compromising validity in similar assessments are outlined.

Jüngste Fortschritte im Gebiet der künstlichen Intelligenz, insbesondere sogenannten Large Language Models (LLMs) wie OpenAIs GPT-4o, bieten innovative Möglichkeiten, die Grenzen traditioneller psychologischer Erhebungen zu überwinden. Diese Arbeit evaluiert die Validität einer LLM-erweiterten Version des Attributional Style Questionnaire (ASQ) und vergleicht dynamisch personalisierte und KI-generierte Vignetten mit traditionellen statischen Szenarien. Unter Verwendung eines Within-Subject-Designs mit 112 Teilnehmer*innen untersucht diese Studie die Vergleichbarkeit mit traditionellen Methoden, die Test-Retest-Reliabilität, das emotionale Engagement und die Benutzererfahrung. Die Attributionsmuster konnten teilweise repliziert werden, unterschieden sich jedoch zwischen den Bedingungen. Das emotionale Engagement wurde durch personalisierte Negativszenarien deutlich verbessert, was die Zufriedenheit der

Teilnehmer und die wahrgenommene Benutzerfreundlichkeit erhöhte, ohne die kognitive Belastung zu erhöhen. Die Ergebnisse deuten auf das Potenzial der Integration strukturierter und standardisierter KI-gesteuerter Personalisierung in psychologische Beurteilungen hin, um die ökologische Validität und die Nutzererfahrung zu verbessern. Es werden Möglichkeiten für künftige Forschungen zur Nutzung generativer KI ohne Beeinträchtigung der Validität in ähnlichen Tests aufgezeigt.

1. Introduction

In recent years, significant advancements in artificial intelligence (AI), particularly in the fields of machine learning (ML) and natural language processing (NLP), have transformed methodologies and tools across numerous scientific disciplines. Among these advancements, the development of Large Language Models (LLMs) such as OpenAI's ChatGPT, Microsoft's Copilot, and Google's Gemini has been particularly influential. These sophisticated computational systems generate coherent and contextually appropriate text by training on vast and diverse datasets, thereby enabling novel approaches in scientific research and practical applications across many fields. The integration of AI technologies into psychological research and mental health interventions has also been rapidly advancing, opening new possibilities for psychological support and diagnostic methods. AI-driven mental health chatbots, for instance, have shown preliminary efficacy in reducing symptoms of anxiety and depression through structured conversational interactions (Fulmer et al., 2018; Gual-Montolio et al., 2022). Additionally, recent research highlights the utility of AI for early detection of mental health symptoms and personalized interventions, improving accessibility and reducing barriers to care (Corcoran & Cecchi, 2020; Currey & Torous, 2023). Rollwage et al. (2023) and Weizenbaum et al. (2024) further demonstrate participant preferences for AI-assisted psychological assessments due to perceived enhancements in clarity, user engagement, and reduced cognitive demand. Despite these promising developments, the integration of AI into psychological assessment methods is accompanied by substantial challenges

and limitations. AI-driven assessments have encountered issues such as unpredictability in generated content, ethical risks concerning participant distress, compromised ecological validity, and varying user acceptance levels due to perceived artificiality or mistrust in AI systems (Rollwage et al., 2023; Fulmer et al., 2018). These challenges underscore the necessity for rigorously controlled and validated implementations that maintain methodological integrity and ethical standards (Demszky et al., 2023; Currey & Torous, 2023). To address these limitations, this thesis proposes a novel methodological approach, integrating generative AI capabilities into a traditional psychological assessment framework. Specifically, the thesis focuses on an LLM-enhanced adaptation of a vignette-based assessment, wherein AI-generated, personalized vignettes are dynamically incorporated into a structured and strictly controlled assessment procedure. Unlike fully autonomous AI-based systems, this study employs a scripted chatbot interface that utilizes generative AI selectively at predetermined interaction points, primarily for scenario content generation. By maintaining strict procedural control and clearly delineated AI interactions, this method aims to ensure more standardized outputs and minimizes the risk of unpredictable or inappropriate AI behavior. Particularly, this study uses the Attributional Style Questionnaire (ASQ) as a preliminary case study to assess the feasibility of such an approach.

1.1. Vignette-based assessments and Attribution Theory

Traditional vignette-based assessments, which present hypothetical scenarios to participants to elicit cognitive and emotional responses, have a long-standing history in psychological research and typically rely on standardized, static vignettes. This standardized structure inevitably limits their relevance to individual participants' unique backgrounds and experiences, thus limiting ecological validity. Moreover, repeated administration of identical vignettes may lead to familiarity bias, diminishing the effectiveness of such tools in longitudinal research designs. Furthermore, it stands to reason that decisions made about hypothetical vignettes may not accurately reflect real-world decisions (Hainmueller et al., 2015). While such assessments are widely used, their reliance on standardized, static vignettes might therefore be a significant limitation.

Within this broader context, attribution theory and the Attributional Style Questionnaire (ASQ) serves as a specific illustrative case. Attribution theory, first introduced by Fritz Heider in 1958, offers a comprehensive framework for understanding how individuals interpret causality across internal-external, stable-unstable, and global-specific dimensions. Attributional styles identified through this theoretical lens have been thoroughly studied and (among other things) are critically associated with resilience, motivation, and susceptibility to conditions like anxiety and depression (Ball, McGuffin, & Farmer, 2008). The ASQ, developed by Martin Seligman et al. (1982), operationalizes this theory through a vignette-based methodology, assessing individuals' attributional styles using hypothetical scenarios. While the ASQ has been extensively validated and widely utilized, its inherent reliance on static, standardized vignettes limits its range of applicability and arguably reduces its effectiveness, particularly in longitudinal designs.

1.2. Purpose and Research Objectives

Given the aforementioned promising theoretical basis for integrating AI technologies into psychological assessments, at this stage there is an evident need for feasible real-world approaches to develop and incorporate such new methodologies into current standards. Such endeavors will first and foremost require empirical validation to establish their efficacy, reliability, and psychometric integrity. This thesis directly addresses these empirical gaps by proposing and systematically evaluating merely one such integrative approach for potential feasibility. Namely, an LLM-based adaptation of the ASQ is suggested and proposed as a case study. The primary aims of this thesis are to:

1. Evaluate whether the LLM-based ASQ generates results comparable to those obtained from traditional ASQ assessments.
2. Investigate the test-retest reliability of LLM-based assessments over repeated administrations.
3. Assess whether personalized, AI-generated vignettes evoke stronger emotional responses compared to static, standardized vignettes.

4. Examine whether personalized formats improve participant engagement, user experience, and overall satisfaction compared to conventional assessment formats.

Consequently, this study is guided by the following hypotheses:

H1: Attributional within-subject scores obtained from the LLM-based ASQ will not significantly differ from those generated by the traditional ASQ.
H2: Attributional within-subject ratings of the LLM-based ASQ will remain consistent across repeated administrations.
H3: Personalized, LLM-generated vignettes will elicit stronger emotional responses from participants compared to standardized static vignettes.
H4: The LLM-based ASQ will significantly enhance participant engagement and overall user satisfaction relative to the traditional ASQ

.

To effectively address these hypotheses and implement such an LLM-based assessment, this thesis develops and introduces a novel approach using a chatbot-interface that integrates AI-generated content within a structured and standardized psychological assessment framework. Unlike fully autonomous AI-driven assessments, this method retains strict procedural control by using a pre-scripted chatbot format that only implements the LLM in highly specific instances, namely, the generation of scenario content for vignettes. This approach allows the study to leverage the creative potential of generative AI while still ensuring that assessments remain largely standardized. The proposed chatbot's role is not to independently conduct assessments but rather to serve as a structured interface through which vignettes are dynamically personalized. By restricting AI involvement to content generation and embedding its outputs within a strictly controlled conversation, this study introduces a safeguard that mitigates the risks commonly associated with AI-based psychological tools. The scripted nature of the chatbot ensures that interactions follow a predefined sequence, maintaining reliability while allowing for the flexible adaptation of scenario content based on individual participant characteristics. This hybrid model aims to introduce a compromise, balancing the adaptability of AI-driven assessments with the methodological integrity of traditional psychometric evaluation. In order to evaluate the established hypotheses, this newly

developed approach will consequently be tested and compared to the “regular” ASQ (Peterson et al., 1982).

Beyond validating the immediate application of this particular LLM-based assessment, this thesis serves to explore broader implications and future opportunities arising from AI integration in psychological research methodologies. Ultimately, this research seeks to contribute foundational evidence regarding the integration of AI tools in psychological assessment.

2. Methods

2.1. Participants

Participants for this study were recruited via Prolific. To be eligible, participants had to be at least 18 years old, possess proficiency in English, be in current employment (at least part-time), and meet the required mental health criteria. These criteria were assessed using the Patient Health Questionnaire-9 (PHQ-9) and the Generalized Anxiety Disorder-7 (GAD-7) during the pre-screening phase. Participants exceeding predetermined cut-off scores indicative of severe depression or anxiety (i.e. a combined PHQ-9 and GAD-7 score >25) were excluded from the study to minimize distress during the assessments. A total of 148 participants were pre-screened, 112 of which (58 female, 54 male) were recruited. The participants’ age ranged from 18 to 79 years (*M*= 36.43). In addition to mental health criteria, further demographic data (profession, and personal interests) were collected. Upon completion of each session, participants received monetary compensation in line with Prolific’s fair pay guidelines.

2.2. Study Design

The study employed a within-subject design, wherein each participant completed assessments under three distinct conditions to examine the impact of LLM-generated vignettes on attributional style assessments. The first condition involved the personalized LLM-ASQ, where participants completed dynamically generated vignettes that were tailored to each participant’s demographic profile, using their profession and personal interests. The goal of this condition was to

assess whether customized scenarios increased ecological validity and engagement while maintaining the psychometric integrity of the ASQ's core attributional dimensions. The second condition, the non-personalized LLM-ASQ, utilized an alternative approach in which participants received AI-generated vignettes that were not tailored to their personal background. Instead, these scenarios were generated using general prompts designed to maintain broad applicability without incorporating user-specific details. This condition allowed for an additional comparative analysis, distinguishing between the effects of vignette personalization and potential effects of the mere modified format of the assessment. In condition 3 the participants received the original, standardized Attributional Style Questionnaire using pre-written, fixed vignettes. This condition served as a baseline to compare the novel LLM-based approaches against. For sample vignettes of each type see *Table 1*.

Each participant completed all three conditions (see *Fig. 1*) over the course of three separate sessions, with a three-day interval between each session. The order of conditions was pseudo-randomized across different participant cohorts to control for potential sequence effects. Due to technical restraints of the Prolific platform a full randomization was not feasible. Each session lasted approximately 30 minutes and consisted of 12 vignettes. In the non-personalized LLM-ASQ condition (Condition 2), participants received a mix of six personalized and six non-personalized vignettes, facilitating preliminary longitudinal analyses of personalized vignette stability and validity over multiple sessions. Prior to beginning the study, participants underwent pre-screening assessments using the PHQ-9 and GAD-7 to ensure that individuals experiencing severe depressive or anxiety symptoms were excluded from participation. During each session, participants not only completed the respective ASQ assessment but also provided ratings on emotional impact and perceived relevance of each vignette, allowing for an evaluation of how different vignette types influenced emotional engagement. At the end of each session, participants completed a usability questionnaire, rating the perceived ease of use, clarity of instructions, level of engagement, satisfaction, and cognitive load) to determine the user experience associated with each condition.

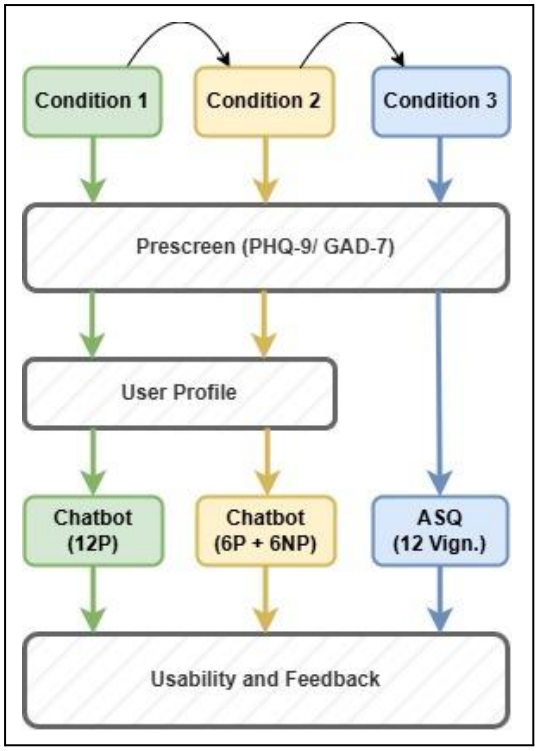

Figure 1. Flowchart of the study design: Participants completed three conditions in pseudo-randomized order in which they either encountered 12 AI-generated personalized vignettes (Condition 1), 6 personalized as well as 6 non-personalized vignettes (Condition 2), or 12 fully standardized vignettes (Condition 3).

Vignette Type	Sample Vignette
Personalized (Cond. 1 & 2)	“In an educational setting, you organize a collaborative project where students create a mural representing historical events. Colleagues praise your innovative approach, and the principal commends the project's success during a staff meeting.”
Non-personalized (Cond. 2)	“You present a project proposal during a team meeting. A senior colleague criticizes your approach, questioning its feasibility, and the discussion becomes heated as others join in."
ASQ (Cond. 3)	"You do a project that is highly praised."

Table 1. Sample vignettes for each condition. Personalized and non-personalized vignettes generated with “Profession” context for a history teacher, as well as standardized ASQ item.

2.3. Materials

2.3.1. Psychometric Instruments

The study utilized established psychometric instruments alongside custom-developed rating scales. The primary measure was the ASQ, a standardized tool designed to assess attributional styles across the dimensions *Locus of control* (internal/external), *Stability* (stable/unstable), and *Globality* (global/specific) dimensions. Depending on the condition, participants either encountered the established traditional version or the experimental LLM-based version of the tool. Custom 10-point Likert-type scales were implemented to assess emotional responses elicited by each vignette. These scales specifically measured emotional impact, positive valence, and negative valence of each vignette. These scales

provided a rudimentary yet consistent means of comparing emotional engagement across different assessment conditions in order to establish a preliminary validation of the approach. Additionally, these scales intended to further encourage participants to vividly imagine each scenario. For improved interpretability these 10-point scales allowed for ratings in decimal steps (i.e. 0.1), effectively allowing for a more continuous and intuitive response behavior.

2.3.2. Usability and Cognitive Load Measures

Post-assessment usability questionnaires were adapted from established usability heuristics, particularly the *NASA Task Load Index* (NASA-TLX). These measures were tailored specifically to reflect the unique aspects of the assessment procedure, evaluating the dimensions clarity of instructions, ease of use, overall satisfaction, level of engagement, and cognitive load. Participants rated their experiences on similar 10-point scales, providing comparative data on the user experience across conditions. These scales also allowed for ratings in 0.1 steps.

2.3.3. Chatbot

The key component of the assessment methodology was the specially developed LLM-integrated chatbot. The chatbot interface was built using Gradio, a Python-based library hosted via Hugging Face. OpenAI's GPT-4o model was employed as the backend language model for the targeted scenario content generation, generation of attributional interpretation options, and basic natural language interpretation of participants' topic choices. Importantly, the chatbot operated within a strictly scripted framework, with clearly outlined interaction stages that minimized the risk of unpredictable AI behavior (see *Fig. 2*). Participants engaged with the chatbot in a structured dialogue format, selecting their preferred topic (based on either their profession or a personal interest) at the start of each vignette. GPT-4o then generated scenario content adhering to explicit prompt-engineering guidelines. These guidelines structured and outlined concise scenarios with clear restrictions regarding content as well as formal structure such as a maximum output length of 50 tokens, second-person language, strictly external event references, and avoidance of personal identifiers or internal emotional content.

LLM usage was thus selectively and strategically integrated into the assessment process and strictly limited to pre-defined interaction points. The chatbot did not independently interact with participants beyond these controlled scenarios. User inputs throughout the assessment primarily involved numeric or discrete choices, which minimized the need for complex natural language processing and thereby improved consistency and reliability in data collection. A full version of the chatbot can be accessed here: https://tiliah-condition2.hf.space/. For the specific prompts, also see Appendix A4. The complete code for each of the three conditions is publicly accessible via https://huggingface.co/Tiliah.

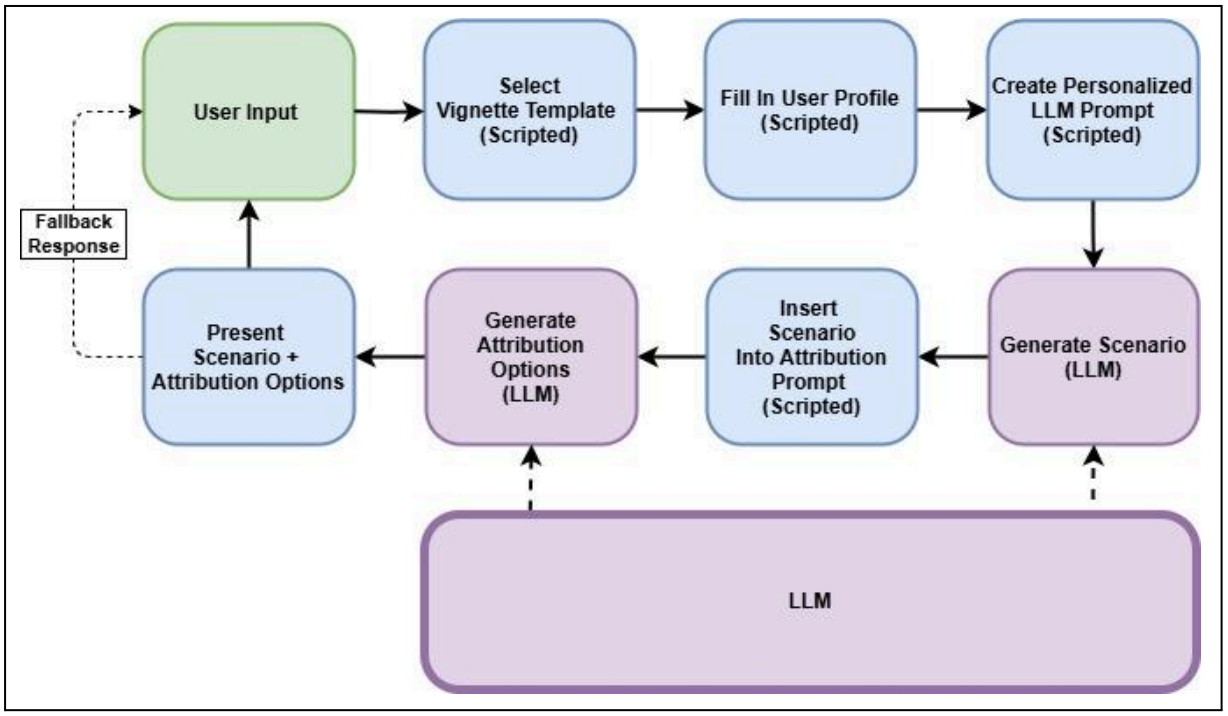

Figure 2. Flowchart of the technical loop depicting the generation of a single vignette: Scripted steps (blue) handle vignette template selection, user-profile insertion, and prompt assembly; the LLM (purple) generates both the personalized scenario and the corresponding attribution response options. The combined scenario and options are presented to the participant (green). A fallback is triggered if content is inappropriate.

2.4. Procedure

Participants completed three online assessment sessions, each approximately 30 minutes in duration, separated by intervals of three days to allow for short-term stability analysis. The sequence of the three assessment conditions (traditional ASQ, personalized LLM-ASQ, and non-personalized LLM-ASQ) was pseudo-randomized across participant cohorts to control for potential order effects. Each session began with an informed consent process, clearly explaining the voluntary nature of

participation, data privacy measures, and participants' right to withdraw at any time without penalty. Following consent, participants underwent a preliminary mental health screening using the PHQ-9 and GAD-7 scales. After screening, participants were asked to provide demographic details, which were used for vignette personalization in conditions 1 and 2. The LLM-ASQ sessions involved participants interacting with the chatbot through a structured conversational interface (*Fig. 3*). At the start of each vignette, participants selected their preferred scenario topic (either their profession, personal interest, or, in the personalized condition exclusively, a "new interest."). The chatbot then generated scenario content dynamically, tailored according to the specific assessment condition. Participants evaluated each vignette's personal relevance using a numeric scale ("*Have you personally experienced something similar to this scenario before? Please rate the similarity on a scale from 1 (Not at all similar) to 10 (Extremely similar).*"). If the relevance rating fell below a threshold (< 6), participants were prompted to briefly explain why the scenario felt inappropriate or irrelevant. This feedback was then incorporated into a revised scenario generation process, after which the participant re-rated the new scenario's relevance.

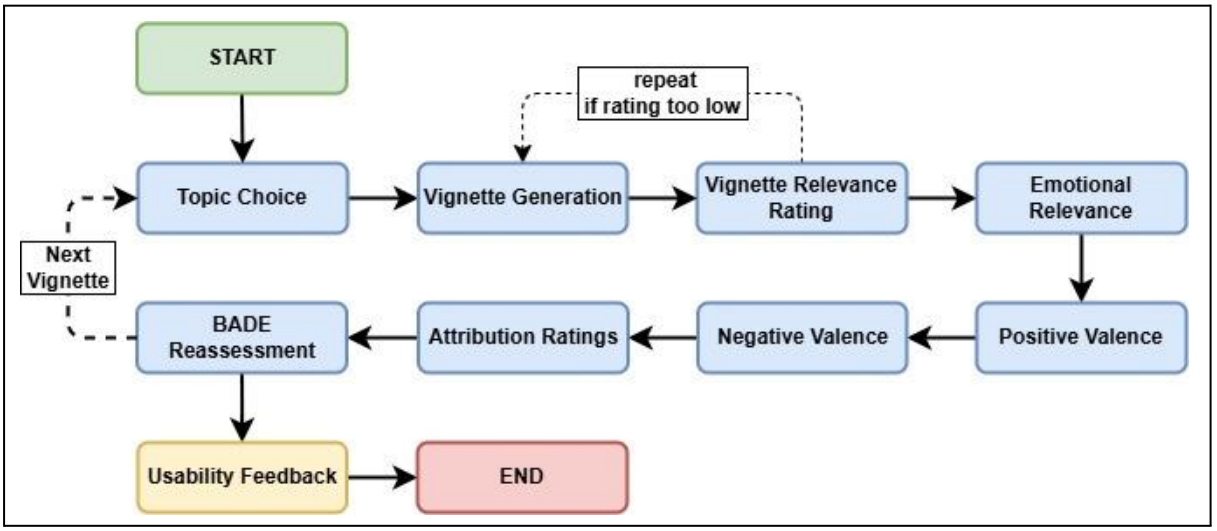

Figure 3. Flowchart of the chatbot interaction loop: Participants select a topic, receive a personalized or generic vignette (with regeneration if relevance is low), rate relevance and emotional valence, complete attribution judgments, and finally provide usability feedback.

Following scenario relevance ratings, participants assessed each vignette's emotional impact, positive valence, and negative valence on separate 10-point Likert scales. Subsequently, participants evaluated attributional dimensions. For each scenario, the chatbot presented two attribution options per dimension (*Locus of*

Control, *Stability*, *Globality*), generated by the *GPT-4o* model. Participants indicated their agreement with the respective attribution options using continuous sliders anchored between clearly opposing attributional poles (see *Fig. 4*). The generation of the presented attribution options was designed in a way that ensured that the options for the *Stability* dimension, as well as the *Globality* dimension were always in line with the participant's chosen internal/external attribution for the situation. After initial ratings, the chatbot "revealed" additional scenario information (i.e. a previously unknown further development of the presented vignette) generated to specifically challenge the participant's initial internal/external attribution. Participants were then asked to reassess their initial internal/external attribution accordingly. This additional step was introduced to measure the latent persistence of the attribution (i.e. stability of attribution style even under disconfirmatory conditions). Following this, participants were asked to pick a new topic for the next vignette from the above mentioned options and were moved on to the next scenario. Each session consisted of a total of 12 scenarios and concluded with participants completing usability and cognitive load questionnaires. Participants were then redirected to Prolific in order to receive their compensation.

2.5. Ethical Considerations

The study received ethical approval from the university ethics committee prior to participant recruitment. Several safeguards were implemented to protect participant well-being and data privacy. Initial mental health screening ensured that individuals experiencing significant depressive or anxious symptoms were excluded from participation to prevent potential distress from the assessment content. Participants were informed explicitly of their right to withdraw at any point without penalty, reinforcing voluntary participation. Additionally, clear contact information for study coordinators was provided, enabling participants to seek support or clarifications as needed. To further minimize risk, vignette content generation was strictly controlled through explicit prompt-engineering rules designed to avoid deeply personal, emotionally distressing, or traumatic content. The chatbot interaction was fully scripted, with AI functionalities confined to pre-determined scenario generation and interpretation tasks, thereby reducing the risk of unpredictable or inappropriate AI behavior. Data confidentiality was maintained through pseudonymization (i.e.

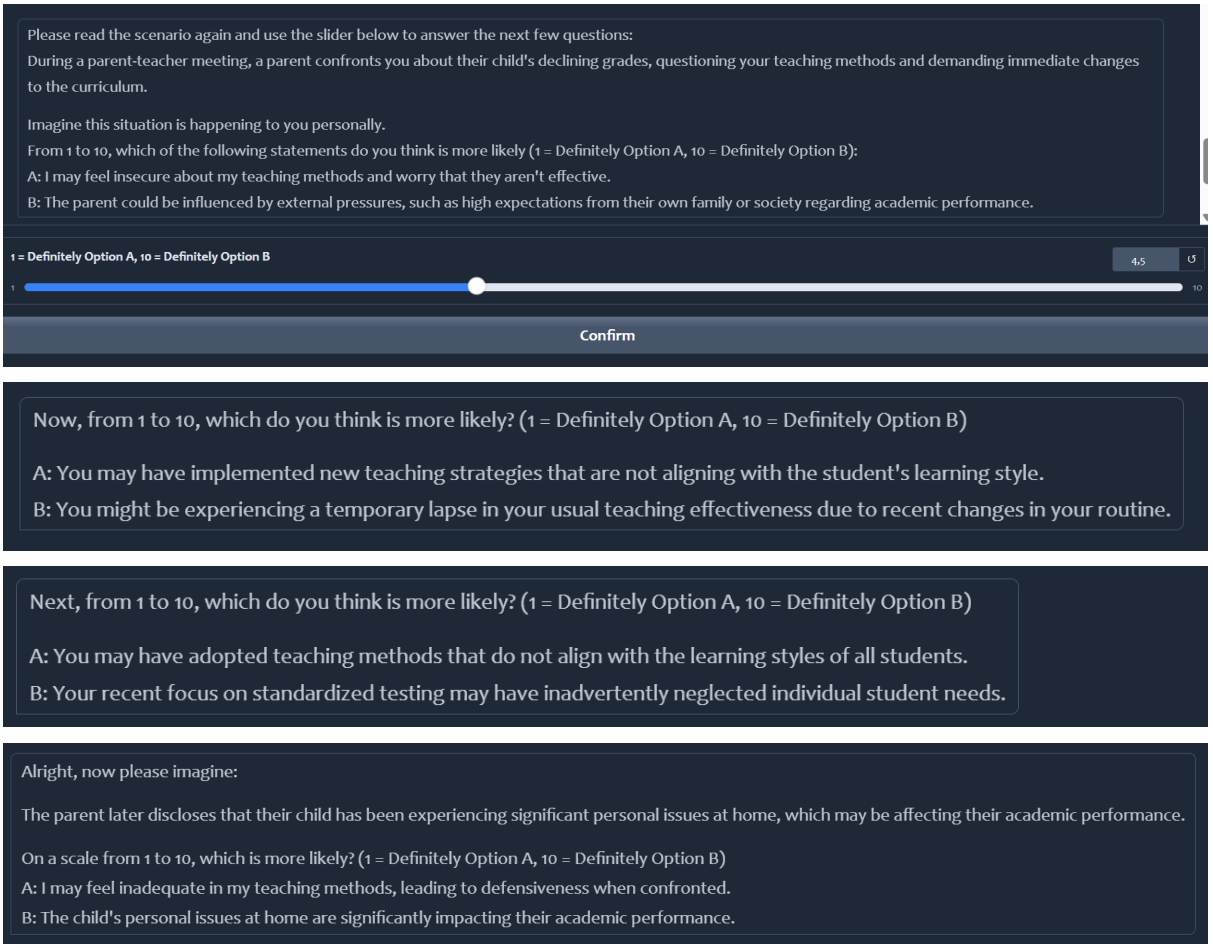

Figure 4. Sample vignette with respective attribution items: Personalized vignette for a school teacher. Top to bottom: 1. Vignette with Internality assessment; 2. Stability assessment; 3. Globality assessment; 4. BADE-reveal with Internality reassessment.

Prolific IDs) and secure storage via private Hetzner-hosted S3 servers. All collected data were managed according to relevant data protection standards, ensuring participants' anonymity and privacy were maintained.

2.6. Validation Process

Given the exploratory nature of integrating LLMs within psychological assessment frameworks, multiple rudimentary validity-checking procedures were employed. Prompt engineering constraints were initially established and pilot-tested to ensure GPT-4o-generated scenarios were comparable to the standardized structure and thematic focus of traditional ASQ vignettes. Preliminary validation was implemented through manual review, by evaluating generated vignettes for consistency, adherence to prompt guidelines, and content validity relative to traditional ASQ scenarios. Participant feedback mechanisms, such as relevance ratings and

qualitative feedback when scenarios were initially rated poorly during pilot-testing, provided additional validation data, confirming or requiring revision to enhance the ecological validity of the scenarios. While emotional impact, valence, and usability scales were not clinically validated beforehand, their consistent application across conditions facilitated valid comparative analyses within the study's scope, allowing presumably meaningful interpretations of condition-specific differences.

2.7. Longitudinal Considerations

Preliminary longitudinal elements were incorporated into the study design to evaluate short-term stability across repeated assessments. Participants completed three separate sessions (two of which included personalized scenarios) to enable initial evaluations of test-retest reliability for the personalized LLM-generated scenarios. The dynamic and real-time generation of unique vignette scenarios for each participant during the LLM-based assessment sessions intended to reduce potential memory effects or familiarity biases. The pseudo-randomized ordering of assessment conditions across different participant cohorts further mitigated potential carry-over or order effects.

2.8. Measures

2.8.1. Attributional Style

In the traditional ASQ condition, participants rated fixed hypothetical scenarios provided by the original questionnaire via Likert-scales and were asked to freely fill in their own subjective interpretation regarding the cause of the scenarios. In the personalized and non-personalized LLM-ASQ conditions, scenarios were dynamically generated by OpenAI's GPT-4o model. In these conditions participants rated each scenario using continuous slider scales anchored between explicitly defined opposing attribution poles (e.g., "*Definitely Option A*" (internal) vs. "*Definitely Option B*" (external)). Additionally, the stability of the attribution style was measured by introducing contrary new information about the scenario and consequently asking participants to reassess their interpretation at the end of every vignette in conditions 1 and 2.

2.8.2. Emotional and Valence Measures

Participants provided ratings regarding the emotional impact and valence of each vignette. Specifically, participants rated emotional impact ("*How much emotional impact did this scenario have on you?*"), positive valence ("*How much joy, excitement, or satisfaction did this scenario evoke in you?*"), and negative valence ("*How much discomfort or tension did you feel while imagining this scenario?*"). Each of these dimensions was assessed using a numeric scale from 1 (None) to 10 (A lot).

2.8.3. Scenario Relevance and Personalization

Participants provided numerical ratings on the perceived personal relevance of each vignette on a scale from 1 to 10. If relevance ratings fell below a specified threshold (< 6), participants were prompted to explain briefly why the scenario felt inappropriate. These qualitative responses were utilized for iterative adjustments to scenario generation, and consequently created new vignettes until sufficient relevance was ensured.

2.8.4. Usability and Cognitive Load

At the conclusion of each session, participants completed usability questionnaires inspired by the NASA-TLX and adjusted specifically for the assessment context. These items included ratings on clarity, ease of use, engagement, satisfaction, and perceived cognitive load, each assessed on a numeric scale from 1 (low) to 10 (high). These items included: "*How clear were the instructions?*", "*How engaging was the experience?*", "*How mentally demanding was the assessment?*", and "*Overall, how satisfied were you with this assessment?*". Lastly, participants were given the option to provide additional comments or feedback at the end of each session before concluding the assessment.

2.8.5. Bias Against Disconfirmatory Evidence (BADE)

Although not part of the traditional ASQ, the Bias Against Disconfirmatory Evidence (BADE) was included as an additional measure to assess the stability of attributional judgments when confronted with new, contradictory information. After the initial attribution rating for each vignette, a brief piece of disconfirmatory information was revealed that was designed to directly contradict the participant's originally chosen

attribution. Participants then provided a second attribution rating on the same slider scale. The BADE effect for each item was calculated as the absolute difference between the initial and revised attribution scores. Larger absolute shifts reflect greater cognitive flexibility in updating attributions, while minimal shifts indicate persistence of the original explanatory style. This procedure allowed examination of how personalized contexts affect the stability of attributional responses. Mean BADE-effect scores were computed separately for negative and positive scenarios to explore whether attributional rigidity varied by emotional valence.

2.8.6. Additional Measures

Initial demographic data were collected, including participants' age, profession, and personal interests. Participants also completed the PHQ-9 and GAD-7.

2.9. Analysis

For Hypothesis 1, which stated that the attributional results from the LLM-based ASQ would not significantly differ from those of the traditional ASQ, paired-sample t-tests were conducted, comparing attributional scores between the LLM-based and traditional conditions. Before analysis, assumptions of normality were verified using the Shapiro-Wilk test. If significant deviations from normality were detected, non-parametric alternatives, such as the Wilcoxon signed-rank test, were employed. Effect sizes (Cohen’s d) were calculated to quantify the magnitude of observed differences. Bayesian approaches were integrated as supplementary analyses for hypothesis testing, particularly given the exploratory nature of comparing traditional and LLM-generated assessment tools. Bayesian paired-samples t-tests provided evidence supporting equivalence (or differences) between the traditional ASQ and LLM-generated conditions, offering interpretative advantages over frequentist statistics, such as directly quantifying support for the null hypothesis. To examine Hypothesis 2 regarding the test-retest reliability of the LLM-based ASQ, repeated-measures analyses were conducted. Intraclass correlation coefficients (ICCs) were used to evaluate test-retest reliability across sessions. Additionally, repeated measures ANOVA was considered to investigate consistency in

attributional responses over multiple sessions, with assumptions such as sphericity tested using Mauchly's test. When the assumption of sphericity was violated, Greenhouse-Geisser corrections were applied. Hypothesis 3 examined whether personalized LLM-generated vignettes evoke stronger emotional responses compared to non-personalized and traditional conditions. A mixed ANOVA approach was employed, incorporating condition (personalized vs. non-personalized vs. traditional ASQ) as a within-subject factor, and emotional response ratings (emotional impact, positive valence, negative valence) as dependent variables. Prior to ANOVA, the assumption of homogeneity of variance was assessed with Levene's test. Hypothesis 4, evaluating whether the interactive and adaptive format of the LLM-based ASQ significantly enhances user engagement and usability, was addressed through descriptive statistics and ANOVA on usability questionnaire scores (clarity, ease of use, engagement, satisfaction, cognitive load). Additionally, thematic qualitative analyses were conducted on open-ended feedback to identify patterns in user experiences and further substantiate quantitative findings.

Finally, descriptive analyses were conducted for demographic data, usability measures, and vignette personalization ratings. Pearson correlations were employed to explore relationships between perceived vignette relevance, emotional engagement, and usability metrics across conditions. These exploratory correlations helped to further elucidate relationships between personalization, emotional engagement, and user experience. Effect sizes (Cohen's d, partial eta squared) were calculated for all significant results to quantify the magnitude of observed effects, providing a clearer understanding of the practical relevance of the study's findings.

3. Results

3.1. Comparability with Traditional Method

The primary hypothesis evaluated whether the Attributional Style Questionnaire (ASQ) scenarios generated by the large language model (LLM) produced attribution scores comparable to those obtained through traditional ASQ methods. Paired-samples t-tests were employed to compare attributional dimension scores (Internal-External, Stable-Unstable, and Global-Specific) between participants'

experiences in the interactive LLM-based condition (Condition 1) and the traditional static vignettes condition (Condition 3). Significant differences emerged in two out of the three attribution dimensions examined (see *Fig. 5*). The Internal-External dimension showed a significant difference between conditions ($p < .001$, $d \sim 0.64$), supported by a compelling Bayes Factor (~53300), strongly suggesting a difference in the ratings. This indicates that participants attributed events differently with regards to *Internality* when engaging with the personalized, AI-generated scenarios compared to the traditional vignettes. Conversely, the Stable-Unstable dimension revealed no significant difference ($p = .79$), reinforced by a negligible Bayes Factor ($BF = 0.13$), thereby strongly supporting that no meaningful differences existed between the conditions in this particular dimension. The Global-Specific dimension demonstrated a significant difference ($p = .002$, $d = -0.36$) with a moderate Bayes Factor ($BF = 13.82$), suggesting moderate evidence for distinct attributional patterns between the LLM-based and traditional assessments. Reliability analyses and exploratory factor analyses were conducted to examine internal consistency and the factor structure of these attributional dimensions. Cronbach's alpha values consistently exceeded the threshold of .70 for all three dimensions across conditions, affirming adequate reliability. Notably, due to the small number of items, these values were lowered ($\alpha = .48 - .73$) after accounting for positive and negative tones separately (see Appendix A5). However, a split-half test furthermore indicated a slightly higher internal consistency for the dimensions Internality and Stability, but not Globality in condition 1, compared to condition 3 (Appendix A5.3, A5.4). Exploratory factor analyses validated a clear and theoretically coherent factor structure aligned closely with the original ASQ constructs (see *Fig. 6*).

3.2. Test-Retest Reliability

To evaluate the temporal stability and reliability of attribution scores generated by personalized, LLM-generated scenarios, test-retest analyses were conducted by comparing participants' attribution scores for the fully personalized vignettes across

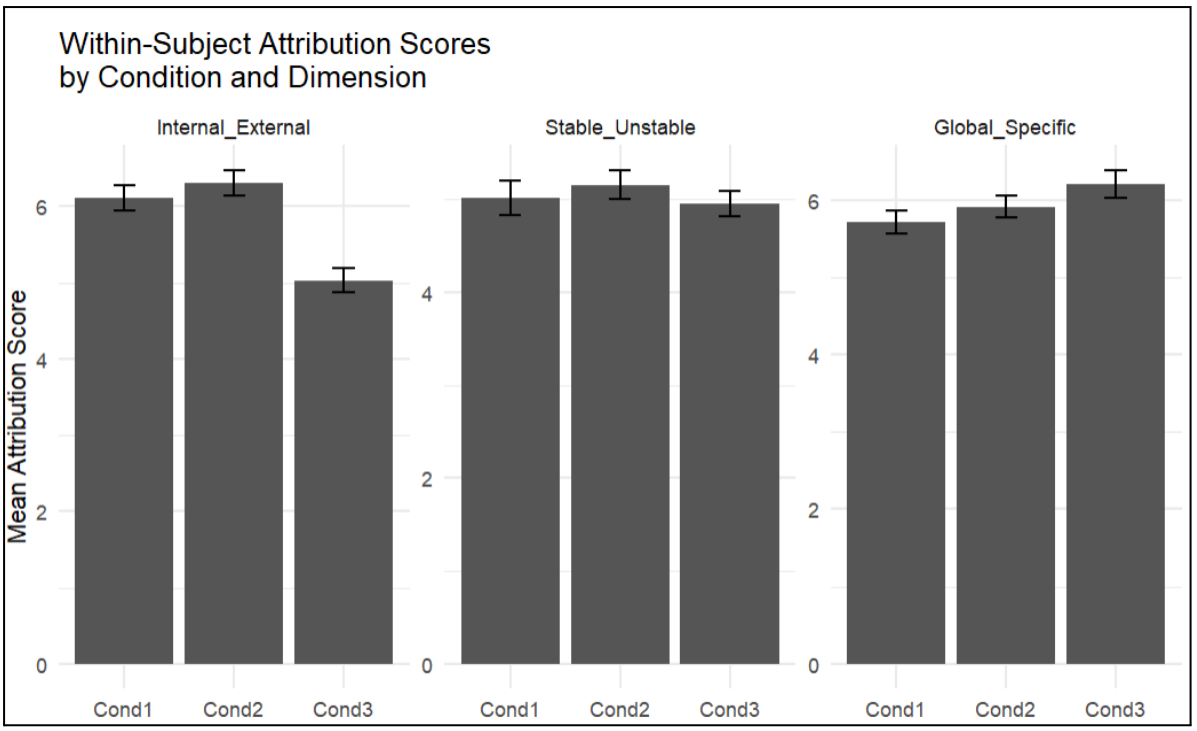

Figure 5. Bar-plot comparing the attributional ratings between conditions: Values closer to zero reflect ratings in line with the first pole (i.e. internal/stable/global), while higher ratings indicate ratings in line with the second pole (external/unstable/specific).

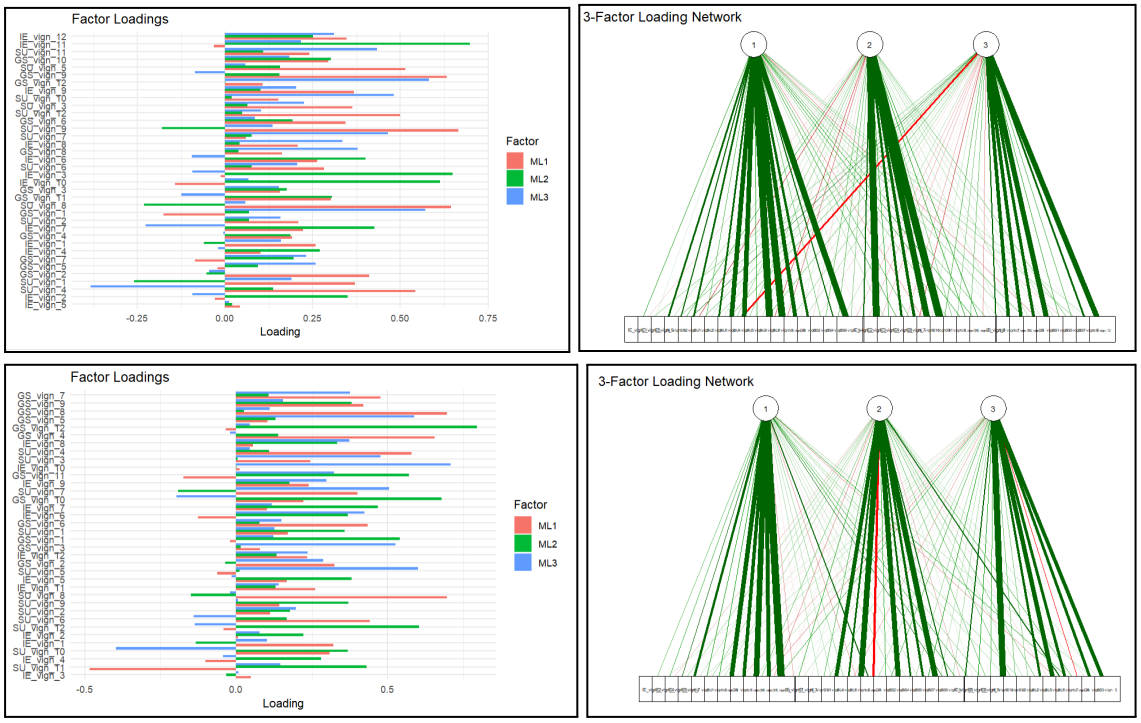

Figure 6. Exploratory factor analysis: Factor loadings and structure for condition 1 (top row) and condition 3 (bottom row). Individual loadings for each attribution rating per vignette (1-12), labeled IE (Internality), SU (Stability) or GS (Globality) respectively.

two sessions (Condition 1 and Condition 2). Intraclass Correlation Coefficients (ICC) indicated moderate temporal stability for attribution scores. The internal-external dimension showed an ICC of .45 ($p < .001$), indicating moderate consistency between the two assessments. Similarly, the stable-unstable dimension yielded an ICC of .47 ($p < .001$), demonstrating comparable stability. The global-specific dimension revealed slightly lower temporal reliability, with an ICC of .39 ($p < .001$), still within the moderate range. Repeated-measures ANOVA examining the three attributional dimensions and assessment conditions (personalized vs. traditional) revealed a highly significant main effect of Dimension ($F(1.77, 128.9) = 23.68$, $p < .001$), indicating clear differentiation between attribution dimensions. However, neither the main effect of Condition ($F(1, 73) = 1.62$, $p = .21$) nor the Dimension - Condition interaction ($F(1.93, 141.2) = 0.33$, $p = .71$) were significant, suggesting that personalization did not uniformly alter overall attribution scores or dimension patterns.

Additionally, composite scores were constructed to summarize each participant's attributional patterns across all personalized scenarios. For each session and valence (positive vs. negative), the three raw dimensions were averaged, yielding one mean score per dimension for positive vignettes and one for negative vignettes. Those three positive-valence means were then averaged into a single "positive attribution" score, and the three negative-valence means were similarly averaged into a single "negative attribution" score. Furthermore, proxies for participants' explanatory style (the difference between internality scores in positive and inverted negative vignettes), as well as optimism (the difference between stability scores in positive versus negative vignettes) were calculated. Positive scores for the explanatory style composite indicates a self-enhancing attribution style, while negative scores indicate a self-deprecating style. Positive scores in the optimism metric indicate a tendency to attribute causes of positive events as more stable than those of negative events, while scores <0 imply more stable cause attributions for negative events. These composites were included to allow preliminary comparison of overall attributional style and its consistency across sessions. Positive composites yielded moderate temporal stability ($r = .45$, $p < .001$), similar to negative composites ($r = .46$, $p < .001$). Notably, the explanatory style composite demonstrated slightly

higher reliability (r = .53, p < .001). Repeated-measures ANOVA further supported these findings, revealing no significant differences between the two sessions across any attribution dimension or composite scores (all p-values > .05). This absence of significant mean differences across sessions indicates (albeit moderate) stability of personalized vignette assessments over short intervals.

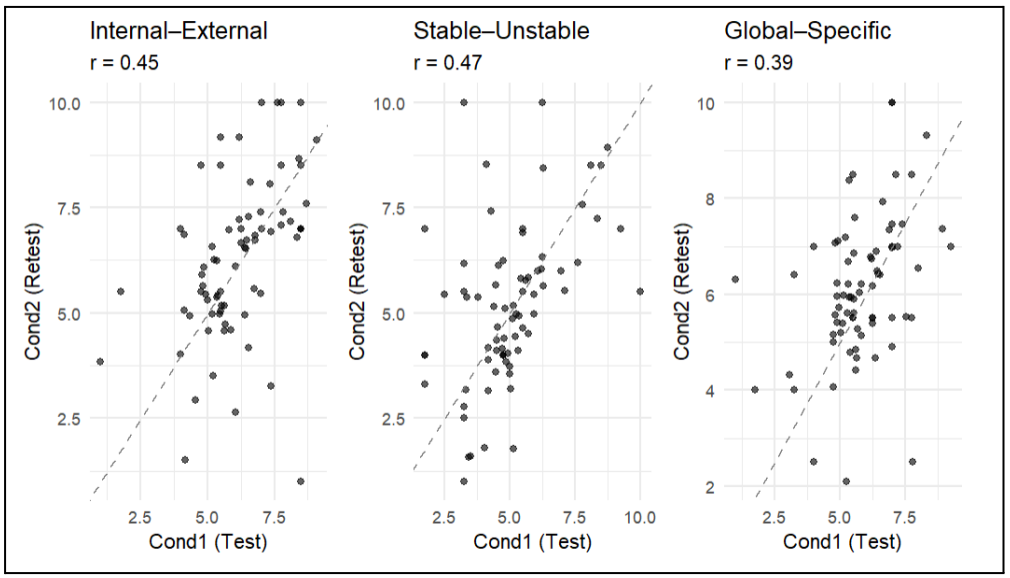

Figure 7. Raw test-retest attributions means: Scatterplots of raw mean slider ratings on the three attribution dimensions in condition 1 and condition 2.

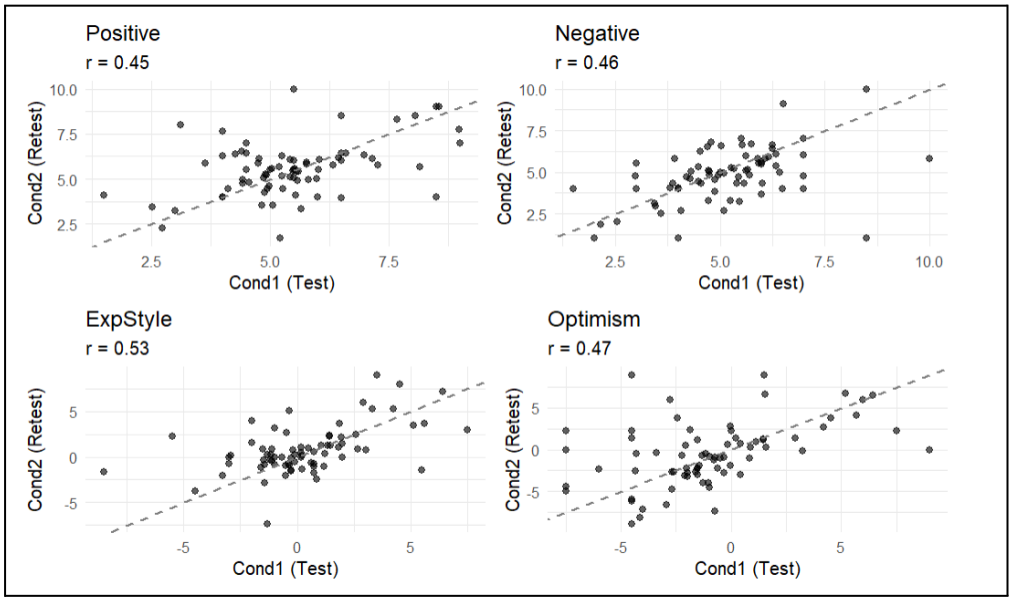

Figure 8. Test-retest reliability of composite attribution scores. Scatterplots depicting the correlations for mean positive (top left), as well as mean negative (top right) scenario ratings along with explanatory style (lower left) and optimism (lower right) proxies.

3.3. Emotional Engagement and Personalization Effects

To investigate emotional engagement and the effects of personalized content, analyses were conducted comparing participants' emotional responses (positive and negative valence ratings) across conditions. ANOVA results demonstrated significant differences in emotional relevance and valence depending on the scenario conditions and their associated tones. Overall, negative scenarios exhibited higher emotional relevance across conditions compared to positive scenarios. Specifically, significant differences emerged in emotional relevance ratings for negative scenarios ($F(2, 3257) = 12.47$, $p < .001$). Post-hoc analyses revealed that both interactive personalized conditions (Condition 1: $M = 7.66$, $SD = 2.19$; Condition 2: $M = 7.31$, $SD = 2.43$) elicited significantly greater emotional relevance compared to the traditional condition (Condition 3: $M = 6.85$, $SD = 2.83$), suggesting enhanced engagement through personalized and dynamically generated scenarios. Valence ratings provided further support for the emotional impact of personalized scenarios. Positive valence ratings for negative scenarios showed significant differences across conditions ($F(2, 3257) = 10.32$, $p < .001$), with Conditions 1 and 2 demonstrating significantly higher positive valence than Condition 3. This might indicate that participants perceived personalized negative scenarios as more emotionally nuanced, possibly due to greater perceived personal relevance.

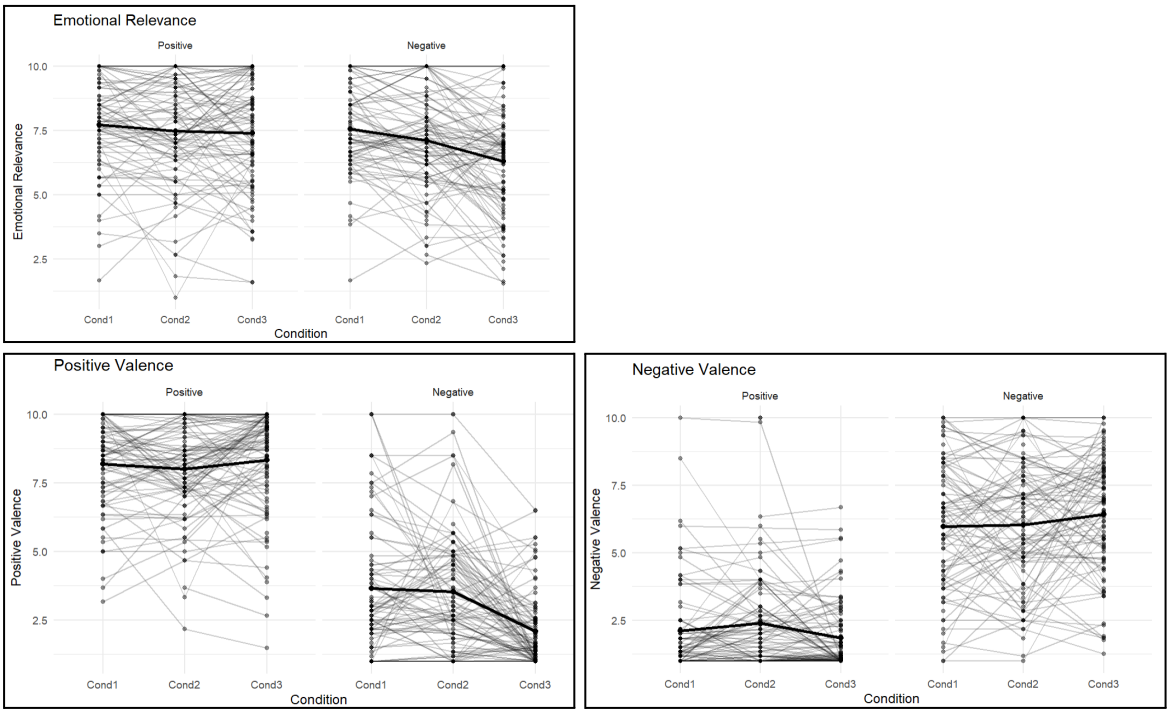

Figure 9. Within-subject ratings of Relevance and Valence: Mean ratings of each participant comparing emotional engagement of positive and negative scenarios between conditions.

Interestingly, emotional relevance ratings for positive-tone scenarios did not differ significantly across conditions, indicating a less pronounced effect of personalization in positive contexts. Similarly, negative valence ratings for positive-tone scenarios exhibited only marginal significance, without clear post-hoc differences. These findings suggest that dynamic personalization particularly enhances emotional engagement in negatively framed scenarios. The distinct emotional reactions observed between personalized and traditional conditions underscore the value of personalized content for increasing ecological validity and participant involvement in psychological assessments.

3.4. Usability and Participant Engagement

Usability and participant engagement ratings were compared across conditions to assess the perceived ease of use, clarity, satisfaction, and cognitive load associated with each assessment method. ANOVA results indicated significant differences across conditions for several usability metrics. Participants rated the ease of use significantly higher in Condition 2 (M = 9.13, SD = 1.19) compared to Condition 3 (M = 8.65, SD = 1.91), as indicated by ANOVA (F(2, 3257) = 4.65, p = .010). Clarity of instructions similarly showed significant differences (F(2, 3257) = 4.33, p = .013), with Condition 2 again rated higher than Condition 3. Satisfaction ratings revealed a comparable pattern, with Condition 2 receiving significantly higher satisfaction scores (M = 9.13, SD = 1.19) compared to Condition 3 (M = 8.65, SD = 1.91; F(2, 3257) = 3.46, p = .032). Engagement scores were marginally significant (F(2, 3257) = 2.95, p = .053), yet without post-hoc differences clearly reaching significance thresholds. Cognitive load did not differ significantly across conditions, suggesting that the increased usability and satisfaction observed in personalized conditions did not come at the expense of increased cognitive demand. These results indicate that participants found the personalized conditions easier to use and clearer in instruction, contributing to higher overall satisfaction without adding further cognitive complexity. Notably, no significant differences occurred between Condition 1 and Condition 3.

Usability ratings (mean ± SD) by condition.

Measure	Cond1	Cond2	Cond3
Ease	8.67 (1.88)	8.71 (1.76)	8.18 (2.21)
Clarity	9.41 (1.19)	9.60 (0.81)	9.07 (1.75)
Engagement	8.94 (1.49)	9.00 (1.29)	8.63 (1.92)
Satisfaction	9.02 (1.33)	9.13 (1.18)	8.65 (1.91)
Cognitive Load	5.64 (2.95)	5.52 (2.85)	5.59 (2.99)

Table 2. Usability rating means per condition.

3.5. Exploratory Analyses

Exploratory analyses were conducted to investigate potential relationships between attribution styles and pre-existing psychological measures. Correlations between pre-screening measures (PHQ-9 and GAD-7) and attributional style dimensions (Internal-External, Stable-Unstable, Global-Specific) revealed small associations, around $r = \pm 0.15$, indicating limited overlap between attributional style and common mental health indicators. Scenario relevance ratings showed modest but consistent correlations with attribution dimensions, highlighting a relationship between the perceived relevance of scenarios and participants' attribution styles (see Appendix A1 for a full correlation matrix). Higher relevance was moderately correlated with more internal and global attributions. Usability ratings provided another exploratory perspective, revealing associations between emotional involvement and usability measures such as satisfaction and engagement. Participants who rated scenarios as more emotionally involving typically reported higher satisfaction and ease of use, suggesting that emotional engagement may enhance perceived usability.

Last but not least, exploratory analyses of the *Bias Against Disconfirmatory Evidence* (BADE) demonstrated notable shifts in participants' attributional judgments following exposure to disconfirmatory evidence. Specifically, negative scenarios prompted participants toward more external attributions (Mean BADE-effect = 0.31, *SD* = 3.68), while positive scenarios resulted in minimal shifts toward internal attributions (Mean BADE-effect = -0.07, *SD* = 4.11). Correlational analyses further indicated that these BADE shifts were not significantly related to participants’ mental health screening

scores, suggesting the BADE effect's independence from baseline psychological distress.

4. Discussion

This study explored whether dynamically personalized, AI-generated vignettes could reproduce and extend the psychometric properties of the traditional ASQ. Across multiple attributional dimensions, the LLM-generated scenarios yielded internality and globality scores that diverged significantly from the standard ASQ ($p < .001$ and $p = .002$, respectively), while stability ratings remained essentially unchanged. However, factor analysis indicated similar factor structures between conditions. This could imply that the differing levels of personalization and emotional engagement might have simply affected the perceived intensities of attribution rather than depicting an altogether different attribution pattern. Test–retest analyses demonstrated moderate temporal consistency (ICCs and Pearson correlations between .39 and .53) for fully personalized vignettes across two sessions, indicating that AI-driven content can achieve some reliability, although not yet quite comparable to established retest benchmarks of the ASQ (~.70). It should be noted that all personalized vignettes used in this study were automatically generated in real-time and could therefore not be validated beforehand, while, on the other hand, the traditional ASQ relied on fully identical standardized vignettes to reach this established higher retest reliability. Additionally, due to the study's design participants received an unequal amount of fully personalized vignettes in condition 1 (12 vignettes) and condition 2 (6 vignettes) which might have skewed the data. Emotional engagement analyses revealed that personalized negative scenarios evoked stronger relevance and positive valence ($F(2, 3257) = 12.47$ and 10.32, both $p < .001$), whereas positive-tone items showed no meaningful differences. It can be assumed that while negative scenarios seem to generally elicit stronger reactions, the improved customization might have increased perceived valence ratings across the board for participants, leading to the counter-intuitive findings such as increased positive valence in negative scenarios. Finally, usability ratings favored the

personalized LLM conditions for ease of use, clarity, and satisfaction (all *p*-values < .05) without increasing cognitive load. Together, these findings support the hypothesis that AI-enhanced vignette generation can potentially maintain, and in some respects improve, the psychometric integrity, engagement, and user experience of traditional ASQ assessments. Last but not least, the effect of the included BADE-element suggests that disconfirmatory evidence may be more effective in altering implicit attribution patterns for negative events than for positive events, which could have implications for potential future interventions.

4.1. Limitations

As already mentioned above, the measured sample sizes for the relevant vignette types differed due to the study's design in between sessions. Additionally, no retest scores for the traditional ASQ method were collected directly within this study, which might limit the interpretability when comparing the temporal stability of assessment types. Furthermore, the reliance on continuous slider scales for attributional and valence ratings yielded rich, fine-grained data, but also assumed linear interpretation of constructs that were originally designed as ordinal categories. Although exploratory factor analyses supported the expected three-factor structure, the psychometric properties of these slider measures have not been fully established, and their fine gradations may have increased random error. The inclusion of BADE trials added an additional layer of cognitive flexibility assessment, yet the single-application format and limited number of disconfirmatory items may have under-sampled this construct. Finally, prompt-engineering rules ensured consistency of vignette length and tone, but could not be validated upfront. Considering the experimental and novel approach of the conducted study, it is to be expected that the psychometric quality of the vignettes is limited by the prompts used for generation and will most likely benefit from additional refinement and further testing.

4.2. Future Research Directions

Future research should tackle these methodological limitations by systematically varying prompt parameters (e.g., token limits, narrative detail), and should validate

slider-based attribution and valence scales against traditional Likert formats and behavioral benchmarks. Importantly, it should be clarified whether the found differences in attribution patterns reflect methodological shortcomings and artifacts or whether the shifted attributions actually indicate more accurate measures. Longer retest intervals and equalized numbers of personalized items per session will also be essential to separate practice or fatigue effects from actual temporal stability. Eventually expanding samples to potentially include clinical populations might also prove insightful. Extending the present work, future studies should evaluate the LLM-driven vignette approach across a broader range of standardized measures and contexts to determine whether dynamically personalized scenarios can faithfully reproduce established psychometric properties in diverse assessment contexts. Finally, embedding AI-personalized vignettes within longitudinal therapeutic or educational interventions could reveal not only measurement validity but also whether such adaptive assessments actively promote cognitive or emotional change over time.

4.3. Conflicts of Interest & Funding

No conflicts of interest are declared. This research received funding through the *Förderstipendium* of the University of Vienna.

References

Ball, H. A., McGuffin, P., & Farmer, A. E. (2008). Attributional style and depression. *The British Journal of Psychiatry*, *192*(4), 275–278. https://doi.org/10.1192/bjp.bp.107.038711

Baumel, A., Fleming, T., & Schueller, S. M. (2020). Digital Micro interventions for Behavioral and Mental Health gains: Core components and conceptualization of digital Micro Intervention care. *Journal of Medical Internet Research*, *22*(10), e20631. https://doi.org/10.2196/20631

Bruehlman-Senecal, E., Hook, C. J., Pfeifer, J. H., FitzGerald, C., Davis, B., Delucchi, K. L., Haritatos, J., & Ramo, D. E. (2020). Smartphone app to address loneliness among college students: pilot randomized controlled trial. *JMIR Mental Health*, *7*(10), e21496. https://doi.org/10.2196/21496

Buschmeyer, K., Hatfield, S., & Zenner, J. (2023). Psychological assessment of AI-based decision support systems: tool development and expected benefits. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1249322

Cameron, G., Cameron, D., Megaw, G., Bond, R., Mulvenna, M., O'Neill, S., Armour, C., & McTear, M. (2019). Assessing the usability of a chatbot for mental health care. In *Lecture notes in computer science* (pp. 121–132). https://doi.org/10.1007/978-3-030-17705-8_11

Corcoran, C. M., & Cecchi, G. A. (2020). Using language processing and speech analysis for the identification of psychosis and other disorders. *Biological Psychiatry Cognitive Neuroscience and Neuroimaging*, *5*(8), 770–779. https://doi.org/10.1016/j.bpsc.2020.06.004

Currey, D., & Torous, J. (2023). Digital phenotyping data to predict symptom improvement and mental health app personalization in college students: Prospective validation of a predictive model. *Journal of Medical Internet Research*, *25*, e39258. https://doi.org/10.2196/39258

De Chiusole, D., Spinoso, M., Anselmi, P., Bacherini, A., Balboni, G., Mazzoni, N., Brancaccio, A., Epifania, O. M., Orsoni, M., Giovagnoli, S., Garofalo, S., Benassi, M., Robusto, E., Stefanutti, L., & Pierluigi, I. (2024). PsycAssist: a Web-Based artificial intelligence system designed for adaptive neuropsychological assessment and training. *Brain Sciences*, *14*(2), 122. https://doi.org/10.3390/brainsci14020122

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*. https://doi.org/10.1038/s44159-023-00241-5

Dosovitsky, Kim, E., & Bunge, E. L. (2021). Psychometric Properties of a Chatbot Version of the PHQ-9 With Adults and Older Adults. *Frontiers in Digital Health*, *3*. https://doi.org/10.3389/fdgth.2021.645805

Dykema, J., Bergbower, K., Doctora, J. D., & Peterson, C. (1996). An Attributional Style Questionnaire for General Use. *Journal of Psychoeducational Assessment*, *14*(2), 100–108. https://doi.org/10.1177/073428299601400201

Elyoseph, Z., Hadar-Shoval, D., Asraf, K., & Lvovsky, M. (2023). ChatGPT outperforms humans in emotional awareness evaluations. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1199058

Gual-Montolio, P., Jaén, I., Martínez-Borba, V., Castilla, D., & Suso-Ribera, C. (2022). Using artificial intelligence to enhance ongoing psychological interventions for emotional problems in Real- or Close to Real-Time: A Systematic review. *International Journal of Environmental Research and Public Health*, *19*(13), 7737. https://doi.org/10.3390/ijerph19137737

Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences*, *112*(8), 2395–2400. https://doi.org/10.1073/pnas.1416587112

Jin, Y., Chen, L., Zhao, X., & Cai, W. (2024). Effects of psychological assessment design with closed-ended questions on user response to open-ended questions within a survey chatbot for mental health. *International Journal of Human-Computer Studies*, *189*, 103290–103290. https://doi.org/10.1016/j.ijhcs.2024.103290

Kaland, N., Møller-Nielsen, A., Smith, L., Mortensen, E. L., Callesen, K., & Gottlieb, D. (2005). The Strange Stories test: A replication study of children and adolescents with Asperger syndrome. *European Child & Adolescent Psychiatry*, *14*, 73–82. https://doi.org/10.1007/s00787-005-0434-2

Ke, L., Tong, S., Cheng, P., & Peng, K. (2024). *Exploring the Frontiers of LLMs in Psychological Applications: A Comprehensive review*. https://doi.org/10.48550/arXiv.2401.01519

Khawaja, Z., & Bélisle-Pipon, J. (2023). Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots. *Frontiers in Digital Health*, *5*. https://doi.org/10.3389/fdgth.2023.1278186

Kjell, O. N., Kjell, K., & Schwartz, H. A. (2023). Beyond rating scales: With targeted evaluation, large language models are poised for psychological assessment. *Psychiatry Research*, *333*, 115667. https://doi.org/10.1016/j.psychres.2023.115667

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (1999). *Patient Health Questionnaire-9 (PHQ-9)* [Database record]. APA PsycTests. https://doi.org/10.1037/t06165-000

Li, H., Zhang, R., Lee, Y., Kraut, R. E., & Mohr, D. C. (2023). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *Npj Digital Medicine*, *6*(1). https://doi.org/10.1038/s41746-023-00979-5

Luxton, D. D. (2016). An introduction to artificial intelligence in behavioral and mental health care. In *Elsevier eBooks* (pp. 1–26). https://doi.org/10.1016/b978-0-12-420248-1.00001-5

Maenhout, L., Peuters, C., Cardon, G., Compernolle, S., Crombez, G., & DeSmet, A. (2021). Participatory development and pilot testing of an adolescent health promotion chatbot. *Frontiers in Public Health*, *9*. https://doi.org/10.3389/fpubh.2021.724779

Mörch, C., Gupta, A., & Mishara, B. L. (2020). Canada protocol: An ethical checklist for the use of artificial Intelligence in suicide prevention and mental health. *Artificial Intelligence in Medicine*, *108*, 101934. https://doi.org/10.1016/j.artmed.2020.101934

Nawaz, S., Lewis, C., Townson, A., & Mei, P. (2023). What does the Strange Stories test measure? Developmental and within-test variation. *Cognitive Development*, *65*, 101289. https://doi.org/10.1016/j.cogdev.2022.101289

Norbury, A., Hauser, T. U., Fleming, S. M., Dolan, R. J., & Huys, Q. J. M. (2024). Different components of cognitive-behavioral therapy affect specific cognitive mechanisms. *Science Advances*, *10*(13). https://doi.org/10.1126/sciadv.adk3222

Ock, J., & An, H. (2021). Machine Learning Approach to Personality Assessment and Its Application to Personnel Selection. *Korean Journal of Industrial and Organizational Psychology*, *34*(2), 213–236. https://doi.org/10.24230/kjiop.v34i2.213-236

Pani, B., Crawford, J., & Allen, K. (2024). Can generative artificial intelligence foster belongingness, social support, and reduce loneliness? A conceptual analysis. In *Springer eBooks* (pp. 261–276). https://doi.org/10.1007/978-3-031-46238-2_13

Peterson, C., Semmel, A., Von Baeyer, C., Abramson, L. Y., Metalsky, G. I., & Seligman, M. E. P. (1982). The attributional Style Questionnaire. *Cognitive Therapy and Research*, *6*(3), 287–299. https://doi.org/10.1007/bf01173577

Podina, I. R., Bucur, A. M., Fodor, L., & Boian, R. (2023). Screening for common mental health disorders: a psychometric evaluation of a chatbot system. *Behaviour & Information Technology*, 1–10. https://doi.org/10.1080/0144929X.2023.2275164

Rollwage, M., Habicht, J., Juechems, K., Carrington, B., Viswanathan, S., Stylianou, M., Hauser, T. U., & Harper, R. (2023). Using conversational AI to facilitate mental health assessments and improve clinical efficiency within psychotherapy services: Real-World observational study. *JMIR AI*, *2*, e44358. https://doi.org/10.2196/44358

Schick, A., Feine, J., Morana, S., Maedche, A., & Reininghaus, U. (2022). Validity of Chatbot use for mental health assessment: experimental study. *JMIR Mhealth and Uhealth*, *10*(10), e28082. https://doi.org/10.2196/28082

Servick, K. (2024). Striving to connect. *Science*, *384*(6694), 375–379. https://doi.org/10.1126/science.adq0080

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine*, *49*(09), 1426–1448. https://doi.org/10.1017/s0033291719000151

Sidnell, J., & Stivers, T. (2014). The handbook of conversation analysis. In *Wiley-Blackwell eBooks*. https://ci.nii.ac.jp/ncid/BB16723846

Smith, K. A., Blease, C., Faurholt-Jepsen, M., Firth, J., Van Daele, T., Moreno, C., Carlbring, P., Ebner-Priemer, U. W., Koutsouleris, N., Riper, H., Mouchabac, S., Torous, J., & Cipriani, A. (2023). Digital mental health: challenges and next steps. *BMJ Mental Health*, *26*(1), e300670. https://doi.org/10.1136/bmjment-2023-300670

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). *Generalized Anxiety Disorder 7 (GAD-7)* [Database record]. APA PsycTests. https://doi.org/10.1037/t02591-000

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *Npj Mental Health Research*, *3*(1). https://doi.org/10.1038/s44184-024-00056-z

Stuart, E. M., Torres, S., & Gutierrez, B. (2023). A - 179 AI-Driven Diagnostics for Child Neuropsychology: Promising but Insufficient. *Archives of Clinical Neuropsychology*, *38*(7), 1352–1352. https://doi.org/10.1093/arclin/acad067.196

Sun, J., Dong, Q.-X., Wang, S.-W., Zheng, Y.-B., Liu, X.-X., Lu, T.-S., Yuan, K., Shi, J., Hu, B., Lu, L., & Han, Y. (2023). Artificial intelligence in psychiatry research, diagnosis, and therapy. *Asian Journal of Psychiatry*, *87*, 103705. https://doi.org/10.1016/j.ajp.2023.103705

Thakkar, A., Gupta, A., & De Sousa, A. (2024). Artificial intelligence in positive mental health: a narrative review. *Frontiers in Digital Health*, *6*. https://doi.org/10.3389/fdgth.2024.1280235

Weisenburger, R. L., Mullarkey, M. C., Labrada, J., Labrousse, D., Yang, M. Y., MacPherson, A. H., Hsu, K. J., Ugail, H., Shumake, J., & Beevers, C. G. (2024). Conversational assessment using artificial intelligence is as clinically useful as depression scales and preferred by users. *Journal of Affective Disorders*, *351*, 489–498. https://doi.org/10.1016/j.jad.2024.01.212

Woodward, T. S., Moritz, S., Cuttler, C., & Whitman, J. C. (2006). The contribution of a cognitive bias against Disconfirmatory Evidence (BADE) to delusions in schizophrenia. *Journal of Clinical and Experimental Neuropsychology*, *28*(4), 605–617. https://doi.org/10.1080/13803390590949511

Yang, Q., Wang, Z., Chen, H., Wang, S., Pu, Y., Gao, X., Huang, W., Song, S., & Huang, G. (2024). PsychoGAT: A Novel Psychological Measurement Paradigm through Interactive Fiction Games with LLM Agents. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14470–14505. https://doi.org/10.18653/v1/2024.acl-long.779

Zulueta, J., Piscitello, A., Rasic, M., Easter, R., Babu, P., Langenecker, S. A., McInnis, M., Ajilore, O., Nelson, P. C., Ryan, K., & Leow, A. (2018). Predicting Mood Disturbance Severity with Mobile Phone Keystroke Metadata: A BiAffect Digital Phenotyping Study. *Journal of Medical Internet Research*, *20*(7), e241. https://doi.org/10.2196/jmir.9775

List of Figures

List of Tables

List of Abbreviations

AI:	Artificial Intelligence
ASQ:	Attributional Style Questionnaire
BADE:	Bias Against Discriminatory Evidence
GAD-7:	Generalized Anxiety Disorder Scale-7
LLM:	Large Language Model
ML:	Machine Learning
NASA-TLX:	NASA Task Load Index
NLP:	Natural Language Processing
PHQ-9:	Patient Health Questionnaire Scale-9

Appendix

A1 Correlation matrix:

	PHQ-9 Score	GAD-7 Score	Final_Relevance	Emotional_Relevance	Positive_Valence	Negative_Valence
PHQ-9 Score	1.00	0.70	0.06	-0.02	0.00	0.07
GAD-7 Score	0.70	1.00	0.03	0.03	0.00	0.08
Final_Relevance	0.06	0.03	1.00	0.43	0.28	0.00
Emotional_Relevance	-0.02	0.03	0.43	1.00	0.29	0.12
Positive_Valence	0.00	0.00	0.28	0.29	1.00	-0.62
Negative_Valence	0.07	0.08	0.00	0.12	-0.62	1.00
Engagement	-0.07	-0.07	0.32	0.20	0.09	0.02
Ease of Use	-0.08	-0.01	0.29	0.21	0.08	0.04
Clarity of Instructions	-0.12	-0.08	0.27	0.15	0.06	-0.01
Satisfaction	-0.12	-0.09	0.37	0.23	0.11	0.04
Cognitive Load	0.08	0.08	-0.07	0.01	0.06	-0.03
Internal_External	0.05	-0.01	0.11	0.06	-0.05	0.10
Stable_Unstable	0.08	0.06	0.02	-0.07	-0.21	0.20
Global_Specific	0.03	0.02	0.07	0.17	0.19	-0.05

	Engagement	Ease of Use	Clarity of Instructions	Satisfaction	Cognitive Load	Internal_External
PHQ-9 Score	-0.07	-0.08	-0.12	-0.12	0.08	0.05
GAD-7 Score	-0.07	-0.01	-0.08	-0.09	0.08	-0.01
Final_Relevance	0.32	0.29	0.27	0.37	-0.07	0.11
Emotional_Relevance	0.20	0.21	0.15	0.23	0.01	0.06
Positive_Valence	0.09	0.08	0.06	0.11	0.06	-0.05
Negative_Valence	0.02	0.04	-0.01	0.04	-0.03	0.10
Engagement	1.00	0.51	0.55	0.79	0.05	0.11
Ease of Use	0.51	1.00	0.49	0.54	-0.12	0.09
Clarity of Instructions	0.55	0.49	1.00	0.63	0.00	0.06
Satisfaction	0.79	0.54	0.63	1.00	0.05	0.10
Cognitive Load	0.05	-0.12	0.00	0.05	1.00	-0.03
Internal_External	0.11	0.09	0.06	0.10	-0.03	1.00
Stable_Unstable	-0.02	-0.03	-0.01	-0.02	-0.01	0.11
Global_Specific	0.09	0.05	0.01	0.06	0.03	0.08

	Stable_Unstable	Global_Specific
PHQ-9 Score	0.08	0.03
GAD-7 Score	0.06	0.02
Final_Relevance	0.02	0.07
Emotional_Relevance	-0.07	0.17
Positive_Valence	-0.21	0.19
Negative_Valence	0.20	-0.05
Engagement	-0.02	0.09
Ease of Use	-0.03	0.05
Clarity of Instructions	-0.01	0.01
Satisfaction	-0.02	0.06
Cognitive Load	-0.01	0.03
Internal_External	0.11	0.08
Stable_Unstable	1.00	-0.04
Global_Specific	-0.04	1.00

A2 Prompt for personalized vignette generation:

```
prompt = f"""
You are tasked with generating a short, personalized vignette for a psychological assessment.
The user is {user_age} years old.

{chosen_topic_text}

Scenario Details:
- The vignette must strictly focus on {topic_context}.
- The tone of the vignette should reflect {tone_context}.
- The vignette should include:
    - A clear context type (e.g., work, social, physical, educational, creative).
    - A specific interaction (e.g., feedback from authority, peer collaboration, group dynamics).
    - A defined event or action (e.g., presenting an idea, resolving a conflict, completing a project).

Guidelines:
- The vignette must describe external events only, without referencing thoughts, emotions, or internal reflections.
- Use second-person language ("You...").
- Keep the vignette concise (under 50 tokens) and neutral/professional.
- Do not include personal identifying information.
- Do not mention both profession and personal interests together; focus exclusively on the chosen topic.
- Do not mix multiple interests or contexts; stay consistent with the single chosen topic.

{feedback_instructions}

Now generate the vignette:
"""
```

A3 Prompt for non-personalized vignette generation:

```
prompt = f"""
You are tasked with generating a short, generic vignette for a psychological assessment.
Scenario Details:
- The topic of the vignette is {topic_context}.
- The tone of the vignette should reflect {tone_context}.
- The vignette should include:
    - A clear context type (e.g., work, social, physical activity, educational, or creative).
    - A specific interaction (e.g., feedback from authority, peer collaboration, or group dynamics).
    - A defined event or action (e.g., presenting an idea, resolving a conflict, or completing a project).

Guidelines:
- The vignette must describe external events only, without referencing thoughts, emotions, or internal reflections.
- Use second-person language ("You...").
- Keep the vignette concise, under 50 tokens.
- Use neutral and professional language suitable for a psychological assessment.
- Do not include personal identifying information or references to the user's actual profile.

Example:
    Positive Example:
        "You collaborate with a colleague on a shared report.
        They praise your clear communication, and together you finalize the work ahead of schedule."

    Negative Example:
        "During a community gathering, you run into a disagreement over how to organize an activity.
        The conversation becomes tense, and participants struggle to find common ground."

Now generate the vignette:
"""
```

A4 Prompts used for generation of attribution options

A4.1 Internality:

```
def generate_internal_external_options(chat_history, vignette, user_state):
    prompt = f"""
        Based on the following scenario:

        "{vignette}"

        Provide two possible causes for why this situation occurred, focusing on **internal** and **external** factors
        that are directly related to the scenario.

        Ensure each explanation is **concise** and captures the essence of an internal or external cause,
        without excessive detail. The causes should be easy for the user to understand,
        without overwhelming them with too many specifics.

        Describe these explanations from the user's point of view, making them relatable and personal.

        Present the options in the following format exactly:

        A: [Specific internal reason]
        B: [Specific external reason]

        Do not include any additional text or explanation.

        Use simple, but neutral and professional language suitable for a psychological assessment.
    """
```

A4.2 Stability:

```
prompt = f"""
Generate exactly two short statements explaining what might have led to this situation, each on its own line.
They must strictly reflect a {attribution_type.lower()} cause,
with no references to outside factors if {attribution_type.lower()} is "internal,"
and no references to personal traits if {attribution_type.lower()} is "external."

1) Provide a stable explanation
2) Provide a more temporary or unstable explanation

Instructions:
- Phrase all statements in second person.
- Do not mention external influences if you're generating an internal cause.
- Do not mention internal traits if you're generating an external cause.
- Do not label the options as stable/unstable/etc. in the text itself.
- Start each line with "A:" or "B:" only.
- Keep the language simple, pragmatic, and relevant to the scenario:
"{vignette}"

- Each statement should be concise, neutral, and easy to understand, without extra commentary.
- Do not include disclaimers, personal info, or text beyond the four statements.
"""
```

A4.3 Globality:

```
prompt = f"""
Generate exactly two short statements explaining what might have led to this situation, each on its own line,
with no additional commentary. They must strictly reflect a {attribution_type.lower()} cause.

- If {attribution_type.lower()} is "internal," do not reference any outside factors such as environment,
other people, or external circumstances.

- If {attribution_type.lower()} is "external," do not reference personal traits,
inherent aptitudes, or individual qualities.

Focus on these two points:
1) A broader/global explanation that applies widely.
2) A more situation-specific explanation.

Instructions:
- Phrase all statements in second person.
- Start each line with "A:" or "B:" only, and provide no other lines.
- Keep the language simple, neutral, and directly relevant to:
  "{vignette}"
- Each statement should be concise and easy to understand, with no disclaimers or extra text.
- Output only those two lines, nothing else.
"""
```

A5 Internal consistency

A5.1 Cronbach's alpha (Cond1):

Dimension <chr>	Tone <chr>	N_items <int>	Cronbach_alpha <dbl>
Internal_External	Positive	6	0.5894673
Stable_Unstable	Positive	6	0.7269271
Global_Specific	Positive	6	0.4816386
Internal_External	Negative	6	0.7043269
Stable_Unstable	Negative	6	0.5814360
Global_Specific	Negative	6	0.5966949

A5.2 Cronbach's alpha (Cond3):

Dimension <chr>	N_items <int>	Cronbach_alpha <dbl>
Internal_External	12	0.6907424
Stable_Unstable	12	0.6595500
Global_Specific	12	0.8110653

A5.3 Split-half test (Cond1):

Dimension <chr>	Tone <chr>	split_half_r <dbl>	spearman_brown <dbl>
Internal_External	Positive	0.5111927	0.6765421
Internal_External	Negative	0.5970187	0.7476665
Stable_Unstable	Positive	0.6163698	0.7626594
Stable_Unstable	Negative	0.4993497	0.6660884
Global_Specific	Positive	0.3751262	0.5455880
Global_Specific	Negative	0.4291427	0.6005596

A5.4 Split-half test (Cond3):

Dimension <chr>	Tone <chr>	split_half_r <dbl>	spearman_brown <dbl>
Internal_External	Positive	0.4052744	0.5767904
Internal_External	Negative	0.4578497	0.6281165
Stable_Unstable	Positive	0.5716768	0.7274738
Stable_Unstable	Negative	0.3656762	0.5355240
Global_Specific	Positive	0.5988579	0.7491071
Global_Specific	Negative	0.5573569	0.7157728

A6 Factor Analysis

A6.1 Condition 1:

```
                        ML1  ML2  ML3
SS loadings            3.86 2.60 2.54
Proportion Var         0.11 0.07 0.07
Cumulative Var         0.11 0.18 0.25
Proportion Explained   0.43 0.29 0.28
Cumulative Proportion  0.43 0.72 1.00

 With factor correlations of
     ML1  ML2  ML3
ML1 1.00 0.17 0.18
ML2 0.17 1.00 0.10
ML3 0.18 0.10 1.00

Mean item complexity =  1.7
Test of the hypothesis that 3 factors are sufficient.

df null model =  630  with the objective function =  16.06 with Chi Square =  1111.08
df of  the model are 525  and the objective function was  10.05


 The root mean square of the residuals (RMSR) is  0.09
 The df corrected root mean square of the residuals is  0.09

 The harmonic n.obs is  83 with the empirical chi square  779.21  with prob <  3.3e-12
 The total n.obs was  83  with Likelihood Chi Square =  674.81  with prob <  1e-05

 Tucker Lewis Index of factoring reliability =  0.6
 RMSEA index =  0.057  and the 90 % confidence intervals are  0.045 0.072
 BIC =  -1645.08
 Fit based upon off diagonal values = 0.77
 Measures of factor score adequacy
                                                    ML1 ML2  ML3
 Correlation of (regression) scores with factors   0.92 0.9 0.88
 Multiple R square of scores with factors          0.85 0.8 0.77
 Minimum correlation of possible factor scores     0.69 0.6 0.54
```

A6.2 Condition 3:

```
                        ML1  ML2  ML3
SS loadings            3.86 2.60 2.54
Proportion Var         0.11 0.07 0.07
Cumulative Var         0.11 0.18 0.25
Proportion Explained   0.43 0.29 0.28
Cumulative Proportion  0.43 0.72 1.00

 With factor correlations of
     ML1  ML2  ML3
ML1 1.00 0.17 0.18
ML2 0.17 1.00 0.10
ML3 0.18 0.10 1.00

Mean item complexity =  1.7
Test of the hypothesis that 3 factors are sufficient.

df null model =  630  with the objective function =  16.06 with Chi Square =  1111.08
df of  the model are 525  and the objective function was  10.05


The root mean square of the residuals (RMSR) is  0.09
The df corrected root mean square of the residuals is  0.09

The harmonic n.obs is  83 with the empirical chi square  779.21  with prob <  3.3e-12
The total n.obs was  83  with Likelihood Chi Square =  674.81  with prob <  1e-05

Tucker Lewis Index of factoring reliability =  0.6
RMSEA index =  0.057  and the 90 % confidence intervals are  0.045 0.072
BIC =  -1645.08
Fit based upon off diagonal values = 0.77
Measures of factor score adequacy
                                                ML1 ML2  ML3
Correlation of (regression) scores with factors   0.92 0.9 0.88
Multiple R square of scores with factors          0.85 0.8 0.77
Minimum correlation of possible factor scores     0.69 0.6 0.54
```

Abstrakt:
Jüngste Fortschritte im Gebiet der künstlichen Intelligenz, insbesondere sogenannten Large Language Models (LLMs) wie OpenAIs GPT-4o, bieten innovative Möglichkeiten, die Grenzen traditioneller psychologischer Erhebungen zu überwinden. Diese Arbeit evaluiert die Validität einer LLM-erweiterten Version des Attributional Style Questionnaire (ASQ) und vergleicht dynamisch personalisierte und KI-generierte Vignetten mit traditionellen statischen Szenarien. Unter Verwendung eines Within-Subject-Designs mit 112 Teilnehmer*innen untersucht diese Studie die Vergleichbarkeit mit traditionellen Methoden, die Test-Retest-Reliabilität, das emotionale Engagement und die Benutzererfahrung. Die Attributionsmuster konnten teilweise repliziert werden, unterschieden sich jedoch zwischen den Bedingungen. Das emotionale Engagement wurde durch personalisierte Negativszenarien deutlich verbessert, was die Zufriedenheit der Teilnehmer und die wahrgenommene Benutzerfreundlichkeit erhöhte, ohne die kognitive Belastung zu erhöhen. Die Ergebnisse deuten auf das Potenzial der Integration strukturierter und standardisierter KI-gesteuerter Personalisierung in psychologische Beurteilungen hin, um die ökologische Validität und die Nutzererfahrung zu verbessern. Es werden Möglichkeiten für künftige Forschungen zur Nutzung generativer KI ohne Beeinträchtigung der Validität in ähnlichen Tests aufgezeigt.