Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AI Persona Design and Perceived Difficulty of Legal Advice Tasks

Armin Klaps

armin.klaps@sfu.ac.at

Sigmund Freud University Vienna https://orcid.org/0000-0002-2344-040X

Zuzana Kovacovsky

Sigmund Freud University Vienna

Birgit Ursual Stetina

Sigmund Freud University Vienna https://orcid.org/0000-0001-8711-1923

Research Article

Keywords: conversational AI, chatbot persona, Big Five, OCEAN, legal advice, perceived task difficulty, human-AI interaction

Posted Date: September 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10850860/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Personality cues may influence how users experience conversational artificial intelligence, but there is currently limited understanding of their role in high-risk situations involving expert advice. We developed and tested an experimental approach in which the occupational character of a legal advice chatbot was varied and looked at how this affected users' perceptions of task difficulty in relation to their Big Five personality traits. Sixteen adults (eight women and eight men; mean age 26.50 years, standard deviation 4.32) were randomly given one of three personas: judge (n = 5), lawyer (n = 5), or mediator (n = 6). Each participant worked through three randomly selected legal scenarios and then rated the difficulty of the tasks. The Big Five traits were measured using the BFI-44; emotion regulation (FEEL-E) was also used as an additional check on neuroticism-related patterns. There were strong differences between the groups in terms of the mean level of subjective difficulty, $F(2, 13) = 33.60$, $p < .001$, $\eta^2 = .84$, $\omega^2 = .80$. The mediator persona was associated with the lowest rating of difficulty ($M = 1.67$, $SD = 0.42$), followed by the lawyer persona ($M = 3.60$, $SD = 0.43$) and then the judge persona ($M = 4.07$, $SD = 0.68$). Within the small subgroup sizes — ranging from five to six — exploratory Spearman correlations indicated different patterns depending on personality trait and persona, such as negative associations for the mediator and a strong positive association between conscientiousness and difficulty for the judge. Since the subgroup sizes were only five or six, these correlations are only hypothesis-generating and not confirmatory. The approach proved feasible and led to specific design improvements for a previously registered, adequately powered study. The evidence shows that using mediator-like cues — particularly high agreeableness and low neuroticism in the chatbot's persona — could improve perceptions of the understandability of complex advice.

1. Introduction

More and more, conversational AI is serving as a means to access information that used to require professional knowledge. In legal situations, users are not only required to obtain an answer but also to understand and assess the explanations, all while dealing while facing uncertainty. As a result, the way the task is perceived becomes an important outcome of the interaction: a technically sound explanation might still fall short if its manner makes the task seem impossible to understand.

Studies into human trust in AI separate mental evaluations of ability and dependability from more affective and relational reactions (Glikson & Woolley, 2020). People's attitudes towards AI also depend on their level of anxiety, prior knowledge, demographic characteristics, and temperament (De Freitas et al., 2023; Kaya et al., 2024). Recent research indicates that the Big Five traits can be linked to people's attitudes towards and performance with generative AI, although the effects differ across contexts and traits (Grassini et al., 2025; Kovbasiuk et al., 2025). These data show that user characteristics may interact with how the system is presented rather than exert only an additive influence.

The choice of chatbot persona is a feasible design decision. Occupational titles such as judge, lawyer, and mediator entail expectations regarding authority, confrontation, empathy, structure, and the ability to

consider others' points of view. We turned these expectations into fixed OCEAN design profiles. These profiles did not represent estimates of professionals' actual personalities; rather, they were based on design criteria obtained from research linking personality with occupational interests and behavior (Barrick & Mount, 1991; Furnham, 2001; Redding & Reppucci, 1999; Schulze & Schmitt, 2004).

The objective of this pilot study was twofold: first, to examine the feasibility and sensitivity of an experimental approach for research into legal-advice chatbots, and second, to detect effect-size patterns that could be used to formulate confirmatory hypotheses. We wanted to observe patterns in users' Big Five personality traits and their perceptions of task difficulty for each chatbot persona, and to determine which hypotheses should be given priority in a subsequent main study. Although this was only an exploratory study, we anticipated that the direction and strength of the associations between personality traits and subjective difficulty would vary across personas.

1.1 Acceptance of and resistance to AI advice

Reactions to AI are not uniformly positive or negative. Whether it is accepted depends on the field in question, the type of information being processed, the consequences of making an error, and how much the system seems to replace human decision. In areas such as law, astronomy, and pharmacology, applications that handle large collections of documents or carry out complex calculations are seen as useful, while those that make intimate social or psychological judgments tend to provoke greater reservations (Schepman & Rodway, 2020). This point is particularly important when it comes to legal advice. Although legal chatbots offer quick and inexpensive access to information, they do so in a field where errors can affect people's rights, obligations, employment, privacy, and financial situations.

Anxiety related to AI can involve concerns about surveillance, loss of privacy, discrimination, and the unclear use of personal data (Li & Huang, 2020). Such concerns go beyond simple general opinions about technology. Someone can find computers useful while remaining skeptical of AI-produced recommendations. To account for this, Schepman and Rodway's (2020) General Attitudes towards Artificial Intelligence Scale distinguishes between perceptions of personal or societal benefit and negative concerns. This two-dimensional approach assists explain why users might appreciate the simplicity of accessing information while still preferring a human professional for important decisions.

De Freitas and colleagues (2023 classify psychological resistance to AI into five types: opacity, emotionlessness, rigidity, autonomy, and group membership. Opacity occurs when users cannot understand how the system arrived at its output. The issue of emotionlessness refers to the idea that AI lacks the affective abilities needed for sensitive tasks. Rigidity stems from the expectation that an AI cannot adapt to individual situations. Autonomy is seen as a threat when users feel that they have lost control or when their meaningful human work is being replaced. The aspect of group membership involves treating an artificial agent as an out-group member and regarding it as categorically different from a human advisor. The design of the persona can affect several of these barriers simultaneously: a

mediator persona can convey flexibility and social sensitivity, while a judge persona can convey authority, distance, and procedural rigidity.

Trying to lower resistance means making various compromises. Although greater transparency can increase cognitive trust, overly simple explanations may come across as patronizing. While anthropomorphic or emotionally responsive behavior can make an interaction more approachable, it can also lead users to think the system understands or is competent to a greater extent than it actually does. Stressing user control might increase acceptance, but could decrease consistency if the system becomes too easy to manipulate. A useful persona, therefore, has to do more than seem friendly; it has to strike a balance among accessibility, predictability, calibrated authority, and the accurate communication of its limitations.

1.2 Personality and the use of conversational AI

The Big Five framework gives a shared set of terms for describing stable individual differences in openness, conscientiousness, extraversion, agreeableness, and neuroticism. Openness is linked to curiosity and a desire to embrace new things; conscientiousness with organization, persistence, and carefully carrying out tasks; extraversion with sociability and assertiveness; agreeableness with trust, a cooperative attitude, and a tendency to act in a prosocial manner; and neuroticism with a high level of emotional reactivity and sensitivity to threats. These traits can influence how users approach a chatbot, how much dialogue they initiate, how critically they examine advice, and whether a socially warm or an authoritative style of communication appears supportive or off-putting.

Evidence of a connection between personality and AI outcomes is appearing, but it is varied. Kovbasiuk et al. (2025) examined how people performed on business-related tasks with and without chatbot assistance and investigated their intentions to use AI in the future. Their findings showed that agreeableness and conscientiousness were related to task performance quality, while neuroticism showed a context-dependent pattern that did not conform to simple assumptions about lower performance. In a larger survey involving 520 participants, Grassini et al. (2025) found that openness predicted positive attitudes towards AI but not towards everyday technology; neuroticism did not significantly predict either kind of attitude, and men and younger participants tended to express more positive attitudes. These results show that the effects of personality depend on the particular outcome and context rather as opposed to reflecting a general disposition towards all technology

The study now changes its focus from seeing personality in the role of a predictor of attitude towards technology to considering personality in the role of a possible moderator of the interaction. This point is significant because a very conscientious user may value structured explanations but could also find a rule-bound persona to be demanding or excessively complicated. On the other hand, a highly agreeable user will likely respond well to a cooperative mediator approach, whereas a more skeptical user may prefer direct, adversarial reasoning. Therefore, the fit between the person and the system could be at least as important as the overall desirability of any individual persona.

1.3 Trust, social cues, and chatbot persona

The cognitive and emotional aspects are both part of trust in AI (Glikson & Woolley, 2020). Cognitive trust involves people's judgments regarding the reliability and competence of the system, and the rational grounds for accepting advice. It is enhanced by transparency, uniform performance, proper task assignment, and explanations that enable the user to understand why a recommendation makes sense. Emotional trust is based on the feeling of social connection, the sense of reassurance, and the feeling of support. It can be affected by anthropomorphism, embodiment, the use of names, responsiveness, and interaction behaviors that mimic active listening.

For virtual AI, immediate responses are especially important. Giving prompt replies, showing awareness of the user's concerns, providing examples, adjusting how one speaks, and trying to see things from the user's point of view can make the system seem attentive. However, the appropriate level of such behavior will vary depending on the user's preferences and the nature of the task. Being extremely human-like can lead to unrealistic expectations, cause discomfort, or make a very capable system seem manipulative. In the context of legal advice, it is particularly dangerous to build emotional trust without a cognitive basis, since a comforting tone might be mistaken for legal accuracy.

Occupational personas offer a concise way to bring together expectations regarding expertise and the way one interacts. Thus, a judge persona can convey neutrality, a focus on the rules, and authority; a lawyer persona can represent strategic advocacy and assertiveness; and a mediator persona can express empathy, balance, and a pledge to cooperative problem-solving. Even if the underlying model's basic legal capabilities have not changed, these labels can still influence user expectations. The personas used here, therefore, acted as a package of design features rather than as statements about the actual personalities of judges, lawyers, or mediators.

1.4 Research questions and pilot logic

This section presents the main research questions and the rationale for the pilot study. The main goal is to evaluate the practical effectiveness of the proposed method. Specifically, the study considers whether the method produces better results than current techniques, assesses user access and safety, and determines its applicability throughout various settings.

The pilot study was arranged to address these questions in real-world contexts. Data were collected from participants and compared with previous datasets to evaluate whether the method met its objectives and to determine areas for potential improvement.

The main research question was: What patterns and effect sizes emerge from the correlations between the five OCEAN traits of participants and their perceptions of the difficulty of expertise-dependent tasks for each AI personality, and what hypotheses can be formulated for future research? As an exploratory working hypothesis, it was suggested that there would be associations between the traits and the difficulty, and that the direction and strength of these associations would differ among the judge, lawyer, and mediator personas.

The study, as a pilot, was not intended to determine a stable interaction effect or to identify a chatbot superior for deployment. Rather, it aimed to assess whether the procedure could produce interpretable variation, identify any problems with measurement and the workflow, and obtain parameters for a larger study. This difference determines how the analysis is carried out: the persona-level ANOVA is given a descriptive account, with effect sizes, while the within-persona correlations are regarded as indications for forming hypotheses rather than as evidence of relationships in the population.

2. Method

2.1 Participants and design

The sample consisted of 16 people (8 women and 8 men) with an average age of 26.50 years (standard deviation = 4.32). The majority of the participants had studied psychology (56.3%) or psychotherapy science (18.3%) at Sigmund Freud Private University in Vienna. They were assigned to one of three different chatbot roles: judge (n = 5), lawyer (n = 5), or mediator (n = 6). Each participant completed three of the six legal situations, yielding 48 task interactions. The throw of a die determined the choice of persona and scenario.

The key element in the design was the chatbot persona. The users' personality characteristics were measured using continuous variables. The main outcome was the average score each participant gave to the experienced difficulty of the three tasks they performed. The thesis defines difficulty as an ordinal item that is obtained by taking the arithmetic mean; the precise response anchors must be included in the confirmatory protocol.

The binary gender was balanced in the data based on participants' self-reports. Eight people were enrolled at the bachelor's level and eight at master's or magister's level. About 62.5% of them worked while studying. Recruitment took place spontaneously within the main university building, so participants were available without waiting for a scheduled appointment. Although this convenient method was suitable for assessing feasibility, it yielded a sample that was relatively young, highly educated, and oriented toward psychology.

The study used a between-participants approach to assigning personas and included repeated tasks for each participant. The three groups of personas had five, five, and six members respectively. Each individual provided a mean difficulty score for three interactions. Even though 48 scenarios were carried out in all, the inferential analyses based on the participant means used a sample size (N) of 16 rather than the full N of 48. The sample would have been artificially inflated if the interactions been been treated as independent, since ratings from the same person are correlated.

2.2 Chatbot personas

The study used a beta version of AMIGA-AI accessed via a browser interface. Participants entered messages and received text replies. Before each session, the experimenter set up one occupational

persona. The personas were defined in terms of the OCEAN dimensions – openness (O), conscientiousness (C), extraversion (E), agreeableness (A), and neuroticism (N). The values used were taken as design parameters rather than as estimates of the actual occupational norms.

Table 1
Prespecified OCEAN design parameters for the three chatbot personas.

Persona	O	C	E	A	N	Intended interaction style
Judge	.55	.90	.30	.60	.25	Structured, rule-oriented, restrained
Lawyer	.70	.80	.70	.45	.35	Strategic, persuasive, more confrontational
Mediator	.75	.85	.60	.80	.20	Perspective-taking, empathic, consensus-oriented

2.3 Materials and procedure

The scenarios dealt with contract formation, negligence, employment status, withdrawal rights, latent defects, and employee monitoring. At the university site, participants first completed the BFI-44 and FEEL-E using the SoSci Survey. They then interacted with the chatbot persona assigned to each of the three scenarios until they were satisfied with their assessment, then made a brief legal conclusion and rated the difficulty. Finally, they responded to questions specific to the study regarding their attitudes towards AI, their intended future use of it, and the adapted technophobia items. They were afterward debriefed and given a small snack that had not been mentioned during the recruitment process.

The OCEAN traits were measured in participants using the BFI-44 (John et al., 2012), and FEEL-E emotion-regulation scores were added as a further check of the interpretation of neuroticism (Grob & Horowitz, 2014). Since the study-specific future-use and AI-attitude scales had low internal consistency ($\alpha = .21$ and .54, respectively), they were not considered main outcomes. The adapted technophobia scale, on the other hand, showed greater promise ($\alpha = .78$).

2.3.1 Legal scenarios

The range of scenarios included various examples of ordinary legal reasoning. Each vignette contained a certain level of ambiguity, which meant that an interpretation was necessary rather than basic factual recall. The participants were asked to provide a probable legal evaluation, along with a rationale, rather than a definite professional opinion. This approach permitted dialogue with the chatbot while lessening the need for participants to have specialist legal training.

Table 2
Overview of the six legal scenarios used in the experiment.

Case	Legal topic	Core question
1	Contract formation	Whether a seller's statement that a bicycle was reserved created an enforceable claim after sale to another buyer.
2	Negligence	Whether a balcony owner was liable when a flowerpot fell during unusually strong wind and damaged a car.
3	Employment status	Whether a delivery-platform worker labelled self-employed could qualify as an employee because of schedules, task allocation, and performance monitoring.
4	Withdrawal right	Whether an online purchaser retained a statutory right to return shoes ten days after delivery despite trying them on.
5	Latent defect	Whether a private buyer had a remedy when a laptop advertised as fully functional had a battery lasting only minutes.
6	Employee monitoring	Whether undisclosed keystroke logging and screenshots on work laptops infringed an employee's privacy.

The distribution of scenario allocation was roughly even among the different personas, though not exactly so. Of the 48 completed tasks, cases 1 through 6 occurred 9, 10, 8, 5, 9, and 7 times, respectively. Case 4 had the greatest imbalance, as it was carried out three times in the judge group and once in both the lawyer and mediator groups. Since the task difficulty was not separately calibrated, this unevenness is a potential confounder, even though the overall distributions were similar.

2.3.2 Personality and emotion-regulation measures

The 44-item Big Five Inventory was chosen as a middle ground between the conciseness of the BFI-10 and the comprehensiveness of the 240-item NEO-PI-R; the items were rated on a five-point scale ranging from 1 (does not apply at all) to 5 (applies very well). The fact that the same OCEAN framework was applied to both participants and system personas enabled the pilot study to address questions about the absolute traits of the personas and the degree of similarity between users and these personas.

The FEEL-E is used to evaluate regulation strategies for anger, fear, and sadness. Examples of adjustment strategies are taking action directed at the problem, acceptance, cognitive problem-solving, reappraisal, boosting one's mood, and forgetting. As for maladaptive strategies, they include withdrawal, self-devaluation, giving up, perseveration, catastrophizing, and blaming others. The instrument wasn't used as a second direct personality scale; rather, it served as convergent evidence for neuroticism, which had yielded counterintuitive results in earlier AI research. The items on the FEEL-E make use of a five-point frequency scale ranging from rarely to almost always.

2.3.3 AI attitudes, use intentions, and adapted technophobia

Three items specific to the study examined reasons people would use AI legal advice in the future: pleasant use, the perception that the advice was competent, and the high cost of hiring a professional. Three more items examined whether respondents saw an AI label as indicating quality, whether they regarded AI applications as dangerous, and whether they, in general, considered computers to be useful. For each of these, a five-point agreement scale was used. The low internal consistency indicates that they reflected different variables rather than tapping coherent underlying scales; therefore, the results at the item level were retained.

A 13-item scale designed to measure technophobia was first developed in connection with automated banking technology (Sinkovics, 2003) and subsequently modified by substituting references to bank machines with references to AI. The items included concerns about anxiety and frustration, a feeling of inferiority when using AI, a preference for human service, opposition to AI's influence, trust in professionals, the displacement of human jobs, comfort, ease of access to specialist information, and convenience. Three of the positively worded items had to be reverse-scored. Although the context had changed significantly, the resulting scale still showed nearly acceptable internal consistency ($\alpha = .78$) and was used descriptively.

2.3.4 Experimental sequence

Before each session, the experimenter used die rolls to select one persona and three scenarios. A participant profile, a random identifier, a survey tab, a chatbot session, and an individual answer document were prepared. At the beginning of the session, the experimenter delivered a standardized explanation of the study, procedure, and data protection. Participants were told that they could ask the chatbot questions until satisfied, that the experimenter could assist only with technical issues, and that the study focused on their reasoning process and experience rather than a single right or wrong response.

Phase 1 consisted of the BFI-44 followed by FEEL-E in the SoSci Survey. In Phase 2, the experimenter opened AMIGA-AI and explained the text interface. The participant received one scenario at a time, conversed with the assigned persona, recorded a short legal conclusion in their own words, and rated the task before receiving the next scenario. In Phase 3, the participant returned to SoSci Survey to complete the AI-attitude, future-use, and technophobia items. Session duration ranged from approximately 15 to 90 minutes, reflecting substantial variation in interaction depth. No time limit was imposed because restricting dialogue could reduce ecological validity.

2.4 Statistical analysis

Given the pilot objective and subgroup sizes of five to six, trait-difficulty associations were summarized with Spearman’s rho. P values for these correlations were not interpreted, and no multiplicity-adjusted confirmatory claims were made. Correlation magnitudes were used only to generate hypotheses. An exploratory one-way ANOVA compared mean reported difficulty across personas; Bonferroni-adjusted pairwise comparisons were reported. Effect sizes included η^2 and ω^2. Fifteen persona-specific correlations were calculated: five user traits crossed with reported difficulty in each of three persona

groups. In a confirmatory analysis, this family would require an explicit multiplicity strategy; a simple Bonferroni correction would yield α ≈ .003. The pilot deliberately did not use such a threshold because significance testing with n = 5–6 per subgroup was deemed uninformative. Coefficients with absolute values of approximately 0.30, 0.50, and 0.70 were used as qualitative markers of small, medium, and large patterns. The possibility of demographic confounding was examined by examining correlations with age and gender, both in the total sample and within the different personas. Partial correlations were not carried out because, for the smallest persona group, there would have been only 2 degrees of freedom for the residuals after considering 1 control variable, leading to very unstable estimates. Nominal variables, including field of study, employment type, and religion, were not suitable for the simple partial-correlation approach. The analysis, therefore, placed greater emphasis on delivering a clear description rather than on overfitting adjustments.

3. Results

3.1 Sample attitudes and measurement performance

The participants showed a high level of acceptance of computers, along with reservations about replacing professional advice. For all three future-use items, the scores were above the midpoint. The average score for deciding to continue using the system because the interaction had been pleasant was 3.81 (SD = 0.91); the average score for continuing to use it because the chatbot seemed competent was 4.31 (SD = 0.60); and the average score for continuing to use it because a professional might be too expensive was 3.50 (SD = 1.55). The item on competence yielded the clearest consensus, with 93.75% of participants selecting one of the two agreement categories. By contrast, the cost item was highly divided, with 68.75% agreeing and 31.25% disagreeing; no one had chosen the neutral option. The 'AI' label was not seen as an indication of quality (M = 2.38, SD = 0.81): 56.25% of people disagreed that the 'works with AI' statement meant the program or application was of good quality, and 37.50% were neutral. Yet, 43.75% agreed that AI applications could be dangerous, 37.50% were neutral, and 18.75% disagreed (M = 3.38, SD = 0.96). Almost all the participants thought that computers were generally useful (M = 4.88, SD = 0.50). This shows that judgments about the utility of technology, AI risk, and trust are separate.The technophobia scale, as adapted, showed a low average score (M = 2.53, SD = 0.54), indicating that the individuals in the sample generally did not feel much fear or discomfort about AI. Yet, 75% of the participants stated that they trusted a human expert more than AI (M = 4.06, SD = 0.93). At the same time, 81.25% said that AI had made it easier to get specialist information (M = 4.19, SD = 0.75), and 75% agreed that using AI was convenient (M = 4.00, SD = 1.10). Thus, the sample did not reject AI; instead, the participants distinguished between trust in professionals and access and convenience.

Table 3
Descriptive results for selected AI attitude and future-use items (N = 16).

Construct or item	M	SD	Interpretive summary
Future use: pleasant interaction	3.81	.91	81.3% agreed
Future use: competent advice	4.31	.60	93.8% agreed
Future use: professional costs	3.50	1.55	Polarised; 68.8% agreed
AI label as quality signal	2.38	.81	56.3% disagreed
AI applications are dangerous	3.38	.96	Mixed responses
Computers are useful	4.88	.50	93.8% strongly agreed
Adapted technophobia total	2.53	.54	Low overall technophobia
Trust professional more than AI	4.06	.93	75.0% agreed
AI eases access to specialist information	4.19	.75	81.3% agreed

The psychometric findings make it impossible to draw any conclusions regarding the custom scales. The three items concerning general AI attitudes yielded $\alpha = .21$, and the three items relating to future use yielded $\alpha = .54$. These coefficients do not justify the use of a single total score. Although the adapted 13-item technophobia scale produced an α of .78, its content validity with respect to contemporary generative AI still needs to be assessed because the original scale was developed in the context of an earlier technology.

3.2 Persona differences in perceived task difficulty

Participants assigned to different personas reported varying levels of perceived task difficulty. Some people found specific tasks straightforward, whereas others found them challenging. This pattern was consistent across most tasks evaluated. For instance, individuals in the "Expert" persona completed the data entry task rapidly. They rated it as simple, while those in the "Novice" persona required more time and rated the task as difficult. These outcomes indicate that the assigned persona can influence users' perceptions of task difficulty.

The level of difficulty varied depending on the persona condition, with $F(2, 13) = 33.60$, $p < .001$, $\eta^2 = .84$, $\omega^2 = .80$. The lowest mean difficulty was observed for the mediator persona ($M = 1.67$, $SD = 0.42$), then followed the lawyer ($M = 3.60$, $SD = 0.43$), and finally the judge personas ($M = 4.07$, $SD = 0.68$). The Bonferroni comparisons showed that the difficulty was lower for the mediator than for the judge (mean difference = − 2.40, 95% CI [− 3.26, − 1.54], $p < .001$) and also lower for the mediator than for the lawyer (mean difference = − 1.93, 95% CI [− 2.80, − 1.07], $p < .001$). There was no difference between the judge and the lawyer (mean difference = 0.47, 95% CI [− 0.44, 1.37], $p = .538$).

3.3 Exploratory user-trait patterns within personas

Table 4
Spearman correlations between user Big Five traits and mean perceived task difficulty.

Persona	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Judge	.11	.32	.97	.00	.11
Lawyer	.05	−.29	.29	.39	−.37
Mediator	−.74	−.74	−.92	−.34	−.55

The mediator condition showed negative correlations throughout all five user traits, with the largest magnitudes for conscientiousness (ρ = −.92), extraversion (ρ = −.74), and agreeableness (ρ = −.74). In the judge condition, conscientiousness correlated positively with difficulty (ρ = .97). The lawyer condition showed only small-to-moderate coefficients. These estimates are extremely unstable because they are based on five. Across the entire sample, neuroticism showed a positive correlation with maladaptive emotion-regulation strategies (ρ = .63) and a near-zero correlation with adaptive approaches (ρ = −.09), which generally supports the intended interpretation of the neuroticism scores. Moreover, the exploratory demographic analyses carried out within the judge subgroup also yielded large coefficients. Still, these were probably influenced by other factors and lacked sufficient power to allow reliable separation.bg3.4 Demographic and emotion regulation checks are underpowered for reliable separation.

3.4 Demographic and emotion-regulation checks

If we ignore the assignment of persona, age showed a small-to-moderate negative relationship with difficulty (ρ = −.40), and the point-biserial correlation with gender was only small (ρ = .17). Among the judges, however, age was negatively correlated with difficulty (ρ = −.76). The gender code (with women coded as 1 and men as 2) was positively correlated with difficulty (ρ = .91). Gender also showed a strong correlation with conscientiousness in this subgroup (ρ = .89). Because these correlation coefficients overlap so greatly, it is not possible to tell whether the pattern of conscientiousness seen in the judge group reflects personality, gender, age, or a small number of influential observations.

The pattern obtained from the FEEL-E check was theoretically coherent. There was a positive correlation between neuroticism and the total score for maladaptive strategies (ρ = .63), as well as with anger-related maladaptive regulation (ρ = .59), fear-related maladaptive regulation (ρ = .34), and sadness-related maladaptive regulation (ρ = .55). The associations with adaptive regulation, on the other hand, were negative or approximately zero: the total correlation was ρ = −.09, that for anger was ρ = −.41, for fear it was ρ = −.32, and for sadness ρ = −.01. Although these correlations do not confirm the validity of the BFI-44 scores within a formal measurement model, they lessen the concern that neuroticism had been reversed or miscoded.

3.5 Scenario allocation and protocol feasibility

Scenarios are assigned based on each site's capacity and prior experience with related studies. Protocol feasibility is assessed by evaluating staff availability, participant population, and the site's ability to

adhere to study requirements. Each site is consulted to discuss possible challenges, and allocations are adjusted as necessary to minimize delays and boost resource utilization.

The 16 participants were almost equally divided among the different roles: 31% were judges, 31% were lawyers, and 38% were mediators. Gender distribution was also roughly equal across conditions (among the judges, three women and two men; among the lawyers, two women and three men; and among the mediators, three women and three men). In the six cases, the number of scenarios ranged from 5 to 10; most proportions involving a particular role and a case were close to the ideal figure of 1/6, but cases 3 and 4 showed the clearest deviations. Such imbalance must be corrected by using computerized blocked randomization in the main study.

The technical elements—namely the computer, the scenario sheets, access to AMIGA-AI, and SoSci Survey—were mostly stable. The biggest procedural disturbance was caused by using separate Word documents for the answers and the difficulty ratings. Since these documents had to be created, the preparation time increased, and switching between the chat and the documents disrupted the interaction. Additionally, three of the participants went beyond the FEEL-E instruction, which asked them to contact the experimenter. These are workflow flaws that can be corrected and do not prove that the paradigm itself is not feasible.

The sessions ranged in length from 15 to 90 minutes. This wide range could be due to differences in reading speed, prior legal knowledge, the depth of the dialogue, and the individual participants' thresholds to consider their queries satisfactorily answered. Although the variation is of substantial interest, it also creates an uncontrolled difference in exposure: those who interact for a longer time get more output from the chatbot and may therefore become more familiar with it or feel more fatigued. Future studies should record the number of messages, the number of words, the latency, the elapsed time, and the number of clarification exchanges, even if a strict time limit is not imposed.

4. Discussion

This pilot provides preliminary evidence that persona framing can materially change how difficult users perceive expertise-dependent chatbot tasks. The mediator persona was associated with markedly lower difficulty than either the judge or lawyer persona. Because persona bundled multiple OCEAN parameters and occupational cues, the study does not identify which component caused the difference. Plausible candidates include the correlations found in the trait-stratified analysis, which indicate that users' personality traits might have a moderating effect on the persona effects. However, their main usefulness lies in guiding design. It is possible that the negative coefficients in the mediator group indicate that users who scored high on several traits found the interaction easier; it is also conceivable that the pattern is due to sampling noise or reflects unmeasured differences in demographic characteristics or user familiarity. Similarly, the positive coefficient for conscientiousness in the judge condition is interesting but is too unstable to allow for any substantial conclusion. The findings support the general idea that both personality traits and social signals are important in interactions between humans and AI.

Previous studies have linked agreeableness and conscientiousness with the quality of tasks carried out with the aid of AI and with people's preferences for technology (Kovbasiuk et al., 2025). In contrast, research on trust draws attention to the importance of cues relating to competence, transparency, reliability, and immediacy (Glikson & Woolley, 2020). A mediator-type personality could provide cognitive support alongside socially reassuring signals, therewith reducing the sense of complexity but not altering actual performance.ly4.1 Understanding the advantage of the mediatoreived complexity absent necessarily changing objective performance.

4.1 Interpreting the mediator advantage

The strongest descriptive finding was not a correlation but rather the separation of the group means. The mediator group found the tasks to be more than 2 scale points easier than the judge group and almost 2 points easier than the lawyer group. The mediator contrasts were large in comparison to the within-group variability, while there was no difference between the judge and lawyer groups. This pattern corresponds to a mediator advantage. However 'mediator' was a composite treatment whose profile included the highest level of agreeableness (.80), the highest level of openness (.75), the lowest level of neuroticism (.20), a moderately high level of extraversion (.60), and a high level of conscientiousness (.85). The treatment also had an occupational label linked to listening, balance, and conflict resolution.

The same observed mean difference could originate from several different mechanisms. One possibility is that high levels of agreeableness led to a more cooperative, polite, and validating style of communication. This type of language helps to reduce interpersonal tension and promotes the asking of explanatory questions. Another possibility is that high openness enabled the use of alternative examples and changes in perspective, which are helpful when dealing with ambiguous legal situations. A third possibility is that low neuroticism decreased the chance of experiencing urgency, defensiveness, or using emotionally charged language. Fourth, the fact that the person had the role of mediator might have altered people's expectations before any response had been read. Lastly, the way the implementation was carried out may have led to differences in content beyond tone, for instance, by producing more structured summaries or more cautious qualifications.

It is not possible for the data to distinguish between these mechanisms, since the five persona dimensions were not varied independently, and the chatbot transcripts were not coded. A future content analysis should measure response length, readability, the number of legal ideas used, the number of examples given, the presence of statements conveying uncertainty, the number of questions asked of the participant, the use of validation phrases, and any explicit recommendations. It will only be after that that researchers will be able to decide if the effect is due to social style, the organization of the information, the quality of the substantive answers, or a combination of these aspects.

Just because the difficulty is perceived as low does not mean it is advantageous. A persona might make a task seem easy by communicating clearly, but it could, at the same time, lead to unjustified confidence or hide the complexity involved. Since legal problems often require caveats, assumptions about

jurisdiction, and an recognition of missing facts, future research needs to therefore combine subjective difficulty ratings with objective ratings of legal quality and calibration measures. The desired result should be an accurate understanding together with appropriately calibrated confidence—not just a pleasant or effortless experience.

4.2 User-persona fit and Big Five patterns

The persona-specific coefficients offer two opposing theoretical approaches. On the one hand, there is an account that operates at an absolute level, according to which certain chatbot characteristics, for example, a high degree of agreeableness or low neuroticism, might strengthen communication for the majority of users. On the other hand, there is a similarity or fit approach, in which advice is seen as more helpful when the system's behavior corresponds to the user's personality or preferred way of interacting. The pilot study does not settle this issue. Since the mediator persona had both high agreeableness and high conscientiousness, its negative correlations with the users' respective traits are consistent with either a general benefit of the persona or with trait congruence.

This example shows the dangers of direct interpretation. The degree to which conscientiousness correlated with greater difficulty was .97, suggesting that highly conscientious participants examined the rule-focused judge more carefully or felt there was greater legal complexity. Yet, in the same group of five people, gender was correlated .89 with conscientiousness and .91 with difficulty. Because there were so few observations, a simple ranking might produce an extremely high Spearman's rank correlation coefficient. Therefore, a bigger study should examine moderation within a single model rather than dividing the sample into such small groups.

Neuroticism warrants close examination, as its effects in AI have been inconsistent. Although traditional descriptions of the trait link neuroticism with threat sensitivity and negative emotions, Kovbasiuk et al. (2025) found performance patterns that a simple deficit model could not explain, and Grassini et al. (2025) found that neuroticism predicted people's attitudes towards AI. The FEEL-E results from this pilot study show that participants' neuroticism scores moved in the direction expected by theory, consistent with maladaptive regulation. The more interesting question to consider might therefore be neuroticism as expressed by the system—namely, whether a calm and emotionally stable personality reduces the perceived burden regardless of the user's level of neuroticism.

4.3 Trust, usefulness, and professional substitution

The findings regarding attitude show a helpful kind of contradiction. While the participants strongly favored computers and believed that AI makes it easy to obtain specialist information, the majority would go on to use the system since it appeared competent, yet most still trusted a professional more than they trusted AI, and the AI label was not seen as an indication of quality. This does not make things inconsistent since cognitive trust can be limited and specific to a task: people may think a chatbot is convenient and informative, yet keep the final decision-making power with a human professional.

The AmigaAI Legal LLM project means that this distinction has design consequences. A responsible legal assistant ought to not present adoption as a way to replace human expertise in all cases; rather, it can help with orientation, providing explanations, preparing questions, and identifying situations that require referral to a professional. Persona cues that reduce the barriers to interaction should be combined with clear statements about the scope, communication of uncertainty, and suggestions for escalation. The objective must be to improve access to and understanding of the information without promoting uncritical dependence.

The way people have reacted to the cost of professional legal advice points to an additional significant route. Since some individuals could resort to AI mainly because traditional legal advice is out of reach, it becomes all the more important, ethically, to secure accuracy and build calibrated trust; those who are most affected by cost will have fewer opportunities to check an answer. A future study should assess people's access to legal services, their previous experience with lawyers, their perception of affordability, their level of AI literacy, and their existing knowledge of the law. These factors are able to account for both the extent of reliance and the preferences regarding the type of AI persona.

4.4 Implications for the main study

Instead, a confirmatory study should examine the various dimensions of personal personality traits rather than viewing occupational personas as fixed, unchangeable packages. A factorial design might be used to compare high levels of agreeableness with moderate levels, and low levels of neuroticism with moderate levels, while keeping the response content, the model version, the temperature, and the legal knowledge constant. The outcomes should include objective answer quality, perceived comprehensibility, cognitive trust, interaction length, and reliance, together with perceived effort.

Power analyses derived from the thesis suggested a sample size of N = 66 for a three-group ANOVA, with an effect size of f = .40, a significance level of α = .05, and a power of .80. While this proposed sample size of 66 serves as a useful lower bound, the actual sample size should be recalculated to take into account the preregistered primary contrast, expected attrition, the clustering that occurs across the repeated scenarios, and more conservative assumptions regarding the effect size. It is not advisable to treat the suggested N = 11 for one-tailed correlations based on |ρ| = .70 as a definite target since pilot correlations obtained from samples of n = 5 to 6 are both highly inflated and uncertain.

A more effective next design should involve the use of a factorial manipulation instead of having three different occupational packages. For instance, the chatbot's agreeableness (compared at moderate and high levels) and its expressed emotional stability (also compared at moderate and high levels) could be crossed, with the occupational labels kept constant or left out. There could then be a separate label condition to see whether referring to the same behavior as ‘mediator’, ‘lawyer’, or ‘judge’ has an effect on people's expectations. All response-generation parameters, system prompts, retrieval sources, and model versions must be archived. Manipulation checks should confirm that the participants actually perceived the intended differences in warmth, assertiveness, flexibility, authority, and emotional stability.

The hierarchy of outcomes should be established in advance. An appropriate primary outcome could be either perceived comprehensibility or an objective comprehension score rather than difficulty alone. The secondary outcomes might be considered difficulty, the quality of the answers as assessed by blind legal experts, confidence, confidence calibration, cognitive trust, emotional trust, willingness to rely on the system, willingness to ask for human review, conversation length, and future-use intention. Wherever possible, multi-item validated measures should be used instead of ad hoc totals.

The analysis of the repeated-scenario structure should be carried out using a multilevel model since the task ratings are associated with individual participants and the scenarios differ in both topic and innate complexity. By including random intercepts for both participants and scenarios, it would be possible to distinguish between stable user tendencies and the difficulty of individual cases. The effects of the persona and users' characteristics could be included as fixed effects, along with the interactions specified on theoretical grounds. Finally, blocked or stratified randomization should be used to ensure that gender, age, prior experience with AI, and legal knowledge are balanced across the conditions.

The figure of 66 assumes a very large overall effect and says nothing about interactions or multilevel uncertainty. To plan, a smaller, justifiable target effect should be used, along with a simulation based on the intended model. When the main question involves an interaction between a user trait and a persona, a number significantly greater than 66 may be needed. The final calculation should be made after the primary estimand has been determined, the number of repeated scenarios has been decided upon, the expected intraclass correlation has been established, and the correction strategy has been set.

4.5 Feasibility lessons

The overall protocol was workable, but the pilot identified several improvements. Difficulty ratings should be collected directly in the survey rather than in separate Word files; participant answers can be marked in the chat interface to reduce workflow disruption. A stronger transition instruction is needed after FEEL-E. The custom AI-attitude items should be replaced or supplemented with a validated measure such as the General Attitudes towards Artificial Intelligence Scale (Schepman & Rodway, 2020). Testing should occur in a standardized quiet room when internal validity is prioritized, while a field setting may be retained for ecological-validity questions.

The choice between internal and ecological validity should be made explicitly. A quiet laboratory improves control over distractions and timing. A moderately busy setting better resembles real-world use but introduces noise that may interact with reading speed or stress. A two-stage programme could first establish causal effects under controlled conditions and then test generalisability in realistic environments.

The main study should use computerised randomisation and automatic data capture. The interface can display a scenario, host the chat, collect a marked final answer, and present the difficulty and comprehension measures without manual tab switching. Logs should be linked through pseudonymous

participant IDs. The system should preserve exact prompts and outputs while protecting personal data and excluding unnecessary identifiers.

4.6 Limitations

The sample was small, non-representative, and drawn largely from students at one university. Randomization was manual, subgroup sizes were unequal, and three of six scenarios were sampled per participant. Scenario difficulty was not independently calibrated, the mediator manipulation combined several design attributes, and the technical model configuration was not fully documented in the thesis. The analysis aggregated three ordinal ratings and did not model repeated scenarios. The large ANOVA effect may therefore reflect persona, wording, scenario allocation, model-output variation, or chance. No causal claim about specific OCEAN dimensions is warranted.

The use of occupational profiles raises questions about construct validity. The numerical OCEAN values were assumed design parameters informed by literature, not measurements of actual judges, lawyers, or mediators. The categories also risk reinforcing stereotypes about professions. Future manuscripts should frame them strictly as interface prototypes and report the exact behaviors produced by each prompt rather than implying that professional groups possess fixed personalities

The outcome was subjective difficulty, averaged across three ordinal ratings. The manuscript could not verify the precise response anchors from the available thesis description, and no independent measure of legal accuracy or comprehension was reported. Because participant responses and chatbot transcripts were not reanalyzed here, the paper cannot determine whether lower difficulty corresponded to better reasoning. The apparent normality of five- or six-person groups provides only weak reassurance, and ANOVA p-values remain sensitive to small samples despite homogeneity-test results.

The custom attitude and future-use scales lacked sufficient internal consistency, and adaptation of the technophobia scale substantially changed its target technology. The convenience sample had high overall computer acceptance, which may limit generalisability to less experienced or more anxious users. Neither previous generative-AI use nor legal knowledge was integrated into the reported central analyses. Finally, ethics approval, model details, data availability, project identifiers, and conflict-of-interest statements must be completed before dissemination.

5. Conclusion

In this small pilot, a chatbot that took the form of a mediator was found to be associated with a considerably lower perceived level of task difficulty than those styled as judges or lawyers. While the study outlines a feasible approach and defines a clear set of follow-up questions, it does not make any claim about general effectiveness. To establish whether the design of the persona can reliably make complex professional advice more understandable, a preregistered and adequately powered experiment that isolates the traits of agreeableness, neuroticism, and occupational framing is require

The main value of the pilot study lies in linking persona design to a more specific understanding of AI acceptance; participants appreciated convenient access to specialist information and saw the AI as competent, yet they still placed more trust in human professionals. It therefore follows that legal conversational AI should be developed as a carefully calibrated tool for explanation and access rather than being offered as a seamless replacement for professional judgment.

The next step for AmigaAI Legal LLM should involve a transparent comparison of individual social and communicative cues, along with a blind evaluation of legal quality, comprehension tests, and trust calibration measurements. This approach will enable the determination of whether the mediator effect results from clearer explanations, a warmer interaction style, the presence of the occupational label, or sampling variation. It will also clarify whether making the task appear easier positively influences users' ability to reach better-informed conclusions.

Declarations

Ethics approval: The study was conducted as part of the research project “AmigaAI” which was reviewed and approved by the Ethics Commission of Sigmund Freud University Vienna (Signature: GCLV2HEFBKU1GD89926).

Consent to participate: Informed consent to participate in the study was obtained from all participants.

Funding.

This work was conducted within the FFG project “AmigaAI Legal LLM.” [Digitale Technologien and FO999915238.]

Ethics. [Ethics Commission, GCLV2HEFBKU1GD89926.]

Data and materials availability. Owing to the small sample and the associated risk of participant re-identification, the participant-level data are not publicly available. A suitably de-identified dataset may be provided by the corresponding author upon reasonable request, subject to applicable ethical, legal, and data-protection requirements.

Competing interests. The authors participated in the FFG-funded AmigaAI Legal LLM project and were involved in the development and/or evaluation of AMIGA-AI. The authors declare that they have no financial ownership, equity interests, patent interests, or other commercial interests related to AMIGA-AI. The funding body had no role in the analysis or interpretation of the data, the preparation of the manuscript, or the decision to publish.

Author contributions.

Armin Klaps: Conceptualization, Methodology, Software, Project administration, Supervision, Visualization, Writing – original draft, and Writing – review and editing. Zuzana Kovacovsky: Conceptualization, Methodology, Project administration, Validation, and Writing – review and editing. Birgit U. Stetina: Conceptualization, Methodology, Supervision, Validation, and Writing – review and editing. All authors reviewed and approved the final manuscript.

References

1. Barrick MR, Mount MK (1991) The Big Five personality dimensions and job performance: A meta-analysis. Pers Psychol 44(1):1–26
2. Bellaiche L, Shahi R, Turpin MH, Ragnhildstveit A, Sprockett S, Barr N, Christensen A, Seli P (2023) Humans versus AI: Whether and why we prefer human-created compared to AI-created artwork. Cogn Research: Principles Implications 8(1):42. https://doi.org/10.1186/s41235-023-00499-6
3. Costa PT, McCrae RR (1985) Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI). Psychological Assessment Resources
4. De Freitas J, Agarwal S, Schmitt B, Haslam N (2023) Psychological factors underlying attitudes toward AI tools. Nat Hum Behav 7(11):1845–1854. https://doi.org/10.1038/s41562-023-01734-2
5. Furnham A (2001) Personality and occupational behaviour. Br J Psychol 92(1):1–24. https://doi.org/10.1348/000712601162276
6. Glikson E, Woolley AW (2020) Human trust in artificial intelligence: Review of empirical research. Acad Manag Ann 14(2):627–660. https://doi.org/10.5465/annals.2018.0057
7. Grassini S, Thorp SO, Ree AS, Sevic A, Cipriani E (2025) Distinct predictors of positive attitudes toward artificial intelligence and general technology: Big Five traits, gender, and age. Behav Inform Technol 1–14. https://doi.org/10.1080/0144929X.2025.2598623
8. Grob A, Horowitz D (2014) Fragebogen zur Erhebung der Emotionsregulation bei Erwachsenen (FEEL-E). Hans Huber
9. John OP, Donahue EM, Kentle RL (2012) Big Five Inventory [Data set]. https://doi.org/10.1037/t07550-000
10. Kaya F, Aydin F, Schepman A, Rodway P, Yetişensoy O, Kaya D, M (2024) The roles of personality traits, AI anxiety, and demographic factors in attitudes toward artificial intelligence. Int J Human–Computer Interact 40(2):497–514. https://doi.org/10.1080/10447318.2022.2151730
11. Kovbasiuk A, Triantoro T, Przegalińska A, Sowa K, Ciechanowski L, Gloor P (2025) The personality profile of early generative AI adopters: A Big Five perspective. Cent Eur Manage J 33(2):252–264. https://doi.org/10.1108/CEMJ-02-2024-0067
12. Li J, Huang J-S (2020) Dimensions of artificial intelligence anxiety based on the integrated fear acquisition theory. Technol Soc 63:101410. https://doi.org/10.1016/j.techsoc.2020.101410
13. Ostendorf F, Angleiter A (2004) NEO-PI-R: NEO-Persönlichkeitsinventar nach Costa und McCrae. Hogrefe

14. Rammstedt B, Kemper CJ, Klein MC, Beierlein C, Kovaleva A (2014) Big Five Inventory (BFI-10). Zusammenstellung sozialwissenschaftlicher Items und Skalen. https://doi.org/10.6102/ZIS76
15. Redding RE, Reppucci ND (1999) Effects of lawyers' and judges' gender and gender-related attitudes on professional behavior. Law Hum Behav 23(1):73–92. https://doi.org/10.1023/A:1022311822171
16. Schepman A, Rodway P (2020) Initial validation of the General Attitudes towards Artificial Intelligence Scale. Computers Hum Behav Rep 1:100014. https://doi.org/10.1016/j.chbr.2020.100014
17. Schulze D, Schmitt M (2004) Beruf und Persönlichkeit: Big-Five-Prädiktoren beruflicher Interessen und Leistungen. Z für Differentielle und Diagnostische Psychologie 25(4):217–226. https://doi.org/10.1024/0170-1789.25.4.217
18. Sinkovics R (2003) Technophobie. Zusammenstellung sozialwissenschaftlicher Items und Skalen. https://doi.org/10.6102/ZIS62

Figures

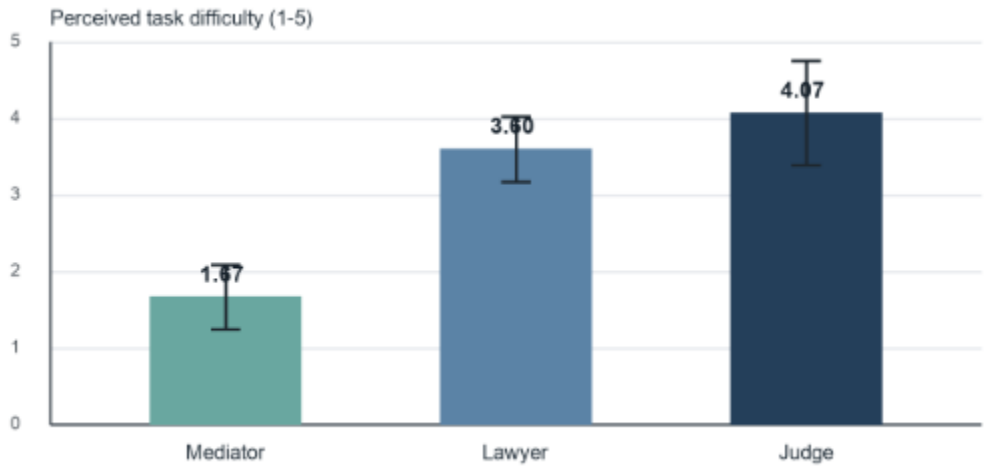


Figure 1

Mean perceived task difficulty by chatbot persona. Error bars show ±1 SD. Descriptive values reproduced from the thesis.