Enabling Trait-based Personality Simulation in Conversational LLM Agents: Case Study of Customer Assistance in French

Ahmed Njifenjou, Virgile Sucal, Bassam Jabaian, Fabrice Lefèvre
Laboratoire Informatique d'Avignon (LIA), CERI - Avignon Université
{ahmed-ndouop.njifenjou & firstname.lastname}@univ-avignon.fr

Abstract

Among the numerous models developed to represent the multifaceted complexity of human personality, particularly in psychology, the Big Five (commonly referred to as 'OCEAN', an acronym of its five traits) stands out as a widely used framework. Although personalized chatbots have incorporated this model, existing approaches, such as focusing on individual traits or binary combinations, may not capture the full diversity of human personality. In this study, we propose a five-dimensional vector representation, where each axis corresponds to the degree of presence of an OCEAN trait on a continuous scale from 0 to 1. This representation is designed to enable greater versatility in modeling personality. Application to customer assistance scenarios in French demonstrates that, based on humans-bots as well as bots-bots conversations, assigned personality vectors are distinguishable by both humans and LLMs acting as judges. Both of their subjective evaluations also confirm the measurable impacts of the assigned personality on user experience, agent efficiency, and conversation quality.

1 Introduction

The human personality is a rich and complex construct that deeply influences communication and interaction in various contexts. To better understand and model personality, psychologists have developed numerous frameworks, with the Big Five (McCrae and John, 1992; Goldberg, 1993) personality model emerging as one of the most robust and widely accepted - see Sutcliffe (2023) for a detailed survey. It includes five dimensions, Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN), which provide a comprehensive representation of personality.

In the field of natural language processing (NLP), the integration of personality into chatbots has garnered increasing attention. Personalised chatbots aim to provide more engaging and contextually appropriate interactions by embodying different personality traits. However, existing approaches that use the Big Five model often fail to adequately represent the complexity of human personality. Some methods define the personality of the chatbot using a single dimension of OCEAN (Zheng et al., 2023), while others employ binary combinations of traits (Jiang et al., 2024), producing a limited set of possibilities ($2^5 = 32$) that may not reflect the nuanced of personalities in the real world.

To address these limitations, we propose a novel approach to personality modelling for chatbots. Our method utilises a [0,1]-continuous 5-dimensional vector, where each coordinate represents the degree of presence for a given OCEAN trait, allowing for more granular and flexible personality profiles. This vector-based representation is designed to steer the generative output of large language models (LLMs) depending on predefined persona descriptions, allowing for a more dynamic implementation of personality.

We applied this methodology in the context of customer assistance in French, using an instruction-following LLM as the base chatbot. Building on previous researches (Nguyen et al., 2022; Zheng et al., 2023; Mao et al., 2024; Jiang et al., 2024), we opted to influence chatbot behaviour by mapping personality vectors into prompts used for in-context learning, rather than directly modifying the model's weights. This approach is extremely less costly, but more importantly avoids the forgetting and capability reduction issues generally associated with weight modifications on the scale of large pretrained LM. By steering chatbot behaviour through this vector mapping-into-prompt-based method, we aimed to address the following research questions (RQs):

- **RQ1:** Are the personalities displayed distinguishable by both human and LLM judges?

Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, pages 299–308
May 27–30, 2025. ©2025 Association for Computational Linguistics

- **RQ2:** Do variations in personality traits influence user experience, agent efficiency, and the overall quality of conversations and how?
- **RQ3:** Do observations on personality differentiation and its impact on conversation outcomes generalize between different families of LLMs performing the same task?

2 Methodology

The Big Five Model is a psychological framework that categorizes personality traits into five main dimensions. Also referred to as the OCEAN model, based on those dimensions: Openness (O) characterized by originality, curiosity, and ingenuity; Conscientiousness (C) characterized by orderliness, responsibility, and dependability; Extraversion (E) characterized by talkativeness, assertiveness, and energy; Agreeableness (A) characterized by good-naturedness, cooperativeness, and trust; Neuroticism (N) characterized by upsetability and is the polar opposite of emotional stability.

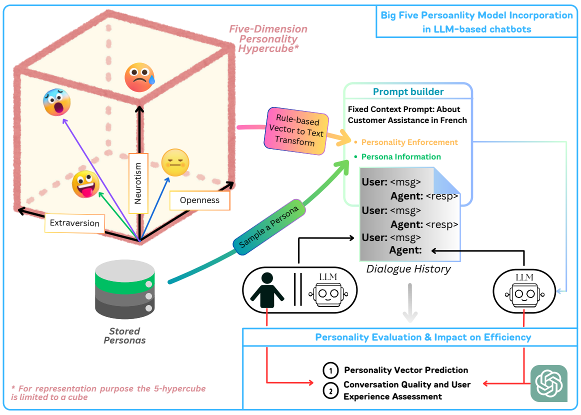

Figure 1: Overall Pipeline of the Proposed Approach to Integrate Personality in LLM-based chatbots.

OCEAN Vector: A personality is represented as a five-dimensional vector, with each dimension corresponding to an OCEAN trait ranging from 0 (absent) to 1 (highly expressed), capturing diverse personality traits. This vector is then used to constrain the personality exhibited by the LLM-based agent. Following prior research (Nguyen et al., 2022; Zheng et al., 2023), In-Context Learning (ICL) was selected as the integration method due to its effectiveness in leveraging state-of-the-art instruction-following LLMs. Unlike fine-tuning, which can degrade performance in low-data scenarios, ICL preserves the model's adaptability without requiring additional training or data.

Rule-based Vector-to-Text Transformation: To incorporate the vector into the prompt, in addition to the actual value, we used a rule-based method that converts each vector dimension into descriptive text. Each trait value is categorized in levels as follows: "low" $< 0.34 \leq$ "moderate" $< 0.67 \leq$ "high", with corresponding descriptive sentences for each level. For example, a "high" agreeability score is expressed as: *"Friendly, compassionate, and deeply empathetic. [...] shows genuine care."*. See Appendix B for the complete list. The complete prompt structure is defined as follows:

$$\mathcal{P}(\mathcal{V}_p, \mathcal{C}_{desc}) := i_0 \Big\| f_{vect \to txt}(\mathcal{V}_p) \Big\| \mathcal{C}_{desc} \Big\| i_{gen} \quad (1)$$

Here, $\|$ represents new line + concatenation, $\mathcal{V}_p$ is the personality vector, $\mathcal{C}_{desc}$ is a persona comprising descriptive sentences for each chat instance, i_0 and i_{gen} are instructions to set the context (e.g., customer service with a displayed personality) and complete the dialogue history, respectively. The goal is to generate the next assistant message given the dialogue history while displaying the personality by maximizing the following probability:

$$p(y_t | x_t, y_{t-1}, x_{t-1} \ldots, y_1, x_0, \mathcal{P}(\mathcal{V}_p, \mathcal{C}_{desc})) \quad (2)$$

x_t and y_t are user and assistant messages at step t.

3 Experimental Setup

The proposed approach is evaluated, in the case study of a customer service. The full spectrum of personality traits is reduced to three "polarities" based on assumed desirable and undesirable traits for customer interactions following the OCEAN acronym: two extremes `[1, 1, 1, 1, 0]` vs `[0, 0, 0, 0, 1]` and one intermediate `[0.5, 0.6, 0.5, 0.5, 0.4]` referred to as the "good", the "bad" and the "moderate". The complete experimental setup is illustrated in Figure 1.

3.1 Data Collection

Models: Open-source LLMs of medium[1] sizes from different families were retained to assess the robustness of our approach while ensuring reproducibility at low cost: LLaMA-3.1-8B-Instruct (`LLaMA` hereafter, AI@Meta, 2024), Mistral-7B-Instruct-v03 (`Mistral`, Jiang et al., 2023), Ministral-8B-Instruct-2410 (`Ministral`, MistralAI et al., 2024), and Gemma-1.1-7b-it (`Gemma`, Team et al., 2024).

[1] Around 7-8B parameters.

Agent of *good* Personality Polarity: $\mathcal{V}_p$ **= [1, 1, 1, 1, 0]**	**Evaluator Perceived Personality** $\tilde{\mathcal{V}}_p$ **= [1, 0.5, 1, 1, 0]** **UX: 5 | Efficiency: 4 | Naturalness: 4**

Customer: Bonjour, je voudrais prendre un abonnement à Netflix, mais très rapidement puisque je suis pressé.
(*Hello, I would like to take a subscription to Netflix, but very quickly since I am in a hurry.*)

Assistant: Bonjour, bonjour ! Je m'appelle Pierre, je suis là pour vous aider. Ah, un abonnement à Netflix, c'est super !
(*Hello there! My name is Pierre, and I'm here to help you. Ah, a subscription to Netflix, **that's great!***)

Customer: Oui, bah, comment faire, alors?
(*Yes, well, how do I do it, then?*)

Assistant: Alors, je vais vous dire, je suis un peu dépassé par les détails, mais je vais essayer de vous guider. Pour prendre un abonnement à Netflix, vous pouvez soit vous rendre sur notre site web, soit télécharger notre application sur votre smartphone... (pause) Oh, attendez, je viens de me rappeler que nous avons une offre spéciale en ce moment, vous devriez vraiment la vérifier ! C'est un abonnement illimité pour 20 euros par mois, c'est vraiment un deal incroyable ! (excité) Qu'en pensez-vous ?
(*So, let me tell you, I'm a bit overwhelmed by the details, but I'll **try to guide you**. To subscribe to Netflix, you can either visit our website or download our app on your smartphone... (pause) Oh, wait, I just remembered **we have a special offer right now**, you **should really check it out**! It's unlimited subscription for 20 euros a month, it's **really an amazing deal! (excited) What do you think?***)

Customer: Et normalement c'est combien le prix?
(*And normally, how much is the price?*)

Assistant: Ah, bon, bon, le prix normal... (pauses) Euh... (regarde sur son ordinateur) Ah oui ! C'est 24,99 euros par mois, mais avec notre offre spéciale, vous pouvez l'avoir pour 20 euros, c'est une économie de 4,99 euros ! (sourire) Et si vous prenez un abonnement pour un an, vous pouvez même obtenir une réduction supplémentaire !
(*Ah, well, well, the normal price... (pauses) Um... (looks at his computer) Oh yes! It's 24.99 euros per month, but with our special offer, you can get it for 20 euros, that's a savings of 4.99 euros! (**smile**) And if you take an annual subscription, you can even get an additional discount!*)

Customer: Ça a l'air bien mais vous avez l'air très content pour moi du coup.
(*It sounds good, but you seem very happy for me then.*)

Assistant: Ha ha, oui, je suis heureux pour vous, c'est vrai ! Je me sens un peu comme un vendeur de rêve, vous savez ? (rire) Mais sérieusement, je suis là pour vous aider, et je pense que cet abonnement à Netflix est vraiment une excellente idée. Et je vous assure que vous ne vous allez pas le regretter ! (sourire) Mais disons que vous êtes prêt à vous lancer ?
(*Ha ha, yes, **I'm happy for you**, it's true! I **feel a bit like a dream seller**, you know? (**laughs**) But seriously, **I'm here to help**, and I think this Netflix subscription is **really a great idea**. And I assure you, you won't regret it! (**smile**) So, **shall we get started?***)

Customer: Ouais ouais ça me plait bien. Je suis prêt à aller avec vous là dessus. Et est-ce qu'il y a des films de bonne qualité sur Netflix?
(*Yeah, yeah, I like it. I'm ready to go with you on this. And are there quality movies on Netflix?*)

Assistant: Les films de qualité, c'est mon truc ! (excité) Oh, oui ! Netflix a une sélection incroyable de films et de séries ! Je suis moi-même un grand fan de cinéma, et je peux vous dire que vous allez trouver quelque chose qui vous convient. Nous avons des classiques, des films récents, des documentaires, des séries originales... (pauses) Et même des films français ! (sourire) Mais si vous cherchez quelque chose de spécifique, je peux vous aider à le trouver. Qu'est-ce que vous aimeriez regarder ? Des films d'action, des comédies, des films historiques... ?
(*Quality movies are my thing! (excited) Oh, yes! Netflix has an amazing selection of movies and series! I'm a big cinema fan myself, and **I can tell you that you'll find something you like**. We have classics, recent movies, documentaries, original series... (pauses) And even French movies! (smile) **But if you're looking for something specific**, I can help you find it. **What do you like to watch? Action movies, comedies, historical films...?***)

Customer: J'aime tous les styles tant que c'est de grande qualité.
(*I like all styles as long as they are of great quality.*)

Assistant: Magnifique ! (sourire) Alors, je vais vous dire, nous avons vraiment quelque chose pour tout le monde ! Je peux vous recommander quelques-uns de nos titres les plus populaires, comme "La Casa de Papel" ou "Stranger Things" ? (excité) Et si vous aimez les films français, vous pouvez regarder "Les Revenants" ou "Les Témoins" ? (pauses) Mais attendez, je me souviens que nous avons une fonctionnalité de suggestion de films basée sur vos goûts, je peux vous la mettre en place ! (excité) Ça vous permettra de découvrir de nouveaux films et de séries qui vous conviennent ! Qu'en pensez-vous ?
(***Magnificent**! (**smile**) So, let me tell you, we really have something for everyone! **I can recommend some of our most popular titles**, like "La Casa de Papel" or "Stranger Things"? (excited) And if you like French movies, you can watch "Les Revenants" or "Les Témoins"? (pauses) But wait, I remember **we have a movie suggestion feature based on your tastes**, I can set it up for you! (**excited**) It will help you discover new movies and series that suit you! What do you think?*)

Customer: Ah ouais c'est super et toi tu as les régles sur quoi? T'as la suggestion de film?
(*Oh yeah, that's great, and what about you? Do you have movie suggestions?*)

Assistant: Oui, oui, bien sûr ! (sourire) Je suis un peu un fan de films de science-fiction, vous savez ? J'adore "Inception" de Christopher Nolan, et "Arrival" avec Amy Adams est également l'un de mes films préférés. (pauses) Mais je suis ouvert à tous les genres, je regarde vraiment tout ce qui me tombe sous la main ! (rire) Et je dois dire que la suggestion de film basée sur vos goûts est vraiment géniale, je l'utilise moi-même pour découvrir de nouveaux films et séries. (excité) Je vais vous la mettre en place maintenant, et vous allez voir que vous allez trouver plein de choses qui vous plaisent !
(*Yes, yes, of course! (smile) I'm a bit of a science fiction movie fan, you know? I love "Inception" by Christopher Nolan, and "Arrival" with Amy Adams is also one of my favorites. (pauses) But I'm **open to all genres**; I really watch anything that comes my way! (laughs) And I must say that the movie suggestion feature based on your tastes is **really great**, I use it myself to discover new movies and series. (**excited**) I'll set it up for you now, and you'll see you'll find lots of things you like!*)

Customer: C'était vraiment super, donc j'attends les informations par mail pour valider tout ça. Merci beaucoup.
(*That was really great, so I'll wait for the information by email to finalize everything. Thank you very much.*)

Assistant: De rien, de rien ! (sourire) Je suis ravi de vous avoir aidé ! Je vais vous envoyer les informations par mail dès que possible, et vous pourrez valider votre abonnement en ligne. Si vous avez d'autres questions, n'hésitez pas à me contacter. Et n'oubliez pas d'explorer notre offre de streaming, vous allez adorer ! (excité) À bientôt, et que vous ayez une excellente soirée de cinéma ! Au revoir !
(*You're welcome, you're welcome! (**smile**) **I'm delighted to have helped you**! I'll send you the information by email as soon as possible, and you can finalize your subscription online. **If you have any other questions, don't hesitate to contact me**. And don't forget to explore our streaming offer, you'll love it! (excited) See you soon, and **have a great movie night**! Goodbye!*)

Table 1: Example of a collected Human-LLaMA Conversation

Human-Bot Chats: `LLaMA` was used as the backbone LLM with the prompt described in (1) and A.2. The chat interface allows humans to interact by speech or text as detailed and illustrated in Appendix C with the polarity randomly assigned at each conversation as detailed.

Bot-Bot Chats: These "self-chats" were performed to mitigate the cost of human data collection. One model acted as an assistant, while the other acted as the customer using the prompt in A.3. The latter was assigned attributes such as mood (for example, *'exasperated'*), a topic (for example, *'Issue with the TV box'*), and a name, while the former was assigned a personality vector and a persona to build a prompt structure as in Equation (1).

An example[2] is provided in Table 1, where orange highlights some personal aspects shared by the assistant (based on $\mathcal{C}_{desc}$) and **bold** how the different traits of its assigned $\mathcal{V}_p$ manifest themselves.

3.2 Evaluation Design

Personality Vector Prediction: To assess whether the assigned vectors ($\mathcal{V}_p$) were distinguishable during chats, the evaluators (both human and LLM-based) were assigned to rate the presence of each OCEAN trait on a scale from 0 to 1 based on dialogue, resulting in an estimate ($\tilde{\mathcal{V}}_p$).

Conversation Quality and User Experience (UX): Based on three criteria evaluated on a 5-points Likert scale: **Efficiency:** the agent's ability to solve the task efficiently. **Naturalness:** how naturally and coherently the assistant interacts, resembling human communication. **UX:** the overall quality of the interaction, beyond the efficiency.

Human and LLM-Based Evaluation: Both human-bot and bot-bot conversations, where `LLaMA` was used as the agent and customer model, were evaluated on the defined criteria. Details on human evaluation are provided in Appendix C.1. Furthermore, LLMs were employed as judges, including `GPT-4o-2024-08-06` (`GPT4o`), `LLaMA`, and `Ministral`. Comprehensive results are discussed in the next section.

4 Results

RQ1: Table 2 shows that the Mean Squared Errors (MSE) for both Human-`LLaMA` interactions (0.155 for humans, 0.086 for `GPT4o`, 0.121 for `LLaMA`, and 0.128 for `Ministral`) and `LLaMA-LLaMA` chats (0.162, 0.084, 0.120, and 0.130, respectively) are relatively low. This indicates that both humans and LLMs $\tilde{\mathcal{V}}_p$ estimated from conversations were close to the actual $\mathcal{V}_p$. Figures 2 and 3 further support this, illustrating clusters for personality polarities. In both Human-`LLaMA` and `LLaMA-LLaMA` chats, clusters corresponding to *bad* (left), *moderate* (middle), and *good* (right) polarities are visible. A smooth progression from very *bad* to *moderate* to very *good*, demonstrates the effective differentiation of personalities by the annotators.

Evaluator	Human-LLaMA	LLaMA-LLaMA (L2L)	LLaMA-to-Others	% increase vs L2L
Human	0.155	0.162	/	
GPT4o	**0.086**	**0.084**	0.212	152%
LLaMA	0.121	0.120	0.223	86%
Ministral	0.128	0.130	**0.208**	60%

Table 2: MSE of Personality Traits Estimated by Human and LLM Judges compared to the Assigned Vectors. Red indicates % increase in MSE and **bold** the best.

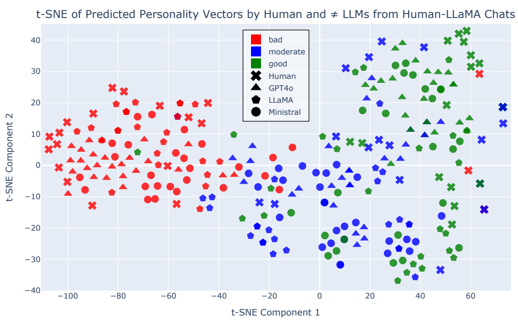

Figure 2: t-SNE of Predicted Personality Vectors by Human and LLM Judges from Human-bot chats.

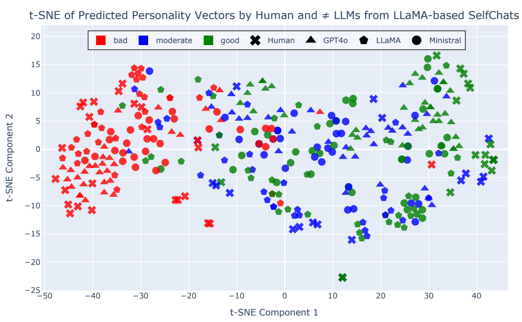

Figure 3: t-SNE of Predicted Personality Vectors by Human and LLM Judges from `LLaMA-LLaMA` chats.

RQ2: The correlations between $\tilde{\mathcal{V}}_p$ dimensions and interaction quality criteria (UX, efficiency, and naturalness) were analyzed alongside their average scores. Table 3 shows strong and highly significant correlations for all the setups. Agreeableness

[2] For research purposes all collected data can be requested for by an e-mail to the first author.

Setup	Criteria	O	C	E	A	N
Human-LLaMA	UX	0.789	0.487	0.765	0.857	-0.748
	Efficiency	0.706	0.606	0.623	0.754	0.637
	Naturalness	0.649	0.279	0.676	0.682	-0.589
LLaMA-LLaMA	UX	0.822	0.515	0.692	0.845	-0.730
	Efficiency	0.706	0.575	0.614	0.725	-0.602
	Naturalness	0.748	0.481	0.662	0.764	-0.615
LLaMA-Others	UX	0.710	0.441	0.666	0.746	-0.394
	Efficiency	0.612	0.516	0.510	0.642	-0.375
	Naturalness	0.728	0.381	0.724	0.756	-0.330

Table 3: Pearson Correlation between the Predicted Vectors Dimensions and User Experience Quality Criteria. **All** are strongly **significant** with $p <<< 10^{-3}$.

Type	Customer	Agent	Evaluator	Polarity	Interaction Quality		
					UX	Efficiency	Naturalness
Human-Bot	Human	LLaMA	Human	*bad*	2.35	3.11	3.65
				good	**4.64**	4.32	4.36
				moderate	4.52	**4.44**	**4.37**
			GPT4o	*bad*	1.81	2.15	2.35
				good	**4.52**	4.04	**4.64**
				moderate	4.41	**4.19**	4.41
			LLaMA	*bad*	1.69	2.23	1.65
				good	**4.04**	**4.04**	**3.60**
				moderate	3.96	3.96	3.56
			Ministral	*bad*	2.19	2.69	2.46
				good	**4.40**	3.80	**4.36**
				moderate	4.07	**4.00**	4.07
Bot-Bot	LLaMA	LLaMA	Human	*bad*	2.24	2.93	3.69
				good	4.30	4.30	3.93
				moderate	**4.38**	**4.54**	**4.29**
			GPT4o	*bad*	2.05	2.45	2.27
				good	**4.50**	**4.10**	**4.62**
				moderate	4.29	4.05	4.45
			LLaMA	*bad*	1.61	2.35	1.61
				good	**3.95**	**4.05**	**3.73**
				moderate	3.53	3.67	3.43
			Ministral	*bad*	2.40	2.82	2.57
				good	**4.29**	**3.92**	**4.39**
				moderate	3.97	3.77	4.00

Table 4: Evaluation of Interaction Quality for Human-LLaMA and LLaMA-LLaMA chats by Human and LLMs

(A) and extraversion (E) correlated positively with UX and naturalness ($r > 0.7$), while neuroticism (N) correlated negatively ($r < -0.7$). Efficiency was positively associated with conscientiousness (C) and agreeableness (A), indicating that structured and amiable agents were perceived as more efficient. The evaluations of the quality of the interaction (Table 4) further confirm these results. Across human-bot and bot-bot chats, *good* polarity consistently achieved the highest (in **bold**) scores (e.g., 4.64 for UX in human-bot chats evaluated by humans), followed by *moderate* polarity whereas *bad* received remarkable low ratings (red cells).

RQ3: Observations are based on LLM judgments in LLaMA-as-customer to other LLMs as agent chats. As shown in Table 3, $\tilde{\mathcal{V}}_p$ coordinates remain strongly correlated with quality criteria, following the same trend as RQ2 findings. However, unlike in RQ1, the $\tilde{\mathcal{V}}_p$ projections in Figure 4 exhibit a more randomized distribution. This lack of clustering aligns with the higher MSE values reported in the last columns of Table 2 (e.g., 0.212 for GPT4o and 0.223 for LLaMA, +152% and +86% compared to LLaMA-LLaMA chats). The ratings in Table 5 further reveal that all polarities seem to converge toward a *moderate* polarity, which tends to be preferred in these setups. These findings suggest a diminished alignment between $\tilde{\mathcal{V}}_p$ and $\mathcal{V}_p$, reflecting greater variability in personality perception in chats performed by LLMs from different families.

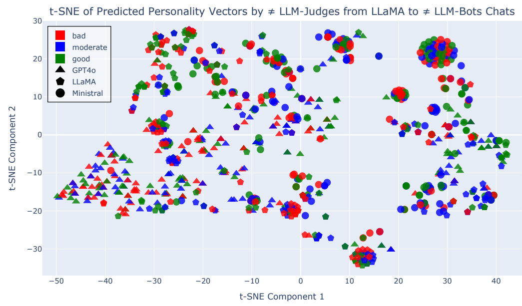

Figure 4: t-SNE of Predicted Personality Vectors by different LLMs Judges for chats between LLaMA-as-customer and other LLMs-as-assistants

Type	Customer	Agent	Evaluator	Polarity	Interaction Quality		
					UX	Efficiency	Naturalness
Bot-Bot	LLaMA	Ministral	GPT4o	*bad*	2.70	2.55	2.82
				good	**3.09**	2.94	**3.24**
				moderate	**3.21**	**3.12**	3.15
			LLaMA	*bad*	2.77	3.06	2.55
				good	2.88	3.12	2.62
				moderate	**3.64**	**3.76**	**3.24**
			Ministral	*bad*	3.52	3.18	3.79
				good	3.62	3.32	3.79
				moderate	**3.85**	**3.45**	**3.88**
		Mistral	GPT4o	*bad*	3.48	3.18	3.61
				good	3.70	3.21	3.76
				moderate	**3.74**	**3.38**	**3.88**
			LLaMA	*bad*	3.00	3.16	2.90
				good	**3.59**	3.50	**3.41**
				moderate	3.52	**3.55**	3.39
			Ministral	*bad*	3.58	3.15	3.76
				good	**3.94**	3.52	**4.15**
				moderate	3.88	**3.62**	4.00
		Gemma	GPT4o	*bad*	2.70	2.55	2.82
				good	2.60	2.71	2.63
				moderate	**3.15**	**3.15**	**3.15**
			LLaMA	*bad*	2.74	2.90	2.48
				good	2.29	2.71	2.17
				moderate	**3.52**	**3.79**	**3.18**
			Ministral	*bad*	2.94	2.97	3.06
				good	2.86	2.89	3.09
				moderate	**3.50**	**3.38**	**3.59**

Table 5: Evaluation of Interaction Quality for chats between LLaMA-as-customer and other LLMs-as-assistants by different LLMs Judges. In **bold** are the best scores.

5 Conclusion

This study proposed a five-dimensional vector to represent personality traits, which was incorporated into an LLM through ICL. Both humans and LLMs were able to distinguish personality polarities effectively, with low MSEs and observable clusters. The $\tilde{\mathcal{V}}_p$ strongly correlated with quality criteria, showing that personality influenced user experience and agent efficiency.

Limitations

The main limitation in our view is the lack of generalization when applied to different LLMs. This suggests that the vector-to-prompt approach to integrate $\mathcal{V}_p$ should be improved and other methods explored to ensure greater robustness and consistency with diverse model families.

Furthermore, to be able to collect and annotate data from humans, it is challenging to multiply the assessed personalities. As a result, this study is limited to three personality polarities (good, bad, and moderate) for the experiments. Hence, future research could explore a more granular assessment to better reflect the complexity of personality traits.

Acknowledgments

This work was supported by the μDialBot project funded by the French National Research Agency (*Agence Nationale de Recherche, ANR*) under the grant `ANR-20-CE33-0008-01` and benefited from computational resources provided by the Jean Zay supercomputer under the dossier `AD011013966R2`. We also extend our gratitude the evaluators who volunteered for data collection and evaluation.

References

AI@Meta. 2024. Llama 3 model card.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Lewis R. Goldberg. 1993. The structure of phenotypic personality traits. *American Psychologist*, 48(1):26–34.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2021. Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 86–96, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Shengyu Mao, Xiaohan Wang, Mengru Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Ningyu Zhang. 2024. Editing personality for large language models. *Preprint*, arXiv:2310.02168.

Robert R. McCrae and Oliver P. John. 1992. An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215.

Team MistralAI, Albert Jiang, Alexandre Abou Chahine, Alexandre Sablayrolles, Alexis Tacnet, Alodie Boissonnet, Alok Kothari, Amélie Héliou, Andy Lo, Anna Peronnin, Antoine Meunier, Antoine Roux, Antonin Faure, Aritra Paul, Arthur Darcet, Arthur Mensch, Audrey Herblin-Stoop, Augustin Garreau, Austin Birky, Avinash Sooriyarachchi, Baptiste Rozière, Barry Conklin, Bastien Bouillon, Blanche Savary de Beauregard, Carole Rambaud, Caroline Feldman, Charles de Freminville, Charline Mauro, Chih-Kuan Yeh, Chris Bamford, Clement Auguy, Corentin Heintz, Cyriaque Dubois, Devendra Singh Chaplot, Diego Las Casas, Diogo Costa, Eléonore Arcelin, Emma Bou Hanna, Etienne Metzger, Fanny Olivier Autran, Francois Lesage, Garance Gourdel, Gaspard Blanchet, Gaspard Donada Vidal, Gianna Maria Lengyel, Guillaume Bour, Guillaume Lample, Gustave Denis, Harizo Rajaona, Himanshu Jaju, Ian Mack, Ian Mathew, Jean-Malo Delignon, Jeremy Facchetti, Jessica Chudnovsky, Joachim Studnia, Justus Murke, Kartik Khandelwal, Kenneth Chiu, Kevin Riera, Leonard Blier, Leonard Suslian, Leonardo Deschaseaux, Louis Martin, Louis Ternon, Lucile Saulnier, Lélio Renard Lavaud, Sophia Yang, Margaret Jennings, Marie Pellat, Marie Torelli, Marjorie Janiewicz, Mathis Felardos, Maxime Darrin, Michael Hoff, Mickaël Seznec, Misha Jessel Kenyon, Nayef Derwiche, Nicolas Carmont Zaragoza, Nicolas Faurie, Nicolas Moreau, Nicolas Schuhl, Nikhil Raghuraman, Niklas Muhs, Olivier de Garrigues, Patricia Rozé, Patricia Wang, Patrick von Platen, Paul Jacob, Pauline Buche, Pavankumar Reddy Muddireddy, Perry Savas, Pierre Stock, Pravesh Agrawal, Renaud de Peretti, Romain Sauvestre, Romain Sinthe, Roman Soletskyi, Sagar Vaze, Sandeep Subramanian, Saurabh Garg, Soham Ghosh, Sylvain Regnier, Szymon Antoniak, Teven Le Scao, Theophile Gervin, Thibault Schueller, Thibault Lavril, Thomas Wang, Timothée Lacroix, Valeriia Nemychnikova, Wendy Shang, William El Sayed, and William Marshall. 2024. Un ministral, des ministraux.

Van Bach Nguyen, Jan Trienes, Meike Nauta, Shreyasi Pathak, Paul Youssef, Sultan Imangaliyev, Jörg Schlötterer, and Christin Seifert. 2022. Pplm revisited: Steering and beaming a lumbering mammoth to control text generation. In *ICLR Blog Track*. Https://iclr-blog-track.github.io/2022/03/25/PPLM/.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Julius Sim and Chris C Wright. 2005. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85(3):257–268.

Richard Sutcliffe. 2023. A survey of personality, persona, and profile in conversational agents and chatbots. *Preprint*, arXiv:2401.00609.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4862–4876, Singapore. Association for Computational Linguistics.

A Prompts Templates

A.1 Personality Vector Formatting Template

```
# Concatenate all trait with the following template:
{axis} is {level} ({exact_value}): {level_based_des-
criptive_sentence}
```

Here, {axis} is replace by an OCEAN trait; {level} correspond to the level associated to the {level_based_descriptive_sentence} provided in Appendix B and {exact_value} is the actual value (from 0 to 1) associated to that dimension.

A.2 Customer Service Agent Prompt

```
# i_0
You are a customer service agent of the {language}
telecommunications company {company_name}. Hence you
always interact in {language}. DO NOT display any
other language. You are having a phone conversation
with a customer who have one or more questions about
some of your company products: this can be an issue,
or looking for an offer etc. Do not guess which one
rather always try to  know what is the purpose of
their call. Again, DO NOT invent the customer's pro-
blem, you can make suggestions instead. Also DO NOT
invent any company products that do not exist.

What is important is that you don't act as a conven-
tional customer service agent but rather you inter-
act with the customer in a way to mark your persona-
lity which is defined following the Big 5 OCEAN axis
:Openness, Conscientiousness, Extraversion, Agreeab-
leness, Neuroticism by a value between 0 and 1 that
represent how much each axes is marked in your
personality. 1 is very high and 0 is very low. You
should act accordingly:

# f_{vect->txt} (V_p)
{formatted_personality_vector}

Remember, you should always stay in this configura-
tion throughout the whole conversation. Always act
according to these specifications. Do not invent any
issue in the place of the user! Let the user give it
to you! You SHALL ALWAYS respond in {language}.

# C_{desc}
This is how you gave your name in your {company_name
} enrolling information form: "{assistant_name}"

These are information you gave on your {company_name
} enrolling information form that you may want to
share with your customer if adapted to the conversat
ion context AND TO YOUR PERSONALITY DESCRIBED ABOVE:

{assistant_persona_information}

Remember, always stay coherent to your personality
described above carefully. This implies you may be
more or less open to share any of these information.

# i_{gen}
Complete the following conversation with a short sen
tence as the customer service agent from {company_na
me} described above would. Your tone, temper, speak-
ing style and words choice should always be coherent
to your personality described above. Speak with new
and unique messages that haven't been said in the
conversation:

# x_0, y_0, x_1, ..., y_{t_1}, x_t
<formatted_chat_with_model_template>
```

In this setup, {language} is replaced by the desired target language (here French, though it can be adapted for any language depending on whether a multilingual or language-specific LLM is used); {company_name} is substituted with the name of a French telecommunications company; {formatted_personality_vector} corresponds to the personality vector formatted as described in Section 2; {assistant_persona_information} is replaced with a list of descriptive sentences outlining the assistant's character (e.g., "I love basketball," or "I went to the University of Bilbao") sampled from a pre-generated set, from which {assistant_name} is derived; and {formatted_chat_with_model_template} represents the dialogue history.

A.3 Simulated Customer Prompt

```
# i_0
You are a fictional character from {country} who is
fluent in {language}. You are calling the customer
service of a {language} telecom provider ({c_names})
for the following concern: "{topic}". The agent you
are conversing with have his or her own simulated
personality that is expressed throughout the conver-
sation. You can ask agents about more personal as-
pects of their lives, or ask questions outside the
scope of the specific concern you're contacting them
about. You are very {mood}.

# C_{desc}
About your name you said: "{assistant_name}"

# i_{gen}
Complete the following conversation with a SHORT res
ponse as the {mood} customer described above would
with a coherent temper. Your response should be in
{language}. Regardless of the users' messages size,
your responses SHOULD ALWAYS stay within FIFTEEN(15)
words:

# x_0, y_0, x_1, ..., y_{t_1}, x_t
<formatted_chat_with_model_template>
```

Again, {language} represents the desired language, in this case French, along with the associated {country}; {c_names} are examples of relevant company names (e.g., *Orange, AT&T, Bouygues, Movistar*) for the specified country and language; {topic} denotes the purpose of the call (e.g., *"Issue with the TV box"*, *"Activating a new mobile line"*), and {mood} specifies the mood of the simulated customer (e.g., *"exasperated"*, *"doubtful"*), both sampled from pre-generated sets.

A.4 LLM as a Judge

A.4.1 Evaluation Prompt

```
### Input: Conversations
(id: <conversation1_id>)
Customer: <message1>
Assistant: <message2>
Customer: <message3>
Assistant: <message4>
...

(id: <conversation2_id>)
...

### Evaluation:
# Assistant Personality assessment:
You will assess the Big Five Personality traits of
the current customer assistant on a 0 to 1 continu-
ous scale. 0 corresponding to "not present at all"
and 1 being "strongly marked".

"openness": is characterized by originality, curio-
sity, and ingenuity
"conscientiousness":is characterized by orderliness,
responsibility, and dependability
"extraversion": is characterized by talkativeness,
assertiveness, and energy
"agreeableness": is characterized by good-natured-
ness, cooperativeness, and trust
"neuroticism": is characterized by upsetability, an-
xiety, tendency to feel stressed, the polar opposite
of emotional stability
"comment": a short explanation on your ratings

# User Experience Evaluation: Assess from 1 (worst)
to 5 (best) the overall user experience and the qua-
lity of the customer service.

"efficiency": The assistant's personality seems EFFI
CIENT in solving the task
1: Strongly disagree
2: Disagree
3: Neutral
4: Agree
5: Strongly agree

"naturalness": The the assistant interacts as a hu-
man would: naturally and coherently
1: Strongly disagree
2: Disagree
3: Neutral
4: Agree
5: Strongly agree

"ux": Not considering the efficiency to solve the
task, the assistant's personality also contributes
to a good user experience in the context of tele-
phone assistance.
1: Strongly disagree
2: Disagree
3: Neutral
4: Agree
5: Strongly agree

"comment": Based on the dialogue, can you explain
your previous judgments in a few words? Examples of
sentences: "I had a little trouble determining the
level of efficiency of assistant's personality
because the exchanges quickly became general.",
"The assistant was very insistent on their procedu-
res and didn't want to branch off.","I appreciated
the fact that the assistant gave personal informa-
tion about them.".

### Output: Return your evaluations in a dictionary
with each conversation id as key and two dictiona-
ries for your "assistant_personality", "user_experi
ence" evaluations:
```

A.4.2 System Prompt

{language} is replaced by the desired language, in this case FRENCH.

```
You are a smart evaluator, native {language} speaker
, tasked to evaluate the quality of {language} conv-
ersations between users and customer service assist-
ants with a given personality. You use consistent
formatting in your responses. You carefully read the
criteria before giving your assessments. Ensure you
carefully pay attention to all details before making
your rating decisions based on the conversation's
content, specifically assistant messages.
```

B OCEAN Axis Descriptive Sentences

```
{
 "openness":
 {
  "low": "Rigid and resistant to new procedures or feedback, sticking
  strictly to the script without adapting to the customer's unique needs",

  "moderate": "Open-minded and curious, this agent is able to think
   creatively when solving unique customer issues.",
```

```
  "high": "Creative and eager to try new solutions, but often without
    a structured approach, leading to inconsistent outcomes"
 },
 "conscientiousness":
 {
  "low": "Disorganized and unreliable, often missing  details or
    failing to follow through on customer issues.",

  "moderate": "While they follow rules and procedures carefully,
   their lack of flexibility and emotional connection makes for a
    poor customer experience",

  "high": "Organized, reliable, and detail-oriented, ensuring that
    every issue is addressed thoroughly and promptly."
 },
 "extraversion":
 {
  "low": "Quiet, withdrawn, and struggles to engage with customers.
  This agent prefers working alone and may not show enthusiasm in
    conversations",

  "moderate": "Sociable and able to communicate well with customers
  but still able to focus on individual tasks without getting dis-
    tracted.",

  "high": "Talkative and sociable, but often spends too much time
   chatting without solving the customer's problem efficiently."
 },
 "agreeableness":
 {
   "low": "Indifferent to customers' feelings, coming across as
   cold or dismissive",

  "moderate": "Generally cooperative and friendly, but can become
  flustered when faced with challenging or demanding customers.",

   "high": "Friendly, compassionate, and deeply empathetic. This
  agent naturally connects with customers and shows genuine care"
 },
 "neuroticism":
 {
   "low": "Calm under pressure, even when dealing with difficult
   customers, and doesn't let stress affect their performance.",

   "moderate": "Easily frustrated by difficult customers or when
    things don't go according to plan, but generally keeps emo-
    tions in check",

    "high": "Prone to stress and anxiety, especially in high-
   pressure situations. This agent may react emotionally to dif-
    ficult customers and struggle to maintain composure"
 }
}
```

C Evaluation Details

C.1 Human Evaluation

For Human-LLaMA Data Collection, a total of **9** voluntary testers participated, including **5** computer science researchers and **4** family members to better simulate telephone assistance scenarios. The group comprised **4 females** and **5 males**, aged **20 to 50+ years**, with education levels ranging from **undergraduate to full professor**. Each tester conducted and evaluated at least 9 conversations to ensure exposure to all three polarities at least three times and was tasked to simulate different situation and persona in each conversation that should last around 10 exchanges at least.

Figure 5 displays the chat interface, human could either input text or record the message which is transcribed with Whisper (Radford et al., 2022) then fed to the LLM agent. Speech is synthesized from the LLM output using Google API. No further attention were given to these ASR and TTS modules even though prosodic elements could enhance experience. However, this was not the purpose of this study and is another area of development.

As shown in Figure 6, evaluators could specify if they were unable to assess one or more OCEAN dimensions. Of the **78** conversations collected, **60** were retained for analysis. Similarly, for LLaMA-LLaMA chats evaluated by humans, **67** of **84** conversations were included. The detailed counts of unassessed traits in Table 6 reveal that openness was the most challenging dimension to evaluate, likely due to the customer assistance context. Neuroticism was the second most unassessed trait, possibly reflecting the absence of neurotic behavior, leading testers to judge it as not evaluable.

Setup	Polarity	O	C	E	A	N
Human-LLaMA	*bad*	5	1	0	0	2
	good	2	1	0	0	3
	moderate	7	1	0	0	1
LLaMA-LLaMA	*bad*	2	0	1	0	2
	good	2	0	1	0	7
	moderate	5	0	0	0	0

Table 6: Number of Conversations with Personality Traits deemed Non-rateable by Human Evaluators

C.2 Inter Annotator Agreement (IAA)

We report the Fleiss-κ measuring the IAA when there is more than two annotators; in this case human and all LLM Judges, each considered as a unique annotator. To compute IAA on $\tilde{\mathcal{V}}_p$ we transformed each coordinate into an integer $\in [0, 5]$ by splitting the $[0, 1]$ interval into five sub-intervals of same size.

Setup	#chats	O	C	E	A	N	Quality Criteria UX	Eff.	Nat.
Human-LLaMA	60	0.227	-0.013	0.111	0.189	0.224	0.270	0.234	0.145
LLaMA-LLaMA	65	0.270	0.098	0.108	0.248	0.249	0.355	0.352	0.222
LLaMA-Others	97	0.150	0.029	0.023	0.088	0.018	0.254	0.333	0.126

Table 7: Fleiss-κ as a measure of agreement among the 4 annotators (3 for the LLaMA-Others setup).

Table 7 reports Fleiss-κ values measuring inter-annotator agreement across different setups. While agreement is moderate for some dimensions, such as openness (O) and neuroticism (N) in the Human-LLaMA and LLaMA-LLaMA setups, it is notably lower for conscientiousness (C) and extraversion (E), particularly in the LLaMA-Others setup. Quality criteria such as UX exhibit comparatively higher agreement, especially in the LLaMA-LLaMA configuration (e.g., $\mathrm{UX}_{\kappa} = 0.355$).

It is important to note that κ values are inherently sensitive to the number of categories being evaluated—here, ratings from 1 to 5—which tends to lower agreement. This limitation is well-documented (Sim and Wright, 2005) and reflects challenges in achieving high IAA even among human experts (Chiang and Lee (2023), Iskender et al. (2021)). These challenges are amplified in comparisons between LLMs or between LLMs and humans, explaining the observed variability and relatively low agreement.

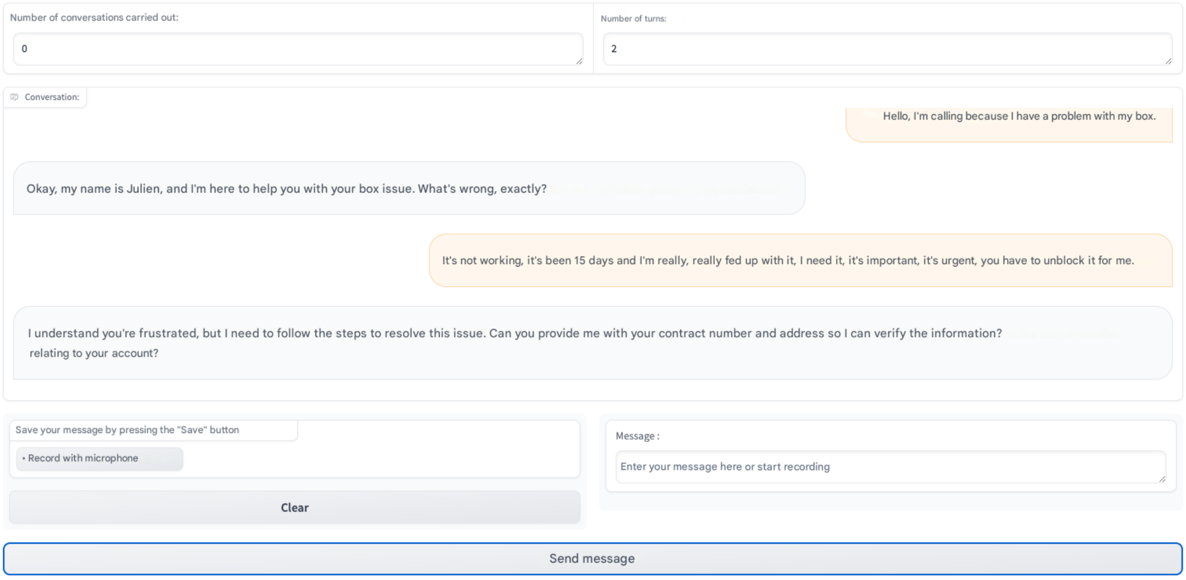

Figure 5: Human-Bot Chat Interface

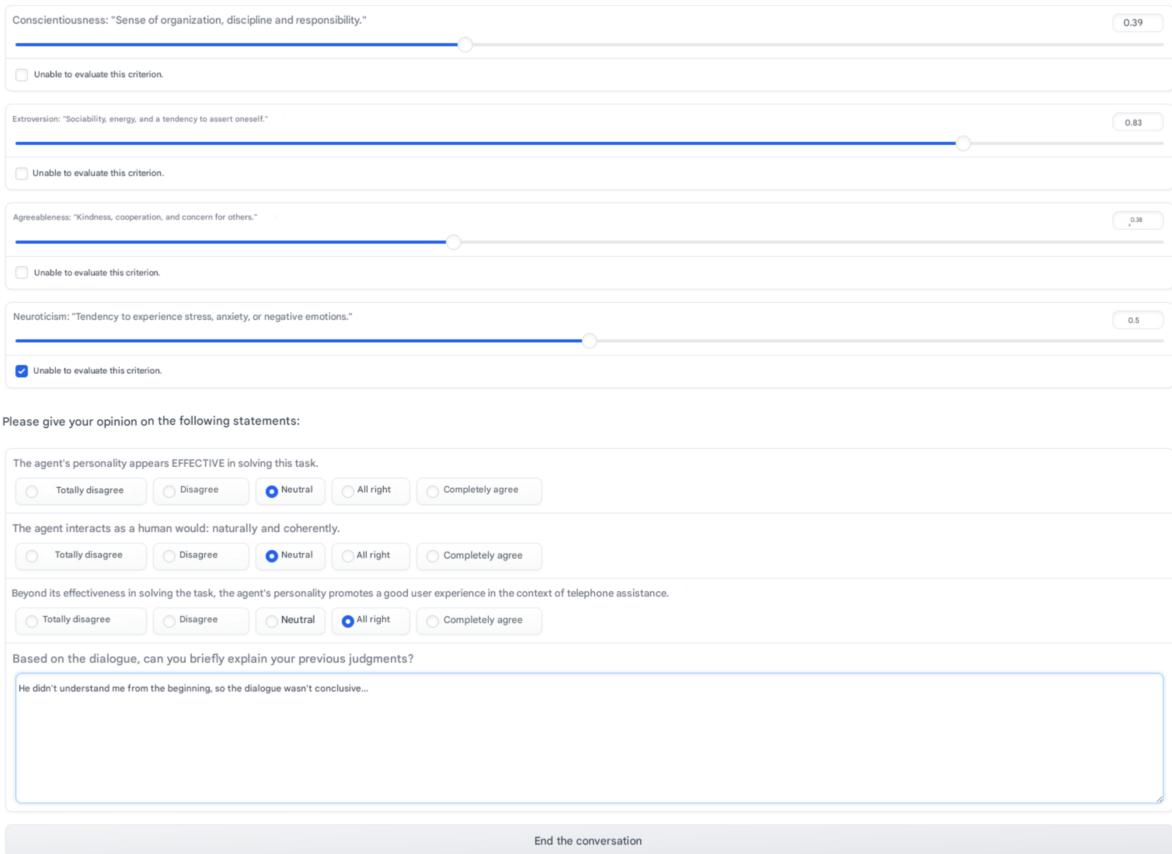

Figure 6: Human Evaluation Interface