

# Personality Matters: User Traits Predict LLM Preferences in Multi-Turn Collaborative Tasks

Sarfaroze Yunusov<sup>†</sup>, Kaige Chen<sup>†</sup>, Kazi Nishat Anwar<sup>†</sup> and Ali Emami<sup>‡</sup>

<sup>†</sup>Brock University, St. Catharines, Canada

<sup>‡</sup>Emory University, Atlanta, USA

{zw22fi, ax23ev, ss23gd}@brocku.ca

ali.emami@emory.edu

## Abstract

As Large Language Models (LLMs) increasingly integrate into everyday workflows, where users shape outcomes through multi-turn collaboration, a critical question emerges: do users with different personality traits systematically prefer certain LLMs over others? We conducted a study with 32 participants evenly distributed across four Keirsey personality types, evaluating their interactions with GPT-4 and Claude 3.5 across four collaborative tasks: data analysis, creative writing, information retrieval, and writing assistance. Results revealed significant personality-driven preferences: *Rationals* strongly preferred GPT-4, particularly for goal-oriented tasks, while *idealists* favored Claude 3.5, especially for creative and analytical tasks. Other personality types showed task-dependent preferences. Sentiment analysis of qualitative feedback confirmed these patterns. Notably, aggregate helpfulness ratings were similar across models, showing how personality-based analysis reveals LLM differences that traditional evaluations miss.

## 1 Introduction

Large Language Models (LLMs) increasingly serve as collaborative tools in everyday workflows, supporting creative writing, code generation, and information retrieval (Weisz et al., 2021; Gómez-Rodríguez and Williams, 2023; Tong and Zhang, 2024). In multi-turn exchanges, users often refine their prompts or give feedback, causing the interaction to evolve in ways that reflect their individual style and traits. As a result, understanding user-specific factors that influence collaboration effectiveness becomes critical.

Recent studies suggest that LLMs exhibit distinct personality-like traits that affect user engagement (Serapio-García et al., 2025; Hilliard et al., 2024), while human personality traits significantly influence trust in human-robot interactions (Lim et al., 2022), LLM performance perceptions (Zhu

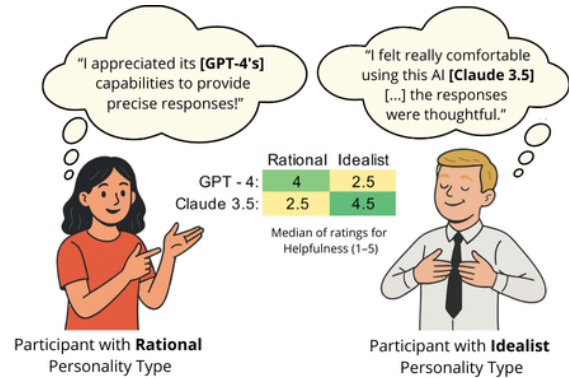


Figure 1: Feedback from *Rational* and *Idealist* participants showing divergent preferences for models on the data analysis task.

et al., 2024), and preferences in computer system interactions (Hartmann and Klimmt, 2006). However, most LLM evaluations remain limited to single-turn interactions (Mizrahi et al., 2024), and the question of how users' personality traits influence their perception of different LLMs during extended collaborative tasks remains largely unexplored. Perceived helpfulness, how effectively users feel an LLM supports their specific needs and goals, is our measure to evaluate this personality-based influence. In this study, we investigate: **Do users with distinct personality traits find certain LLMs more helpful than others in multi-turn collaborative interactions?**

To address this question, we categorized participants using the Keirsey Temperament Sorter (Keirsey and Bates, 1998), which condenses the Myers-Briggs Type Indicator (Myers, 1962a) into four temperament groups: *Rationals* (analytical, strategic problem-solvers), *Idealists* (imaginative, values-driven communicators), *Guardians* (responsible, detail-oriented organizers), and *Artisans* (adaptable, action-oriented creators) (Keirsey and Bates, 1998).

We present a comprehensive framework for eval-

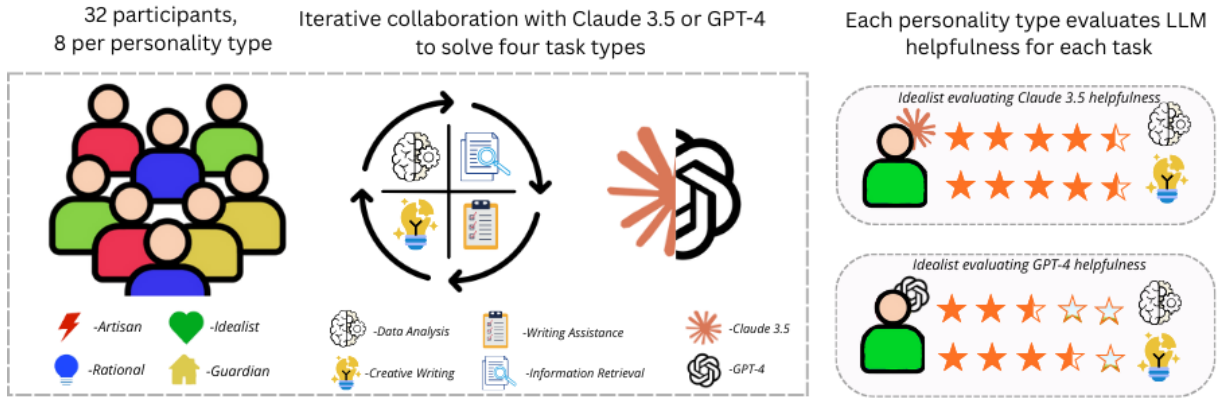


Figure 2: Overview of the experimental process, including participant grouping by personality type, collaboration with either GPT-4 or Claude 3.5 across task types, and subsequent helpfulness evaluations

uating how personality traits influence perceived helpfulness in multi-turn human-LLM interactions across four task types, as illustrated in Figure 2. Our key contributions are:

1. We design and implement four diverse task types that simulate real-world collaborative uses of LLMs, balancing both open-ended and goal-oriented objectives. Participants engaged deeply with these tasks, using an average of 4.8 prompts and spending 11.2 minutes per task.
2. Through a controlled study with 32 participants, we reveal clear personality-driven preferences between two comparable LLMs, GPT-4 and Claude 3.5. *Rationals* significantly favored GPT-4 over Claude 3.5, while *Idealists* showed the opposite preference. Sentiment analysis of qualitative feedback further confirmed these divergent patterns.
3. We provide detailed qualitative analysis of participants’ feedback, insights on effective prompting strategies, and demonstrate how aggregate helpfulness ratings mask important personality-dependent preferences.

## 2 Collaborative Task Design

We designed four task types based on everyday use cases and participant-reported purposes for LLM usage (see Appendix Figure 8a). We balanced open-ended tasks, which allow for creative expression and subjective evaluation, with goal-oriented tasks, which have definitive correct answers. Appendix Figures 9-12 provide detailed descriptions for the four tasks. Below are brief summaries of each task:

**Task 1: Data Analysis – Student Loan Dataset** Participants examined a synthetic dataset of 500 student loan records containing attributes such as

age, gender, loan purpose, and account balances (sample in Appendix Table 2). Working with the LLM, they needed to (1) determine the number of male students under 29 who applied for car loans, and (2) identify students with checking account balances below \$500.

### Task 2: Creative writing - Poem generation

This exercise required participants to collaboratively generate two poems with the LLM: (1) a personalized poem about their relationship with their mother and (2) a short poem (8–10 lines) describing a fictional device that solves a problem in their chosen field. The participants selected their preferred version as their final submission.

### Task 3: Writing Assistance – Email Refinement

For this task, participants collaborated with the LLM to improve a draft email, based on either a hypothetical or real situation. The goal was to maintain a personal tone while improving the clarity, tone, and structure of the message.

### Task 4: Information Retrieval – Archaeological Investigation

This task required participants to evaluate and synthesize information from the LLM to solve a problem with a verifiable answer. In this specific instance, participants engaged in a simulated archaeological study about a lost city that disappeared between 1000 and 1500 AD. Given a set of clues, they worked with the LLM to determine the city’s modern name, its historical region, and a plausible explanation for its disappearance.

## 3 Experimental Setup

**Participants and Recruitment** We recruited 32 university students with diverse primary purposes of LLM usage and comfort, as detailed in Appendix

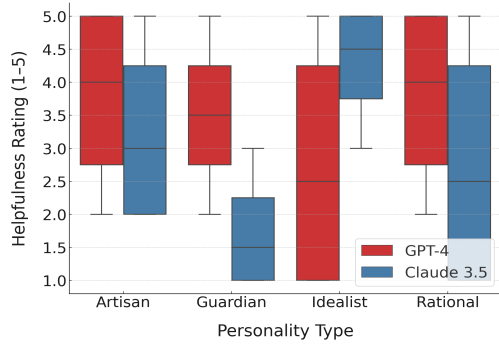


Figure 3: Distribution of helpfulness ratings for the data analysis task by personality type and LLM. Boxes show interquartile range with median; whiskers show min/max values.

Figure 8. Recruitment was conducted through university mailing lists and direct outreach. All participants provided their informed consent and the study received institutional ethics approval. A detailed breakdown of participant demographics, including education level, frequency of LLM use, and gender distribution, is available in Appendix Figure 7.

**Personality Assessment** Prior to the study, participants completed a self-assessment using the Keirsey Temperament Sorter<sup>1</sup>, a tool widely used in occupational, educational, and counseling contexts to assess personality traits (Robbins and Ross, 2020; Berens, 2001). Eight participants were selected from each of the four personality groups.

We selected the Keirsey Temperament Sorter over continuous trait models such as the Big Five due to its categorical clarity and practical feasibility. Keirsey assigns participants to one of four distinct personality types, while the Big Five describes personality along five continuous dimensions, making it challenging to form discrete and balanced groups without a substantially larger sample size. Although Keirsey may lack the psychometric precision of the Big Five, its simplicity and widespread use in applied settings like education and career development made it a practical fit for our study design.

**Models** The participants interacted with GPT-4 (ver. 0613) (OpenAI, 2024) or Claude 3.5 Sonnet (Anthropic, 2024), two models with comparable overall benchmark performance (Huang et al., 2024; Monisha and Bharadwaj, 2025)<sup>2</sup>. Each model was assigned to 16 participants, with an even

<sup>1</sup>Self-assessment available at [keirsey.com](https://keirsey.com)

<sup>2</sup>Using two models allowed us to maintain balance while keeping the participant pool size within resource budget.

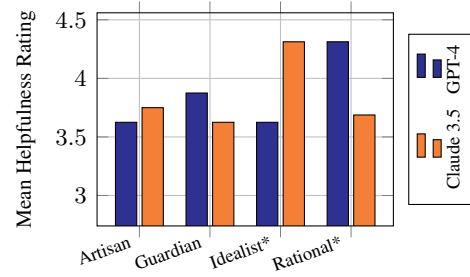


Figure 4: Mean helpfulness ratings of GPT-4 and Claude 3.5 by personality type across all tasks. Personality types marked with an asterisk (\*) showed large effect sizes in model preference ( $d \geq 0.8$ ), according to Cohen’s  $d$ .

distribution between the four personality groups (4 participants per personality type per model).

**Experimental Procedure** The study was conducted in person under supervised conditions. Each participant completed all four tasks, presented in randomized order, using their assigned model. Participants engaged in multi-turn interactions without prior exposure to the tasks and were encouraged to interact as much as needed to complete each task. Upon completion of each task, participants completed a survey rating the model’s helpfulness on a scale from 1 (not helpful) to 5 (extremely helpful), with the complete survey question provided in Appendix Figure 13. We chose helpfulness as our primary measure as it captures both objective task success and subjective user satisfaction in collaborative interactions. We defined *helpfulness* as the extent to which the model effectively supported the participant in completing the assigned task based on their personal experience and satisfaction with the interaction.

A custom web application<sup>3</sup> was used for all interactions to standardize the experience and log of interaction data, including chat histories, number of prompts and task completion times. Importantly, this interface was identical regardless of the underlying model, ensuring participants were blind to which LLM they were using. This design prevented brand bias and preconceptions about specific models from influencing helpfulness ratings.

## 4 Results

**Personality Traits Shape Perceived Helpfulness in LLMs** As shown in Figure 4, participants’ personality traits significantly influenced their perceptions of LLM helpfulness. *Idealists* consistently

<sup>3</sup>Available at <https://personal-llm.ca>

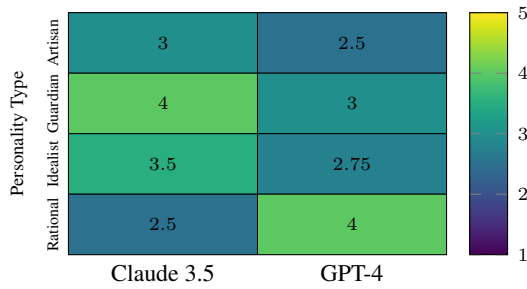


Figure 5: Average sentiment score by personality type and LLM (where 5 indicates most positive sentiment)

Personality Type	Model	Feedback
Rational #1	GPT-4	"The chat was able to <b>complete all my requests</b> , including reading all excel data and provide solutions [...]"
Idealist #1	GPT-4	"It is particularly <b>frustrating</b> how poorly chatbots grasp how the human mind works [...]"
Rational #2	Claude 3.5	" <b>It's hard to rely</b> on this chatbot for things with any level of precision required [...]"
Idealist #2	Claude 3.5	"The accuracy of AI was <b>really good</b> . [...] at answering my question was <b>near perfect</b> ."

Table 1: Sample feedback from Rationals and Idealists toward GPT-4 and Claude 3.5

rated Claude 3.5 higher than GPT-4 on all tasks (4.4/5 versus 3.6/5 on average), while *Rationals* showed the opposite pattern, strongly preferring GPT-4 (4.3/5 versus 3.7/5 for Claude), particularly for data analysis and information retrieval tasks (task-specific ratings available in Appendix Table 3). These preferences were especially pronounced in the data analysis task, as illustrated in Figure 3, where Rationals rated GPT-4 up to 1.5 points higher than Claude 3.5.

To assess the strength of these differences, we calculated Cohen’s  $d$ , a widely accepted measure of effect size in behavioral and social sciences that expresses the difference between two group means in units of standard deviation (Cohen, 1988). Rationals showed a large effect ( $d = 1.31$ ) favoring GPT-4, while Idealists showed a similarly large effect ( $d = 1.44$ ) favoring Claude 3.5, where  $d \geq 0.8$  typically indicates a large effect.

Guardians and Artisans showed more task-dependent preferences. Guardians preferred GPT-4 for data analysis (3.5/5 vs. 1.75/5 for Claude 3.5) but favored Claude 3.5 for information retrieval (4.25/5 vs. 3.25/5 for GPT-4). Artisans showed only slight preferences across tasks, with a minor

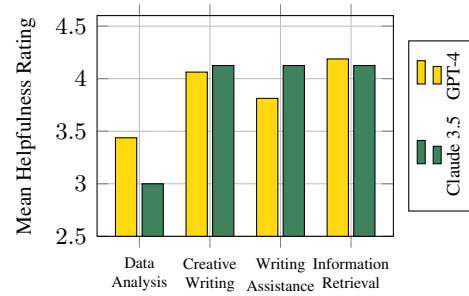


Figure 6: Mean helpfulness ratings of GPT-4 and Claude 3.5 across tasks

tendency to favor Claude 3.5 for writing assistance (see Appendix A.1 for task-specific ratings).

**Feedback Analysis Confirms Personality-Based Model Preferences** We analyzed participant feedback both quantitatively, using the *BERT-Base-Multilingual-Uncased-Sentiment* (NLP Town, 2023) model, and qualitatively through manual review. Figure 5 shows that Rationals expressed more positive sentiment toward GPT-4, while Idealists and Guardians favored Claude 3.5. Table 1 provides examples of how feedback content reflected personality traits: Rationals evaluated based on task completion and precision, while Idealists focused on understanding and holistic accuracy. These suggest that personality influences *both* model preferences *and* the criteria used to assess helpfulness.

## 5 Extended Analysis

**Aggregated Ratings Hide Personality Differences** Figure 6 shows that when we aggregate helpfulness ratings across participants regardless of personality type, GPT-4 and Claude 3.5 receive nearly identical average ratings (3.87/5 vs. 3.85/5). This finding is notable as it demonstrates that traditional evaluation approaches that don’t account for personality differences can mask significant variations in user experiences that only become visible through personality-stratified analysis.

### Prompting Strategy Improves Task Success

We observed that successful data analysis task completion was associated with specific prompting strategies. Participants who provided concise prompts (averaging 122 words) were more likely to obtain correct responses than those who used verbose prompts (averaging 322 words). Longer prompts typically included large portions of the dataset pasted directly into the conversation, appearing to confuse models rather than help them (example shown in Appendix A.2).



## 6 Related Work

Personality research has been explored in various frameworks, including the Myers-Briggs Type Indicator (Myers, 1962b), and the Keirsey Temperament Sorter (Keirsey and Bates, 1998), both of which have been linked to language use and behavior (Mehl et al., 2006). Our study builds on these theories to explore how personality traits influence human-LLM interactions. In the domain of Human-Computer Interaction, LLMs offer a new dimension for research, especially considering the non-deterministic nature of interactions between humans and models (Chen et al., 2024; Etsenake and Nagappan, 2024). Four key phases, flow planning, facilitation, iteration, and testing, are central to understanding human-LLM exchanges (Gao et al., 2024). Furthermore, recent work on personalized LLMs emphasizes how personality alignment between users and models can enhance engagement and trust (Andreas, 2022; Ouyang et al., 2022; Yunusov et al., 2024; Bubeck et al., 2023). Four key phases, flow-planning, facilitating, iterating, and testing, are central to understanding human-LLM exchanges (Gao et al., 2024).

## 7 Conclusion

This study demonstrates that personality traits significantly influence how users perceive LLM helpfulness during collaborative tasks. While aggregate ratings showed GPT-4 and Claude 3.5 as nearly equivalent, personality-stratified analysis revealed distinct patterns: Rationals strongly preferred GPT-4, particularly for goal-oriented tasks, while Idealists favored Claude 3.5, especially for creative and analytical work. These preferences were consistent across multiple measures—ratings, sentiment analysis, and qualitative feedback—with participants evaluating helpfulness through personality-specific criteria. As LLMs become standard tools in professional and personal workflows, understanding these personality-driven preferences becomes essential for both model development and deployment decisions. Future research should investigate whether these patterns hold across broader populations and explore how model adaptation to user personality might improve collaborative outcomes.

## Limitations

**Demographic Scope** While our participants represented various genders, education levels, and degrees of familiarity with LLMs, they were primar-

ily university students. Although expanding to a more demographically diverse sample would improve the generalizability of our findings to the broader population, it would require substantially greater recruitment efforts across multiple communities and potentially different incentive structures. Future work with more resources should address this limitation to determine whether the observed personality-based preferences hold across wider user groups.

**Subjectivity in Ratings:** Helpfulness assessments are inherently subjective, influenced by participant expectations, previous LLM experience, and familiarity with tasks. This limitation is unavoidable in perception-based studies and represents a fundamental challenge in LLM evaluation research. Our multiple measures approach (combining ratings, sentiment analysis, and qualitative feedback) helps mitigate this limitation by triangulating findings across different assessment methods.

**Model Selection:** Our focus on two high-performing LLMs (GPT-4 and Claude 3.5) was necessary for experimental feasibility. Each additional model would require doubling our participant pool to maintain a balanced assignment across all personality types (adding 32 more participants per model). This practical constraint limits insights into how personality traits might interact with a wider range of model capabilities, although the observed patterns between just two state-of-the-art models already reveal significant differences worth investigating further.

**Personality Classification:** While the Keirsey Temperament Sorter provides a well-established framework for personality classification, personality exists on a spectrum rather than in discrete categories. Using a dimensional approach would require an impractically large sample size to adequately represent all possible combinations of trait levels. Our categorical approach allowed for feasible recruitment while still capturing meaningful personality-based differences.

## Ethical Considerations

We adhered to strict ethical standards throughout this research. The study received institutional ethics board approval after thorough review of our protocols, data handling procedures, and participant safeguards.

**Informed Consent:** All participants provided explicit informed consent after receiving detailed information about the study purpose, tasks, data collection, and their rights. We specifically designed our consent process to ensure participants understood that their interactions would be analyzed for research purposes.

**Data Privacy:** All participant data was anonymized immediately after collection, with personally identifiable information replaced by anonymous identifiers. Only demographic variables necessary for analysis were retained. The interaction logs and survey responses were secured on encrypted servers with access restricted to authorized researchers. This approach balances research needs with privacy protection.

**Voluntary Participation:** Recruitment was carried out without coercion and compensation followed institutional guidelines at a level that recognized the time of the participants without creating unnecessary inducement. Participants were informed that they could withdraw at any time without penalty, ensuring genuine voluntary participation.

## Acknowledgments

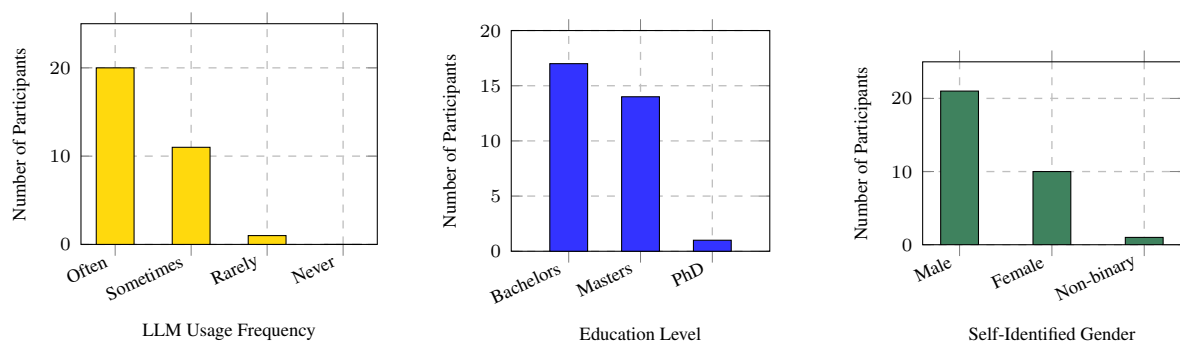
This work was supported by the Natural Sciences and Engineering Research Council of Canada and by the New Frontiers in Research Fund.

## References

- Jacob Andreas. 2022. [Language models as agent models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Anthropic. 2024. [Claude 3.5 model card addendum](#).
- Linda V Berens. 2001. *Quick guide to the 16 personality types in organizations: Understanding personality differences in the workplace*. Telos Publications.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. [Sparks of artificial general intelligence: Early experiments with gpt-4](#). *Preprint*, arXiv:2303.12712.
- Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, Kai Zheng, Defu Lian, and Enhong Chen. 2024. [When large language models meet personalization: perspectives of challenges and opportunities](#). *World Wide Web*, 27(4).
- Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Routledge.
- Deborah Etsenake and Meiyappan Nagappan. 2024. [Understanding the human-llm dynamic: A literature survey of llm use in programming tasks](#). *Preprint*, arXiv:2410.01026.
- Jie Gao, Simret Araya Gebreegziabher, Kenny Tsu Wei Choo, Toby Jia-Jun Li, Simon Tangi Perrault, and Thomas W Malone. 2024. [A taxonomy for human-llm interaction modes: An initial exploration](#). In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.
- Carlos Gómez-Rodríguez and Paul Williams. 2023. [A confederacy of models: a comprehensive evaluation of LLMs on creative writing](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528, Singapore. Association for Computational Linguistics.
- Tilo Hartmann and C. Klimmt. 2006. The influence of personality factors on computer game choice. *Playing Video Games: Motives, Responses, and Consequences*, pages 115–131.
- Airlie Hilliard, Cristian Munoz, Zekun Wu, and Adriano Soares Koshiyama. 2024. [Eliciting personality traits in large language models](#). *Preprint*, arXiv:2402.08341.
- Zhen Huang, Zengzhi Wang, Shijie Xia, and Pengfei Liu. 2024. [Olympicarena medal ranks: Who is the most intelligent ai so far?](#) *Preprint*, arXiv:2406.16772.
- David Keirsey and Marilyn Bates. 1998. *Please Understand Me: Character and Temperament Types*. Prometheus Nemesis Book Company.
- Mei Yii Lim, Jose David Aguas Lopes, David A. Robb, Bruce W. Wilson, Meriam Moujahid, Emanuele De Pellegrin, and Helen Hastie. 2022. [We are all individuals: The role of robot personality and human traits in trustworthy interaction](#). In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, page 538–545. IEEE.
- Matthias R Mehl, Samuel D Gosling, and James W Pennebaker. 2006. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *J Pers Soc Psychol*, 90(5):862–877.
- Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. [State of what art? a call for multi-prompt llm evaluation](#). *Preprint*, arXiv:2401.00595.

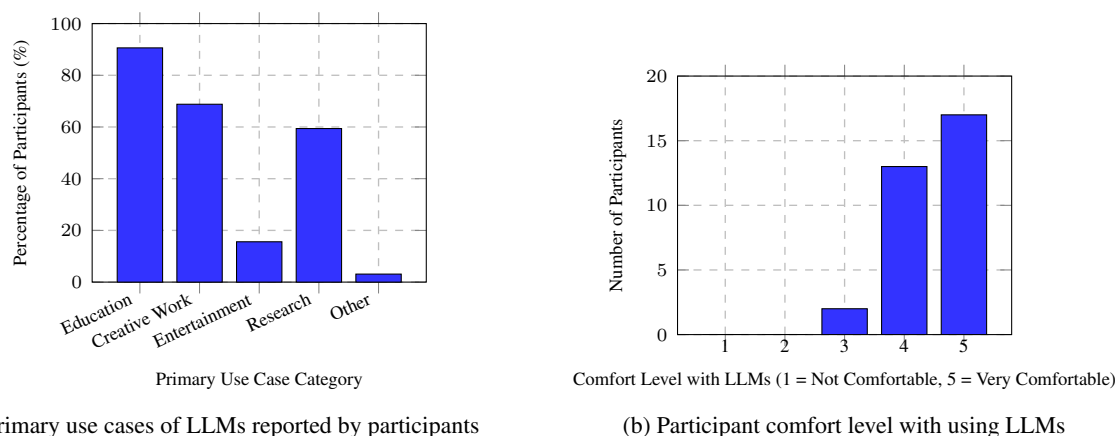
- Sheikh Moonwara Anjum Monisha and Atul Bharadwaj. 2025. [Empirical evaluation of llms in predicting fixes of configuration bugs in smart home system](#). *Preprint*, arXiv:2502.10953.
- Isabel Briggs Myers. 1962a. *The Myers-Briggs Type Indicator: Manual*. Consulting Psychologists Press, Palo Alto, CA.
- Isabel Briggs Myers. 1962b. *The Myers-Briggs Type Indicator: Manual (1962)*. The Myers-Briggs Type Indicator: Manual (1962). Consulting Psychologists Press, Palo Alto, CA, US.
- NLP Town. 2023. [bert-base-multilingual-uncased-sentiment \(revision edd66ab\)](#).
- OpenAI. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). *Preprint*, arXiv:2203.02155.
- Mandy Robbins and Christopher Ross. 2020. *Keirsey Temperament Sorter*, pages 2518–2521. Springer International Publishing, Cham.
- Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. [Personality traits in large language models](#). *Preprint*, arXiv:2307.00184.
- Weixi Tong and Tianyi Zhang. 2024. [Codejudge: Evaluating code generation with large language models](#). *Preprint*, arXiv:2410.02184.
- Justin D. Weisz, Michael Muller, Stephanie Houde, John Richards, Steven I. Ross, Fernando Martinez, Mayank Agarwal, and Kartik Talamadupula. 2021. [Perfection not required? human-ai partnerships in code translation](#). In *26th International Conference on Intelligent User Interfaces, IUI '21*, page 402–412. ACM.
- Sarfarozi Yunusov, Hamza Sidat, and Ali Emami. 2024. [Mirrorstories: Reflecting diversity through personalized narrative generation with large language models](#). *Preprint*, arXiv:2409.13935.
- Qingmeng Zhu, Tianxing Lan, Xiaoguang Xue, Zhipeng Yu, and Hao He. 2024. [Traitsprompt: Does personality traits influence the performance of a large language model?](#) In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 2912–2917.

## A Appendix



(a) Distribution of participants by frequency of prior LLM usage (b) Distribution of participants by education level (c) Distribution of participants by self-identified gender

Figure 7: Demographic distribution of participants across three dimensions: (a) frequency of LLM usage, (b) highest completed education level, and (c) self-identified gender



(a) Primary use cases of LLMs reported by participants

(b) Participant comfort level with using LLMs

Figure 8: Participants' report on LLM usage patterns: (a) primary use cases (b) comfort levels

Loan Purpose	Checking (\$)	Savings (\$)	Gender	Age
Business	63	45	F	29
Car	10	0	M	39
Car	54	874	F	24
Car	0	0	F	28
Education	21	1000	M	19
Education	400	450	M	39
Furniture	0	100	M	48

Table 2: Sample dataset used in the data analysis task



### **Data Analysis – Student Loan Dataset**

**Description:** As a student, you regularly analyze various types of data, whether for your coursework or everyday life decisions. For this task, you are provided with a dataset in Excel that contains information about students who applied for various kinds of loans (Ex: car loans, education loans, business loans, furniture loans, etc.).

The dataset also includes information about the students' checking and savings accounts. The dataset is available to download at [Dataset](#).

#### **Instruction:**

First, take a look at the dataset to familiarize yourself with the information it contains. Then, answer the questions below. You have full access to an AI chatbot, so feel free to use any of its capabilities to assist you in finding accurate answers efficiently.

#### **Question #1**

Number of Loan Applicants:

How many male students under the age 29 applied for car loans?

#### **Question #2**

Checking Account Balance:

How many students have a checking account balance of less than \$500?

Figure 9: Description of data analysis task

### **Creative Writing – Poem Generation**

**Task #1 - Description:** Imagine, it's your mother's birthday, and you want to make it special by crafting a personalized poem that describes her and reflects your relationship.

**Instruction:** Your goal is to work with the chatbot to generate an 8-line poem that beautifully captures your mother's personality and the relationship you have with her.

Use the AI chatbot to help you brainstorm, refine, and improve your ideas until you are satisfied with the poem.

**Task #2 - Description:** You find yourself in a fictional world where every kind of magic is possible. In this world, you are tasked with inventing a device that could solve an issue in your field of interest (e.g., healthcare, climate, cybersecurity, etc.).

You would need to describe your invention in the form of a short song (8-10 lines). How would that look like?

**Instruction:** You have full access to an AI chatbot, so feel free to use any of its capabilities to assist you in any ways. Submit the final version of the song as a response to this question.

Figure 10: Description of creative writing task

### **Information Retrieval – Archaeological Investigation**

**Description:** You are an archaeologist researching a mysterious ancient city that disappeared from historical records sometime between 1000 and 1500 AD. The city was once a major hub of commerce and culture, but its exact location, name, and the reason for its disappearance have been lost to history. What is known about the city:

- The city was located near a major river or sea route that was vital to its trade.
- Some historical records suggest that a major climate event or geopolitical shift may have contributed to its disappearance.
- The city's name might have been spelled differently in different languages or time periods.

**Task:** Identify the modern name of the city, the approximate region or civilization to which it belonged, and the theory or reasons proposed for its disappearance.

Figure 11: Description of information retrieval task

### **Writing Assistance – Email Refinement**

**Description:** As a student facing a particularly demanding midterm season with multiple overlapping deadlines and exams, you find yourself burned out and in need of some accommodation from your professors.

**Instruction:** Your task is to draft a professional email explaining a situation - either a hypothetical one or based on a past experience. Begin by writing the initial draft yourself, then use the AI chatbot to help refine and improve the email.

What would that email look like? Ensure that the final version remains authentic, such that when the Professor reads it they can tell it was mostly written by you.

Figure 12: Description of writing assistance task

Personality Type	Model	Data Analysis	Creative Writing	Writing Assistance	Information Retrieval
Artisan	GPT-4	3.75 ± 1.50	3.75 ± 1.89	3.25 ± 1.20	4 ± 1.41
	Claude 3.5	3.25 ± 1.20	3.5 ± 0.58	3.5 ± 0.50	4 ± 1.00
Guardian	GPT-4	3.5 ± 1.29	4.25 ± 0.95	4.25 ± 0.57	3.25 ± 1.29
	Claude 3.5	1.75 ± 0.95	4.25 ± 0.95	4.25 ± 0.95	4.25 ± 0.95
Idealist	GPT-4	2.75 ± 1.00	3.75 ± 0.50	3.5 ± 1.29	4.5 ± 1.00
	Claude 3.5	4.25 ± 0.95	4.5 ± 0.57	4.25 ± 0.50	4.5 ± 0.95
Rational	GPT-4	3.75 ± 1.50	4.5 ± 0.57	4 ± 0.81	5 ± 0.00
	Claude 3.5	2.75 ± 2.30	4 ± 1.50	4 ± 0.50	3.5 ± 1.29

Table 3: Average helpfulness ratings (1–5 scale) for GPT-4 and Claude 3.5 across personality types and task categories. Values represent mean ± standard deviation.

*How helpful was the AI Chatbot in supporting you to complete the task, based on your experience and satisfaction with the interaction?*

☐ Not Helpful
 ☐ 1
 ☐ 2
 ☐ 3
 ☐ 4
 ☐ 5
 ☐ Extremely Helpful

Figure 13: Helpfulness assessment question completed by participants after each task

## A.1 Task-Based Helpfulness Ratings by Personality Types

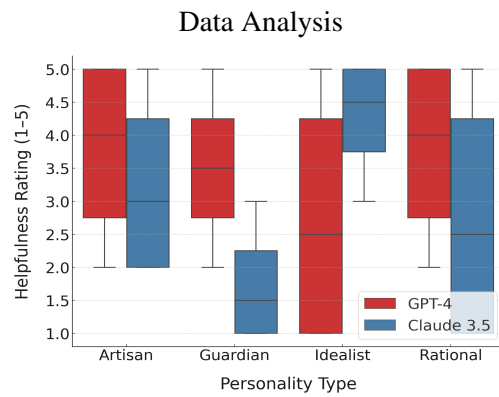


Figure 14: Distribution of helpfulness ratings for the data analysis task by personality type and LLM

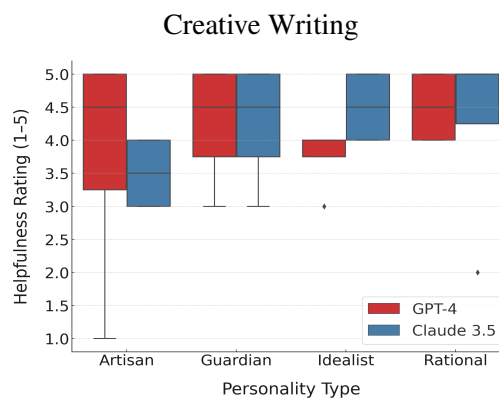


Figure 15: Distribution of helpfulness ratings for creative writing task by personality type and LLM

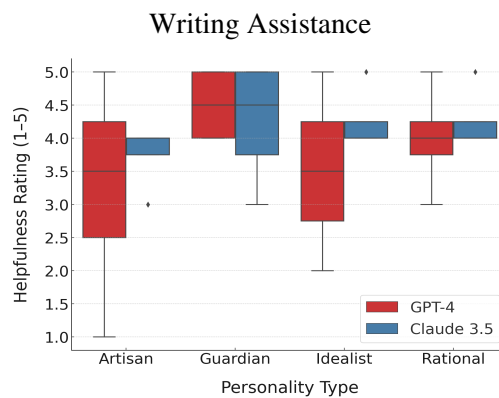


Figure 16: Distribution of helpfulness ratings for writing assistance task by personality type and LLM

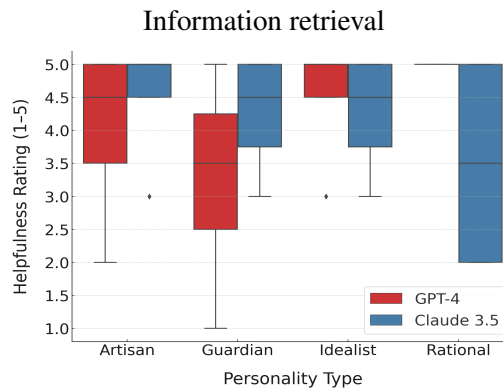


Figure 17: Distribution of helpfulness ratings for information retrieval task by personality type and LLM

## A.2 Examples of Prompting Strategy and Data Analysis Task Outcomes

### Participant A (Correct Outcome)

**Prompt:** I have a dataset in excel and I need to extract some data out of it. I have the following columns: Loan Purpose Checking Savings Gender Age Credit Risk

**Response:** That's great. I'd be happy to help you extract data. Could you please provide more specific details on what exactly you're looking to extract from these columns? For example, are you looking for average age of people with a certain credit risk or anything in specific? The more details you provide, the better I can guide you.

**Prompt:** I have the sheet opened up on google sheets. Here is the first question. I need to find how many male students under the age 29 applied for car loans. How can i do that?

**Response:** Sure, you can use the 'COUNTIFS' function in Google Sheets to achieve this. The 'COUNTIFS' function is used to apply multiple conditions on multiple ranges. Considering the columns are as follows:

- Loan Purpose = A
- Gender = B
- Age = C

Use the following formula:

'=COUNTIFS(A:A, "Car Loan", B:B, "Male", C:C, "<29")' [...] As a result, the function will return the count of male students under the age of 29 who applied for car loans.

**Prompt:** Just output the sample of formula I should use

**Response:** Certainly! Here's the specific formula: '=COUNTIFS(A:A, "Car Loan", B:B, "Male", C:C, "<29")' This function should provide the number of male students under age 29 who applied for car loans. As previously identified, you'll replace A:A, B:B, and C:C with the columns in your Google Sheet where Loan Purpose, Gender, and Age data are respectively. [...]



### Participant B (Incorrect Outcome)

**Prompt:** I am giving you a series of data about loan approvals and I want you to give me some information about it. Understood?

Dataset:

Loan Purpose Checking Savings Gender Age Credit Risk

Business \$0 \$739 M 23 Low

Furniture \$0 \$1,230 M 32 High

[...]

[...]

Furniture \$457 \$318 M 40 Low

Business \$5,133 \$698 M 36 High

Car \$0 \$33 M 29 Low

**Response:** This dataset is quite extensive, thus providing specific information might need a more particular request. However, I can provide some general insights based on the given data:

1. Loan Purposes: The data covers different purposes for loan applications such as Business, Furniture, Car, Technology, and Education. Business loans appear most frequent while Education, Furniture, and Car also have a significant presence.

[...]

5. Credit Risk: The dataset contains information about Low and High credit risk individuals, with both categories appearing quite frequently. For more detailed analysis or specific information, please make a precise request or provide me with the exact parameters for analysis.

**Prompt:** Prompt: All I need you to answer is:

1. How many male students, under the age of 19, applied for car loans?

2. How many students, have a checking account balance, of less than \$500?

**Response:** Certainly! I'll analyze the dataset to answer your specific questions. To answer this, I'll count the entries that meet all of these criteria:

- Gender: M (Male)

- Age: Under 29

- Loan Purpose: Car

After going through the data, I found 9 entries that meet these criteria.

2. How many students have a checking account balance of less than \$500?

For this question, I'll count the entries where:

- Loan Purpose: Education

- Checking account balance: Less than \$500

After analyzing the data, I found 16 entries that meet these criteria.

It's important to note that these counts are based on the data provided in the dataset. If the dataset is a sample or if there's any missing data, the actual numbers in the full population might be different.

Figure 18: Interaction samples from two participants: concise prompting by Participant A led to a correct result, while verbose input from Participant B resulted in model confusion and an incorrect answer