The Generation Gap: Age Differences in Human–LLM Interaction

Anonymous ACL submission

Abstract

As society's reliance on LLMs becomes stronger with each passing day, the concerns over its treatment of less-tech-savvy and other discriminated demographics grow and grow. We examine the effects of age on LLM use on both the input and output sides, over two tasks representing an assisted creativity scenario and an information request scenario. The results of our human study suggest that LLMs do not significantly discriminate against younger or older adults, despite their clear stylistic differences. We do, however, find that assisting creativity is easier for models than providing information. We also find that LLM-as-a-judge should be interpreted with caution, as LLM evaluators tended to assign systematically higher scores than both human experts and users' self-evaluations. At the same time, agreement across evaluation sources was more consistent for interactions with older adults.

1 Introduction

Human language undergoes notable changes during healthy aging, affecting both spoken and written communication (Biran et al., 2023; Cho et al., 2021). These changes, driven by age-related sensory and cognitive alterations, are reflected across multiple language domains, including lexical-semantic processing, phonology, grammar, comprehension, and discourse abilities (Burke and Shafto, 2011; Baciu and Roger, 2024). Age affects vocabulary and syntax in digital communication, with younger adults using more diverse and descriptive language than their older counterparts (Alshahrani, 2023). In addition, older adults make more typing errors and provide fewer corrections (Kalman et al., 2015). These findings highlight the distinct linguistic styles that different age groups may employ in written communication and how they interact with AI systems.

The development of large language models (LLMs), such as ChatGPT (Brown et al., 2020; Naveed et al., 2025), have made AI tools widely accessible across various domains (Pack and Maloney, 2023; Burger et al., 2023). However, despite their impressive capabilities, these systems are not neutral. LLMs are known to exhibit demographic biases, including systematic age bias (Zhang et al., 2021; Chu et al., 2023). Beyond these structural biases, LLM behavior can also vary substantially depending on how prompts are phrased (Sclar et al., 2024; Hackmann et al., 2024).

In this study, we focus on different age groups interacting with LLMs. We suspect that, given the way wording and style of user input directly shape model responses, differences in how younger and older adults formulate prompts for LLMs may influence the quality and effectiveness of the output. While prior work shows that sociodemographic cues in prompts can influence how LLMs simulate demographic groups (Lutz et al., 2025) and that implicit personalization may occur when models infer user traits from conversational signals (Neplenbroek et al., 2025), no studies to date have examined how naturally-produced prompts by different demographic groups differ in structure, nor how such differences influence model outputs. This leaves open the question of how user characteristics, prompt formulation, and LLM performance interact in real-world settings. We address this problem in a human-subject study, comparing older and younger adults in their interaction with an LLM in two distinct use modes: an *assisted creativity* scenario and an *information request* scenario.

Along three axes of comparison, we find somewhat surprising results. Although older and younger adults addressed the LLM with distinct linguistic styles, age-related differences in the model's responses were limited across surface measures and perceived quality. This pattern was reflected in users' self-evaluations and in human expert and LLM judges, despite a systematic tendency for LLM judges to score responses more generously.

We do, however, find consistent differences between the two task scenarios: the models struggle to provide information but excel at fostering creativity.

2 Method

The study employed a 2 × 2 mixed design, with age group (younger adults, ages 20–45, vs. older adults, ages 65+) as a between-subjects factor and task type (assisted creativity vs. information request) as a within-subjects factor.

2.1 Participants

Our sample consisted of 90 participants recruited via Prolific,[1] of whom 45 were older adults and 45 were younger adults.[2] All participants were native English speakers residing in the UK with no prior experience with AI systems such as ChatGPT or Gemini. Participants completed the study in an average of 22:13 minutes (range 13:46–48:02). See full details for recruitment, eligibility criteria, and screening procedures in Appendix A.1.

2.2 Design and Procedure

Participants first received an explanation of GPT and completed a brief training phase to become familiar with its interactions. They then completed two experimental tasks, which were presented in counterbalanced order to control for sequence effects. All interactions were open-ended, unconstrained in both length and structure, and routed through the G4R system (Kim, 2025), using the GPT-4o-mini model, with a default temperature setting of 1.0. To maintain the model's standard conversational profile, no custom system prompts were applied, and each session was initiated with the system's default opening message ("What can I help with?"). Complete instructions for each stage are provided in Appendix A.2.

Information request task (trip planning). Following travel-planning tasks used in prior work (Knoth et al., 2024), participants were instructed to interact with the system to obtain a complete plan for a multi-day trip to Andorra.

Assisted creativity task (greeting writing). Participants were instructed to interact with the AI to generate a personalized greeting-card message for a close friend or family member.

(**R**)	To what extent did the response address what you asked for?
(**C**)	How clear and understandable was the response?
(**U**)	How useful was the response for your purpose?
(**S**)	If this were the only answer you received, would it be sufficient?

Table 1: Self-evaluation questionnaire items: **R**elevance, **C**larity, **U**sefulness, **S**ufficiency.

2.3 Measures

Surface quantities. We extracted from the conversations the response times for both model and participant, measured in seconds; as well as text characteristics from user input and model output: total word count and spelling error rate. These measures were used to index response length, verbosity, and linguistic quality.

Lexical features. We measured lexical differences between the groups in terms of word frequency counts (raw and Older/Younger ratio), as well as Positive Pointwise Mutual Information (PPMI). Full details in Appendix B.1.

Self-evaluation. After each task, participants completed a questionnaire assessing their satisfaction with the model's output across four dimensions: relevance, clarity, usefulness, and sufficiency (see Table 1 for the exact formulation). Each dimension was rated on a 5-point Likert scale (1 = Not at all, 5 = Completely).

Human expert evaluations. A set of independent human coders also evaluated the conversations according to relevance, clarity, and response accuracy. These evaluations provide a human-grounded assessment of response quality in a more objective manner than self-evaluation. Full details of the evaluation procedure are provided in Appendix B.2.

LLM as a judge. We solicited LLMs different than the task's (Google Gemini, Meta Llama, and Anthropic Claude) for ratings mirroring those of the self-evaluation, as well as ratings for response accuracy. Full details of the evaluation procedure are provided in Appendix B.3.

3 Results

Surface quantities. The main effects of age and task on statistical measures in the conversations are

[1] https://prolific.com/

[2] Older adults: $M = 70.53$, $SD = 4.69$, range = 65–84; 31 female, 14 male; younger: $M = 31.89$, $SD = 5.20$, range = 21–40; 29 female, 14 male, two participants who selected "prefer not to say".

Measure	Task	Age	Task × Age	Direction
User Input Characteristics				
# of messages	–	–	–	–
Word count	*	–	–	Trip > Greet
Error rate	*	–	–	Trip > Greet
GPT Output Characteristics				
Response time	***	–	–	Trip > Greet
Output length	***	–	*	↑ Trip (YA)
Error rate	***	–	–	Trip > Greet

Table 2: Summary of significant effects for user input and GPT output characteristics. Asterisks indicate statistically significant effects (* $p < .05$, ** $p < .01$, *** $p < .001$). ↑ indicates a larger task-related increase in a task for an age group.

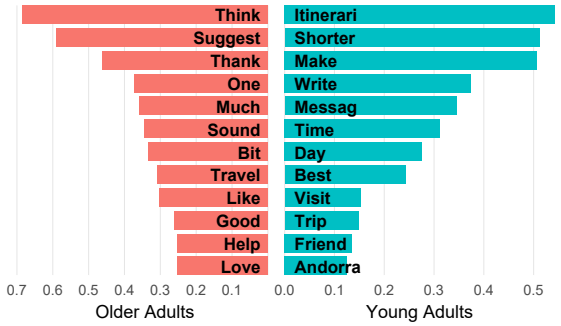

Figure 1: Top word stems per age group in participants' prompts to the language model by age group, ranked by PPMI, on both tasks. Higher absolute PPMI values indicate stronger-than-expected associations between a stem and a given age group.

shown in Table 2, and full statistical analyses are presented in Appendix C. Overall, we found that task type was the primary source of variation across all measures in both user input and model output. On the information request task (trip planning), participants in both age groups produced longer inputs than in the assisted creativity task (greeting), and GPT generated substantially longer, slower responses. This pattern was consistent across all measures, reflecting the increased cognitive and computational demands of information request interactions.

Age-related differences were limited and specific: the only variation in age-group output was observed on the information request task, with longer responses generated for younger adults. These results may suggest a bias congruent with the assumption that AI systems target more dominant user demographics.

	Assisted Creativity		Information Request		Age Group
Measure	YA	OA	YA	OA	Diff
Relevance	4.53	4.36	4.11	4.40	*
Usefulness	4.18	4.22	4.02	4.31	–
Clarity	4.53	4.73	3.96	4.47	**
Sufficiency	4.13	4.16	3.69	3.91	–

Table 3: Mean self-evaluation ratings by task type and age group. The final column indicates significant differences between age groups within a task condition (* $p < .05$, ** $p < .01$).

Lexical features. We analyzed the conversation texts based on relative frequency and PPMI, identifying systematic age-related differences in both participant prompts and model responses. When prompts were combined across tasks, core vocabulary (i.e., the most frequent words) was similar across age groups, but systematic differences emerged at the extremes of the distribution. Consistent with this pattern, as shown in Figure 1, PPMI revealed consistent differences in lexical emphasis. Prompts from younger adults contained more action- and production-oriented language, whereas those from older adults more frequently included reflective, evaluative, and interpersonal terms. These patterns persisted after stemming, indicating stable interaction styles (full results and analysis in Appendix D.1). PPMI analyses of the model's responses, provided in detail in Appendix D.2, revealed complementary age-related differences in its outputs, particularly in the information request task. Overall, textual analysis indicates expected effects of the social grounding across the two age groups, with each emphasizing different aspects of life in ways consistent with their life-stage priorities and everyday experiences.

Self-evaluation. Participants rated the model's output across four dimensions: relevance, usefulness, clarity, and sufficiency. Full analyses are provided in Appendix E. In Table 3, we present the aggregated ratings across all dimensions. We find that task type and age group influenced self-evaluations selectively rather than uniformly: responses in the greeting task received higher ratings for clarity and sufficiency for task completion compared to those in the trip planning task, but not much difference was found on relevance and usefulness. This may suggest that users' expectations for a creativity use case are lower than for obtaining information with respect to specificity of the

Comparison	Older Adults Greeting	Older Adults Trip	Younger Adults Greeting	Younger Adults Trip
Expert vs. Self				
Relevance	**.36***	.07	.22	−.02
Clarity	−.04	.04	.18	.08
Expert vs. LLM				
Relevance	**.47****	.29	**.54*****	.05
Clarity	**.42****	**.31***	.26	−.01
Accuracy	**.37***	.05	**.32***	.29
Self vs. LLM				
Relevance	**.31***	−.20	−.05	.02
Clarity	**.33***	**.41****	.00	.13
Usefulness	.29	−.06	−.03	.13
Sufficiency	.20	−.08	−.03	.02

Table 4: Pearson correlations between evaluation sources by age group and task. * $p < .05$, ** $p < .01$, *** $p < .001$.

answer, or that the model is less successful in producing reliable information than free text.

Age-related differences were most evident in clarity and relevance: older adults consistently rated the model's responses as clearer than those of younger adults across all tasks. For relevance, younger adults were more sensitive to task type, whereas older adults provided consistently high ratings across contexts.

In summary, participants' subjective evaluations of the model's responses were primarily shaped by task characteristics rather than by age-related evaluation patterns. The observed age-related differences seem to reflect selective sensitivity to task structure rather than overall evaluation tendencies.

Agreement between evaluation sources. We examined the agreement between participants' self-evaluations, expert human raters, and LLM evaluations (see Appendix F for full analyses).

The correlations across the three evaluation sources, shown in Table 4, suggest stronger agreement between expert raters and the LLM than with participants' self-evaluations. This agreement appears more consistently in responses generated for older adults and in the greeting-writing task, possibly reflecting the greater complexity of evaluating response quality in the information request task.

We observed significant correlations between self-evaluations and LLM ratings only for responses produced for older adults. This pattern may reflect a closer alignment between older (inexperienced) participants and LLMs in their implicit criteria for evaluating response quality.

To examine whether evaluation sources differed not only in agreement but also in rating levels, we

Dimension	Task	Older adults	Younger adults
Relevance	Greeting	L > S > E	L ≈ S > E
	Trip	L ≈ S > E	L > S ≈ E
Clarity	Both	L > S > E	L > S ≈ E
Accuracy		L > E	L > E

Table 5: Pairwise differences between evaluation sources (**L**LM, **S**elf, **E**xpert). Evaluators are ordered by mean rating (> significant; ≈ not significant).

compared mean scores across evaluators for relevance and clarity (see Appendix F for full analyses). As summarized in Table 5, LLM evaluators tended to assign the highest ratings, suggesting a more lenient evaluation criterion relative to expert human judges, consistent with prior work on LLM-based evaluation (e.g., Balog et al., 2025).

Self-evaluations generally fell between LLM and human judgments, with some variation across tasks and age groups for relevance. In contrast, only older adults rated the responses they received as *clearer* than expert human judges did. This pattern may reflect differences in expectations or interpretations of clarity, suggesting that older adults perceived responses as sufficiently clear even when external evaluators judged them more critically.

Both human judges and LLMs were asked to assess the accuracy of the end-result of the conversation. The comparison between these two types of judges, shown in Table 5, suggests that LLMs are more forgiving towards their counterparts' outputs, leading us to recommend careful use of this technique (Thakur et al., 2025; Chen et al., 2024).

4 Conclusion

Within the context of ongoing concern over the "technological gap" exacerbated by AI systems and their biased behavior towards already-discriminated groups, we find our results point towards a cautiously optimistic prospect: a popular system with no knowledge of the user's age did not, as a whole, present age-related bias when faced with users picked to not have prior experience with such possible bias.

At the same time, our findings suggest that LLM-based evaluations should be interpreted with caution. Across several evaluation dimensions, LLM judges tended to assign higher ratings than external human evaluators, indicating a more lenient evaluation criterion. In addition, evaluator agreement on response quality was more consistent for responses generated for older adults.

Limitations

This study has several limitations.

First, this study focused only on GPT-4o-mini. Although no significant age-related discrimination was found, results may not generalize to other architectures (e.g., Claude or Gemini), which use different training data and safety alignments. As generative AI evolves rapidly, newer or larger models might show biases that this "mini" version does not. Future work should benchmark more models to ensure technological gaps do not widen as AI advances. Second, this study examined only two task types, each needing different cognitive and communicative skills. Thus, observed age-related language differences may partly reflect task-specific demands and may not generalize to other interactions. The selected tasks reflect common scenarios—assisted creativity and information requests—rather than more complex problem-solving or cognitively demanding activities. This design lets us see if age differences appear in everyday interactions, but also limits the conclusions we can draw. Future research should examine language use across a wider range of task types, including more cognitively demanding or domain-specific interactions, to clarify which age-related patterns are consistent and which are task-dependent.

Another limitation is that the observed age-related differences may reflect thematic content and life-stage priorities rather than purely psycholinguistic or structural changes in language. For example, distinctions in travel preferences (e.g., "family-friendly" vs. "comfort-oriented") or communication styles (e.g., "action-oriented" vs. "relational framing") are often driven by evolving social roles and life circumstances. Consequently, GPT outputs might be capturing stereotypical concerns associated with different ages rather than the subtle linguistic shifts typically studied in healthy aging. Future research should disentangle content from form to clarify whether AI models distinguish between age groups based on what is said or how it is said.

Finally, all participants were from the UK. As English-speaking Western contexts are well represented in LLM training, findings may not generalize to underrepresented linguistic or cultural regions. Future work should examine human–AI interaction across more diverse populations.

Ethical Considerations

The study was approved by the ethics committee of the authors' institution.

References

Abdulaziz Alshahrani. 2023. Analyzing the impact of generational variations on vocabulary and syntax in instagram posts of older and younger adults. *Eurasian Journal of Applied Linguistics*, 9(1):202–212.

Monica Baciu and Elise Roger. 2024. Finding the words: How does the aging brain process language? a focused review of brain connectivity and compensatory pathways. *Topics in Cognitive Science*.

Krisztian Balog, Donald Metzler, and Zhen Qin. 2025. Rankers, judges, and assistants: Towards understanding the interplay of llms in information retrieval evaluation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 3865–3875. ACM.

Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.

Michal Biran, Aviah Gvion, and Shira Shmuely-Samuel. 2023. Language in healthy ageing: A comparison across language domains. *Folia Phoniatrica et Logopaedica*, 75(2):90–103.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bastian Burger, Dominik K Kanbach, Sascha Kraus, Matthias Breier, and Vincenzo Corvello. 2023. On the use of ai-based tools like chatgpt to support management research. *European journal of innovation management*, 26(7):233–241.

Deborah M. Burke and Meredith A. Shafto. 2011. Language and aging. In Fergus I. M. Craik and Timothy A. Salthouse, editors, *The Handbook of Aging and Cognition*, 3 edition, pages 373–443. Psychology Press.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement biases. *Preprint*, arXiv:2402.10669.

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.

Sunghye Cho, Naomi Nevler, Sanjana Shellikeri, Natalia Parjane, David J Irwin, Neville Ryant, Sharon Ash, Christopher Cieri, Mark Liberman, and Murray Grossman. 2021. Lexical and acoustic characteristics of young and older healthy adults. *Journal of Speech, Language, and Hearing Research*, 64(2):302–314.

Charlene H. Chu, Simon Donato-Woodger, Shehroz S. Khan, Rune Nyrup, Kathleen Leslie, Alexandra Lyn, Tianyu Shi, Andria Bianchi, Samira Abbasgholizadeh Rahimi, and Amanda Grenier. 2023. Age-related bias and artificial intelligence: a scoping review. *Humanities and Social Sciences Communications*, 10(1):510.

Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–191.

Stefan Hackmann, Haniyeh Mahmoudian, Mark Steadman, and Michael Schmidt. 2024. Word importance explains how prompts affect language model outputs. *Preprint*, arXiv:2403.03028.

Yoram M Kalman, Gitit Kavé, and Daniil Umanski. 2015. Writing in a digital world: self-correction while typing in younger and older adults. *International Journal of Environmental Research and Public Health*, 12(10):12723–12734.

Jin Kim. 2025. How to capture and study conversations between research participants and chatgpt: Gpt for researchers (g4r.org). *Preprint*, arXiv:2503.18303.

Nils Knoth, Antonia Tolzin, Andreas Janson, and Jan Marco Leimeister. 2024. Ai literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, 6:100225.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23212–23237, Suzhou, China. Association for Computational Linguistics.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Wajid Nasser. 2026. Evaluative fingerprints: Stable and systematic differences in llm evaluator behavior. *Preprint*, arXiv:2601.05114.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. 2025. Reading between the prompts: How stereotypes shape LLM's implicit personalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20378–20411, Suzhou, China. Association for Computational Linguistics.

Austin Pack and Jeffrey Maloney. 2023. Using generative artificial intelligence for language education research: Insights from using openai's chatgpt. *TESOL Quarterly*, 57(4):1571–1582.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *Preprint*, arXiv:2406.12624.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Sheng Zhang, Xin Zhang, Weiming Zhang, and Anders Søgaard. 2021. Sociolectal analysis of pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4581–4588. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Detailed Experimental Setup

A.1 Participants

The final sample comprised 90 participants, including 45 younger adults (ages 20–45) and 45 older adults (ages 65+), all native English speakers residing in the UK. Participants were recruited via Prolific[3] across two independent trials, one for each age group. Participants were required to have completed at least 50 previous studies on Prolific with an approval rate of 98% or higher, to report regular internet usage, and to have no prior experience with AI systems such as ChatGPT, Gemini, or Claude.

A rigorous multi-stage screening process was implemented. Initially, recruitment was limited to Prolific users who had previously selected "None of these" in response to the platform's standard pre-screening question about interaction with AI chatbots. To address potential recent changes in experience not yet reflected in Prolific's database, a custom screening question was introduced: "Do you have experience using artificial intelligence (AI) systems such as ChatGPT, Gemini, or Claude?" Only individuals who responded negatively were allowed to proceed.

To verify the integrity of self-reported data, two additional validation questions were embedded within the experiment. Participants were again asked to confirm their lack of experience; those indicating prior use were prompted to describe the nature and frequency of their interactions. Responses were manually reviewed, and any individual with prior AI experience was excluded from the final analysis.

A total of 880 individuals initiated the screening process. Of these, 790 were excluded: 785 due to prior AI experience identified at various stages, and 5 due to technical issues. The final sample comprised 90 participants.

Since slightly older adults were initially recruited, the younger adults were subsequently matched to ensure equal group sizes. The planned sample size ($n = 84$) was determined through an a priori power analysis using G*Power (Faul et al., 2007) ($\alpha = .05$, power = .95, $f = .20$). The final sample ($n = 90$), therefore, exceeded the required sample size.

The older adult group ($n = 45$) had a mean age of 70.53 years ($SD = 4.69$; range = 65–84) and included 31 females and 14 males, whereas the younger adult group ($n = 45$) had a mean age of 31.89 years ($SD = 5.20$; range = 21–40) and included 29 females, 14 males, and two participants who selected "prefer not to say." Participants completed the study in an average of 22 min 13 s ($SD = 5$ min 46 s), with completion times ranging from 13 min 46 s to 48 min 02 s.

All participants provided informed consent. They were informed of their right to withdraw from the trial at any time without penalty. Participants were also notified that they would receive credit for their participation time and that their data would be used exclusively for research purposes. Compensation for all participants was provided in accordance with Prolific's standard payment policies. Participants who completed the study received £3.75 for an average duration of 25 minutes (approximately £12/hour). Participants who were screened out after only completing the initial screening question received £0.10 for that stage.

A.2 Instructions Given to Participants

Below is the full text of the instructions provided to participants for each of the three tasks: GPT explanation, the training phase, the information request task, and the assisted creativity task.

GPT explanation: Introduction to ChatGPT
Thank you for agreeing to take part in this study. In the study, you will be asked to complete two different tasks that involve interacting with a tool called ChatGPT — a type of Artificial Intelligence (AI) that can understand and generate human-like language. You do not need any previous experience with AI. The tasks will guide you through how to use GPT, and you'll be asked to type a few questions or instructions and observe how it responds.

What is ChatGPT?
ChatGPT stands for Generative Pre-trained Transformer. It is a powerful AI system that was exposed to a large amount of text from the internet, including books, articles, and conversations, which allows it to respond in a natural, human-like way.

When you write a message (called a prompt), ChatGPT uses what it has learned to generate a reply. But it's not just a question-and-answer system. GPT is designed to carry out a conversation, so you can continue the interaction, clarify what you mean, refine your request, or build on what it said previously — much like you would in a real discussion with a person.

People use ChatGPT for many different pur-

[3] https://prolific.com/

poses, such as:

- Asking for information or explanations (e.g., "What causes climate change?")
- Getting ideas (e.g., "What can I cook for dinner with only pasta and spinach?")
- Planning (e.g., "Help me plan a 3-day trip to Barcelona.")
- Improving writing (e.g., "Can you make this message sound more polite?")
- Creating things (e.g., poems, stories, recipes, or even computer programs)

In this study, we are interested in how people interact with GPT. There are no right or wrong answers — just be yourself and try your best to communicate naturally.

Training Phase: first short task Before moving on to the main exercises, you'll now have a chance to explore how ChatGPT works briefly. Start by asking ChatGPT for something simple — for example, a recipe. Once you receive the response, ask ChatGPT to adjust it based on a specific preference or need — for instance, removing one ingredient because you want a version without it. Try to notice how the chatbot responds and adapts. This task is designed to help you become familiar with interacting with ChatGPT.

You have 4 minutes for this short task. You may complete it in less time if you feel you have reached a satisfactory result. After a few minutes, a "continue" button will appear, allowing you to move on at any point. However, it is recommended to use the full 4 minutes to work on the task. If you need more time to finalize your answer, it is also fine to continue a little beyond the 4 minutes. On the next screen, you will see the task interface. At the bottom of the page, a timer will show the remaining time.

Information Request Task [Trip Planning] Trip to Andorra Task: Traveling can be a wonderful way to discover new places, relax and learn about new cultures. But planning a trip can often be challenging, especially if you're traveling to a new country or if you're unsure of everything you want to do and perhaps traveling alone for the first time. In such cases, it can be helpful to turn to the assistance of chatbots. We'll now look at whether and how ChatGPT can help you plan trips. These sample prompts (input or instructions you type for the AI) might help you with your trip planning:

- "You are a tour guide. I'm very interested in theater in Naples, please tell me more about what places and buildings I should visit and in what order."
- "What is the cheapest destination for a 3-day city trip in Europe? My budget is around 1000 euros."
- "List me free museums in Amsterdam. I am primarily interested in modern art."

Your task now is to plan a 4-day trip to Andorra in September. Whether you travel alone or with others, where you stay, whether you travel around, what activities you do, etc., are entirely up to you. Please plan your trip as concretely as possible. Use your own language, do not copy-paste from the example prompts above. Ask the chatbot to plan the itinerary as if you had no prior knowledge, relying entirely on its suggestions. Try to create the itinerary as "automated" as possible using (almost exclusively) the chatbot.

You have 7 minutes for the task "Travel to Andorra." You may complete it in less time if you feel you have reached a satisfactory result. After a few minutes, a "continue" button will appear, allowing you to move on at any point. However, it is recommended to use the full 7 minutes to work on the task. If you need more time to finalize your answer, it is also fine to continue a little beyond the 7 minutes. On the next screen, you will see the task interface. At the bottom of the page, a timer will show the remaining time.

Assisted Creativity Task [Greeting Writing] Birthday Greeting Task: Sending a thoughtful greeting can be a wonderful way to show someone you care. But finding the right words can often be difficult, especially when you're trying to strike the right tone: warm, kind, and appropriate for the occasion. Whether it's a birthday, a wedding, the birth of a child, or a retirement celebration, many people turn to chatbots like ChatGPT to help them write meaningful and well-phrased messages. We'll now look at whether and how ChatGPT can help you write greetings. These sample prompts (input or instructions you type for the AI) might help you with your greeting:

- "Write a warm and heartfelt message to congratulate a friend on the birth of their first child."
- "Create a short and cheerful wedding message suitable for a greeting card."
- "Generate a respectful and appreciative note for someone retiring after many years of work."

Your task now is to write an instruction for ChatGPT to generate a birthday message for someone close to you. Whether it's a family member, a friend, or someone else you care about is entirely up to you. The tone, length, and level of formality are also yours to decide. Please write your prompt as clearly and concretely as possible. Use your own language, do not copy-paste from the example prompts above. However, avoid "unnecessary" personal contributions in the form of your own wording or personal details. Try to formulate the instruction as "automated" as possible using (almost exclusively) the chatbot.

You have 7 minutes for the task "Birthday Greeting." You may complete it in less time if you feel you have reached a satisfactory result. After a few minutes, a "continue" button will appear, allowing you to move on at any point. However, it is recommended to use the full 7 minutes to work on the task. If you need more time to finalize your answer, it is also fine to continue a little beyond the 7 minutes. On the next screen, you will see the task interface. At the bottom of the page, a timer will show the remaining time.

A.3 Self-Evaluation Questionnaire

After each task, participants completed a self-evaluation questionnaire assessing their satisfaction with the model's output across four dimensions: Relevance, Usefulness, Clarity, and Sufficiency. Each dimension was rated on a 5-point Likert scale (1 = Not at all, 5 = Completely).

Instructions and Items Participants were presented with the following prompt:

> "Based on the most recent response you received from ChatGPT, please evaluate the performance of the model by answering the following questions:"

The questionnaire consisted of the following four items:

1. **Relevance:** To what extent did the response address what you asked for?
2. **Usefulness:** How useful was the response for your purpose?
3. **Clarity:** How clear and understandable was the response?
4. **Sufficiency:** If this were the only answer you received, would it be sufficient?

All items were answered using the following 5-point scale: (1) Not at all, (2) Slightly, (3) Somewhat, (4) Mostly, and (5) Completely.

B Detailed Analysis Setups

B.1 Lexical Analysis Method Details

To examine systematic differences in language use between age groups and to characterize the linguistic distinctiveness of the model's responses, we conducted a quantitative text analysis using R and the `quanteda` package (Benoit et al., 2018).S The analysis was applied to two distinct corpora: (1) participants' prompts to the model, and (2) the model's generated responses.

Data Aggregation and Preprocessing. To capture stable interaction styles and reduce noise from idiosyncratic utterance-level variations, textual data were aggregated at the participant level. This process yielded one concatenated document per participant for each analysis target. Before analysis, all texts underwent a standard preprocessing pipeline. First, the text was tokenized into individual tokens, with punctuation, numbers, and symbols removed. Second, standard English stopwords were removed. Third, tokens shorter than 2 characters were filtered out to focus on content-bearing vocabulary. Finally, we applied the Snowball stemming algorithm implemented in the `quanteda` package (Benoit et al., 2018) to reduce words to their root forms (e.g., merging "planning" and "planned" into "plan"). This ensured that morphological variations did not obscure broader lexical patterns.

Relative Frequency Ratio. To visualize broad differences in word usage, we constructed a Document-Feature Matrix (DFM). To ensure robustness, features were filtered to retain only words that appeared in at least 2 participants' documents ($docfreq \geq 2$). For each retained word, we calculated its relative frequency within each age group

(count divided by the group's total token count). We then computed the ratio of these frequencies (Older Adults / Younger Adults) to identify words disproportionately used by one group relative to the other, following standard approaches for lexical comparison (Monroe et al., 2008).

Lexical Distinctiveness (PPMI). To identify specific vocabulary markers that statistically distinguish between age groups, we computed Positive Pointwise Mutual Information (PPMI) (Turney and Pantel, 2010). For this analysis, the participant-level DFM was aggregated by Age Group. To mitigate the known PMI bias toward low-frequency terms, we applied a secondary frequency threshold, retaining only words with a minimum total frequency of 30 across the entire corpus.

PMI quantifies the discrepancy between the observed joint probability of a word (w) and a group (g) and their expected probability under the assumption of independence. The metric was calculated as:

$$\text{PMI}(w, g) = \log_2 \left(\frac{P(w, g)}{P(w) \times P(g)} \right) \quad (1)$$

We retained only positive values to focus on characteristic associations:

$$\text{PPMI}(w, g) = \max(\text{PMI}(w, g), 0) \quad (2)$$

Higher PPMI scores indicate that a specific word stem appears in a group's text significantly more often than would be expected by chance, given its overall frequency in the corpus.

B.2 Expert Human Evaluation

Expert human evaluations. To obtain an additional human-grounded benchmark for response quality, we conducted an independent human evaluation study via Prolific. Participants were recruited from Prolific's pool of *Qualified AI Taskers*, a group of workers who have successfully passed Prolific's AI task assessment and demonstrated advanced reasoning, fact-checking, and annotation skills.

A total of 27 independent evaluators participated in the annotation process. Their ages ranged from 20 to 69 years (*M* = 45.04, *SD* = 12.46); 8 identified as female and 19 as male. The average evaluation time was 89.6 minutes (*SD* = 27.0), ranging from 51.1 to 148.7 minutes. Evaluators were compensated at approximately £16.5 per hour.

Each evaluator rated 20 conversations sampled from the experimental dataset. The conversations were balanced across experimental conditions: 10 generated by younger adults and 10 by older adults, with equal distribution across the two task types (assisted creativity and information request), yielding five conversations per age $\times$ task combination.

Evaluators were presented with anonymized conversation transcripts in which explicit age-related information had been masked prior to evaluation. A preprocessing step removed or redacted textual references to age (e.g., "I am 70 years old", "turning 30", "in my twenties") and replaced them with a neutral token (###). As a result, evaluators were not provided with any demographic or identifying information about the participants and were unable to infer their age group. They were instructed to evaluate the quality of the AI responses solely in terms of the task.

Evaluators assessed the AI responses along three dimensions: *Relevance* (the extent to which the responses addressed the user's request), *Clarity* (how clear and understandable the responses were), and *Accuracy* (the extent to which the responses were factually correct and free from errors or misleading information). Ratings were provided on a 5-point Likert scale ranging from 1 (*Not at all*) to 5 (*Completely*).

Each conversation was evaluated by three independent raters, and the final score for each dimension was computed as the mean of the three ratings.

Inter-rater reliability was assessed using Krippendorff's α (Krippendorff, 2018), a reliability coefficient designed for situations involving multiple raters and incomplete rating designs. This measure is particularly suitable for annotation settings in which different subsets of raters evaluate different items, as it accounts for varying rater assignments and chance agreement.

Reliability was computed across all human evaluations in the dataset. In total, the analysis included 540 ratings across 180 conversations, each evaluated by three raters drawn from a pool of 27 evaluators.[4] Krippendorff's α was computed separately for each evaluation dimension (Relevance, Clarity, and Accuracy).

[4] Agreement estimates may be influenced by variability between evaluators, as each conversation was rated by a different subset of raters. We did not assign all 180 conversations to each evaluator to avoid excessive evaluator fatigue and to preserve annotation quality.

B.3 LLM-as-a-Judge: Method Details

Recent studies have shown that large language models can serve as evaluators of natural language generation quality when provided with structured prompts and explicit scoring criteria (Zheng et al., 2023; Liu et al., 2023). Building on this approach, the present study implemented an LLM-as-a-Judge framework to assess the quality of AI-generated responses in the Assisted Creativity task (Greeting writing) and the Information Request task (Trip planning). This framework uses advanced large language models as impartial evaluators to score complete conversational transcripts against predefined evaluation criteria.

Models and Configuration. To ensure robust and diverse evaluation, three distinct large language models from different model families were selected as judges. The models included:

- **Google Gemini**: Specifically, the `gemini-2.5-pro` version.
- **Anthropic Claude**: Using the `claude-sonnet-4-5-20250929` model.
- **Meta Llama (via Groq)**: Using the `llama-3.3-70b-versatile` model.

This selection facilitates comparisons across different architectures and providers, thereby reducing reliance on the inductive biases of any single model. To maximize reproducibility and minimize variability in evaluation scores, the temperature parameter for all models was set to 0.0.

Evaluation Metrics and Prompting. Similar to the human evaluators, the LLM judges received anonymized conversation transcripts in which age-related and identifying information had been masked prior to evaluation. Each judge was provided with the complete transcript along with task-specific instructions.

A comprehensive system prompt was developed to guide the evaluation process, directing the LLMs to act as "expert evaluators of human–AI interactions." The models were instructed to produce a structured JSON response containing both reasoning and a numerical score for each evaluation dimension, as defined below. We defined five specific evaluation dimensions to capture different aspects of the interaction quality. The judges were provided with the following definitions for each metric:

1. **Task Relevance**: Assesses the extent to which the AI's responses directly addressed the specific requests made by the user throughout the conversation.
2. **Usefulness**: Evaluates the practical value of the responses, specifically considering how useful the AI's suggestions and content were for the user's stated purpose.
3. **Clarity**: Measures how understandable, coherent, and clear the AI's responses were to read and comprehend.
4. **Sufficiency**: Determines whether the responses provided enough information to complete the task (e.g., plan the trip or write the greeting) without the user needing to seek further external input.
5. **Accuracy**: Checks the factual correctness of the information provided and ensures the responses were free from errors or misleading statements.

Scoring Scale. The models were instructed to score each dimension using a 5-point Likert scale, ranging from 1 ("Not at all") to 5 ("Completely"). To ensure rigorous evaluation, the prompt explicitly directed the models to interpret the scale conservatively and reserve high scores for responses that clearly exceeded the standard of quality.

Aggregation of Judge Scores. To reduce bias from individual models, ratings from three judges were used. For each evaluation dimension, the scores from Gemini, Claude, and Llama were averaged to yield a single judge-averaged score for each participant and task. All five dimensions were analyzed separately, and no overall score was computed. These judge-averaged scores were used as dependent variables in subsequent statistical analyses to examine the effects of age group and task type.

Similar to the human evaluation, inter-rater reliability among the three LLM judges was assessed using Krippendorff's α.

C Surface Quantities

We begin by reporting the results of the **surface quantities analysis**. This analysis examines the structural and quantitative properties of the interaction, providing a baseline characterization of the

dialogue flow. We present findings regarding volume (word count, message count), technical accuracy (spelling errors), and timing (latency) for both participants' prompts and the model's responses.

C.1 User Input Characteristics

Number of messages to GPT A 2 × 2 mixed-design ANOVA was conducted to examine whether the number of messages sent to GPT differed by age group (younger adults vs. older adults; between-subjects) and task type (assisted creativity task [greeting writing] vs. information request task [trip planning]; within-subjects) (see Figure 2)

The main effect of age group was not significant, $F(1, 88) = 0.60$, $p = .44$, $\eta_g^2 = .005$, indicating that older adults ($M = 4.84$, $SE = 0.27$) and younger adults ($M = 5.14$, $SE = 0.27$) sent a similar number of messages. The main effect of task type was also not significant, $F(1, 88) = 0.38$, $p = .54$, $\eta_g^2 = .000$, with a similar number of messages for the assisted creativity task ($M = 5.06$, $SE = 0.21$) and the information request task ($M = 4.93$, $SE = 0.23$). The interaction between age group and task type was not significant, $F(1, 88) = 0.25$, $p = .62$, $\eta_g^2 = .000$, indicating that system interaction was similar across age groups and task types. These results indicate that the number of messages sent to the system was similar across age groups and task types, suggesting that both younger and older adults engaged with the model to a comparable extent regardless of the interaction context.

Number of words sent to GPT A 2 × 2 mixed-design ANOVA was conducted to examine whether the total number of words sent to GPT differed by age group (younger adults vs. older adults; between-subjects) and task type (assisted creativity task [greeting writing] vs. information request task [trip planning]; within-subjects) (see Figure 3)

The main effect of age group was not significant, $F(1, 88) = 0.61$, $p = .44$, $\eta_g^2 = .005$, indicating that older adults ($M = 74.6$, $SE = 5.0$) and younger adults ($M = 69.1$, $SE = 5.0$) generated a similar number of words in their prompts. A significant main effect of task type was found, $F(1, 88) = 9.23$, $p = .003$, $\eta_g^2 = .013$, with participants in both age groups producing more words in the information request task ($M = 76.0$, $SE = 3.4$) than in the assisted creativity task ($M = 67.6$, $SE = 4.1$). The interaction between age group and task type was not significant,

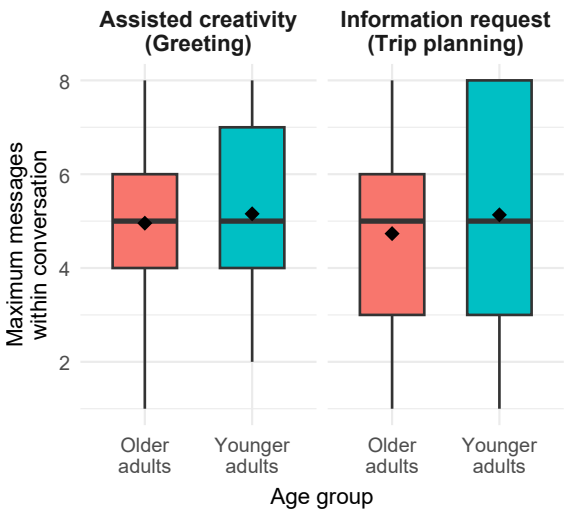

Figure 2: Maximum number of messages sent to GPT by age group and task type. Boxes represent interquartile ranges, horizontal lines indicate medians, and diamonds denote means.

$F(1, 88) = 0.07$, $p = .79$, $\eta_g^2 = .000$, indicating that the amount of text sent to GPT was similar across both age groups and task types. These results indicate that word production was similar across age groups, while varying by task type, suggesting that both younger and older adults produced comparable amounts of text, with longer prompts generated in the information request task regardless of interaction context.

Spelling Error Rate in Messages Sent to GPT Spelling error rate was operationalized as the number of misspellings per message, identified automatically using the English LanguageTool model (`language-tool-python` v3.2.1). Only matches explicitly labeled as misspellings were included, excluding suggestions related to grammar, style, or punctuation. For each message, the total number of detected misspellings was computed and used as a raw count measure of spelling errors. Messages containing no textual content were assigned a spelling error rate of zero.

A 2 × 2 mixed-design ANOVA was conducted to examine whether the spelling error rates (misspellings per message) sent to GPT differed by age group (younger adults vs. older adults; between-subjects) and task type (assisted creativity [greeting writing] vs. information request [trip planning]; within-subjects) (see Figure 4)

The main effect of age group was not significant, $F(1, 88) = 0.12$, $p = .73$, $\eta_g^2 = .000$, indicating

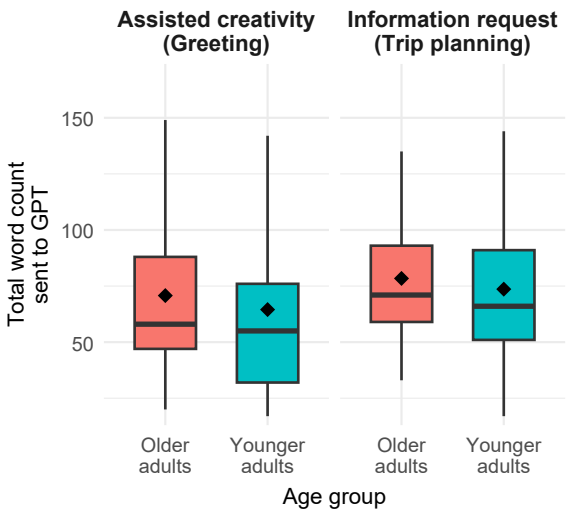

Figure 3: Total word count sent to GPT by age group and task type. Boxes represent interquartile ranges, horizontal lines indicate medians, and diamonds denote means.

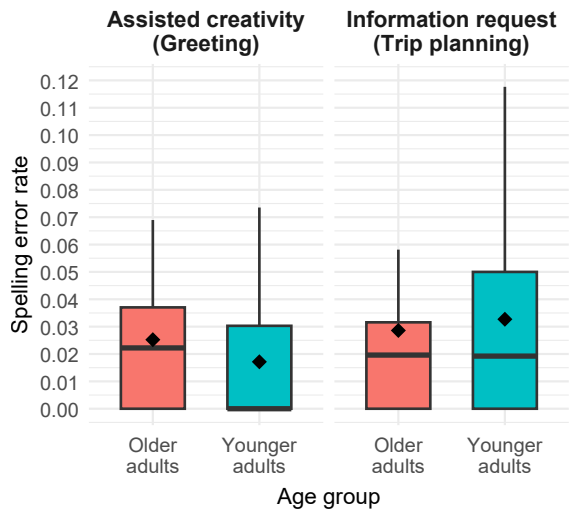

Figure 4: Spelling error rate in messages sent to GPT by age group and task type. Boxes represent interquartile ranges, horizontal lines indicate medians, and diamonds denote means.

that older adults ($M = 0.027$, $SE = 0.004$) and younger adults ($M = 0.025$, $SE = 0.004$) had a similar spelling error rate. A significant main effect of task type was found, $F(1, 88) = 6.91$, $p = .010$, $\eta_g^2 = .022$, indicating higher spelling error rates in the information request task ($M = 0.0307$, $SE = 0.00372$) than in the assisted creativity task ($M = 0.0212$, $SE = 0.00304$), averaged across age groups. The interaction between age group and task type did not reach statistical significance, $F(1, 88) = 2.86$, $p = .094$, $\eta_g^2 = .009$, indicating that the spelling error rate sent to GPT was similar across both age groups and task types. These results indicate that spelling error rates were similar across age groups, while varying by task type, suggesting that both younger and older adults exhibited comparable levels of spelling errors, with higher error rates observed in the information request task regardless of interaction context.

C.2 GPT Output Characteristics

GPT response time A 2 × 2 mixed-design ANOVA was conducted to examine whether GPT response time differed by age group (younger adults vs. older adults; between-subjects) and task type (assisted creativity [greeting writing] vs. information request [trip planning]; within-subjects) (see Figure 5)

The main effect of age group was not significant, $F(1, 88) = 0.32$, $p = .58$, $\eta_g^2 = .002$, indicating that older adults ($M = 7.14$, $SE = 0.41$) and younger adults ($M = 6.81$, $SE = 0.41$) elicited similar response times from GPT. A significant main effect of task type was found, $F(1, 88) = 347.12$, $p < .001$, $\eta_g^2 = .524$, indicating that GPT responded more slowly in the information request task ($M = 10.35$, $SE = 0.43$) than in the assisted creativity task ($M = 3.59$, $SE = 0.23$), averaged across age groups. The interaction between age group and task type was not significant, $F(1, 88) = 0.00$, $p = .95$, $\eta_g^2 = .001$, indicating that the task-related increase in GPT response time was similar across both age groups and task types. These results indicate that GPT response times were similar across age groups but varied by task type. Specifically, responses were consistently slower in the information request task than in the assisted creativity task, irrespective of the user's age or interaction context.

Number of words generated by GPT A 2 × 2 mixed-design ANOVA was conducted to examine whether the total number of words generated by GPT differed by age group (younger adults vs. older adults; between-subjects) and task type (assisted creativity [greeting writing] vs. information request [trip planning]; within-subjects) (see Figure 6)

The main effect of age group was not significant, $F(1, 88) = 2.64$, $p = .11$, $\eta_g^2 = .020$, indicating that older adults ($M = 1,014$, $SE = 61.7$) and younger adults ($M = 1,155$, $SE = 61.7$) received

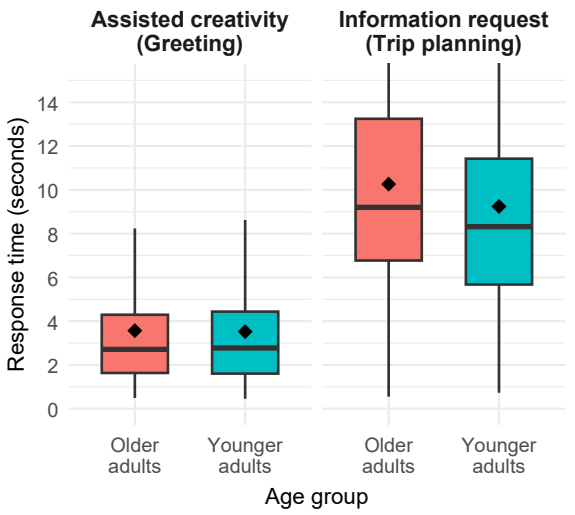

Figure 5: GPT response time (in seconds) by age group and task type. Boxes represent interquartile ranges, horizontal lines indicate medians, and diamonds denote means.

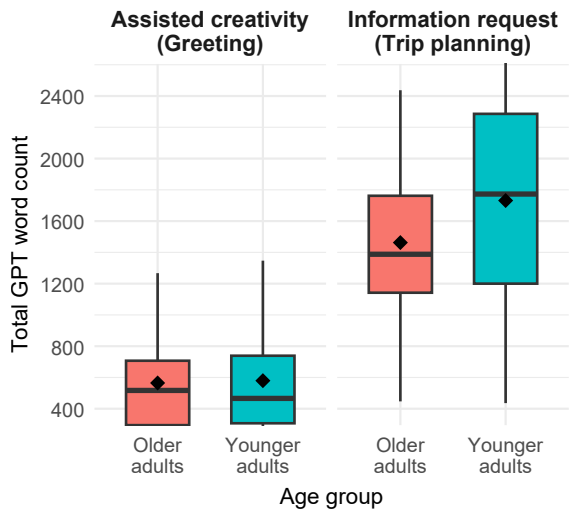

Figure 6: Total number of words generated by GPT by age group and task type. Boxes represent interquartile ranges, horizontal lines indicate medians, and diamonds denote means.

similar amounts of text overall. A significant main effect of task type was found, $F(1, 88) = 287.43$, $p < .001$, $\eta_g^2 = .515$. GPT generated more words in the information request task ($M = 1,597$, $SE = 65.1$) than in the assisted creativity task ($M = 572$, $SE = 37.3$). The interaction between age group and task type was also significant, $F(1, 88) = 4.42$, $p = .038$, $\eta_g^2 = .016$. Follow-up comparisons indicated that the task-related increase in GPT output was large in both age groups but more pronounced among younger adults (assisted creativity: $M = 579$; information request: $M = 1,731$) than among older adults (assisted creativity: $M = 565$; information request: $M = 1,463$). However, within each task, the two age groups did not differ significantly (all ps $> .17$). These results indicate that GPT's output length was primarily determined by task demands, with age exerting only a modest influence, reflected in the magnitude of the task effect rather than in absolute differences between age groups within each task.

Spelling Error Rate in Messages from GPT A 2×2 mixed-design ANOVA was conducted to examine whether spelling error rates in messages generated by GPT differed by age group (younger adults vs. older adults; between-subjects) and task type (assisted creativity [greeting writing] vs. information request [trip planning]; within-subjects) (see Figure 7)

The main effect of age group was not significant, $F(1, 88) = 1.45$, $p = .23$, $\eta_g^2 = .007$, indicating that older adults ($M = 0.012$, $SE = 0.001$) and younger adults ($M = 0.014$, $SE = 0.001$) had a similar spelling error rate in GPT-generated messages. A significant main effect of task type was found, $F(1, 88) = 235.83$, $p < .001$, $\eta_g^2 = .580$, indicating higher spelling error rates in GPT-generated messages in the information request task ($M = 0.025$, $SE = 0.001$) than in the assisted creativity task ($M = 0.001$, $SE = 0.000$). The interaction between age group and task type was not significant, $F(1, 88) = 3.83$, $p = .053$, $\eta_g^2 = .022$, indicating that spelling errors in GPT-generated text were similar by age groups and task types. These results indicate that spelling errors in GPT-generated text were primarily influenced by task demands rather than participant age. Both age groups exhibited similar increases in error rates in the information request task compared to the assisted creativity task, regardless of interaction context.

Association Between Participant Input Length and GPT Output Length Following the identification of differences in participant input and GPT output across tasks and age groups, the relationship between these two variables was systematically examined. Pearson correlations were calculated separately for each age group to assess the relationship between the number of words participants wrote to GPT and the number of words GPT returned. Significant positive associations were observed in both

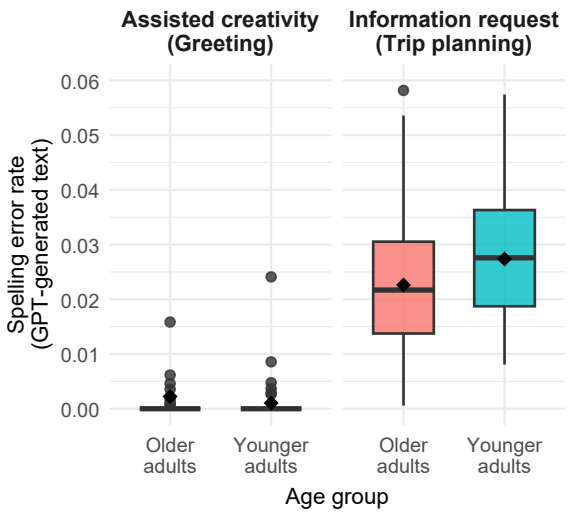

Figure 7: Spelling error rate in messages generated by GPT by age group and task type. Boxes represent interquartile ranges, horizontal lines indicate medians, and diamonds denote means.

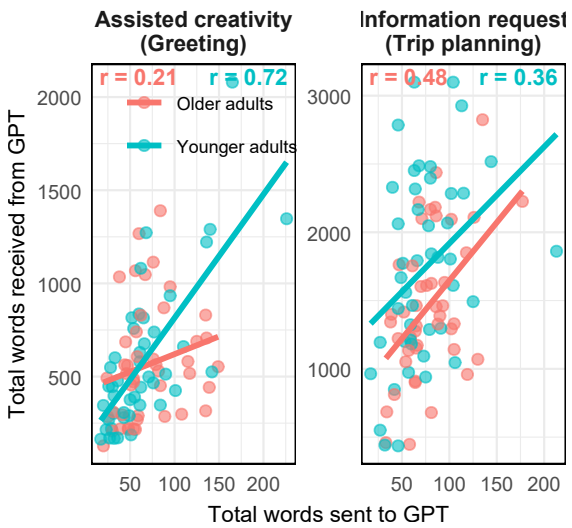

Figure 8: Relationship between the number of words participants sent to GPT and the number of words generated by GPT, shown separately by task type and age group. Lines represent linear fits within each group; correlation coefficients are reported for descriptive purposes.

groups: among younger adults, $r = .41$, 95% CI $[.22, .57]$, $p < .001$; among older adults, $r = .33$, 95% CI $[.13, .50]$, $p = .002$. A Fisher's z test for independent correlations indicated no significant difference between the two coefficients, $z = 0.64$, $p = .52$, and the confidence interval for the difference included zero ($-.17$ to $.34$). This finding suggests that the association's strength was similar across age groups.

Correlations were also computed separately for each task within each age group. For older adults, the association between participant input length and GPT output length was $r = .21$ in the assisted creativity task [greeting writing] and $r = .48$ in the information request task [trip planning]. For younger adults, the corresponding correlations were $r = .72$ in the assisted creativity task and $r = .36$ in the information request task. As these correlations were derived from repeated measurements within individuals, no statistical comparisons were conducted among the four coefficients. However, the descriptive pattern indicates that the strength of the association between user input and GPT output varies by task type.

To further investigate whether GPT output length varied as a function of participant input length, age group, and task type, a linear mixed-effects regression model was fitted with participant included as a random intercept. Participant input length did not significantly predict GPT output length overall, $\beta = 1.78$, $p = .36$. However, a significant interaction between input length and task type was observed, $\beta = 7.16$, $p = .008$, indicating a stronger association between input and output length in the information request task compared to the assisted creativity task. No reliable interactions involving age group and input length were observed (all ps $> .07$). In contrast, a significant interaction between age group and task type was found, $\beta = 695.80$, $p = .014$. After controlling for participant input length, GPT produced longer responses for younger adults than for older adults in the information request task. No comparable age-group difference was observed in the assisted creativity task. Collectively, these findings indicate that GPT's responsiveness to participant input length is primarily determined by task demands rather than by age group. Age-related differences are reflected in the overall magnitude of GPT output in the information request task, rather than in the input–output relationship.

D Lexical Differences Across Age Groups

D.1 Lexical Differences in User Prompts

To examine age-related differences in interaction style, we analyzed participants' model prompts. These analyses were conducted across both task types, Assisted Creativity and Information Request, rather than by contrasting them. Aggregating data

from multiple contexts enabled the identification of stable, group-specific characteristics in how participants formulate requests, independent of specific task constraints or topic content. We present findings using two complementary metrics: relative frequency differences (Figure 9) and lexical distinctiveness measured by PPMI (Figure 1). Details regarding data preprocessing, frequency calculations, and metric definitions are provided in Appendix B.1.

Relative Frequency Differences. Figure 9 summarizes age-related differences in lexical usage in participants' prompts to the model as a function of both relative frequency and overall prevalence. Most words clustered around equal relative frequency, which indicates broadly similar lexical choices in user prompts across age groups. Systematic differences emerged at the extremes of the distribution. Words more strongly associated with younger adults reflected action-oriented, task-focused language (e.g., *try*, *activities*, *things*), while words characteristic of older adults reflected reflective, communicative, or explanatory language (e.g., *think*, *suggest*, *mention*, *ideas*). These differences were most pronounced for words of moderate to high overall frequency, suggesting stable group-level preferences in how participants address and engage with the model rather than idiosyncratic usage.

Descriptive analyses of the most frequent words in each group further illustrate qualitative differences in prompt formulation. Younger adults' most frequent words were dominated by directive and execution-oriented terms (e.g., *can*, *make*, *please*), alongside task-relevant vocabulary, suggesting a concise and goal-directed style of prompting the model. In contrast, older adults more frequently used words associated with politeness, social exchange, and interpersonal framing (e.g., *thank*, *like*, *best*, *help*), indicating a more conversational and relational approach to structuring requests to the model.

To evaluate robustness, analyses were repeated using stemmed word forms (see Appendix B.1). The resulting patterns closely paralleled those for unstemmed words: younger adults' stems consistently highlighted action and production, while older adults' stems emphasized planning and assistance. This convergence confirms that age-related differences primarily reflect stable interaction styles rather than superficial morphological variations.

Lexical Distinctiveness (PPMI). To identify specific vocabulary markers disproportionately associated with each age group, we examined Positive Pointwise Mutual Information (PPMI) scores calculated across both task types. Figure 1 displays the most distinctive words for each group.

For older adults, the most lexically distinctive words included *think*, *suggest*, and *thank*, followed by functionally lighter or interpersonal terms such as *one*, *much*, *sound*, *bit*, *travel*, and *like*. Additional distinctive words included *good*, *help*, *love*, *need*, *card*, and *year*. Together, these words reflect a pattern of more reflective, evaluative, and relational language use in prompts produced by older participants.

In contrast, younger adults were characterized by a different set of lexically distinctive words. The highest PPMI values were observed for action- and production-oriented terms such as *add*, *itinerari*, *shorter*, *make*, and *write*. Other distinctive words included *messag*, *time*, *stay*, *day*, *best*, *place*, *funni*, *visit*, *trip*, and *friend*, suggesting a stronger focus on task execution, content manipulation, and concrete planning in how younger participants formulated their prompts to the model.

D.2 Lexical Distinctiveness in GPT Responses

To examine whether the language model's responses differed systematically by users' age group, we conducted PPMI analyses on the model's outputs. We first examined responses within each task type individually to identify context-specific adaptations, followed by an aggregated analysis across all interactions.

Information Request Task. In the trip-planning task, the model's responses exhibited clear age-related lexical distinctiveness (Figure 10).

For older adults, the most lexically distinctive words in the model's responses were dominated by transportation, infrastructure, and procedural travel information, including *London*, *Paris*, *train*, *airline*, *ticket*, *UK*, *station*, *platform*, and *journey*. Additional distinctive terms such as *current*, *save*, *communicate*, and *since*, further reflect an emphasis on logistical coordination and informational completeness. Together, these patterns suggest that when responding to older users' trip-planning requests, the language model tended to generate more structured, transit-focused, and procedurally oriented information, emphasizing routes, transportation systems,

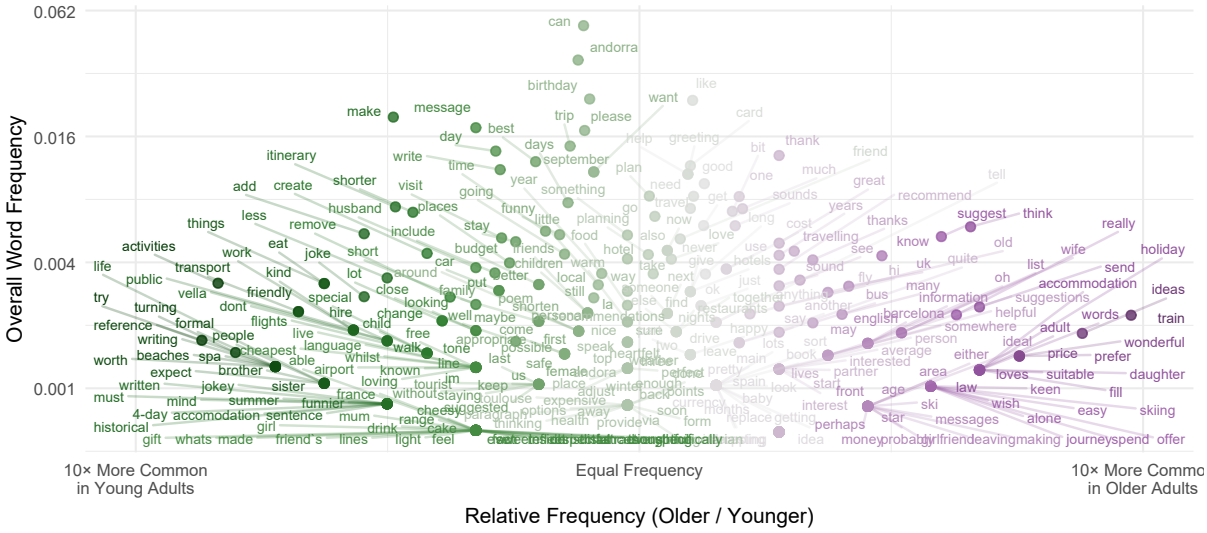

Figure 9: Lexical differences in participants' prompts to the model across age groups. Each point represents a word aggregated at the participant level, plotted by its relative frequency ratio between older and younger adults (Older / Younger; x-axis, log-scaled) and its overall frequency across participants (y-axis, log-scaled). Data are pooled across assisted creativity and information request tasks.

and formal travel logistics.

In contrast, responses generated for younger adults were characterized by a different set of lexically distinctive words, reflecting experiential, social, and lifestyle-oriented content. The highest PPMI values were observed for terms such as *children*, *family-friendly*, *beach*, *boutique*, *café*, *snack*, *water*, and *dietary*, alongside leisure- and activity-related words such as *rest*, *late*, and *crowd*. Additional distinctive terms included *EUR* and *driver*, suggesting a focus on local experiences and practical on-the-ground considerations. These patterns indicate that, when addressing younger users, the language model more frequently emphasized experiential aspects of travel, amenities, and contextual lifestyle information, rather than formal transportation or infrastructural details.

Taken together, these results suggest that age-related differences are reflected not only in how participants formulate prompts to the language model, but also in how the model structures its responses. Even when restricted to the same task type (information request) and controlling for overall word frequency, the model's outputs show systematic lexical differences across age groups: responses to older adults emphasize logistical and transport-related information, while responses to younger adults emphasize experiential and lifestyle-oriented travel content.

Assisted Creativity Task. Age-related differences extended to the creative context (Figure 11).

For older adults, the most lexically distinctive words in the model's responses included *wife*, *daughter*, *proud*, *heart*, *milestone*, and *grow*, alongside evaluative and relational terms such as *glad*, *care*, *reach*, and *offer*. These words suggest that, when responding to older users in a creative context, the model tended to emphasize emotional significance, interpersonal relationships, and life-course framing.

In contrast, responses generated for younger adults were characterized by a different set of lexically distinctive words, including *brighter*, *awesome*, *joke*, *cake*, *little*, and *deserve*, as well as relational terms such as *husband*, *brother*, and *partner*. Additional distinctive words such as *go*, *get*, *now*, *make*, and *challenge* indicate a more action-oriented, upbeat, and immediate tone in responses to younger users. Together, these results indicate

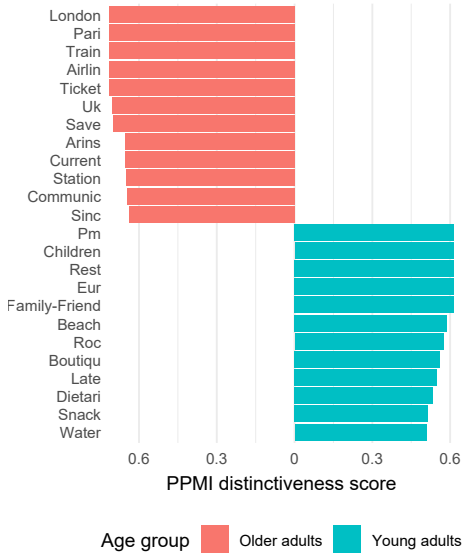

Figure 10: Top word stems per age group in the language model's responses, ranked by PPMI, on the information request task (trip planning).

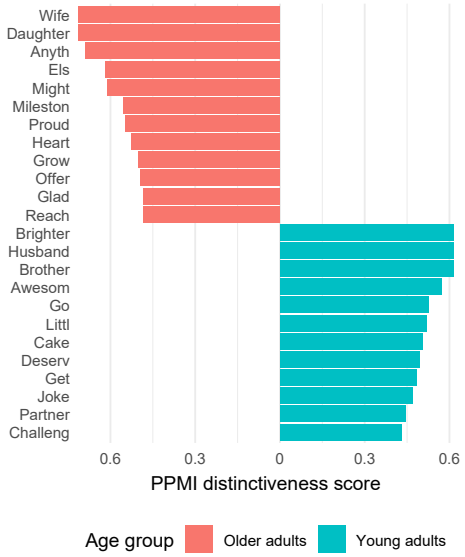

Figure 11: Top word stems per age group in the language model's responses, ranked by PPMI, on the assisted creativity task (greeting writing).

that age-related differences in the language model's outputs extend beyond informational content and persist in creative tasks. Specifically, the model appears to adapt not only the content but also the emotional framing and communicative tone of its responses to the user's age group.

Aggregated Analysis (Across Tasks). Finally, to assess stable response strategies independent of task content, we analyzed lexical distinctiveness across the entire corpus (Figure 12).

For responses addressed to older adults, GPT exhibited higher lexical distinctiveness for words related to formal planning, transportation infrastructure, and institutional contexts, including *son*, *London*, *Paris*, *train*, *airline*, *ticket*, *station*, and *platform*. In addition, several family-related terms (e.g., *wife*, *daughter*) and functionally oriented words (e.g., *communicate*, *current*, *save*) were more strongly associated with responses to older users. Together, these patterns suggest that GPT adopts a more structured, informational, and service-oriented style when generating responses for older adults.

In contrast, responses to younger adults were characterized by lexically distinctive words reflecting experiential, lifestyle-oriented, and informal content. These included terms related to leisure and environment (e.g., *beach*, *sea*, *rest*), social relations (*friend*, *brother*, *kid*), consumption and mobility (*euro*, *boutique*, *driver*), and affective or stylistic framing (e.g., *brighter*, *theme*). This pattern indicates that GPT's responses to younger users emphasize experiential framing and flexible, lifestyle-focused information.

Notably, these age-related lexical differences emerged despite collapsing across task types, indicating that they cannot be attributed solely to differences in task demands or topical content. Instead, the results suggest that the language model systematically adapts its lexical choices based on user age, reflecting age-sensitive response strategies in addition to differences observed in users' prompting behavior.

E Self-Evaluation of Model Responses

To examine how participants subjectively evaluated the quality of the model's responses, we analyzed self-reported ratings along four evaluation dimensions: relevance, usefulness, clarity, and sufficiency. Each dimension was analyzed using a 2×2 mixed-design ANOVA with age group (younger adults, older adults) as a between-subjects factor and task type (assisted creativity [greeting writing] vs. information request[trip planning]) as a within-subjects factor. This approach enabled us to identify potential interactions between task demands and age-related evaluation patterns, thereby allowing us to assess overall perceived response quality.

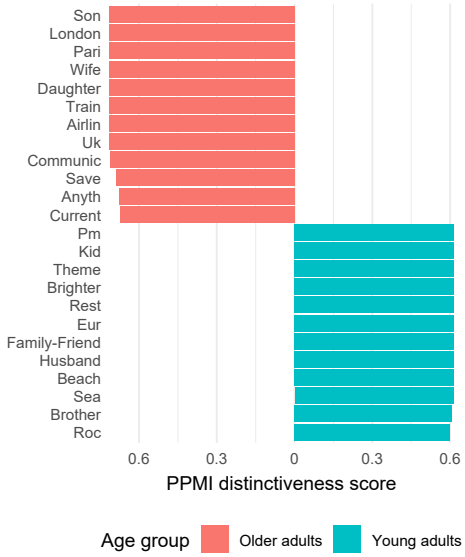

Figure 12: Top word stems per age group in the language model's responses, ranked by PPMI, pooled across task types.

Perceived Relevance Relevance was first examined by assessing participants' self-reported extent to which the model's responses addressed their requests. The main effect of age group was not significant, $F(1, 88) = 0.19$, $p = .664$, $\eta_g^2 = .001$, indicating that older adults ($M = 4.38$, $SE = 0.09$) and younger adults ($M = 4.32$, $SE = 0.09$) provided similar relevance ratings overall. The main effect of task type was also not significant, $F(1, 88) = 3.04$, $p = .085$, $\eta_g^2 = .015$, with similar relevance ratings for the assisted creativity task ($M = 4.44$, $SE = 0.08$) and the information request task ($M = 4.26$, $SE = 0.09$).

The interaction between age group and task type was significant, $F(1, 88) = 4.64$, $p = .034$, $\eta_g^2 = .022$. Follow-up comparisons showed that the effect of task type on relevance ratings differed by age group. Among younger adults, relevance ratings were significantly higher in the assisted creativity task ($M = 4.53$, $SE = 0.11$) than in the information request task ($M = 4.11$, $SE = 0.12$), $t(88) = 2.76$, $p = .007$. In contrast, older adults showed no significant difference in relevance ratings between the assisted creativity ($M = 4.36$, $SE = 0.11$) and information request ($M = 4.40$, $SE = 0.12$) tasks ($p = .77$).

These results indicate that relevance evaluations depended on the interaction between age and task demands. Younger adults were more sensitive to task type, with higher relevance ratings for assisted creativity. Older adults gave consistently high relevance ratings across tasks. As a result, older adults' relevance evaluations in the information request task were higher than those of younger adults.

Perceived Usefulness Usefulness was then examined by assessing participants' self-reported extent to which the model's responses were for their intended purpose. The main effect of age group was also not significant, $F(1, 88) = 1.10$, $p = .298$, $\eta_g^2 = .008$, as older adults reported usefulness ratings ($M = 4.27$, $SE = 0.11$) similar to those of younger adults ($M = 4.10$, $SE = 0.11$). The main effect of task type was not significant, $F(1, 88) = 0.08$, $p = .781$, $\eta_g^2 = .000$, indicating that usefulness ratings were highly similar for the assisted creativity task ($M = 4.20$, $SE = 0.10$) and the information request task ($M = 4.17$, $SE = 0.10$) by age groups. The interaction between age group and task type was not significant, $F(1, 88) = 1.04$, $p = .310$, $\eta_g^2 = .004$, indicating a similar pattern of task-related usefulness ratings across the two age groups. These results indicate that the usefulness evaluation of the model's responses was similar across age groups and task types, suggesting that both younger and older adults found the model's output equally useful regardless of the interaction context.

Perceived Clarity Clarity was next examined by assessing participants' self-reported extent to which the model's responses were clear and understandable. The analysis identified a significant main effect of age group, $F(1, 88) = 7.13$, $p = .009$, $\eta_g^2 = .050$, such that older adults assigned higher clarity ratings overall ($M = 4.60$, $SE = 0.09$) compared to younger adults ($M = 4.24$, $SE = 0.09$). A significant main effect of task type was found, $F(1, 88) = 18.15$, $p < .001$, $\eta_g^2 = .068$, with the assisted creativity task receiving higher clarity ratings ($M = 4.63$, $SE = 0.06$) than the information request task ($M = 4.21$, $SE = 0.10$) by both age groups. The interaction between age group and task type was not significant, $F(1, 88) = 2.46$, $p = .12$, $\eta_g^2 = .009$, indicating that task type did not affect clarity ratings across age groups. These results indicate that clarity evaluations were influenced by both age and task demands: older adults rated the model's responses as clearer than younger adults did. Across age groups, responses in the assisted creativity task were rated as clearer than those in the information request task, regardless of interaction context.

Perceived Sufficiency Sufficiency was subsequently examined by assessing participants' self-report of the adequacy of the model's responses, assuming no additional responses were provided. The main effect of age group was not significant, $F(1, 88) = 0.44$, $p = .508$, $\eta_g^2 = .003$, with similar overall sufficiency ratings for older adults ($M = 4.03$, $SE = 0.13$) and younger adults ($M = 3.91$, $SE = 0.13$). The main effect of task type was significant, $F(1, 88) = 5.82$, $p = .018$, $\eta_g^2 = .024$, indicating higher sufficiency ratings for the assisted creativity task ($M = 4.14$, $SE = 0.11$) than for the information request task ($M = 3.80$, $SE = 0.12$) by age groups. The interaction between age group and task type was not significant, $F(1, 88) = 0.49$, $p = .485$, $\eta_g^2 = .002$, indicating that the difference in sufficiency ratings between tasks was consistent across age groups. These results indicate that sufficiency evaluations were influenced by task demands rather than age, with responses perceived as more sufficient in the assisted creativity task than in the information request task across both age groups, regardless of the interaction context.

Across the four evaluation dimensions, age-related differences appeared selectively. Older adults rated the model's responses as clearer than those of younger adults. Differences in relevance evaluations were also observed: older adults maintained high relevance ratings across tasks, whereas younger adults' ratings varied by task type. No age-related differences were observed in usefulness or sufficiency, indicating similar perceptions among younger and older adults. However, sufficiency ratings demonstrated a main effect of task type, with responses in the assisted creativity task perceived as more sufficient than those in the information request task for both age groups. These findings indicate that task demands influenced perceptions of response completeness, regardless of age.

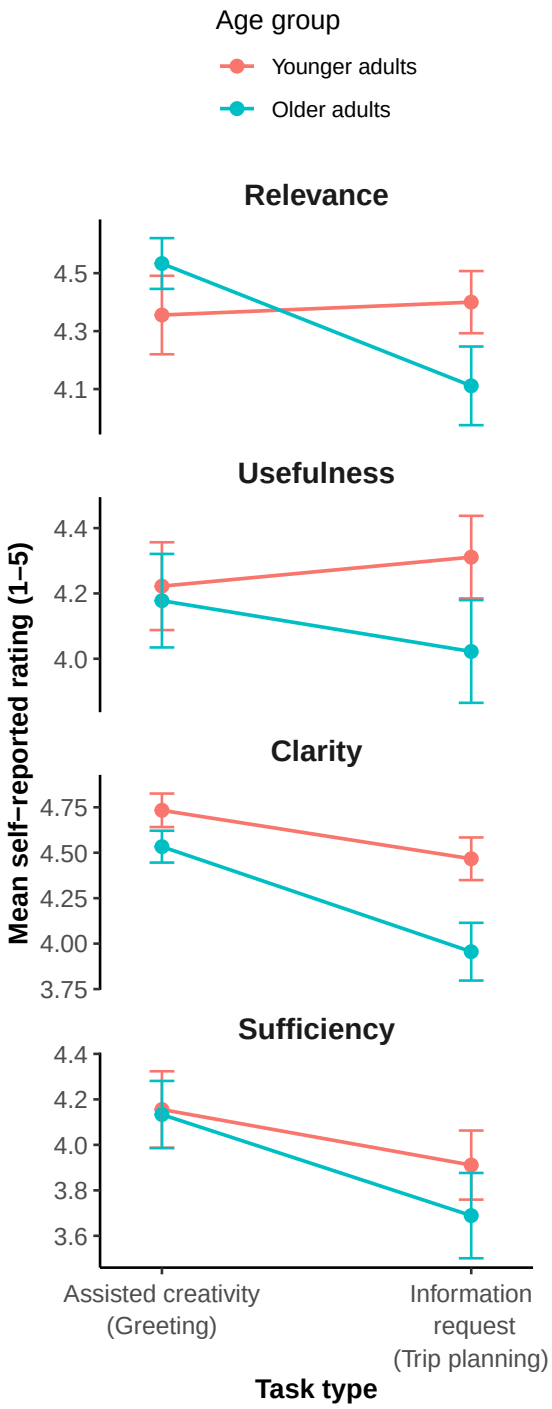

Figure 13: Age-related differences in self-reported evaluations of model responses across task types (Assisted creativity vs. Information request). Error bars represent standard errors.

F Agreement Between Self, Expert Human, and LLM Evaluations of Response

To assess agreement across evaluation sources, we computed Pearson correlations among three evaluation types: participant self-evaluations, external human raters, and LLM-based evaluations. Detailed analyses of each evaluation method separately (LLM judges and expert human raters) are reported in Appendix G. Here we focus on the degree of agreement between the different evaluation sources.

Correlations were calculated for each evaluation dimension shared across the evaluation methods. We first report overall correlations across all tasks to assess general agreement between evaluation

Comparison	r
Expert vs. Self	
Relevance	.13
Clarity	.08
Expert vs. LLM	
Relevance	**.31*** **
Clarity	**.27*** **
Accuracy	**.25** **
Self vs. LLM	
Relevance	.04
Clarity	**.21** **
Usefulness	.07
Sufficiency	.11

Table 6: Overall Pearson correlations across all tasks between expert human external evaluations, participant self-evaluations, and LLM-based evaluations. Rows indicate evaluation dimensions. Bold values indicate statistically significant correlations (** $p < .01$, *** $p < .001$).

Comparison	Metric	Greeting	Trip
Expert vs. Self			
	Relevance	**.27***	−.00
	Clarity	.09	.02
Expert vs. LLM			
	Relevance	**.50*** **	.20
	Clarity	**.31** **	**.22***
	Accuracy	**.31** **	.15
Self vs. LLM			
	Relevance	.18	−.09
	Clarity	.16	**.21***
	Usefulness	.13	.02
	Sufficiency	.11	−.04

Table 7: Pearson correlations between expert human external evaluation, participant self-evaluation, and LLM-based evaluation across tasks. Bold values indicate statistically significant correlations (* $p < .05$, ** $p < .01$, *** $p < .001$).

sources.

Next, we examined whether agreement patterns differed across age groups by computing correlations separately for younger and older participants. Finally, task-specific analyses were conducted within each age group to examine whether agreement patterns differed between the assisted creativity (greeting writing) and information request (trip planning) tasks.

Overall Agreement Across all tasks combined, moderate agreement was observed between human raters and LLM evaluations. Human and LLM ratings were significantly correlated for *Relevance* ($r = .31$, $p < .001$), *Clarity* ($r = .27$, $p < .001$), and *Accuracy* ($r = .25$, $p < .001$).

In contrast, correlations between human ratings and participants' self-evaluations were small and not statistically significant for *Relevance* ($r = .13$, $p = .090$) and *Clarity* ($r = .08$, $p = .292$). Similarly, self-evaluations showed little agreement with LLM ratings. The only significant association emerged for *Clarity* ($r = .21$, $p = .005$), whereas correlations for *Relevance*, *Usefulness*, and *Sufficiency* were not significant ($ps > .13$).

Overall, these results indicate stronger alignment between external human evaluations and LLM-based evaluations than between external evaluations and participants' self-assessments.

Task-Specific Agreement To examine whether agreement patterns differed across task contexts, correlations were computed separately for the assisted creativity (greeting writing) and information request (trip planning) tasks.

For the greeting task, moderate agreement was observed between human raters and LLM evaluations. Human and LLM ratings were significantly correlated for *Relevance* ($r = .50$, $p < .001$), *Clarity* ($r = .31$, $p = .003$), and *Accuracy* ($r = .31$, $p = .003$).

A smaller but significant correlation was also found between human ratings and participants' self-evaluations for *Relevance* ($r = .27$, $p = .011$), whereas correlations for *Clarity* were not significant. Self-evaluations did not significantly correlate with LLM evaluations across the measured dimensions ($ps > .08$).

In contrast, agreement among evaluation sources was generally weaker for the trip-planning task. Human and LLM ratings were significantly correlated only for *Clarity* ($r = .22$, $p = .038$), whereas correlations for *Relevance* and *Accuracy* were not significant.

No significant correlations were observed between human ratings and participants' self-evaluations. A small correlation emerged between self-evaluations and LLM ratings for *Clarity* ($r = .21$, $p = .048$), but no other associations reached significance.

Agreement by Age Group When examining correlations separately by age group (see Table 8), a similar overall pattern emerged across groups. Among older adults, human and LLM evaluations were moderately correlated for *Relevance* ($r = .37$, $p < .001$), *Clarity* ($r = .36$, $p < .001$), and *Accuracy* ($r = .26$, $p = .013$).

	Older Adults	Younger Adults
Expert vs. Self		
Relevance	**.23***	.03
Clarity	.05	.12
Expert vs. LLM		
Relevance	**.37*****	**.24***
Clarity	**.36*****	.15
Accuracy	**.26***	**.29****
Self vs. LLM		
Relevance	.04	.04
Clarity	**.40*****	.07
Usefulness	.07	.08
Sufficiency	.11	.12

Table 8: Pearson correlations between evaluation sources by age group across all tasks combined. Rows indicate evaluation dimensions (Relevance, Clarity, Accuracy, Usefulness, and Sufficiency). Bold values indicate statistically significant correlations (* $p < .05$, ** $p < .01$, *** $p < .001$).

Among younger adults, human and LLM evaluations were also significantly correlated for *Relevance* ($r = .24$, $p = .023$) and *Accuracy* ($r = .29$, $p = .006$), whereas the correlation for *Clarity* was not significant ($r = .15$, $p = .148$).

In contrast, correlations between human ratings and participants' self-evaluations were generally small and not statistically significant in either age group. The only significant association emerged among older adults for *Relevance* ($r = .23$, $p = .028$), whereas correlations for *Clarity* were not significant in either group.

Similarly, correlations between self-evaluations and LLM ratings were generally weak. A significant association was observed only among older adults for *Clarity* ($r = .40$, $p < .001$), whereas correlations for *Relevance*, *Usefulness*, and *Sufficiency* were not significant in either age group.

Overall, LLM evaluations showed moderate alignment with human judgments in both age groups. This alignment was somewhat more consistent among older adults, where significant correlations were observed across all three evaluation dimensions, whereas among younger adults, significant associations emerged only for relevance and accuracy. In contrast, participants' self-evaluations showed generally weak correspondence with external evaluations and LLM evaluations.

Agreement by Age Group and Task To examine whether agreement between evaluation sources varied across tasks and age groups, Pearson correlations were computed separately for each age group and task (see Table 4).

Among older adults, human and LLM evaluations showed consistent moderate agreement in the greeting task, with significant correlations for *Relevance* ($r = .47$, $p = .001$), *Clarity* ($r = .42$, $p = .004$), and *Accuracy* ($r = .37$, $p = .012$). In the trip planning task, a significant association was observed only for *Clarity* ($r = .31$, $p = .041$), whereas correlations for *Relevance* ($r = .29$, $p = .052$) and *Accuracy* ($r = .05$, $p = .735$) were not significant.

Among younger adults, a different pattern emerged. In the greeting task, human and LLM evaluations were significantly correlated for *Relevance* ($r = .54$, $p < .001$) and *Accuracy* ($r = .32$, $p = .030$), whereas the correlation for *Clarity* was not significant ($r = .26$, $p = .089$). In contrast, no significant correlations were observed between human and LLM evaluations in the trip-planning task.

Correlations between human ratings and participants' self-evaluations were generally small and non-significant across both tasks and age groups, with the only significant association emerging among older adults in the greeting task for *Relevance* ($r = .36$, $p = .014$).

Similarly, correlations between self-evaluations and LLM ratings were mostly weak. Among older adults in the greeting task, significant correlations were observed for *Relevance* ($r = .31$, $p = .037$) and *Clarity* ($r = .33$, $p = .028$), and in the trip planning task for *Clarity* ($r = .41$, $p = .005$). In contrast, no significant associations were observed between self-evaluations and LLM ratings among younger adults in either task.

Overall, correlations were stronger in the assisted creativity (greeting writing) task than in the information request (trip planning) task. In particular, correlations between external human raters and LLM evaluations were consistently significant in the greeting task for both age groups, whereas in the trip planning task, significant associations were limited. This pattern was somewhat more consistent among older adults, where significant correlations between human and LLM evaluations were observed across more evaluation dimensions. Correlations between participants' self-evaluations and LLM ratings were generally weak and were observed only among older adults, whereas no significant associations emerged for younger adults.

While the correlation analyses examine the degree of agreement between evaluation sources, they

do not indicate whether different evaluators systematically assign higher or lower ratings. To examine differences in rating levels across evaluators, we conducted a series of mixed ANOVAs.

Differences Between Self, Expert Human, and LLM Evaluations of Response Relevance A mixed ANOVA examined relevance ratings as a function of task (greeting vs. trip planning), age group (younger vs. older adults), and rater type (human, self, and LLM). The analysis revealed a small but significant main effect of task, $F(1, 88) = 4.05$, $p = .047$, $\eta_g^2 = .007$, indicating slightly different relevance ratings across tasks. A strong main effect of rater type was also observed, $F(1.73, 152.54) = 51.94$, $p < .001$, $\eta_g^2 = .152$, reflecting substantial differences across evaluators. The main effect of age group was not significant, $F(1, 88) < .001$, $p > .999$.

Importantly, these effects were qualified by a significant three-way interaction between task, age group, and rater type, $F(1.70, 149.57) = 5.69$, $p = .006$, $\eta_g^2 = .017$. Follow-up comparisons therefore examined each task separately within each age group.

Among older adults, in the greeting task, relevance ratings increased from human evaluations ($M = 3.97$) to self-evaluations ($M = 4.36$) and further to LLM evaluations ($M = 4.73$), with all pairwise differences significant ($ps \leq .01$). In the trip-planning task, both self ($M = 4.40$) and LLM ($M = 4.42$) evaluations were significantly higher than human evaluations ($M = 3.89$), whereas self and LLM ratings did not differ.

Among younger adults, in the greeting task, both self-evaluations ($M = 4.53$) and LLM evaluations ($M = 4.72$) were significantly higher than human evaluations ($M = 3.80$), while self and LLM ratings did not differ. In the trip-planning task, LLM evaluations ($M = 4.54$) were significantly higher than both human ($M = 4.06$) and self ($M = 4.11$) evaluations, whereas human and self ratings did not differ.

Overall, LLM evaluations tended to yield the highest relevance ratings across conditions, whereas the relationship between human and self-evaluations varied depending on task type and age group. In addition, responses in the greeting task received higher relevance ratings overall than those in the trip-planning task.

Differences Between Self, Human, and LLM Evaluations of Response Clarity A mixed ANOVA examined clarity ratings as a function of task, age group, and rater type (human, self, LLM). The analysis revealed a significant main effect of task, $F(1, 88) = 22.76$, $p < .001$, $\eta_g^2 = .032$, and a strong main effect of rater type, $F(1.48, 130.13) = 101.65$, $p < .001$, $\eta_g^2 = .261$. The main effect of age group was not significant, $F(1, 88) = 1.85$, $p = .178$.

These effects were qualified by interactions between rater type and age group, $F(1.48, 130.13) = 8.29$, $p = .001$, $\eta_g^2 = .028$, and between rater type and task, $F(1.51, 133.16) = 6.14$, $p = .006$, $\eta_g^2 = .019$. The three-way interaction between task, age group, and rater type was not significant, $F(1.51, 133.16) = 2.90$, $p = .073$.

Across conditions, LLM evaluations yielded the highest clarity ratings ($M = 4.87$), followed by self-evaluations ($M = 4.42$) and human evaluations ($M = 4.06$), with all pairwise differences significant ($ps < .001$).

Follow-up comparisons showed that in the greeting task, clarity ratings increased from human evaluations ($M = 4.13$) to self-evaluations ($M = 4.63$) and further to LLM evaluations ($M = 4.90$), with all pairwise differences significant ($ps < .001$). In the trip-planning task, LLM ratings ($M = 4.85$) remained significantly higher than both human ($M = 3.99$) and self-evaluations ($M = 4.21$), whereas human and self-evaluations did not differ from each other.

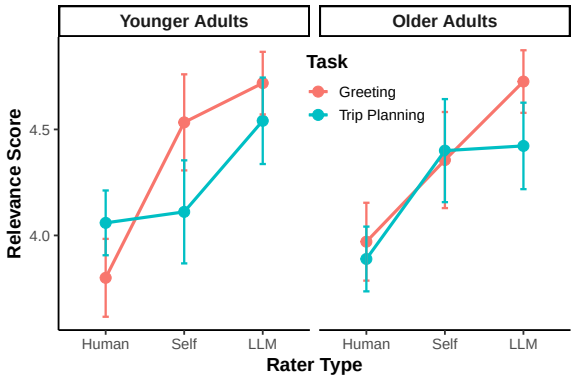

Figure 14: Mean relevance ratings as a function of rater type (human, self, and LLM), task (greeting vs. trip planning), and age group. Error bars represent 95% confidence intervals. Across both age groups, LLM evaluations tended to assign the highest relevance ratings, whereas human evaluators assigned the lowest. The pattern of differences between human and self-evaluations varied across tasks and age groups, consistent with the significant three-way interaction reported in the ANOVA.

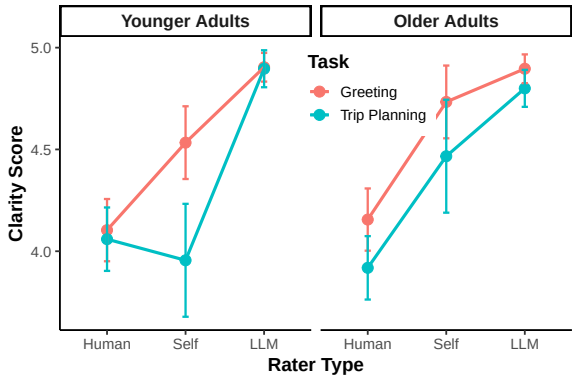

Figure 15: Mean clarity ratings as a function of rater type (human, self, and LLM), task (greeting vs. trip planning), and age group. Error bars represent 95% confidence intervals. Across both age groups, LLM evaluations consistently assigned the highest clarity ratings, followed by self-evaluations and human evaluations. Clarity ratings were also higher in the greeting task than in the trip-planning task. Differences between human and self-evaluations varied across age groups, with older adults rating their responses as clearer than human evaluators, whereas among younger adults self- and human evaluations were more similar, consistent with the significant interactions observed in the ANOVA.

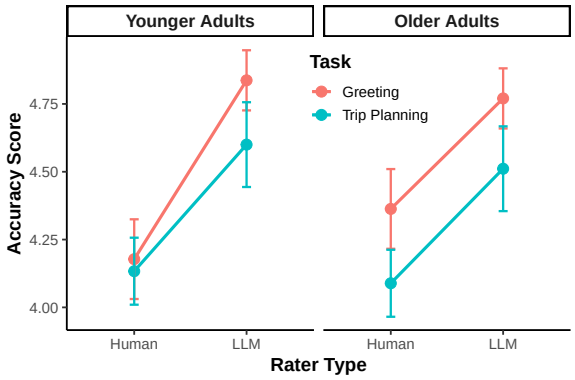

Figure 16: Mean accuracy ratings as a function of rater type (human vs. LLM), task (greeting vs. trip planning), and age group. Error bars represent 95% confidence intervals. Across both age groups, LLM evaluations assigned higher accuracy ratings than human evaluations. In addition, responses in the greeting task were rated as more accurate than those in the trip-planning task, consistent with the significant main effects reported in the ANOVA.

The interaction with age group indicated that among older adults, self-evaluations were significantly higher than human evaluations ($M = 4.60$ vs. $M = 4.04$), whereas among younger adults, self and human evaluations did not differ significantly. In both age groups, LLM ratings were higher than both human and self-evaluations.

Overall, LLM evaluations consistently produced the highest clarity ratings, while differences between self- and human evaluations varied across age groups and tasks.

Differences Between Human and LLM Evaluations of Response Accuracy A mixed ANOVA examined accuracy ratings as a function of task (greeting vs. trip planning), age group (younger vs. older adults), and rater type (human vs. LLM). The analysis revealed a significant main effect of task, $F(1, 88) = 18.02$, $p < .001$, $\eta_g^2 = .048$, indicating that responses in the greeting task ($M = 4.54$) were rated as more accurate than those in the trip-planning task ($M = 4.33$).

There was also a strong main effect of rater type, $F(1, 88) = 112.12$, $p < .001$, $\eta_g^2 = .226$, such that LLM evaluations ($M = 4.68$) yielded higher accuracy ratings than human evaluations ($M = 4.19$).

No other effects were significant, including the main effect of age group ($p = .950$) and all interaction terms (all $ps \geq .112$), indicating that these differences were consistent across both age groups and tasks.

Overall, the results indicate that LLM evaluators systematically assigned higher accuracy ratings than human evaluators, while responses in the greeting task were perceived as slightly more accurate than those in the trip-planning task.

G Expert Human and LLM Evaluations of Model Responses

G.1 LLM-Judged Evaluation of Model Responses

Differences in AI performance across tasks and age groups were evaluated using an LLM-as-a-Judge framework, in which three independent language models rated each interaction across five evaluation dimensions: Relevance, Usefulness, Clarity, Sufficiency, and Accuracy. Ratings were averaged across models to obtain a single score for each participant, task, and dimension. Full methodological details are provided in Appendix B.3.

Inter-rater Reliability of LLM Judges Inter-rater reliability among the three LLM judges (Gemini, Claude, and Llama) was assessed using Krippendorff's α, a reliability coefficient suitable for multiple raters and incomplete rating designs. Reliability was computed separately for each evaluation dimension.

Measure	Overall	Greeting	Trip
Relevance	0.306	0.401	0.218
Usefulness	0.341	0.474	0.192
Clarity	0.136	0.072	0.174
Sufficiency	0.483	0.377	0.166
Accuracy	0.123	0.202	0.026

Table 9: Inter-rater reliability (Krippendorff's α) across the three LLM judges, reported overall and separately for the greeting-writing (assisted creativity) and trip-planning (information request) tasks.

Overall, agreement among the LLM judges was low to moderate across the evaluation dimensions, with α values ranging from .12 to .48. The highest agreement was observed for *Sufficiency* ($\alpha = .483$), whereas lower agreement was observed for *Clarity* ($\alpha = .136$) and *Accuracy* ($\alpha = .123$). Table 9 presents the full set of reliability estimates for each evaluation dimension.

To examine whether agreement among the LLM judges varied across task contexts, reliability was also computed separately for each task (assisted creativity [greeting writing] vs. information request [trip planning]). As shown in Table 9, agreement varied across tasks and evaluation dimensions. For example, agreement was relatively higher for the assisted creativity task for *Usefulness* ($\alpha = .474$), whereas lower agreement was observed for several dimensions in the information request task, including *Accuracy* ($\alpha = .026$).

Low inter-rater agreement among LLM evaluators has also been reported in prior studies of LLM-based evaluation (Chiang and Lee, 2023; Nasser, 2026), suggesting that the relatively modest agreement observed in our results is consistent with earlier findings. The variability in model judgments underscores the need to interpret LLM-based evaluations with caution, particularly for dimensions such as factual accuracy.

Effects of Task Type and Age Group To determine whether these evaluations varied by task type and age group, separate 2×2 mixed-design ANOVAs were conducted for each evaluation dimension, with age group (younger adults vs. older adults) as the between-subjects factor and task type (assisted creativity [greeting writing] vs. information request [trip planning]) as the within-subjects factor.

LLM-judged Relevance Score Relevance reflects the extent to which the AI's responses addressed users' requests. The main effect of age group was not significant, $F(1, 88) = 0.31$, $p = .578$, $\eta_g^2 = .002$, indicating that conversations initiated by older adults ($M = 4.38$, $SE = 0.09$) and younger adults ($M = 4.32$, $SE = 0.09$) were judged as similarly relevant overall. A significant main effect of task type was found, $F(1, 88) = 9.41$, $p = .003$, $\eta_g^2 = .039$, indicating that AI responses were judged as more relevant in the assisted creativity task ($M = 4.72$, $SE = 0.05$) than in the information request task ($M = 4.48$, $SE = 0.07$) (see Figure 17). The interaction between age group and task type was not significant, $F(1, 88) = 0.64$, $p = .424$, $\eta_g^2 = .003$, indicating that the difference in Relevance ratings between tasks was similar across age groups. These results indicate that LLM judges evaluated the AI's responses as more relevant to task demands than to age, with higher relevance ratings for the assisted creativity task across both age groups, regardless of the interaction context.

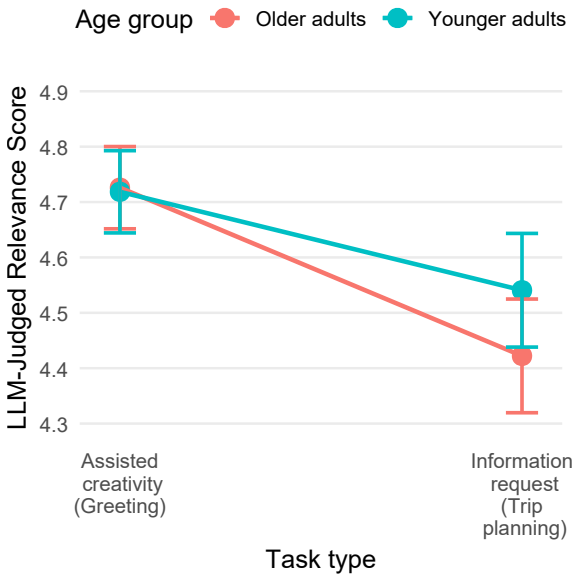

Figure 17: LLM-judged relevance scores by task type and age group. Points represent estimated marginal means; error bars indicate ±1 SE.

LLM-judged Usefulness Score Usefulness reflects the extent to which the AI's responses were perceived as useful for the user's purpose. The main effect of age group was not significant, $F(1, 88) = 0.31$, $p = .581$, $\eta_g^2 = .002$, indicating that conversations initiated by older adults ($M = 4.50$, $SE = 0.08$) and younger adults ($M = 4.56$, $SE = 0.08$) were judged as simi-

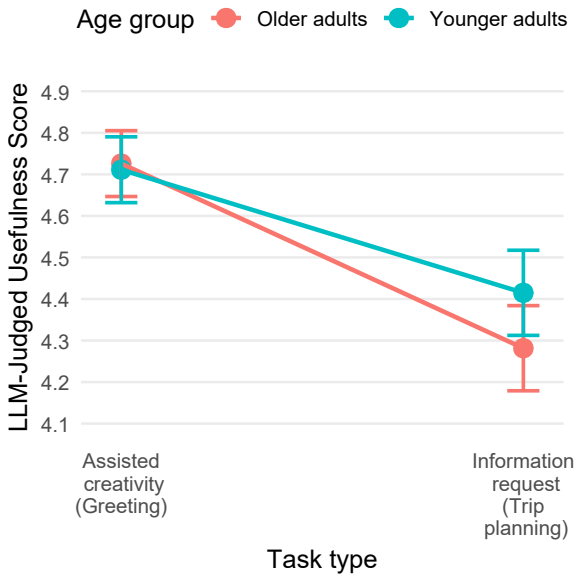

Figure 18: LLM-judged usefulness scores by task type and age group. Points represent estimated marginal means; error bars indicate ±1 SE.

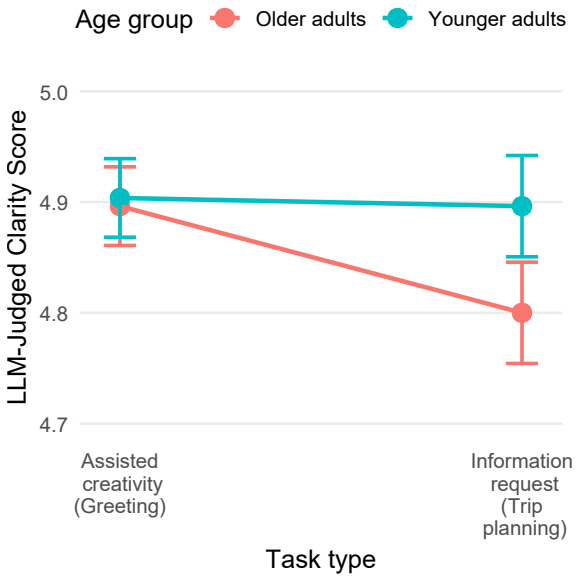

Figure 19: LLM-judged clarity scores by task type and age group. Points represent estimated marginal means; error bars indicate ±1 SE

larly useful overall. A significant main effect of task type was found, $F(1, 88) = 25.84$, $p < .001$, $\eta_g^2 = .085$, indicating that AI responses were judged as more useful in the assisted creativity task ($M = 4.72$, $SE = 0.06$) than in the information request task ($M = 4.35$, $SE = 0.07$) (see Figure 18). The interaction between age group and task type was not significant, $F(1, 88) = 1.03$, $p = .312$, $\eta_g^2 = .004$, indicating that the difference in usefulness ratings between tasks was similar by age groups. These results indicate that LLM judges evaluated the AI's responses as more useful as a function of task demands rather than age, with higher usefulness ratings for the assisted creativity task than in the information request task across both age groups, regardless of the interaction context.

LLM-judged Clarity Score Clarity reflects the extent to which the AI's responses were clear and understandable. The main effect of age group was not significant, $F(1, 88) = 1.49$, $p = .225$, $\eta_g^2 = .009$, indicating that conversations initiated by older adults ($M = 4.85$, $SE = 0.03$) and younger adults ($M = 4.90$, $SE = 0.03$) were judged as equally clear overall. The main effect of task type was also not significant, $F(1, 88) = 1.74$, $p = .191$, $\eta_g^2 = .009$, indicating that conversations initiated by the creativity task ($M = 4.72$, $SE = 0.06$) and information request task ($M = 4.35$, $SE = 0.07$) were judged as equally clear overall. The interaction between age group and task type was not significant, $F(1, 88) = 1.28$, $p = .261$, $\eta_g^2 = .007$, indicating that the difference in clarity ratings between tasks was similar by age groups. These findings indicate that LLM judges perceived the AI's responses as highly clear across both tasks and age groups, with no evidence for systematic differences attributable to task type or user age.

LLM-judged Sufficiency Score Sufficiency reflects whether the AI's responses provided enough information for users to complete the task without seeking additional sources. The main effect of age group was not significant, $F(1, 88) = 1.23$, $p = .271$, $\eta_g^2 = .009$, indicating that conversations initiated by older adults ($M = 4.13$, $SE = 0.08$) and younger adults ($M = 4.25$, $SE = 0.08$) were judged as similarly sufficient overall. A significant main effect of task type was found, $F(1, 88) = 144.72$, $p < .001$, $\eta_g^2 = .377$, indicating that AI responses were judged as more sufficient in the assisted creativity task ($M = 4.70$, $SE = 0.06$) than in the information request task ($M = 3.69$, $SE = 0.08$) (see Figure 20). The interaction between age group and task type was not significant, $F(1, 88) = 1.63$, $p = .205$, $\eta_g^2 = .007$, indicating that the difference in sufficiency ratings between tasks was similar by age groups. These results indicate that LLM judges evaluated the AI's responses as more sufficient as a function of task

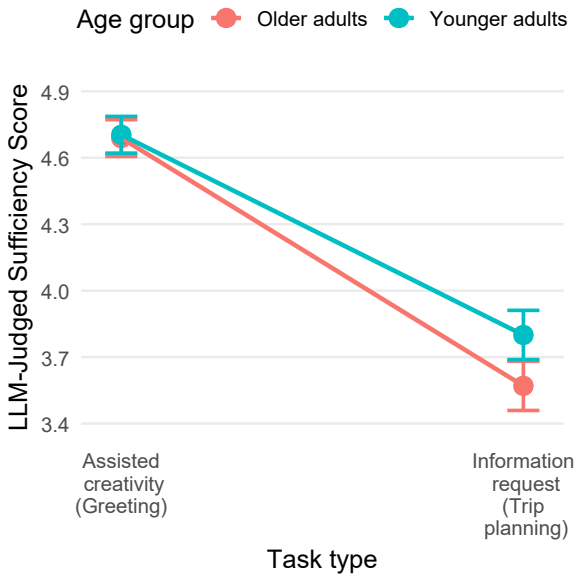

Figure 20: LLM-judged sufficiency scores by task type and age group. Points represent estimated marginal means; error bars indicate ±1 SE.

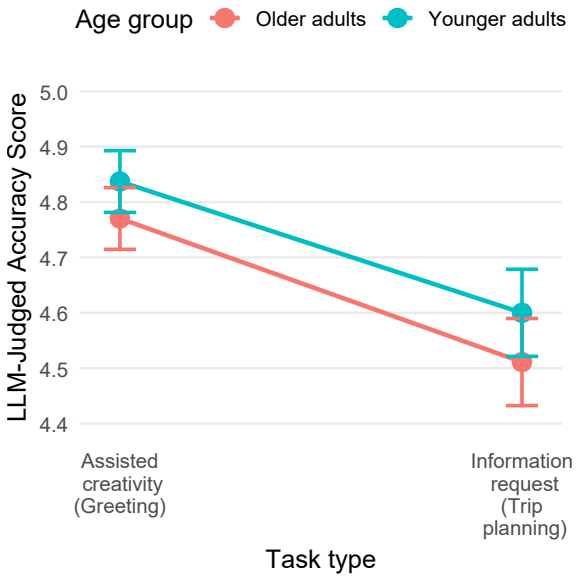

Figure 21: LLM-judged accuracy scores by task type and age group. Points represent estimated marginal means; error bars indicate ±1 SE.

demands rather than age, with higher sufficient ratings for the assisted creativity task than in the information request task across both age groups, regardless of the interaction context.

LLM-judged Accuracy Score Accuracy reflects the extent to which the AI's responses were judged to be factually correct and free from errors or misleading information. The main effect of age group was not significant, $F(1, 88) = 0.99$, $p = .322$, $\eta_g^2 = .007$, indicating that conversations initiated by older adults ($M = 4.72$, $SE = 0.06$) and younger adults ($M = 4.64$, $SE = 0.06$) were judged as similarly accurate overall. A significant main effect of task type was found, $F(1, 88) = 19.33$, $p < .001$, $\eta_g^2 = .070$, indicating that AI responses were judged as more accurate in the assisted creativity task ($M = 4.80$, $SE = 0.04$) than in the information request task ($M = 4.56$, $SE = 0.06$) (see Figure 21). The interaction between age group and task type was not significant, $F(1, 88) = 0.04$, $p = .844$, $\eta_g^2 < .001$, indicating that the difference in accuracy ratings between tasks was similar by age groups. These results indicate that LLM judges evaluated the AI's responses as more accurate as a function of task demands rather than age, with higher accurate ratings for the assisted creativity task than in the information request task across both age groups, regardless of the interaction context.

Measure	Overall	Greeting	Trip
Relevance	0.139	0.154	0.127
Clarity	0.051	0.100	0.004
Accuracy	0.018	0.044	-0.042

Table 10: Inter-rater reliability (Krippendorff's α) across human evaluators, reported overall and separately for the greeting-writing and trip-planning tasks.

G.2 Expert Human Evaluation of Model Responses

Inter-rater Reliability Inter-rater reliability was assessed using Krippendorff's α, a reliability coefficient suitable for multiple raters and incomplete rating designs. Across all ratings, we obtained $\alpha = .14$ for inter-rater agreement across 540 ratings (180 items × 3 raters drawn from a pool of 27 annotators).

As shown in Table 10, agreement among human evaluators was low across all three evaluation dimensions. Across all tasks combined, the highest agreement was observed for *Relevance* ($\alpha = .139$), whereas agreement was lower for *Clarity* ($\alpha = .051$) and *Accuracy* ($\alpha = .018$).

Agreement also varied across tasks. For the greeting-writing task, agreement was somewhat higher across all three dimensions, including *Relevance* ($\alpha = .154$) and *Clarity* ($\alpha = .100$). In contrast, agreement in the trip-planning task was lower, particularly for *Accuracy* ($\alpha = -.042$), in-

dicating substantial variability across raters.

The relatively low agreement may partly reflect the partially-crossed annotation design, in which not all raters evaluated all items, potentially increasing rating variance. Similar levels of low agreement have been reported in prior work on human and evaluation of AI-generated text (Chiang and Lee, 2023; Thakur et al., 2025).

Effects of Task Type and Age Group We report the results of the human external evaluation, in which independent raters assessed the model responses on three dimensions: *Relevance*, *Clarity*, and *Accuracy*. For each dimension, a mixed ANOVA was conducted with task type (Assisted Creativity – greeting vs. Information Request – trip planning) as a within-subject factor and age group (younger vs. older adults) as a between-subject factor. Figure 22 illustrates the mean ratings across conditions.

Human-judged Relevance Score For relevance ratings, there was no significant main effect of task, $F(1, 88) = 1.12$, $p = .292$, $\eta_g^2 = .006$, and no significant main effect of age group, $F(1, 88) < .001$, $p = 1.00$, $\eta_g^2 < .001$. However, the interaction between task and age group was significant, $F(1, 88) = 4.12$, $p = .045$, $\eta_g^2 = .022$. Follow-up comparisons showed that among older adults, relevance ratings did not differ between the greeting task ($M = 3.97$) and the trip-planning task ($M = 3.89$), $t(88) = 0.69$, $p = .49$. In contrast, among younger adults, responses in the trip-planning task ($M = 4.06$) were rated as more relevant than responses in the greeting task ($M = 3.80$), $t(88) = 2.18$, $p = .032$. This pattern suggests that task type influenced relevance ratings for younger adults but not for older adults.

Human-judged Clarity Score For clarity ratings, no significant main effect of task was observed, $F(1, 88) = 3.12$, $p = .081$, $\eta_g^2 = .018$. There was also no significant main effect of age group, $F(1, 88) = 0.35$, $p = .558$, $\eta_g^2 = .002$, and the interaction between task and age group was not significant, $F(1, 88) = 1.46$, $p = .230$, $\eta_g^2 = .009$. Overall, clarity ratings were similar across both task types and age groups.

Human-judged Accuracy Score For accuracy ratings, a significant main effect of task emerged, $F(1, 88) = 5.84$, $p = .018$, $\eta_g^2 = .030$. Responses generated in the assisted creativity (greeting) task were rated as more accurate ($M = 4.27$) than responses in the information request (trip-planning) task ($M = 4.11$). There was no significant main effect of age group, $F(1, 88) = 0.99$, $p = .322$, $\eta_g^2 = .006$, and the interaction between task and age group was not significant, $F(1, 88) = 3.03$, $p = .085$, $\eta_g^2 = .016$.

Taken together, the human external evaluation revealed limited differences across tasks and age groups. Accuracy ratings were higher for responses in the assisted creativity (*greeting*) task than in the information request (*trip planning*) task, while *Clarity* did not differ across tasks. For *Relevance*, a significant interaction between task type and age group was observed: younger adults received higher relevance ratings in the trip-planning task than in the greeting task, whereas no task differences were observed among older adults. No overall differences were found between younger and older adults across the three evaluation dimensions.

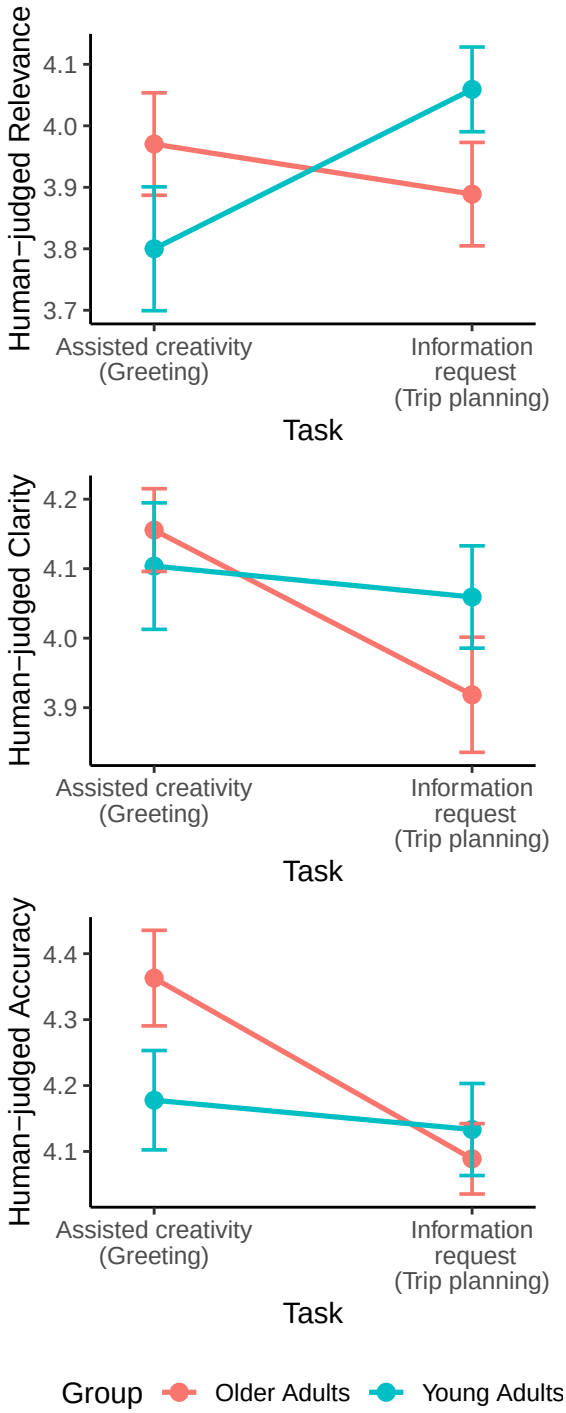

Figure 22: Human external evaluation scores across tasks and age groups. Panels show ratings for *Relevance* (top), *Clarity* (middle), and *Accuracy* (bottom). Younger adults (blue) and older adults (red) are shown separately for the assisted creativity (greeting writing) and information request (trip planning) tasks. Error bars represent standard errors.