

Seeking Information with RAG-Assistants: Does Model Size Matter in Human-AI Collaborations?

Lennard C. FROMA ^{a,1}, Tom KOUWENHOVEN ^a, Maaïke H.T. DE BOER ^b,
Catholijn M. JONKER ^{a,c}, and Max J. VAN DUIJN ^a

^a *Leiden University - The Netherlands*

^b *TNO - The Netherlands*

^c *TU Delft - The Netherlands*

ORCID ID: Lennard C. Froma <https://orcid.org/0009-0006-2061-9208>, Tom
Kouwenhoven <https://orcid.org/0000-0003-2480-4073>, Maaïke H.T. de Boer
<https://orcid.org/0000-0002-2775-8351>, Catholijn M. Jonker
<https://orcid.org/0000-0003-4780-7461>, Max J. van Duijn
<https://orcid.org/0000-0003-0798-9598>

Abstract. Much research on LLMs has focused on increasing benchmark performance. However, the evaluation of such models in real-world collaborative human-AI workflows has stayed behind. This work evaluates a chatbot-style assistant based on Retrieval-Augmented Generation (RAG) in a realistic multi-turn information-seeking scenario inspired by workplace settings where compliance with local legislation and secure handling of sensitive data are often key. Specifically, we examine the performance of humans (N=112) assisted by RAG-assistants compared to LLM-only or LLM+RAG baselines. In this setting, we investigate how underlying model size (3B, 8B, and 70B) shapes the human-AI collaborative dynamic and how it influences perceived usability and satisfaction. Results show that the performance gain of human-AI collaboration over the model-only baselines is significant, irrespective of model size, suggesting that hybrid systems are beneficial in information-seeking scenarios. Interestingly, however, perceived usability and satisfaction among participants showed little difference across model sizes. This demonstrates a nuanced trade-off between model size, performance, and user perception. Our work highlights the added value of evaluating AI applications in actual multi-turn interactions with human users, looking at usability and satisfaction besides accuracy, rather than focusing on benchmark performance only.

Keywords. Human-AI Interaction, User Experiment, Multi-Turn Interaction, Large Language Model, Retrieval Augmented Generation

1. Introduction

Large Language Models (LLMs) are increasingly used in professional areas where information processing is essential [1]. In this paper, we focus on context-aware question

¹Corresponding Author: Lennard Froma, l.c.froma@liacs.leidenuniv.nl

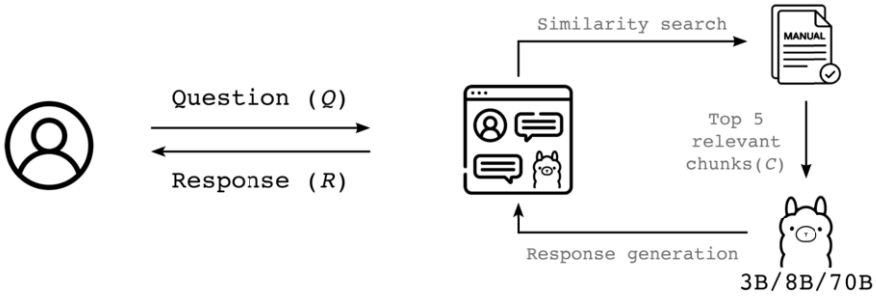


Figure 1. An overview of the RAG pipeline.

answering (QA) and information seeking in realistic workplace scenarios. Professionals are often tasked with analysing large volumes of documentation. While LLMs hold potential to offer support in this process, their deployment is constrained in cases where data is sensitive and compliance with, for example, the EU AI Act is required. To address such constraints, open, transparent, privacy-respecting, and locally deployable language models are needed [2]. This poses the question of how small such models can be in human-AI collaboration, while maintaining adequate performance and usability.

Evaluation of LLM applications, including in QA, often focuses on established benchmarks, relying mainly on quantitative metrics such as accuracy, precision, recall, or similarity scores to assess performance [3, 4, 5]. For many of these problems, benchmark evaluation is well-suited as responses typically fall into discrete categories [6]. However, benchmark approaches often favour larger commercial models instead of more versatile, locally deployable ones [7, 8].

Qualitative assessments, while less common, often highlight dimensions such as fluency, reliability, and perceived single-turn quality, where interactions consist of one question as input, followed by the assistant’s response [9]. However, professional users typically engage in collaborative task solving spanning multiple turns between model output and human input [10]. This means that interpretive nuance and interactive dynamics over multiple turns become significantly more important, and that usability and user perception become key factors influencing the outcome. The disconnect between benchmark evaluations and real-world LLM applications offers an incomplete picture.

Here, we address this disconnect by combining quantitative and qualitative assessments in an actual hybrid setting. We set up an experiment in which human participants (N=112) had to answer a set of detailed questions about a large document (400+ pages), resembling a workplace information-seeking task. Participants had direct access to a PDF of the document and a chatbot-style assistant that could be questioned about the contents. This assistant, henceforth called the RAG-assistant, used Retrieval Augmented Generation (RAG) in combination with open-weight generator LLMs of different sizes, as shown in Figure 1. This technique resembles a realistic workplace scenario where information is kept up-to-date through external databases. Participants could choose to answer questions either by retrieving information directly from the document, via the RAG-assistant, or via a combination of both methods. Financial rewards (a bonus on top of a fixed remuneration) were given for correct answers to the questions, motivating participants to complete the task accurately. This setup allows us to examine three main questions. Firstly, we gauge the added value of human-AI interaction, and ask whether the

human-AI setting outperforms the LLM-only and LLM+RAG (baseline) settings (**RQ1**). Secondly, within the human-AI setting, we ask whether model size yields higher task accuracy (**RQ2**) and whether the perceived usability and satisfaction increase with model size **RQ3**. Our contributions include:

1. a systematic evaluation/comparison of RAG-assistants based on small, open-weight LLMs of three different sizes;
2. a realistic scenario simulating real-world contexts;
3. an experiment in which participants complete an information-seeking task over multiple turns, measuring accuracy, usability, and satisfaction.

2. Related work

2.1. RAG

LLMs are limited to the knowledge encoded during training. This means they cannot recall information absent from their training data or fine-tuning procedures, which poses a significant obstacle to the application of LLMs when domain-specific knowledge is required [11]. To address this constraint, models have been equipped with the ability to consult external sources during inference, as demonstrated by WebBrain [12] and WebGPT [13]. This functionality is now present in many state-of-the-art commercial systems. However, working with such systems can be undesirable for a variety of reasons, including when working with sensitive data that has to be processed within local environments. In such cases, RAG [14] can be used to augment the knowledge encoded in the LLM with external, offline knowledge bases.

At their core, RAG-assistants retrieve relevant document excerpts during inference and pass them to an LLM, which uses the provided information to formulate an answer, producing more up-to-date and/or task-specific answers [14, 15]. We adopt a sequential retrieval-augmented generation approach [9], in which relevant documents are first retrieved and then used by the generator (an LLM) to produce a response, which we found to be well-suited to our use case.

An approach to leveraging LLM capabilities in RAG is to have the model reformulate the original query to better align with the task's requirements [16, 17]. Instead of rewriting the query, LLMs can also be used to provide feedback on the returned documents in terms of match with the posed question, and retrieve new documents if there is no match, making it self-reflective [3]. A related concept is interactive information retrieval [18], in which rewriting is part of a broader set of techniques known as query reformulation [19]. Query reformulation is an important part of conversational/interactive RAG as it condenses the evolving and complex dialogue context into a more precise query. In this work, we do not apply this technique by adding another LLM inference call, but rather let humans reformulate the queries.

Recent work has explored the viability of integrating RAG pipelines with smaller LLMs [2]. Smaller LLMs combined with RAG can meet many AI application needs more efficiently and economically [20]. MiniRAG [21] introduces a lightweight, graph-based retrieval system tailored for smaller LLMs that compensates for the limited reasoning capacity of small generators while reducing the overall memory footprint by up to 75%. Similarly, [22, 23] show that smaller models often benefit disproportionately from high-

quality retrieval, partially closing the performance gap to larger LLMs at the application level. Furthermore, [24] found that larger models can sometimes be less favourable to work with in RAG applications. Similarly, under some conditions, smaller models show better general alignment than their larger counterparts [25].

Besides these advantages, smaller language models in RAG applications face notable challenges. Their smaller parameter capacity limits their ability to integrate complex or conflicting evidence retrieved from context, which can lead to hallucinations or superficial responses [26]. Moreover, they often struggle in zero-shot generalisation and require more carefully curated prompts to perform competitively [27]. Even under favourable conditions, models remain dependent on the placement of relevant information within the provided context [28], especially for smaller models. However, [29] shows that under the right circumstances (in terms of, e.g., architecture, retriever size, data structure), smaller models can outperform bigger ones. This shows that smaller language models augmented with RAG hold promise for applications where local deployment is desirable.

Evaluating RAG RAG applications are generally evaluated using benchmarks that assign quantitative scores, such as ROUGE [30], BLEU [31], accuracy, precision, recall, and error rates to compare performance. Examples of evaluation benchmarks are CRAG, RAGBench, TrustLLM, and RAGAs. CRAG [7] is a large-scale, multi-domain factual QA benchmark designed to evaluate and stress-test RAG applications across diverse question types. RAGBench [32] is a benchmark designed to evaluate RAG applications across five industry-specific domains and to address the lack of standardised, actionable evaluation methods for real-world RAG applications. RAGAs [33], an evaluation framework for RAG pipelines that measures retrieval quality, faithfulness, and answer relevance without human annotations, enabling faster and more scalable assessment of RAG applications. TrustLLM [34] is a benchmark and analysis framework for evaluating the trustworthiness of LLMs across multiple dimensions, including truthfulness, safety, and fairness, revealing notable trade-offs between openness, reliability, and utility.

Evaluation, including a human perspective in RAG evaluation, has been studied. [35] involved human evaluation through a 5-point scale (0, 0.25, 0.5, 0.75, 1), rating the quality of the answer. Although this includes human feedback, we believe it does not evaluate the complex dynamics that human-AI collaboration entails. Others introduced an evaluation framework for assessing RAG outputs from a human perspective [36]. However, the framework does not involve direct human-AI interaction; instead, interactions are observed and evaluated. We add a dimension to the evaluation by assessing not only objective performance but also collecting extensive feedback from participants.

2.2. Human-AI collaboration

Experiments have shown the added value of human-AI collaboration as it can outperform humans working in isolation [37]. A meta-analysis, however, showed that this is context-dependent [38]. For decision-making tasks, human-AI collaboration tends to decrease performance, whereas for more open-ended tasks, human-AI collaboration increases performance. Earlier work has studied collaborative LLM applications across fields, from which a taxonomy of human-LLM interaction was created [39]. Examples of such systems are recommender systems [40] and strategic planning [41]. Others focus on specific domains, such as medical [42], legal [43, 44], mathematics [45] or coding [46].

In such systems, trust and perceived usefulness are key indicators for collaborative success [47, 48]. The extent to which people trust LLMs is strongly influenced by perceived expertise [49]. LLM usage is also heavily task and person-dependent, as variations can lead to different strategies. [50]. In our setup, as part of Contribution 2 formulated above, we hypothesize that (part of the) shortcomings of using smaller language models will be mitigated through interaction with humans, in line with the general advantages of hybrid approaches [10].

Evaluating Human-AI collaboration Measuring user satisfaction combines elements like trust, perceived usefulness, and expertise as they correlate heavily [51], and is a research field with a long tradition [52]. Recent work focuses on satisfaction estimation in conversations. SPUR [53] is successful in using LLMs to estimate user satisfaction in conversational systems. This estimation method assigns binary satisfaction values to prompts based on patterns learned from labelled data, crucially missing the point that user satisfaction is often nuanced and context-dependent. Similar approaches estimate satisfaction using sentiment analysis in conversations between human and AI applications [54]. [55] took a broader empirical approach by putting out a questionnaire with general user satisfaction for all-day LLM use.

User satisfaction in a hybrid RAG context, however, is less well studied. Prior work in conversational RAG (LLM-LLM conversations) explored models that are too resource-intensive for local deployment [56]; besides that, human judgment is only used for annotation, not for interaction. Our work addresses this gap by specifically targeting QA/hybrid task-solving and user satisfaction for differently sized, locally deployable models.

3. Methodology

This study investigates the effect of model size on task performance, perceived usability, and perceived satisfaction. We compare three conditions across two settings. The conditions refer to the model sizes of three instruction-tuned, smaller-sized LLMs: 3B (Llama3.2), 8B (Llama3.1), and 70B (Llama3.3) from the Llama3 family [57]. The settings are LLM+RAG and human-AI. In the LLM+RAG setting, the original question as posed to the participant is used as a single-shot input for the RAG-assistant. We consider this setting to be the baseline, consistent with how benchmark performance is established across the literature, as discussed in Section 2.1. In the human-AI setting, participants could interact with the RAG-assistant to come up with an answer. LLM-only (i.e., querying LLMs from all conditions without RAG functionality) is also evaluated to demonstrate the suitability of the tasks presented to participants.

The generator LLM parameter size serves as the independent variable, while all other technical parameters are held constant. The range of sizes we chose represents practical trade-offs encountered in real-world deployments: while 70B models approximate the performance of state-of-the-art commercial models [57], their computational requirements are substantial; in contrast, 3B and 8B models are so much less demanding that they can be deployed in edge computing supporting full control over privacy-sensitive data handling. Additionally, 3B and 8B models contribute substantially to policies set to reduce the environmental footprint of AI usage in workplace routines. The parameter count ranges from small (3B) to medium (8B), to large (70B), allowing us to systemati-

cally explore how model size impacts both objective performance and subjective human evaluation. Since our baseline method matches the process of widely used benchmark evaluations (i.e., single-turn interactions), we can gauge the added value of human-AI interaction over multiple turns by comparing the baseline with our participants' performance.

3.1. Task

Participants have to answer specific questions about a publicly available flight manual of over 400 pages². The manual serves as a realistic example of domain-specific materials due to its length, structural complexity, and mix of technical and descriptive content, mirroring the language, information density, and domain focus of texts used in many different professional areas.

Based on the manual, nine questions were created. These questions varied in difficulty, determined by the amount of information required and how dispersed the information is. The questions varied from procedural *'What should a pilot do when a door opens mid-flight?'* to practical *'In the document, there are many different conventional grades of gasoline for aeroplanes mentioned. Can you tell how many there are?'*. All nine questions and the data used can be found in the full experimental setup².

We ensured that participants actually retrieved information from the document to answer the questions in three ways. Through pre-selection, we excluded workers with domain experience, making it unlikely that they knew the answers to the questions off the top of their heads. We excluded participants post-hoc who indicated that they had used external sources or tools to answer the questions (see Subsection 3.3). We also ruled out the possibility that the Llama models, which lack RAG functionality, could answer questions correctly through data memorisation alone. This was done by prompting each model three times using the original questions without any further instructions or information. The models generally failed to produce correct answers based on their internal knowledge base only. This indicated that the questions are suitable for our experiment and once again exposed the constraints of using such models without an additional domain-specific knowledge base. These results are added as LLM-only in Section 4.

Answers to all questions could be found in the manual, meaning there were no unanswerable or 'trick' questions. To arrive at the right answer, participants were free to choose whether to use the source document, the RAG-assistant, or both. Participants could use any preferred wording of their question to prompt the RAG-assistant.

3.2. RAG-assistant

Communication between the human and the RAG-assistant follows the architecture shown in Figure 1. First, the source document of 406 pages is chunked (*chunksize* = 1024) into 2497 chunks by the embedder *intfloat/multilingual-e5-large-instruct* [58] because of its strong retrieval performance on the Massive Text Embedding Benchmark [59]. Upon receiving a user question (*Q*), it is embedded. The cosine similarity between the embedded question and embedded chunks is calculated, after which the top 5 (similar to [21], [60]) most relevant chunks *C* are selected. The top 5 is determined by the highest similarity score between the question *Q* and the text chunks *C*. These retrieved

²For our full experimental setup, questions, and code, visit <https://tinyurl.com/peerRepository>

chunks are then passed to the generator model together with a RAG system prompt ², giving the LLM context and instructions.

The retriever selection was based on a pre-study. For each model size, we tested three different *intfloat-e5* retrievers on their LLM+RAG performance and compared the average over three runs. The results of evaluating multiple retrievers across all three generator models showed that the *small* retriever model obtained the best overall performance with only marginal improvements over the other retrievers. Although the *small* embedding model achieved slightly stronger performance on our questions, we opted for a larger embedding model *intfloat-e5-large* due to its greater representational capacity.

The generator model is assigned to the participant out of the three possible models. The response (*R* in Figure 1) is based on the information provided earlier in the conversation, the user's initial question, and the provided documents. This response is then returned to the user.

All generator models were trained on data with the same cut-off (December 2023). It should be noted that larger models inherently require more training data, meaning they most likely utilise more of the dataset than smaller models. Combined, this isolates the effect of model scale while minimising confounding factors related to model origin or alignment technique. The model was always configured with a temperature of 1. This setting controls the randomness of the model's responses, with 1 providing a balanced level of variation and coherence.

3.3. Experimental Procedure

For each condition, we recruited 40 participants via Prolific (www.prolific.com), yielding $N = 120$ ($M_{age} = 39.4$, $SD_{age} = 10.3$, females = 60, and males = 60). From this group of participants, 8 were removed after the experiment as they indicated external tools were used to answer the questions, leaving a total of 112 participants who met the inclusion criteria. The participants were distributed across the three conditions as follows: 37 in condition 1 (3B), 37 in condition 2 (8B), and 38 in condition 3 (70B). Ethics approval was obtained from the university's HREC (2025-011). The average reimbursement (due to varying bonus payments) was £12.88, which is both above the prolific minimum (£6-£8) and the ethically recommended wage (£9-£12) according to Prolific ³.

The experiment consisted of randomised controlled trials followed by cross-sectional surveys. After obtaining informed consent, participants received instructions for the experiment and were kindly asked not to engage in other activities during the experiment. Then participants answered two experience-related questions: one on experience with LLMs and one on experience with flying aeroplanes (real life or simulators), given the nature of the data set.

The participants were directed to a Streamlit ⁴ hosted website. This website provided access to the flight manual via a conventional chat interface, i.e., the RAG-assistant (left side) and a PDF viewer (right side). The chat interface also showed the relevant passages from the source document used by the assistant for answer generation. The questions were presented in the same order for all participants and could be answered using the PDF directly, the RAG-assistant, or a combination of both. Participants were randomly assigned one of three model conditions distributed equally and randomly, making this a

³<https://researcher-help.prolific.com/en/articles/445230-prolific-s-payment-principles>

⁴streamlit.io

between-subject study. Participants were not informed which model their RAG-assistant was using. After answering the question, they were asked how they arrived at this answer: through the RAG-assistant, the source data, a combination of both, or neither. After the experiment, participants filled out a questionnaire, which we detail below.

3.4. Evaluation

The answers to open questions are categorised as correct, partially correct (only applicable to multi-part questions), or incorrect. This was done by two independent raters, one being the main author of this paper, and the other a colleague independent of this research. The multiple-choice questions had only one correct answer, resulting in either correct or incorrect classification. Accuracy was calculated by assigning 1 point for correct, 0.5 for partially correct, and 0 for incorrect answers, and dividing the total score by the total possible points. A one-sample t-test is used to compare the baseline performance (result from the base strategy) to the human-AI performance.

To analyse participants' performance across experimental conditions, we fitted a linear mixed-effects model [61]. Each participant responded to multiple questions, and each question was seen by multiple participants, introducing non-independence in the data. To address this, we included random effects for both participant (hashed Prolific ID for anonymisation) and question ID, controlling for individual differences and variability in question difficulty. Model size was included as a fixed effect to estimate its influence on the outcome variable: the scoring of each response. This approach allowed us to estimate the effect of experimental conditions on performance while appropriately compensating for possible random effects that influence performance. Combined, this led to the following model: $score \sim model_size + (1|participant_{ID}) + (1|question_{ID})$, with Tukey's p-value adjustment for comparing a family of three (all conditions) estimates.

Our evaluation of usability and satisfaction is inspired by the ASA framework as described in [62]. While this framework originally comprises 19 questions, we restricted our analysis to the 10 questions that best fit our specific use case and align with the scope and objectives of this study. The questions were rated on a 5-point Likert scale used to calculate ASA scores as described in the framework. These ASA scores allow us to holistically compare the evaluation of the differently sized RAG-assistants.

To compare participants' responses on the questionnaire across the three experimental conditions at the per-question level, we conducted a one-way analysis of variance for ordinal data [63]. This test assessed whether there were statistically significant differences in average scores between conditions. When the Kruskal-Wallis test indicated a significant effect, we performed post hoc pairwise comparisons using Dunn's post hoc analysis [64] with Bonferroni correction [65] to identify which specific conditions differed from each other while controlling for multiple comparisons. The exact phrasing of questions to the RAG-assistant was logged ². The logs were inspected in a qualitative evaluation by the experimenter. Additionally, the average completion time is calculated and used to compare performance across different groups, providing a quantitative measure of efficiency.

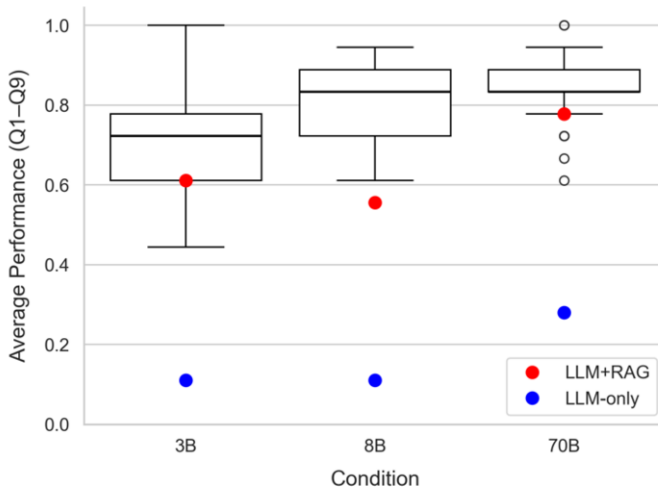


Figure 2. Performance as the accuracy of answers to the 9 test questions across conditions, compared to base-strategy (LLM+RAG) and knowledge base only setting (LLM-only).

4. Results

Comparing the question's evaluation from both independent raters yields a Cohen's Kappa of .898. This high Cohen's Kappa score shows little difference between raters' judgments, which is unsurprising given the unambiguous answers to the questions.

Figure 2 shows average accuracy scores over all questions per model. The box plots show the performance of the human-AI collaboration. The blue and red dots indicate LLM-only and the LLM+RAG base strategy setting, respectively. Adding RAG boosts performance, as indicated by the red dots, confirming the added value of including domain-specific knowledge bases. When comparing the base setting (LLM+RAG) to the human-AI setting through a t-test, the difference in performance is significantly higher for all models ($p < .001$). Returning to **RQ1**, this means that for all models, the human-AI setting outperformed the LLM+RAG setting.

Turning to **RQ2**, we assess performance differences across models using a mixed linear model with random effects for participants and questions. The model showed convergence and a statistically significant increase between the 3B and 8B models ($p < .001$). Differences between the 70B and 8B models were not significant ($p = .354$), but performance increased significantly between the 70B and 3B models ($p < .001$). For question answering, participants favoured the AI assistant (66%), followed by a combination of the AI assistant and the source document (25%), the source document alone (7%), and none (2%). This shows that 91% of the questions were generated in collaboration with the AI assistant. No meaningful differences were found at the per-question level among groups.

Across all conditions, participants spent $25 (\pm 13)$ minutes on the task. An ANOVA test showed no significant ($p = .075$) differences between the time spent on completing the experiment. The amount of experience with LLMs did not influence performance across conditions: $p = .612$, $p = .592$, $p = .344$ for 3B, 8B, 70B, respectively. Experience

Table 1. Participant ratings (mean ± standard deviation) for each question from the questionnaire across three AI models: 3B, 8B, and 70B. Bold values are the best scores. ASA score is calculated as per [62], where higher means better, except for question 8.

Question	3B	8B	70B
1. The AI assistant was good enough to assist me	4.4±0.6	4.1±1.1	4.6±0.6
2. The AI assistant has a human-like manner	2.9±1.2	2.6±0.9	2.7±1.4
3. It was easy to interact with the AI assistant	4.6±0.8	4.2±1.1	4.7±0.6
4. I can rely on the AI assistant	3.8±1.0	3.4±1.1	4.2±0.9
5. I see the interaction with the AI assistant as something positive	4.3±0.7	3.9±0.9	4.6±0.6
6. I like the AI assistant	4.2±0.9	3.8±1.0	4.2±0.9
7. The interaction with the AI assistant captured my attention	4.2±1.0	3.8±1.3	4.1±1.0
8. The AI assistant has no clue of what it is doing	1.6±0.9	2.2±1.1	1.8±1.1
9. The AI assistant is a social entity	2.5±1.2	2.8±1.1	2.4±1.1
10. The AI assistant completed tasks without noticeable delays	4.4±0.8	4.2±1.0	4.3±0.8
ASA	7.0 ± 5.0	4.4 ± 5.7	7.7 ± 5.1

with flying aeroplanes (real life or simulation) showed a positive influence on performance in the 3B condition with $p = .031$; Dunn’s post hoc test did not confirm this.

Perceived performance The subjective results of the experiment can be found in [Table 1](#). Note that for question 8, lower scores indicate better results, as the question is phrased so that agreement corresponds to a less desirable response. For all other questions, higher indicates better results. An ANOVA test comparing participants’ ASA scores indicated significant differences among groups: $F = 4.15, p = .018$. Tukey’s Honestly Significant Difference showed that the difference lies between 8B and 70B, favouring the 70B model. Because of this difference, we zoomed in on the individual questions for both (8B and 70B) groups. Overall, the 70B model was rated highest in helpfulness, usability, and reliability, and users had the most positive attitude towards it. The 3B scored best on the subjective engagement and responsiveness scores and shares the highest rank in likeability with the 70B model. The 8B model was perceived most as a social entity. To test for differences across all conditions, a Kruskal-Wallis test was used. Significant differences were found across all three conditions for questions 4, 5, and 8. Yet, after Dunn’s post hoc test, only the difference in reliance between the 8B and 70B conditions remained, showing the 70B model was perceived as more reliable. Hence, we find no appreciable differences in perceived usability and satisfaction between model sizes (**RQ3**).

Additional qualitative analysis Inspection of the log of the exact phrasings that participants used in the chat window revealed that the most prevalent strategy was copy-pasting the question to the RAG-assistant and pressing return. For some questions, this yields an accurate result; however, other questions needed another strategy to obtain the correct answer. When this strategy did not work, users diverted to other, more interactive strategies. Following what [50] have termed ‘distribution’, participants split the question into sub-questions, which were asked separately. The most prevalent follow-up was providing more context to the assistant. Contextualisation often occurred after participants themselves found and provided relevant excerpts to the assistant. Task evolution [50] was observed when participants started with a base question and progressively sought the right

answer. A collaboration strategy that occurred less often but stood out was, for example, ‘backtracking’ from the answer options provided in the three multiple-choice questions.

Zooming in on performance per question, we observed that participants scored poorest on Question 5: ‘How long does it take to turn a plane around in the air at standard rate?’ A likely source of difficulty was ambiguity in the interpretation of ‘turning around’. Analysis of the interactions revealed that the AI assistant frequently returned information regarding a 360° turn, rather than a 180° turn. Conversely, even when participants were presented with information explicitly referring to a 360° turn, they often failed to recognise that this did not correspond to ‘turning around’, and consequently reported incorrect answers.

5. Discussion

Our results highlight that human-AI collaboration yields significantly higher performance than LLM-only and LLM+RAG systems. The 3B and 8B models benefit more from human-AI interaction than the 70B model. Within the Human-AI setting, task accuracy increased to the degree that the performance gap between the 8B and 70B models was closed. While size matters for LLM+RAG settings, human-AI collaboration mitigates this difference, corroborating the synergetic nature of hybrid intelligence [10].

The ASA scores reflected greater appreciation among participants for the 70B model than for the 8B model. Subjective ratings showed significant differences only in reliability and affective evaluation, in favour of the 70B model. For all other measures, participants rated the assistants similarly. A similar pattern was reported by [25], who found that the smaller Llama model in their setup demonstrated similar or even better general alignment with users than its larger counterpart. Individual preferences played a role as well: for some questions, 3B scored higher than 70B and vice versa, visible in Table 1.

Although the 8B model outperforms the 3B model on benchmarks involving language understanding and reasoning [66], the subjective ratings for the 8B model were lower than those for the 3B and 70B models. Specifically, users found the 8B model less reliable than the 70B model. This might be attributed to its worst performance on instruction-following benchmarks [67], or because of the version difference between the two Llama versions. Consistent with [62], we observed correlations between certain questionnaire items (e.g., agent’s attitude and agent’s likeability). Such correlations stem from lexical ambiguity or conceptual similarity between the items. While this does not invalidate our findings, it suggests that some responses might capture overlapping aspects of participants’ attitudes. In conclusion, participants indicate differences in some areas, but these do not amount to a decisive preference for one of the three models.

Overall, the 8B and 70B models yield the highest objective accuracy. We observe a disconnect between accuracy and perceived usability and satisfaction, as the better-performing models were not rated higher than the 3B model in most areas. Understanding this disconnect could improve not only evaluation frameworks but also the design of user-facing systems that prioritise trust, interpretability, and perceived usefulness alongside performance. This suggests a need for deeper exploration of subjective user evaluations.

In relation to previous work, [38] showed that the added value of human-AI collaboration is task-dependent. While a case for the value of hybrid setups has been made for classification [68] and deliberation [69], our work suggests that human-AI collabora-

tion also adds value for QA settings. Our hybrid setup can be seen as multi-turn *human* rephrasing, a hybrid variant of the computationally costly strategy of automated rephrasing that is effective [70]. We suggest that a hybrid setup is a viable and efficient alternative, especially in sensitive, high-stakes, or domain-specific contexts where human involvement is already required. Especially because with this setup, one could combine the human with a smaller LLM. Still, several small open-source models could be embedded in a system simulating multi-turn rephrasing. This could test whether small models can effectively handle sensitive real-world data, making it an interesting future direction. As a final point, we observed that the standard deviation for the 3B model is larger than for the 8B and 70B models in the human-AI collaboration setting. Smaller models are more sensitive to input changes than larger models [22, 23]. Lastly, ensuring these results generalise broadly, it has to be tested with models from different families.

As for the limitations of our setup, the fixed retriever returns chunks in similar orders for similar queries. Given that LLMs exhibit primacy and recency effects [28], chunk position in the context may have influenced performance. This was partly mitigated by participants generating their own queries, which resulted in chunks presented in different orders. The analyses presented in Section 4 also account for such kinds of variability by incorporating both participants and questions as random effects in the mixed-effect models. Additionally, participants did not receive feedback on the correctness of their answers: while automated evaluation would have been feasible for multiple-choice questions, open-ended responses required manual assessment. This lack of feedback may have shaped perceptions of the models' helpfulness. On the other hand, in real workplace settings, the correct answer would often not be available either. Additionally, participants could review and reflect on their own answers and were incentivised to do so through potential bonus payments. As recruitment was via Prolific, where financial compensation is a primary motivator, some participants may have prioritised speed over accuracy despite these incentives, potentially reducing answer quality and increasing susceptibility to hallucinations. In our current setup, we deliberately chose to use only one document that contains a mixture of technical and procedural information. In future follow-ups, multiple documents with varying content types could be used. Perceived slowness in response times may have resulted from participants' internet connections or latency on the Hugging Face API servers, potentially impacting their experience and perceptions of the AI tools.

6. Conclusion

This paper examined human-AI interaction in a realistic information-seeking task mirroring real-world information processing contexts. It focused on evaluating multiple perspectives: accuracy, usability, and satisfaction across three different conditions, using a combination of performance and participant evaluations. The results underscore the importance of human-AI collaboration, which consistently outperformed LLM-only and LLM+RAG in our experiment. Specifically, smaller models benefitted more from human collaboration to the degree that this closed the performance gap between the 8B and 70B models. Importantly, participants did not perceive larger models as more capable than smaller models, revealing a disconnect between benchmark performance and human user experience. This asks for looking beyond pure benchmark performance, assessing applications in realistic, multi-turn scenarios that take user perspectives into account.

References

- [1] Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*. 2024;15(3):1-45.
- [2] Barbereau T, Dom L. GPT-NL: Towards a Public Interest Large Language Model. In: *PI-AI@ KI*; 2024. p. n/a.
- [3] Asai A, Wu Z, Wang Y, Sil A, Hajishirzi H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In: *The Twelfth International Conference on Learning Representations*; 2024. p. n/a. Available from: <https://openreview.net/forum?id=hSyW5go0v8>.
- [4] Yu Y, Ping W, Liu Z, Wang B, You J, Zhang C, et al. RankRAG: Unifying Context Ranking with Retrieval-Augmented Generation in LLMs. In: Globerson A, Mackey L, Belgrave D, Fan A, Paquet U, Tomczak J, et al., editors. *Advances in Neural Information Processing Systems*. vol. 37. Curran Associates, Inc.; 2024. p. 121156-84. Available from: https://proceedings.neurips.cc/paper_files/paper/2024/file/db93ccb6cf392f352570dd5af0a223d3-Paper-Conference.pdf.
- [5] Wang Z, Wang Z, Le L, Zheng S, Mishra S, Perot V, et al. Speculative RAG: Enhancing Retrieval Augmented Generation through Drafting. In: *The Thirteenth International Conference on Learning Representations*; 2025. p. n/a. Available from: <https://openreview.net/forum?id=xgQfWbV6Ey>.
- [6] Arslan M, Ghanem H, Munawar S, Cruz C. A Survey on RAG with LLMs. *Procedia Computer Science*. 2024;246:3781-90.
- [7] Yang X, Sun K, Xin H, Sun Y, Bhalla N, Chen X, et al. CRAG - comprehensive RAG benchmark. In: *Proceedings of the 38th International Conference on Neural Information Processing Systems*. NIPS '24. Red Hook, NY, USA: Curran Associates Inc.; 2024. p. n/a.
- [8] Lyu Y, Li Z, Niu S, Xiong F, Tang B, Wang W, et al. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *ACM Transactions on Information Systems*. 2025;43(2):1-32.
- [9] Li X, Jin J, Zhou Y, Zhang Y, Zhang P, Zhu Y, et al. From matching to generation: A survey on generative information retrieval. *ACM Transactions on Information Systems*. 2025;43(3):1-62.
- [10] Akata Z, Balliet D, De Rijke M, Dignum F, Dignum V, Eiben G, et al. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*. 2020;53(8):18-28.
- [11] Kandpal N, Deng H, Roberts A, Wallace E, Raffel C. Large language models struggle to learn long-tail knowledge. In: *International Conference on Machine Learning*. PMLR; 2023. p. 15696-707.
- [12] Qian H, Zhu Y, Dou Z, Gu H, Zhang X, Liu Z, et al. Webbrain: Learning to generate factually correct articles for queries by grounding on large web corpus. *arXiv preprint arXiv:230404358*. 2023.
- [13] Nakano R, Hilton J, Balaji S, Wu J, Ouyang L, Kim C, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:211209332*. 2021.

- [14] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*. 2020;33:9459-74.
- [15] Ram O, Levine Y, Dalmedigos I, Muhlgaay D, Shashua A, Leyton-Brown K, et al. In-Context Retrieval-Augmented Language Models. *Transactions of the Association for Computational Linguistics*. 2023;11:1316-31. Available from: <https://aclanthology.org/2023.tacl-1.75/>.
- [16] Ma X, Gong Y, He P, Zhao H, Duan N. Query rewriting in retrieval-augmented large language models. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*; 2023. p. 5303-15.
- [17] Dhole KD, Agichtein E. Genqrensemble: Zero-shot llm ensemble prompting for generative query reformulation. In: *European Conference on Information Retrieval*. Springer; 2024. p. 326-35.
- [18] Ruthven I. Interactive information retrieval. *Annual review of information science and technology*. 2008;42:43-92.
- [19] Mo F, Mao K, Zhao Z, Qian H, Chen H, Cheng Y, et al. A Survey of Conversational Search. *ACM Trans Inf Syst*. 2025 Aug. Just Accepted. Available from: <https://doi.org/10.1145/3759453>.
- [20] Belcak P, Heinrich G, Diao S, Fu Y, Dong X, Muralidharan S, et al. Small Language Models are the Future of Agentic AI. *arXiv preprint arXiv:250602153*. 2025.
- [21] Fan T, Wang J, Ren X, Huang C. Minirag: Towards extremely simple retrieval-augmented generation. *arXiv preprint arXiv:250106713*. 2025.
- [22] Huang Y, Huang JX. A Survey on Retrieval-Augmented Text Generation for Large Language Models. *ACM Comput Surv*. 2026 Apr. Just Accepted. Available from: <https://doi.org/10.1145/3805774>.
- [23] Shao R, He J, Asai A, Shi W, Dettmers T, Min S, et al. Scaling retrieval-based language models with a trillion-token datastore. *Advances in Neural Information Processing Systems*. 2024;37:91260-99.
- [24] Shi Q, Jimenez CE, Yao S, Haber N, Yang D, Narasimhan KR. When Models Know More Than They Can Explain: Quantifying Knowledge Transfer in Human-AI Collaboration. In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems*; 2025. p. n/a. Available from: <https://openreview.net/forum?id=9V2SVE11vP>.
- [25] Bona FBD, Dominici G, Miller T, Langheinrich M, Gjoreski M. Evaluating Explanations Through LLMs: Beyond Traditional User Studies. *CoRR*. 2024;abs/2410.17781. Available from: <https://doi.org/10.48550/arXiv.2410.17781>.
- [26] Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans Inf Syst*. 2025 Jan;43(2). Available from: <https://doi.org/10.1145/3703155>.
- [27] Xu J, Dong Q, Liu H, Li L. Go-tuning: Improving zero-shot learning abilities of smaller language models. *arXiv preprint arXiv:221210461*. 2022.
- [28] Liu NF, Lin K, Hewitt J, Paranjape A, Bevilacqua M, Petroni F, et al. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*. 2024;12:157-73. Available from: <https://aclanthology.org/2024.tacl-1.9/>.

- [29] Lepagnol P, Gerald T, Ghannay S, Servan C, Rosset S. Small Language Models Are Good Too: An Empirical Study of Zero-Shot Classification. In: Calzolari N, Kan MY, Hoste V, Lenci A, Sakti S, Xue N, editors. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italia: ELRA and ICCL; 2024. p. 14923-36. Available from: <https://aclanthology.org/2024.lrec-main.1299/>.
- [30] Lin CY. Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out; 2004. p. 74-81.
- [31] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics; 2002. p. 311-8.
- [32] Friel R, Belyi M, Sanyal A. n/a, editor. RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems. n/a; 2024. Available from: <https://arxiv.org/abs/2407.11005>.
- [33] Es S, James J, Espinosa Anke L, Schockaert S. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In: Aletras N, De Clercq O, editors. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. St. Julians, Malta: Association for Computational Linguistics; 2024. p. 150-8. Available from: <https://aclanthology.org/2024.eacl-demo.16/>.
- [34] Huang Y, Sun L, Wang H, Wu S, Zhang Q, Li Y, et al. Position: TrustLLM: Trustworthiness in Large Language Models. In: Salakhutdinov R, Kolter Z, Heller K, Weller A, Oliver N, Scarlett J, et al., editors. Proceedings of the 41st International Conference on Machine Learning. vol. 235 of Proceedings of Machine Learning Research. PMLR; 2024. p. 20166-270. Available from: <https://proceedings.mlr.press/v235/huang24x.html>.
- [35] Meyer S, Singh S, Tam B, Ton C, Ren A. A comparison of llm finetuning methods & evaluation metrics with travel chatbot use case. arXiv preprint arXiv:240803562. 2024.
- [36] Mangold A, Hoffmann K. Human-Centered Evaluation of RAG outputs: a framework and questionnaire for human-AI collaboration. arXiv preprint arXiv:250926205. 2025.
- [37] Noy S, Zhang W. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*. 2023;381(6654):187-92.
- [38] Vaccaro M, Almaatouq A, Malone T. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*. 2024;8(12):2293-303.
- [39] Gao J, Gebreegziabher SA, Choo KTW, Li TJJ, Perrault ST, Malone TW. A Taxonomy for Human-LLM Interaction Modes: An Initial Exploration. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. CHI EA '24. New York, NY, USA: Association for Computing Machinery; 2024. p. n/a. Available from: <https://doi.org/10.1145/3613905.3650786>.
- [40] Schemmer M, Hemmer P, Kühl N, Benz C, Satzger G. Should I Follow AI-based Advice? Measuring Appropriate Reliance in Human-AI Decision-Making; 2022.
- [41] Li Z, Zhu H, Lu Z, Xiao Z, Yin M. From Text to Trust: Empowering AI-assisted Decision Making with Adaptive LLM-powered Analysis. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems; 2025. p. 1-18.

- [42] Tang L, Li J, Fantus S. Medical artificial intelligence ethics: A systematic review of empirical studies. *Digital health*. 2023;9:20552076231186064.
- [43] Choi JH, Schwarcz D. Ai assistance in legal analysis: An empirical study. *J Legal Educ*. 2024;73:384.
- [44] Vereschak O, Bailly G, Caramiaux B. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction*. 2021;5(CSCW2):1-39.
- [45] Xu Y, Liu X, Liu X, Hou Z, Li Y, Zhang X, et al. ChatGLM-Math: Improving Math Problem-Solving in Large Language Models with a Self-Critique Pipeline. In: Al-Onaizan Y, Bansal M, Chen YN, editors. *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics; 2024. p. 9733-60. Available from: <https://aclanthology.org/2024.findings-emnlp.569/>.
- [46] Zhang K, Li J, Li G, Shi X, Jin Z. CodeAgent: Enhancing Code Generation with Tool-Integrated Agent Systems for Real-World Repo-level Coding Challenges. In: Ku LW, Martins A, Srikumar V, editors. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics; 2024. p. 13643-58. Available from: <https://aclanthology.org/2024.acl-long.737/>.
- [47] Chae HS, Yoon C. Factors Affecting Human-Generated AI Collaboration: Trust and Perceived Usefulness as Mediators. *Information*. 2025;16(10):856.
- [48] Afroogh S, Akbari A, Malone E, Kargar M, Alambeigi H. Trust in AI: progress, challenges, and future directions. *Humanities and Social Sciences Communications*. 2024;11(1):1568.
- [49] Ramrath M, Scharmann A, Ridder A, Kühn T, Weller S, Krämer N. Trust in AI Chatbots: The perceived expertise of CHATGPT in subjective and objective tasks. In: *HHAI 2024: hybrid human ai systems for the social good*. IOS Press; 2024. p. 264-80.
- [50] Zhou Z, Gao W, Li Y, Yu J. Developing an interaction framework for human-large language models collaboration in creative tasks: Insights from UX professionals' communication with ChatGPT. Available at SSRN 4853257. 2024.
- [51] Davis FD, et al. Technology acceptance model: TAM. Al-Suqri, MN, Al-Aufi, AS: Information seeking behavior and technology adoption. 1989;205(219):5.
- [52] Xie C, Wang Y, Cheng Y. Does artificial intelligence satisfy you? A meta-analysis of user gratification and user satisfaction with AI-powered chatbots. *International Journal of Human-Computer Interaction*. 2024;40(3):613-23.
- [53] Lin YC, Neville J, Stokes J, Yang L, Safavi T, Wan M, et al. Interpretable User Satisfaction Estimation for Conversational Systems with Large Language Models. In: Ku LW, Martins A, Srikumar V, editors. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics; 2024. p. 11100-15. Available from: <https://aclanthology.org/2024.acl-long.598/>.
- [54] Song K, Kang Y, Liu J, Li X, Sun C, Liu X. A speaker turn-aware multi-task adversarial network for joint user satisfaction estimation and sentiment analysis. In: *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intel-*

- ligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'23/IAAI'23/EAAI'23. AAAI Press; 2023. p. n/a. Available from: <https://doi.org/10.1609/aaai.v37i11.26592>.
- [55] Wang W, Jiao W, Hao Y, Wang X, Shi S, Tu Z, et al. Understanding and Improving Sequence-to-Sequence Pretraining for Neural Machine Translation. In: Muresan S, Nakov P, Villavicencio A, editors. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics; 2022. p. 2591-600. Available from: <https://aclanthology.org/2022.acl-long.185/>.
- [56] Katsis Y, Rosenthal S, Fadnis K, Gunasekara C, Lee YS, Popa L, et al. MTRAG: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. Transactions of the Association for Computational Linguistics. 2025;13:784-808.
- [57] Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, et al. The llama 3 herd of models. arXiv preprint arXiv:240721783. 2024.
- [58] Wang L, Yang N, Huang X, Yang L, Majumder R, Wei F. Multilingual e5 text embeddings: A technical report. arXiv preprint arXiv:240205672. 2024.
- [59] Muennighoff N, Tazi N, Magne L, Reimers N. MTEB: Massive Text Embedding Benchmark. In: Vlachos A, Augenstein I, editors. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Dubrovnik, Croatia: Association for Computational Linguistics; 2023. p. 2014-37. Available from: <https://aclanthology.org/2023.eacl-main.148/>.
- [60] Nanni F, Chan R, Lazauskas T, Geddes J. n/a, editor. Why we still need small language models – even in the age of frontier AI. n/a; 2025. <https://www.turing.ac.uk/blog/why-we-still-need-small-language-models>.
- [61] Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982;963-74.
- [62] Fitrianie S, Bruijnes M, Li F, Abdulrahman A, Brinkman WP. The artificial-social-agent questionnaire: establishing the long and short questionnaire versions. In: Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents; 2022. p. 1-8.
- [63] Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. Journal of the American statistical Association. 1952;47(260):583-621.
- [64] Dunn OJ. Multiple comparisons using rank sums. Technometrics. 1964;6(3):241-52.
- [65] Abdi H, et al. Bonferroni and Šidák corrections for multiple comparisons. Encyclopedia of measurement and statistics. 2007;3(01):2007.
- [66] Suzgun M, Scales N, Schärli N, Gehrmann S, Tay Y, Chung HW, et al. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Findings of the Association for Computational Linguistics: ACL 2023. Toronto, Canada: Association for Computational Linguistics; 2023. p. 13003-51. Available from: <https://aclanthology.org/2023.findings-acl.824/>.
- [67] Zhou J, Lu T, Mishra S, Brahma S, Basu S, Luan Y, et al. Instruction-following evaluation for large language models. arXiv preprint arXiv:231107911. 2023.

- [68] Ma S, Lei Y, Wang X, Zheng C, Shi C, Yin M, et al. Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems; 2023. p. 1-19.
- [69] Van Der Meer M, Liscio E, Jonker CM, Plaat A, Vossen P, Murukannaiah PK. HyEnA: A Hybrid Method for Extracting Arguments from Opinions. In: Schlobach S, Perez-Ortiz M, Tielman M, editors. HHAI2022: Augmenting Human Intellect – Proceedings of the 1st International Conference on Hybrid Human-Artificial Intelligence. vol. 354 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2022. p. 17-31.
- [70] Wang L, Yang N, Wei F. Query2doc: Query Expansion with Large Language Models. In: Bouamor H, Pino J, Bali K, editors. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics; 2023. p. 9414-23. Available from: <https://aclanthology.org/2023.emnlp-main.585/>.