Why does in-context learning fail sometimes? Evaluating in-context learning on open and closed questions.

Anonymous ACL submission

Abstract

This study investigates the impact of context relevancy on the performance of in-context learning. To quantify that impact, we created a novel database of open-form questions, each paired with different contexts of various relevancy. Next, we perform manual grading (introducing six-fold redundancy to minimize the impact of individual graders), measuring the quality of generated responses in several dimensions. We show that counterintuitively, in many cases, less relevant contexts can perform as well as, or even better than, more relevant ones. By controlling for task novelty and question difficulty, we demonstrate that this phenomenon is particularly pronounced for open-form questions and questions with high perceived novelty or difficulty. This result reveals a fundamental difference in how large language models process closed-form and open-form questions. Furthermore, our findings raise critical questions about optimal context selection for large language models, particularly in open-response scenarios – a question critical when building Retrieval-Augmented Generation (RAG) systems.

1 Introduction

Large Language Models (LLMs), despite their indisputable successes (Bommasani et al., 2021; Drori et al., 2022; Chang et al., 2024), often struggle to answer challenging questions (Rawte et al., 2023). While LLMs can achieve superhuman accuracy on many benchmarks (Luo et al., 2024), they also suffer from hallucinations (Ye et al., 2023; Azamfirei et al., 2023), lack of coherence (Xie et al., 2023b), and are prone to cognitive errors (Jones and Steinhardt, 2022; Hagendorff et al., 2023). To make the difficult situation even worse, it is not always easy to detect mistakes committed by LLMs since their responses are often presented in a way that emulates correct and coherent answers (Bender et al., 2021; Scheurer et al., 2023). Therefore, for practical reasons, many existing benchmarks only test the ability to answer either closed (Chang et al., 2024) or easy-to-verify questions, e.g., regarding common knowledge (Bisk et al., 2020; Clark et al., 2018) or questions that can be algorithmically verified (Srivastava et al., 2024).

Another challenge concerning LLMs is the problem of updating or adding new factual knowledge. Re-training or fine-tuning is both expensive (Bender et al., 2021; Luccioni et al., 2023) and technically challenging (Kandpal et al., 2023; Gaspers et al., 2022). While some techniques like Low-Rank Adaptation (LoRa) can reduce the cost of adaptation to new tasks (Hu et al., 2021), it does not solve the main issue, namely, how to allow LLMs to leverage new pieces of information that were not a part of the initial training corpus (Liu, 2017) in a sustainable way. In that regard, one of the promising approaches is in-context learning (Brown et al., 2020). By embedding examples in the prompt, LLMs can demonstrate an improved performance without the need to update the model parameters (Brown et al., 2020). Despite the progress, the question of how LLM accesses and processes novel information from the context remains vital. A lot is at stake, as this ability of LLMs to extract novel information from the context is central for the correct operations of popular Retrieval-Augmented Generation (RAG) systems (Gao et al., 2024) that build the majority of modern commercial chatbot assistants (Vakayil et al., 2024).

Given the growing popularity of RAG-based systems, in this paper, we examine the role of context relevancy in *in-context learning*. Our findings reveal an intriguing and somehow counterintuitive behavior: *In-context learning performance does not always improve with increasing context relevance*. In addition, the relationship between the quality of the generated answer and the context relevance level seems to depend on several factors, including whether the question is of the open or closed format and whether LLM had prior knowledge of the subject of the question.

In the following sections, we introduce our novel dataset, which we created to capture the mentioned behavior. Next, we compare (and contrast) our findings with the results of Min et al. (2022b). As an additional validation, we then partly replicate the results of Min et al. (2022b), and to extend our analysis, we repeat our experiments on two additional close-formed question datasets, MetaICL (Min et al., 2022a) and NephSAP (Wu et al., 2023). By comparing all our results and by contrasting the diverging behaviors on closed- and open-ended question-banks, we came to the conclusion that the main impact comes from the format of the question (whether open or closed), with additional effects related to the difficulty or novelty of those questions, revealing how essential is the prior knowledge of LLM on the subject for effective utilization of the in-context learning.

2 Related Work

Large Language Models. LLMs have shown remarkable capabilities in various tasks, including code generation (Kojima et al., 2022; Siddiq and Santos, 2023), text summarization (Sahu et al., 2023), and database query optimization (Li et al., 2023). They demonstrate a surprising ability to perform in-context learning (Brown et al., 2020), where an LLM "learns" to perform a task simply by conditioning on a prompt containing some input-output examples. However, there has been little understanding of how the model leverages the context and what makes in-context learning work. In addition, their performance significantly depends on the contextual information provided and, as discussed in this paper, on the form and type of the queries.

In-Context Learning. Unlike traditional fine-tuning methods, in-context learning adapts models to unseen tasks by incorporating examples directly into the input context (Brown et al., 2020). Xie et al. (2022) discussed how in-context learning can be understood as implicit Bayesian inference, where models infer latent concepts to generate coherent responses. Techniques such as chain-of-thought prompting (Wei et al., 2022; Press et al., 2023; Wang et al., 2022; Zhou et al., 2023; Imani et al., 2023; Besta et al., 2023) have shown significant improvements in reasoning tasks. Recent frameworks like OpenICL (Wu et al., 2023) have further streamlined the implementation of in-context learning by providing unified and flexible tools for integrating various retrieval and inference methods.

Many recent research studies focus on the example selection strategies of in-context learning. One of the most common strategies is to select examples for demonstration based on similarity in the embedding space (Liu et al., 2022; Qin et al., 2023; Gao et al., 2021). In-context learning seems robust to label-noise, as indicated by work of Min et al. (2022b), in which authors show that demonstrations, even one with randomly shuffled or replaced labels, can still significantly improve LLM's performance.

Evaluation Benchmarks. Benchmarking is essential for understanding LLM performance across different domains. Existing benchmarks like AGIEval (Zhong et al., 2023), ChenLLMBench (Guo et al., 2023), SCIEval (Sun et al., 2023), PIXIU (Xie et al., 2023a), and MME (Fu et al., 2024) provide comprehensive datasets for evaluating LLMs. While these benchmarks are useful for understanding the general capabilities of LLMs, they do not fully capture the complexity of more open-ended and context-sensitive queries. Here, the added value of our work, as we believe the novel open-form question bank accompanied by the context of variable relevance that we created and shared through this paper, will help to at least partly fill that gap.

3 Originality, Impact, and Model Choice

Originality. While Min et al. (2022b) have shown that context significantly affects LLM performance, they have not quantified how different levels of context relevancy impact the quality of generated responses. The authors also neither controlled for the "openness" nor the difficulty of the questions. Our research addresses this by focusing on open, challenging questions and bt explicitly controlling for the context relevancy level. Another difference is that instead of using automated evaluation metrics like BLEU score, METEOR, or BERTScore, we choose a labor-intensive approach, where we manually grade the generated outputs by a panel of experts. This allowed us to better capture the settled differences present in generated answers and score them in several dimensions (completeness, relevancy, logic, reasoning, etc.) while capturing problems related to hallucination, omission, irrelevancy, etc.

The impact of the paper. Our work suggests areas for improvement regarding generated output evaluation methodologies and the overall Retrieval-Augmented Generation (RAG) systems design principles. Current RAG studies focus on providing context during model inference, and the most common evaluation frameworks used to tune RAG hyper-parameters utilize various automated benchmarks. Given our observation of the inconsistent relationship between the relevance of context and model performance for different question types (open-form vs. closed-form), we believe that the context retrieved by comparing vector similarity using RAG may not always correlate with the most helpful context for enhancing LLM inference performance.

Model Choice. Due to the manual and labor-intensive process of grading generated answers, in this study, we restricted ourselves only to ChatGPT (based on GPT-4) from OpenAI. We are aware of that shortcoming and urge the other researchers to present results of similar analyses for other model versions and architectures. Nevertheless, ChatGPT is arguably the most widely popular (at least in the eyes of the popular audience) LLM-based conversational assistant, so the results presented in this paper should still be relevant for many.

4 Methodology

Novel question bank. To investigate the relationship between the relevance of context and the performance of LLM, we created an original open-form questions dataset comprising physics and computer science questions of varying difficulty levels (*easy*, *medium*, and *hard*) and originality (*known*, *paraphrased*, and *original*). The question's originality was related to how it was harvested. *Known* questions come from popular textbooks – a source likely known to the model of our interest (OpenAI's GPT-4). *Paraphrased* questions were rewritten and modified versions of known questions. *Original* questions were handcrafted by the authors of this paper.

For each question, we created a ground truth answer for scoring reference and four context types with different levels of relevance. The four context types were (1) "no context" to serve as a control group, (2) "irrelevant context", which consists of text on topics that do not match the subject of the question, (3) "vague context", which incorporates some topics or keywords related to the question, and (4) "relevant context", which provides reasoning context for the question, or answer to a highly related question. Next, for each unique pair of question-context, we generated a response employing the OpenAI's *gpt-4-1106-preview* model.

After retrieving the responses, we constructed 160 question-response pairs, each accompanied by the corresponding ground truth. Aware that human grading can be subjective in their judgement, we decided that each question would be evaluated by six independent graders using a pre-defined scoring sheet. Our volunteer graders were composed of five students (aged 20–25) and one faculty (aged 35-40) at our university, whose expertise ranged from physics to computer science. This gave us 960 evaluation responses in total.

While all annotators were residing in the United States at the time of the study, they had international backgrounds, originating from either Europe or Asia. This diverse demographic contributed a range of perspectives to the evaluation process.

The Supplementary Material includes examples of the questions and context types, as well as the evaluation sheets.

Evaluation. Our evaluation system comprised three main categories: *Completeness and Relevancy* (5 points), *Logic and Reasoning* (5 points), and *Truthfulness (understood as lack of hallucination)* (5 points). In addition, graders had the option to identify specific problems in the responses, such as *hallucinations*, *omissions*, *irrelevancy*, *calculation errors*, and *logic errors*. The graders could also highlight portions of the responses as *incorrect*, *correct*, or *irrelevant*. In addition, an open response section was provided for graders to give comments and feedback about the generated responses. Finally, graders were asked to rate how confident they felt in giving those grades. These options allowed us to gain deeper insights into the grading process and to assess the quality of the generated responses in detail.

All question-response pairs were presented to graders in random order and without information about the type of context used to generate the responses. For the purpose, we used *potato annotation system* (Pei et al., 2022). To enhance the accuracy and reliability of our evaluation, we ensured that all graders assessed all 160 questions. This uniform evaluation approach significantly simplified the subsequent statistical analysis, while by involving multiple graders for each response, we reduced the impact of individual biases and other statistical errors. As demonstrated later, this approach significantly improved the accuracy and consistency of our findings.

5 Results

Context Relevancy. In Fig. 1, we illustrate the relation between the context types and the quality of the corresponding generated responses. In panel A, we show the average raw score for each grader and each context type. Note that the difference between the graders is likely due to their individual tolerance for different types of errors. However, after the grades are standardized and average (panel B), a clear trend emerges, as illustrated in panel C. Contrary to what one could expect, we see the best performance for *no context* case and the lowest score associated with the *relevant* context case, indicating that adding relevant context does not help (contrary, it hurts the performance of the model).

To further investigate how the difficulty of questions affects the quality of generated responses, we compared the results across three difficulty levels (*easy*, *medium*, and *hard*) for each of the four context types, as presented in Fig. 2. In panel A, we can observe a clear trend of decreasing scores as the difficulty of the questions increased from medium to hard (consistent result for each context type), indicating that GPT-4's performance declines with greater question difficulty. This result also indicates that the questions that a human perceived as difficult were, in fact, correlated with the factual difficulty experienced by GPT-4, a result interesting on its own. However, for easy and medium-difficulty problems, GPT-4 generated responses with similar scores, indicating that the alignment between the human-perceived and machine-perceived difficulty might be highly nonlinear, though mostly monotonic. One could potentially leverage this to map human-perceived difficulty to machine-perceived one, but given that nonlinear relation, the creation of such a map would require careful calibration.

In panel B of Fig. 2, we compare the aggregated standardized score for the different levels of originality for each context type. It is evident that GPT-4 scores highest for known questions, likely because these questions were part of its training data. Interestingly, the score for known questions given irrelevant context is twice as high as for relevant context. This suggests that an irrelevant context might be more helpful than a relevant context for known questions, at least for the open type of question, as measured here.

Analyzing those results, we can see that responses generated with no additional context or with the help of irrelevant context are, on average, of higher quality than responses generated for queries incorporating highly relevant context. This result is in striking difference to results of Min et al. (2022b). To further understand this discrepancy, in the next section, we replicate the key findings of Min et al. (2022b) and discuss what might cause the difference in the observed behavior.

Comparison with existing studies. Min et al. (2022b) demonstrates that in-context learning allows us to achieve significantly better results than in the "no context" case. Moreover, the authors show that in-context learning is robust to irrelevant context. Specifically, they demonstrate that the quality of responses for closed-form questions, such as multiple-choice and true/false questions, remains largely unaffected as long as the *structure* of the context is preserved, even if its *content* is irrelevant to the question.

To ensure a meaningful comparison between our and their results and to eliminate the effect of different versions of ChatGPT playing a potential role here (Min et al. (2022b) used GPT-3, while our study focuses on GPT-4), we decided to replicate the key results from Min et al. (2022b) using precisely the same framework and the same model as described in the previous section. For this replication, we utilized two existing benchmarks, MetaICL under Attribution-NonCommercial 4.0 International license (Min et al., 2022a) and a dataset from NephSAP under Apache license 2.0. For the MetaICL dataset, we took a subset of 10 different tasks, each containing multiple-choice questions. For the NephSAP dataset, we focused on multiple-choice questions, choosing among 20 subjects. We share details about tasks, subjects, and sample questions in the Supplementary Materials.

We conducted an 80-20 train test split for both datasets. Next, for each multiple-choice question in the test set, we generated a response using the *gpt-4-1106-preview* model. We did it three times: once without any context (*no-context* control group), once with a randomly sampled demonstration from *a different* task or subject from the training set of the dataset, and once with a randomly sampled demonstration with *the same* subject or task from

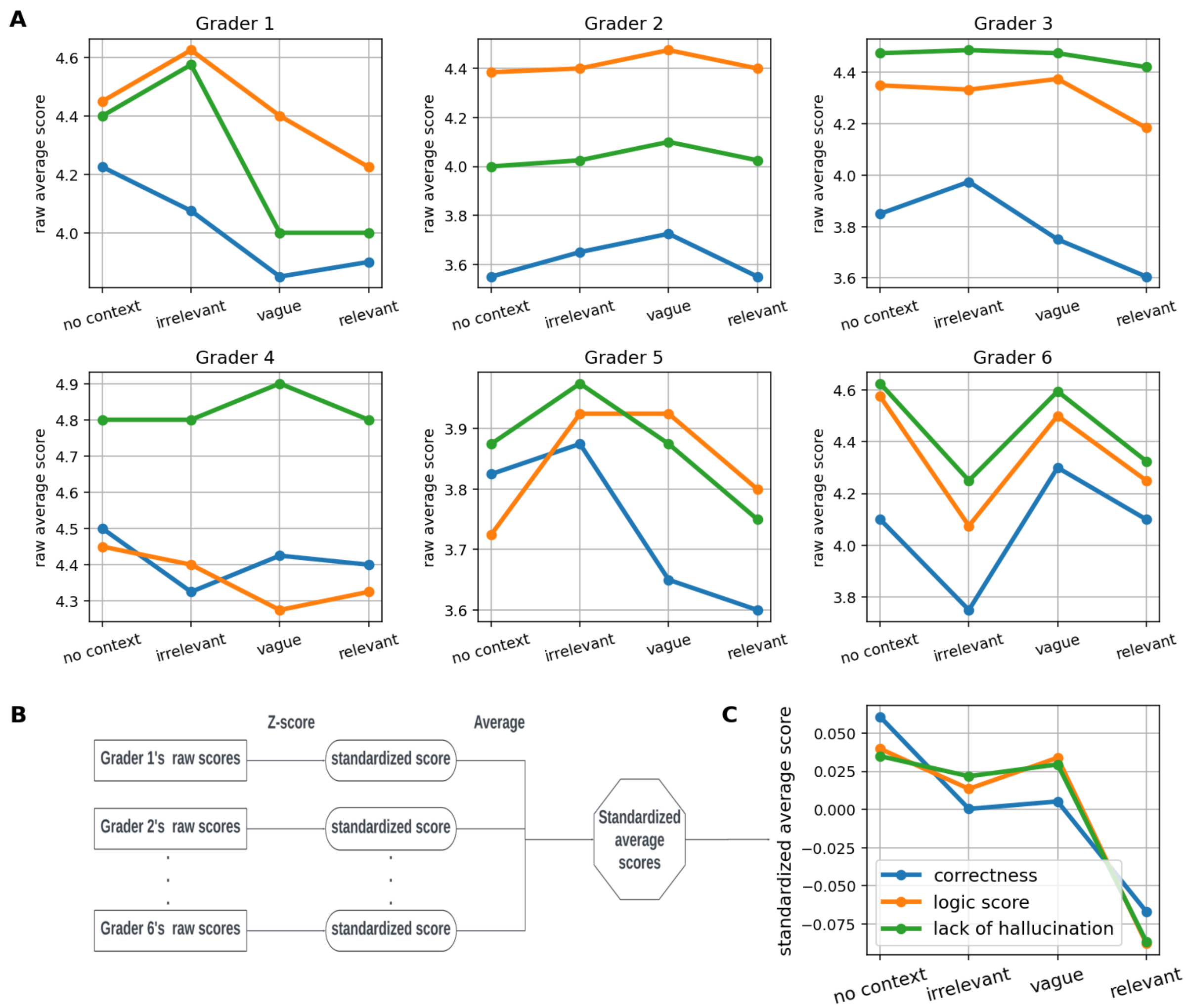

Figure 1: (A) Raw average scores of generated responses for each context type (*no context*, *irrelevant context*, *vague context*, and *relevant context*) evaluated for *Completeness and Relevancy (Correctness)*, *Logic and Reasoning (Logic Score)*, and *Truthfulness (lack of hallucination)*, assessed by six different graders. (B) The process of standardizing raw scores from each grader to calculate the overall standardized average scores. (C) Standardized average scores of generated responses for each context type aggregated across all graders.

the training set. To quantify the context relevancy, we computed its cosine similarity to the question in the embedded space. That allowed us to separate the context into classes of relevancy (denoted as similarity bins in Fig. Figs. 3–5). Next, treating the results of the control group (with no context) as a reference point, we record the general score improvement of the response within each embedding similarity bin.

Analyzing results in Figs. 3–4, note how context similarity is positively correlated with the mean score improvement for both tested closed-question datasets (MetalICL and NephSAP). Note also that in both closed-question datasets, the context with the lowest levels of similarity scores has a tendency to have a negative mean improvement (meaning, adding irrelevant context hurts the results). As contexts with low levels of similarities are more likely to be contexts with a different subject or task, this result is consistent with the findings in Liu et al. (2021), where it was reported that irrelevant demonstrations hurt the performance of LLM.

Next, we repeated the same procedure in the open-form questions scenario, leveraging our original dataset. In Fig. 5, we show the results. Our open-form question results display a negative correlation between context similarity and mean improvement, meaning that context with a lower level of similarity can be more helpful in improving the quality of the response, whereas context with a higher level of similarity can actually hurt the quality of the response. This stands in a striking con-

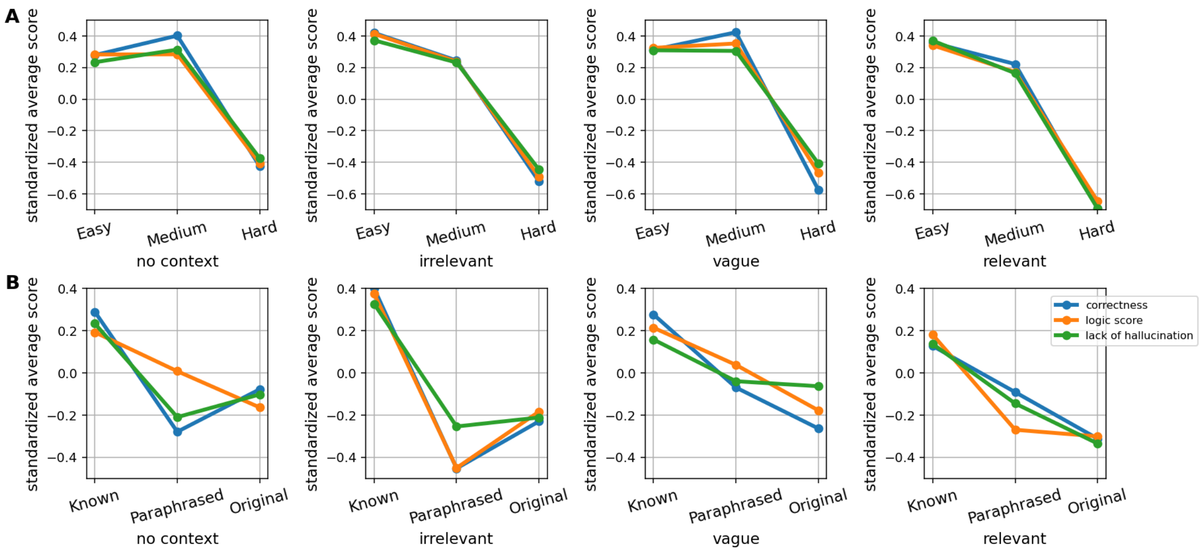

Figure 2: (A): Standardized average scores of generated responses for each context type (*no context*, *irrelevant context*, *vague context*, and *relevant context*), categorized by three levels of question difficulty (*easy*, *medium*, and *hard*) for correctness, logic errors, and lack of hallucination. (B): Standardized average scores of generated answers for each context type, subdivided into *known*, *paraphrased*, and *original* categories, evaluated for correctness, logic score, and lack of hallucination.

trast to the results for closed-form questions (cf. again Figs. 3–4).

6 Discussion

Impact. Our findings indicate that context might play fundamentally different roles depending on the format of the question (whether open or closed-format), as evidenced by the reversed relationship between context similarity and performance improvement in these two cases (see Fig.5 versus Figs.3–4). This result carries two significant implications. First, choosing an optimal context for in-context learning might be highly task-dependent, and factors such as the question's "openness", perceived difficulty, or novelty might play a significant role. Second, selecting context based solely on the minimal embedding distance to the question may not always be optimal, particularly for tasks involving open-form questions. This insight has profound implications for Retrieval-Augmented Generation (RAG) systems, many of which address open-form scenarios and complex tasks.

Interpretation. The different behaviors exhibited in the open-form and closed-form answer generation scenarios seem to stem from a different treatment of the context in those two cases. We provide a hypothetical interpretation of that mechanism. In closed-form multiple-choice questions, the evaluated language model is treated as a classification model. A relevant demonstration provided as a context can improve the LLM's performance by aligning it with the correct choice. However, in the open-form answer generation tasks, the evaluated language model is treated as a generative model. Instead of being either correct or incorrect, an open-form response can be anywhere between. A relevant context provides alignment with one way of approaching the question, but it can also introduce bias, leading at the end to performance degradation instead of improvement.

Furthermore, for highly complex and difficult questions (which are common among many open-form questions), the evaluated language model can have difficulty learning from the logic and methodology applied in the relevant context. Without properly understanding the reasoning behind the relevant context, the relevant context can hardly provide any more help. However, LLMs can still benefit from the provided examples by mimicking the correct *style* of the answer, as it happened in Min et al. (2022b), thus the improvements over the baseline when *vague* or *irrelevant* context was present.

Implications The results discussed in this work have a profound impact on strategies regarding the implementation of RAG-based systems. Our work points out at the difference between the impact of context relevancy on the model performance

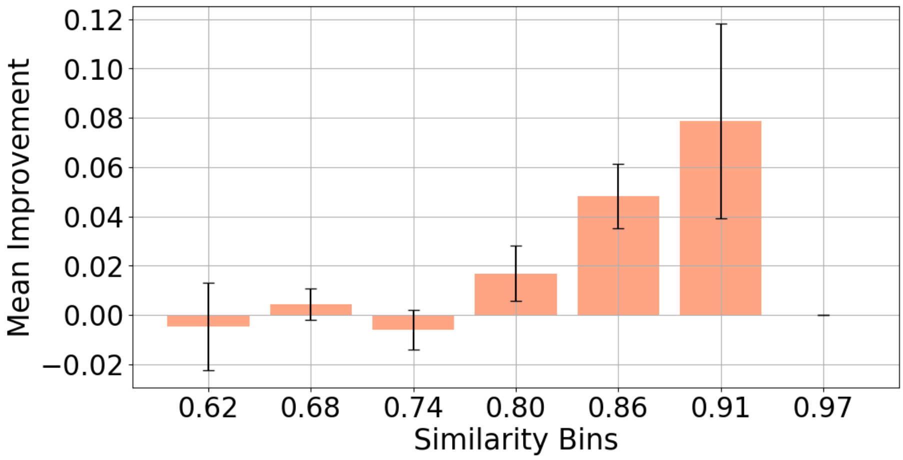

Figure 3: The MetaICL dataset contains close-form questions. The last bin is insignificant as it contains only seven samples of data. Note that the relationship between similarity and score improvement is positively correlated.

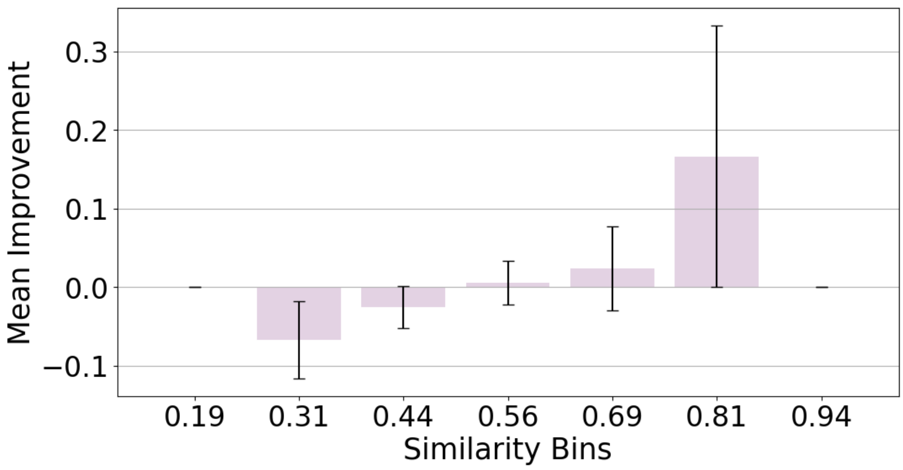

Figure 4: The NephSAP dataset contains close-form questions. The first bin and the last bin are insignificant as they contain only 1 sample each. Note that the relationship between similarity and score improvement is positively correlated.

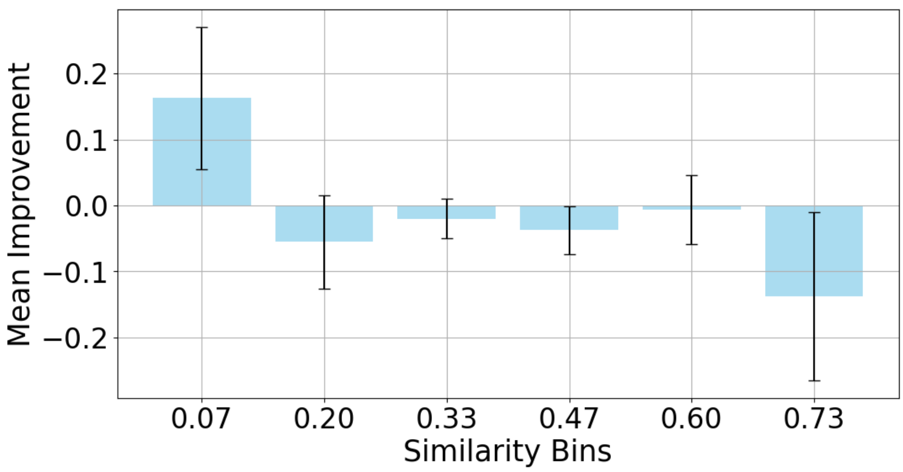

Figure 5: Using the same methodology for our original dataset of open-form questions, we see that the relationship between similarity and score improvement is anti-correlated. Thus, the trend recorded here is in a striking difference with results for closed-form questions (cf. Figs. 3–4).

for open-form and closed-form questions. This suggests that the performance of in-context learning, and as a result, RAG-based systems, might be highly application-dependent and related to many factors, such as the form of the question, its difficulty, "openness", etc. In particular, we suggest that the strategy for context retrieval for open-form applications should be different from those employed in closed-form scenarios. This also creates additional complexity when tuning the hyperparameters of RAG systems. Employing automated frameworks utilizing both closed-form benchmarks and automated scoring mechanisms might be shortsighted. A good performance on a close-question benchmark might not indicate similarly good operation in open-form scenarios. Therefore, more complex and therefore expensive evaluation methods might be required.

One final, practical remark touches the strategies for context retrieval for RAG-based systems. Especially in open-form applications, when retrieving context it seems important to include some other factors than only its embedding distance. For example, instead of selecting passages that simply lie near a certain point in the embedding space (e.g., representing the query), a better choice could be to include more complex retrieval protocols that promote passages from some intermediate distance. The logic would be that a piece of context that is a bit further in the embedding distance to the question might still provide valuable information while not reinforcing the hidden bias inside the question.

7 Limitations

We measured the quality of the generated answers manually to quantify their quality. This restricted our ability to check different architectures and versions of the model. It also limited our ability to test different prompt versions (prompt engineering). For practical reasons, we limited ourselves to the most popular model at the time (provided by OpenAI). To partially assess the impact of prompt engineering, we explored the effects of different prompt formats on the generated outputs (see Supplementary Materials). These experiments reassured us that the results gathered with the help of our prompt were representative. However, a dedicated study would be valuable to examine how other LLMs respond to contexts of differing quality.

This particular study is limited to English only. Whether the described behavior generalizes to other languages is open and requires dedicated research.

We also recognize the potential for the evaluation of open-form questions on a much larger scale using automatic methods. We experimented with BLEURT, an automatic evaluation method, on our open-form data (see Supplementary Materials). The results show little to no correlation between the scores of manual evaluation and the scores generated with this automated method. This leads us to believe that carefully designed manual evaluation, even conducted on a smaller scale, is still important. Aligning automatic evaluations to better represent manual ones can be a great direction for future research, and new automatic evaluation methods can provide more scalable solutions for the evaluation.

Risks Regarding Potential Societal Impact

The findings of this work highlight the nuanced role of context relevancy in *in-context learning*, which could inadvertently reinforce biases or lead to unintended outcomes when applied in real-world systems. Specifically, the observed tendency for less relevant contexts to sometimes outperform more relevant ones in open-form scenarios might be misused to justify the use of less precise or contextually mismatched information, potentially amplifying misinformation, perpetuating biases, or producing unreliable outputs in critical applications such as legal, medical, or educational systems.

Code

Code that can be used to replicate all results of this work is available at https://github.com/Context-matters-research/Context-matters.

Data

The dataset of open-form questions with accompanying contexts of varying relevancy is intended strictly for research purposes and can be found at https://github.com/Context-matters-research/Context-matters/tree/main/open_dataset. Its primary use is to advance understanding of large language models' behavior in generative tasks.

Use of AI Assistants

The use of AI Assistants was limited only to the following activities: grammar and spelling correction, and synonym search.

References

Razvan Azamfirei, Sapna R. Kudchadkar, and James Fackler. 2023. Large language models and the perils of their hallucinations. *Critical Care*, 27(1).

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21. ACM.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2023. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *arXiv preprint*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. *arXiv preprint*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint*.

Iddo Drori, Sarah Zhang, Reece Shuttleworth, Leonard Tang, Albert Lu, Elizabeth Ke, Kevin Liu, Linda Chen, Sunny Tran, Newman Cheng, Roman Wang, Nikhil Singh, Taylor L. Patti, Jayson Lynch, Avi Shporer, Nakul Verma, Eugene Wu, and Gilbert Strang. 2022. A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proceedings of the National Academy of Sciences*, 119(32).

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Judith Gaspers, Anoop Kumar, Greg Ver Steeg, and Aram Galstyan. 2022. Temporal generalization for spoken language understanding. In *NAACL 2022*.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2023. What Can Large Language Models Do in Chemistry? A Comprehensive Benchmark on Eight Tasks. *arXiv preprint*.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. MathPrompter: Mathematical Reasoning using Large Language Models. In *Proceedings of the The 61st Annual Meeting of the Association for Computational Linguistics: Industry Track, ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 37–42. Association for Computational Linguistics.

Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. In *Advances in Neural Information Processing Systems*, volume 35, pages 11785–11799. Curran Associates, Inc.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *NeurIPS*.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Rongyu Cao, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin C. C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023. Can LLM Already Serve as A Database Interface? A BIg Bench for Large-Scale Database Grounded Text-to-SQLs. *arXiv preprint*.

Bing Liu. 2017. Lifelong machine learning: a paradigm for continuous learning. *Frontiers of Computer Science*, 11(3):359–361.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *CoRR*, abs/2101.06804.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good incontext examples for gpt-3? In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research*, 24(253):1–15.

Xiaoliang Luo, Akilles Rechardt, Guangzhi Sun, Kevin K. Nejad, Felipe Yáñez, Bati Yilmaz, Kangjoo Lee, Alexandra O. Cohen, Valentina Borghesani, Anton Pashkov, Daniele Marinazzo, Jonathan Nicholas, Alessandro Salatiello, Ilia Sucholutsky, Pasquale Minervini, Sepehr Razavi, Roberta Rocca, Elkhan Yusifov, Tereza Okalova, Nianlong Gu, Martin Ferianc, Mikail Khona, Kaustubh R. Patil, Pui-Shee Lee, Rui Mata, Nicholas E. Myers, Jennifer K Bizley, Sebastian Musslick, Isil Poyraz Bilgin, Guiomar Niso, Justin M. Ales, Michael Gaebler, N Apurva Ratan Murty, Leyla Loued-Khenissi, Anna Behler, Chloe M. Hall, Jessica Dafflon, Sherry Dongqi Bao, and Bradley C. Love. 2024. Large language models surpass human experts in predicting neuroscience results.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022a. MetaICL: Learning to learn in context. In *NAACL-HLT*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022b. Rethinking the role of demonstrations: What makes in-context learning work? In *EMNLP*.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. Potato: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. *Preprint*, arXiv:2210.03350.

Chengwei Qin, Aston Zhang, Anirudh Dagar, and Wenming Ye. 2023. In-context learning with iterative demonstration selection. *CoRR*, abs/2310.09881.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S. M Towhidul Islam Tonmoy, Aman Chadha, Amit P. Sheth, and Amitava Das. 2023. The Troubling Emergence of Hallucination in Large Language Models – An Extensive Definition, Quantification, and Prescriptive Remediations. *arXiv preprint*.

Gaurav Sahu, Olga Vechtomova, and Issam H. Laradji. 2023. Enchancing Semi-Supervised Learning for Extractive Summarization with an LLM-based pseudolabeler. *arXiv preprint*.

Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. 2023. Technical Report: Large Language Models can Strategically Deceive their Users when Put Under Pressure. *arXiv preprint*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Mohammed Latif Siddiq and Joanna C. S. Santos. 2023. Generate and Pray: Using SALLMS to Evaluate the Security of LLM Generated Code. *CoRR*, abs/2311.00889.

Saurabh Srivastava, Annarose M B, Anto P V, Shashank Menon, Ajay Sukumar, Adwaith Samod T, Alan Philipose, Stevin Prince, and Sooraj Thomas. 2024. Functional benchmarks for robust evaluation of reasoning performance, and the reasoning gap.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2023. SciEval: A Multi-Level Large Language Model Evaluation Benchmark for Scientific Research. *arXiv preprint*.

Sonia Vakayil, D. Sujitha Juliet, Anitha. J, and Sunil Vakayil. 2024. Rag-based llm chatbot using llama-2. In *2024 7th International Conference on Devices, Circuits and Systems (ICDCS)*, pages 1–5.

Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively prompt pre-trained language models for chain of thought. *Preprint*, arXiv:2203.08383.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*.

Zhenyu Wu, YaoXiang Wang, Jiacheng Ye, Jiangtao Feng, Jingjing Xu, Yu Qiao, and Zhiyong Wu. 2023. Openicl: An open-source framework for in-context learning. *Preprint*, arXiv:2303.02913.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023a. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *Preprint*, arXiv:2306.05443.

Qiming Xie, Zengzhi Wang, Yi Feng, and Rui Xia. 2023b. Ask again, then fail: Large language models' vacillations in judgment.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. *Preprint*, arXiv:2111.02080.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. *arXiv preprint*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. *Preprint*, arXiv:2205.10625.

Supplementary Material

A Sample question

A.1 Sample Question for Open Dataset

Question: Given the wavelength of an electron is $0.364 \cdot 10^{-9}$ m, calculate the speed of the electron.

Ground Truth for Grading:
$\lambda = 0.364 \times 10^{-9}$ m
Mass of electron, $m = 9.1 \times 10^{-31}$ kg
Planck's Constant, $h = 6.62607015 \times 10^{-34}$ Js
The de Broglie wavelength is given by $\lambda = \frac{h}{mv}$
Velocity of the electron, $v = 2 \times 10^6$ ms^{-1}

Relevant Context
The De Broglie states that $\lambda = \frac{h}{mv}$. The mass of an electron is about $9.109 \cdot 10^{-31} kg$

Vague Context
Wave-particle duality is the concept in quantum mechanics that quantum entities exhibit particle or wave properties according to the experimental circumstances.

Irrelevant Context
Quantum physics is the study of matter and energy at the most fundamental level. At very small scale, classical theories may not be applicable any more. That is where quantum theories come into play.

A.2 Sample question for MetaICL dataset

Test Input: Bird feet can also vary greatly among different birds. Some birds, such as gulls and terns and other waterfowl, have webbed feet used for swimming or floating (Figure below). Other birds, such as herons, gallinules, and rails, have four long spreading toes, which are adapted for walking delicately in the wetlands (Figure below). You can predict how the beaks and feet of birds will look depending on where they live and what type of food they eat. Flightless birds also have long legs that are adapted for running. Flightless birds include the ostrich and kiwi. Some birds, such as gulls and terns and other waterfowl, have what type of feet used for swimming or floating?

Test Output: webbed

Test Options:

1. lobed
2. quad toed
3. bipedal
4. webbed

For our task selections from the MetaICL dataset, please visit our GitHub repository, where the task category selections and code are presented.

A.3 Sample question for NephSAP dataset

A 54-year-old man with ESRD is admitted for management of presumed catheter–related bacteremia. He had no pre–ESRD nephrology care and recently started maintenance hemodialysis on an urgent basis for symptomatic uremia. Two days ago, he developed acute onset of fever to 40 C, chills, and rigors during dialysis. After obtaining blood cultures, he received intravenous vancomycin and ceftazidime at the outpatient dialysis center before admission. His tunneled dialysis catheter is removed because of erythema and purulent exudate at the exit site. His fever is now resolved. He does not have back pain, cough, dental complaints, or rash. On physical examination, he is now afebrile. The BP is 136/88 mmHg. There is a 1-cm area of erythema at the tunneled dialysis catheter removal site. He has no stigmata of emboli. He has no obvious dental issues on gross inspection of the oropharynx. The cardiac examination shows a grade 2 early peaking systolic crescendo-decrescendo murmur along the right sternal border without radiation. The remainder of the examination is unremarkable. The blood cultures drawn before implementation of antibiotic therapy grew *S. aureus* within 12 hours. The results of sensitivity testing and repeat blood cultures are pending. Laboratory studies show sodium 138 mEq/L, potassium 3.8 mEq/L, chloride 102, total CO2 25 mmol/L, BUN 45 mg/dl, and creatinine 6.4 mg/dl.

In addition to continued antibiotic therapy, which ONE of the following is the next BEST step in management?

A. Arrange for placement of a tunneled dialysis catheter today

B. Transesophageal echocardiography

C. Magnetic resonance imaging of the spine

D. Bone scintigraphy (bone scan)

E. Examination for an occult dental source of bacteremia

B Generated Answer

B.1 Sample Response for Question in A.1

Response with No Context: To calculate the speed of the electron, we can utilize de Broglie's equation that relates the momentum of a particle to its wavelength. Here is the formula:

$$\lambda = \frac{h}{p}$$

Where:

- λ is the wavelength
- h is Planck's constant
- p is the momentum.

If we rearrange the terms to express momentum:

$$p = \frac{h}{\lambda}$$

The momentum of a particle can also be described as its mass times its velocity:

$$p = m \cdot v$$

We know that the mass of an electron (m) is 9.11×10^{-31} kg. And Planck's constant (h) is 6.626×10^{-34} Js.
So the velocity (v) is given as:

$$v = \frac{h}{m\lambda}$$

Substituting the known values, we get:

$$v = \frac{6.626 \times 10^{-34}}{9.11 \times 10^{-31} \cdot 0.364 \cdot 10^{-9}}$$

After solving this, the velocity of the electron comes out to be approximately 7.24×10^5 m/s.

Response with Irrelevant Context: The speed of an electron given its wavelength can be calculated using the de Broglie's equation:

$$\lambda = \frac{h}{mv}$$

where:

- λ is the wavelength,
- h is Planck's constant, which is roughly equals to 6.626×10^{-34} Js,
- m is the mass of the electron, which is roughly equals to 9.109×10^{-31} kg,
- v is the velocity of the electron.

We can rearrange the equation to solve for the velocity:

$$v = \frac{h}{m\lambda}$$

Substitute the given values into the equation:

$$v = \frac{6.626 \times 10^{-34}\,\text{Js}}{9.109 \times 10^{-31}\,\text{kg} \times 0.364 \times 10^{-9}\,\text{m}}$$

and calculate the velocity.

Response with Vague Context: The wavelength of a matter wave, such as an electron, can be found using De Broglie's equation, which is given by:

$$\lambda = \frac{h}{p}$$

where:

- λ is the wavelength
- h is Planck's constant (6.626×10^{-34} Js)
- p is the momentum of the particle, which can be expressed as $p = mv$

Since we want to find the speed of the electron (v), we can rearrange this equation in terms of v:

$$v = \frac{h}{m\lambda}$$

Given that $\lambda = 0.364 \times 10^{-9}$ m and the mass of an electron is $m = 9.109 \times 10^{-31}$ kg, we can substitute these values into the equation:

$$v = \frac{6.626 \times 10^{-34}}{9.109 \times 10^{-31} \times 0.364 \times 10^{-9}}$$

Evaluating this will give the speed of the electron in m/s.

Response with Relevant Context: The De Broglie equation relates the wavelength of a particle to its momentum. According to De Broglie's principle, every object in motion has a wave character. The De Broglie wavelength is given by the formula:

$$\lambda = \frac{h}{mv}$$

where:

- λ is the wavelength,
- h is Planck's constant ($6.626 \times 10^{-34}\,\text{m}^2 \cdot \text{kg/s}$),
- m is the mass of the particle,
- v is the velocity (or speed) of the particle.

We are given that $\lambda = 0.364 \times 10^{-9}$ m, and the mass m of an electron is 9.11×10^{-31} kg. We can find the speed v by rearranging the above formula:

$$v = \frac{h}{m\lambda}$$

Substitute the known values into the rearranged equation to find v, the speed of the electron:

$$v = \frac{6.626 \times 10^{-34}\,\text{m}^2 \cdot \text{kg/s}}{(9.11 \times 10^{-31}\,\text{kg})(0.364 \times 10^{-9}\,\text{m})}$$

Solving the above expression will give us the speed of the electron.

C Grading Interface

To better evaluate the open response to our physics questions, we modified the *potato annotation system* (Pei et al., 2022) and applied it as our evaluation system. Our evaluation system not only allows users to select numeric grades for each response but also enables the user to highlight parts of the response, apply labels, and write descriptions to justify their grading. In addition, the system randomly shuffles the order of the responses for each grader to mitigate any potential bias in grading as a result of the ordering of responses. The instructions given to the graders are shown in table 1, and a short video tutorial is provided at the beginning page to provide guidance and alignment in grading.

A screenshot of the interface of the evaluation system is shown in Fig. 6.

D Sanity Check

To check whether our context relevancy is well defined, we compute the embedding of the questions and their respective contexts for both our open-form question dataset and the two closed-form question datasets we use. We then calculate the cosine distance between the embedding of each question and the different contexts associated with them. We show the results for the open question dataset in Fig. 7.

We computed the embedding of each question and each context using OpenAI's "text-embedding-3-large" model. For the no-context part, we used a space as a placeholder instead of an empty string. As expected, the results show that more relevant contexts, as perceived by us when designing the dataset, receive a higher mean similarity score with their respective questions. Different question types can result in a large standard deviation in similarity scores in different contexts. We show the details breakdown of those results in Fig. 8.

All question types except hard paraphrased questions display the same trend, confirming the relationship between context types and embedding similarities.

For the closed datasets, the similarity score between context and question is shown in Table 2. For both datasets, the same task/subject demonstrations possess a higher mean similarity score than the different task/subject demonstrations. To further verify this relationship, we have also plotted the similarity score of the same task demonstrations and different task demonstrations for each task in the MetaICL dataset in Fig. 9. The results confirm that the same task demonstration displays higher mean similarity than the different task demonstration in every task in the dataset.

E BLEURT score vs Manual Score

We used BLEURT (Sellam et al., 2020) to grade the response of GPT-4 to questions in our open dataset. We then plot the BLEURT score against the results of our standardized manual grading. The results are shown in Fig.10. Since no obvious correlation is found, BLEURT fails to capture the trend of our manual grading, suggesting potential limitations for this automatic evaluation technique. At the same time, those results validate the sensibleness of our labor-intensive, manual approach.

F Generalizability and prompt adaptability

We test a different prompt template and repeat our experiment with our open-form dataset and MetaICL (Min et al., 2022a) with GPT-4. The sample prompt template is shown below:

> I want you to act as an expert in physics, math, and computer science. I will provide problems and sometimes some context, and it will be your job to answer them in easy-to-understand terms. This could include providing step-by-step instructions for solving a problem, demonstrating various techniques with visuals or suggesting online resources for further study.

For MetaICL, the plot of performance improvement against the embedding distance of demonstration is shown in Fig. 12. It is clear that the results obtained through the new template effectively capture the trend in our finding in Fig. 11. This result suggests that the relationship between embedding similarity of demonstration and performance improvement might be independent of prompt in closed-form benchmark.

Since conducting another round of manual evaluation is very costly, we choose to apply faithfulness metrics (Es et al., 2024) instead to evaluate responses to the new prompt. Treating the original responses to the same question (four for each question since we have four different levels of context) as a reference, we plot the faithfulness score of our new response to each of the original responses to the same question. The results are shown in Fig. 13. The plot suggests no obvious difference in faithfulness mapping between the same context type and different context types in the new responses and the original ones.

One possible explanation is that the faithfulness measure fails to capture differences in objective truth and mistakes in different responses in a human-like manner. However, prompt can potentially be a more important factor here as we need to recognize the fundamental difference between open-form benchmarks and closed-form benchmarks. This only further supports the case for manual evaluation in the case of open-form answer generation.

Category	Details
Procedures	You will be presented with 160 short document-question pairs and tasked to answer the question by highlighting part of the document.
Risks	Risks are no greater than those ordinarily encountered in daily life, such as when surfing the internet.
Benefits	There may be no personal benefit, but the knowledge gained may have academic or industrial value.
Confidentiality	Researchers may disclose your consent form, data, or personally identifiable information as required by law. Otherwise, confidentiality will be maintained by: (1) assigning participants a number; (2) recording data by number, not name; (3) storing recordings and files in a secured location accessed only by authorized researchers.
Voluntary Participation	Participation is voluntary. You may discontinue at any time.

Table 1: Potato grading instructions.

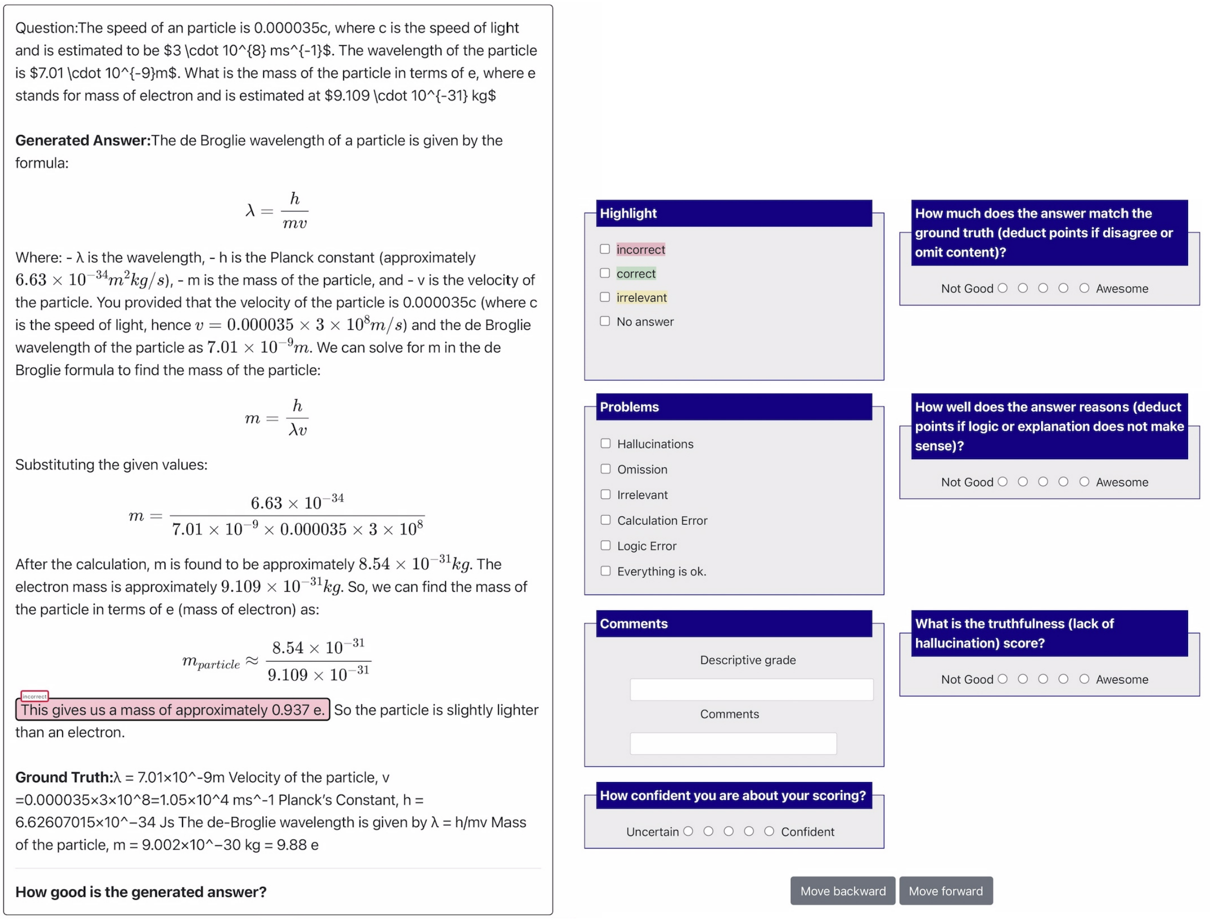

Figure 6: The potato grading interface used in evaluation.

Dataset	Average Different Task Similarity	Average Same Task Similarity
MetaICL	0.719	0.787
NephSAP	0.443	0.557

Table 2: Mean context similarity for closed datasets.

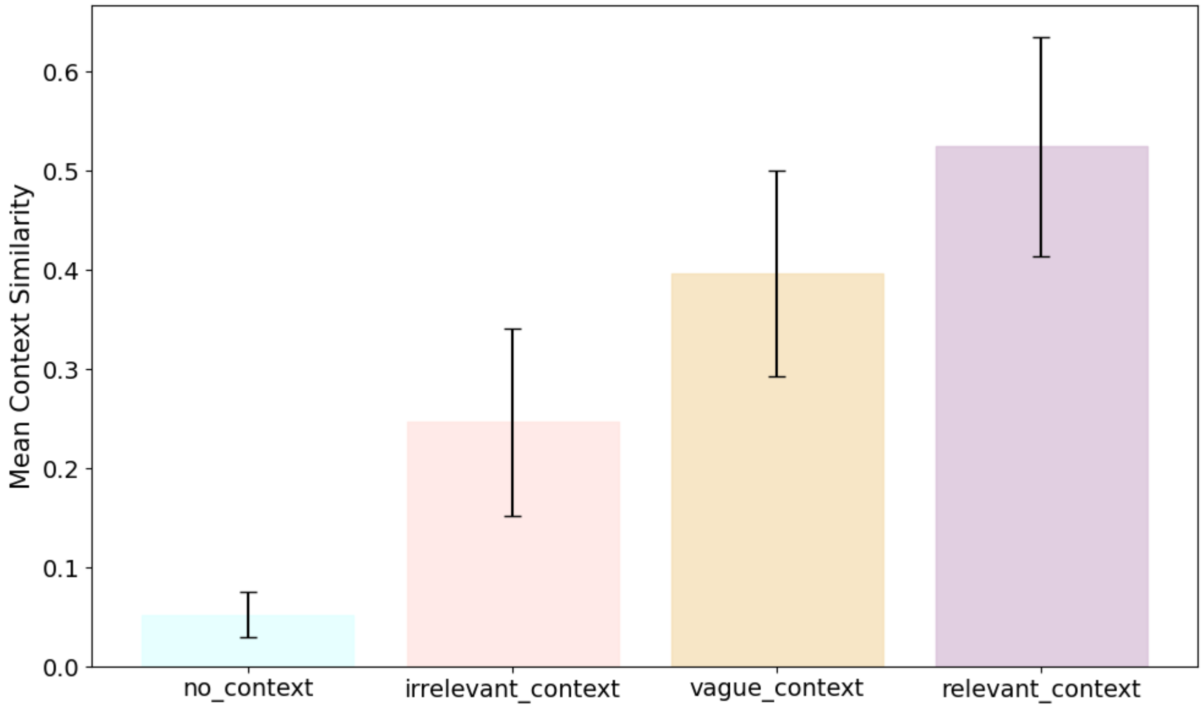

Figure 7: Mean context similarity by context type for open-form questions' context. the context of no context is replaced by a space as a placeholder, as the embedding of an empty string cannot be computed.

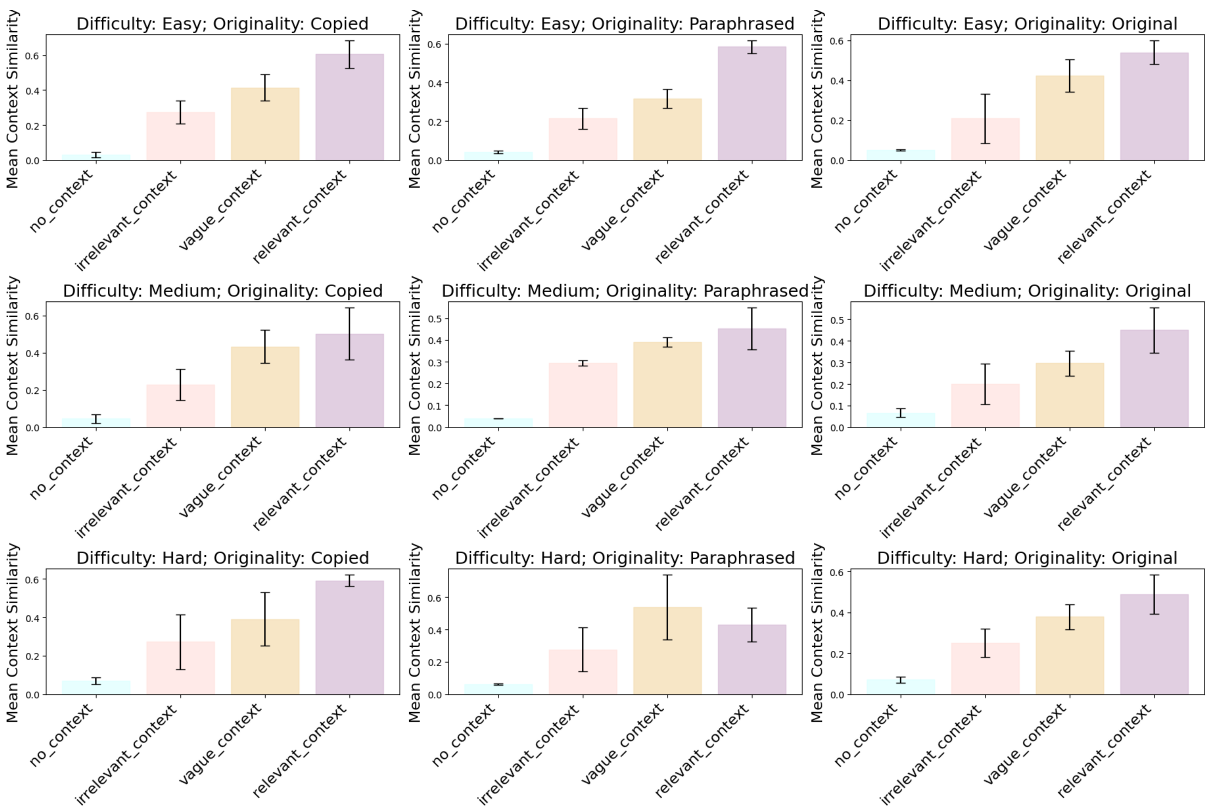

Figure 8: Mean context similarity by context type across different originality and difficulty for open-form questions' context.

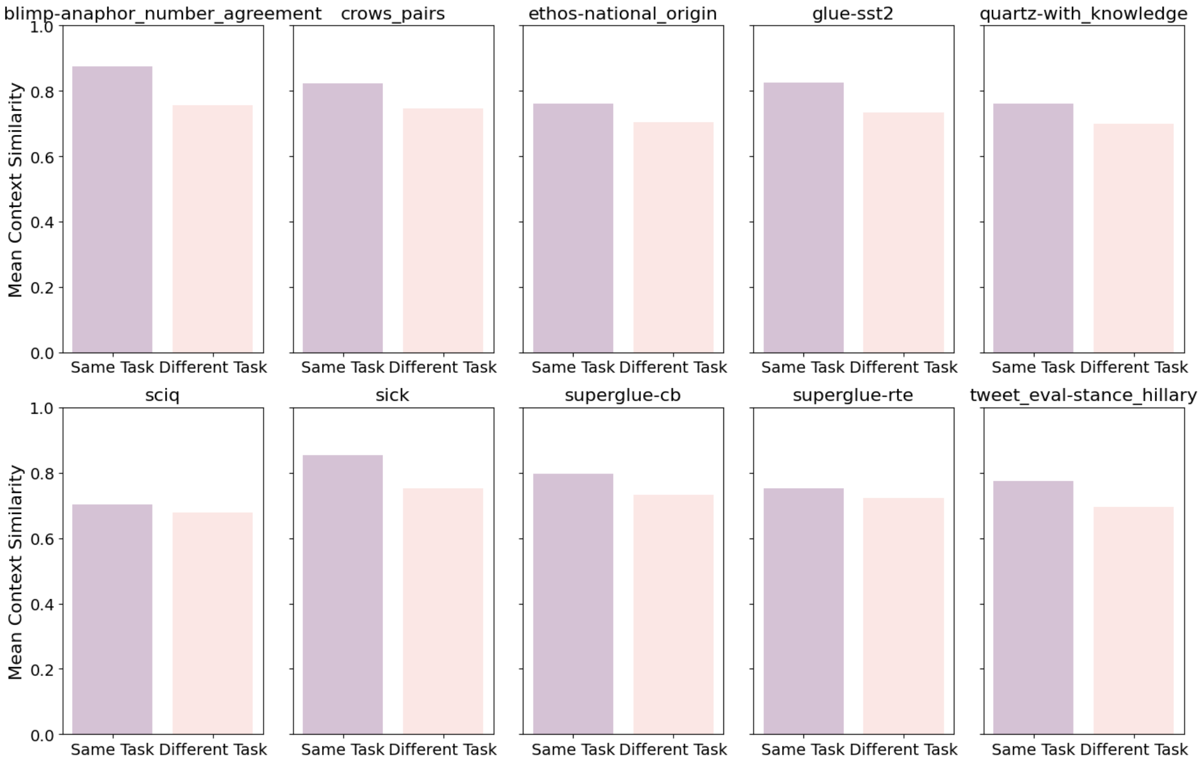

Figure 9: Mean context similarity by demonstration types across different MetaICL tasks.

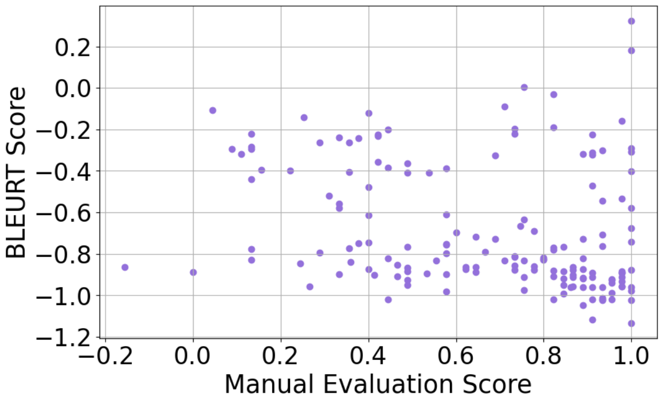

Figure 10: BLEURT score vs manual score for open-form questions.

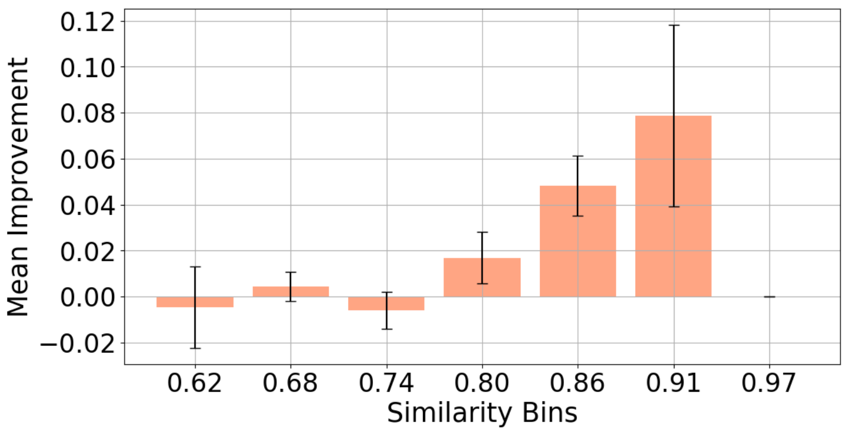

Figure 11: MetaICL performance improvement against context similarity using the original prompt template as in figure 3 in the paper.

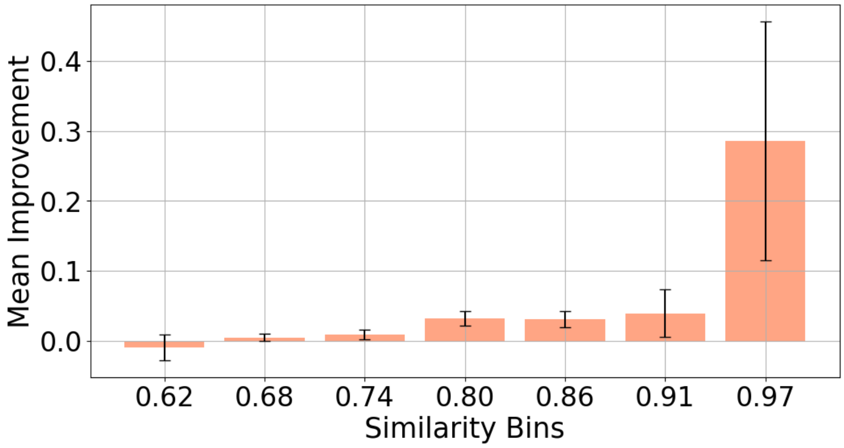

Figure 12: MetaICL performance improvement against context similarity using the new prompt template for comparison with the original template.

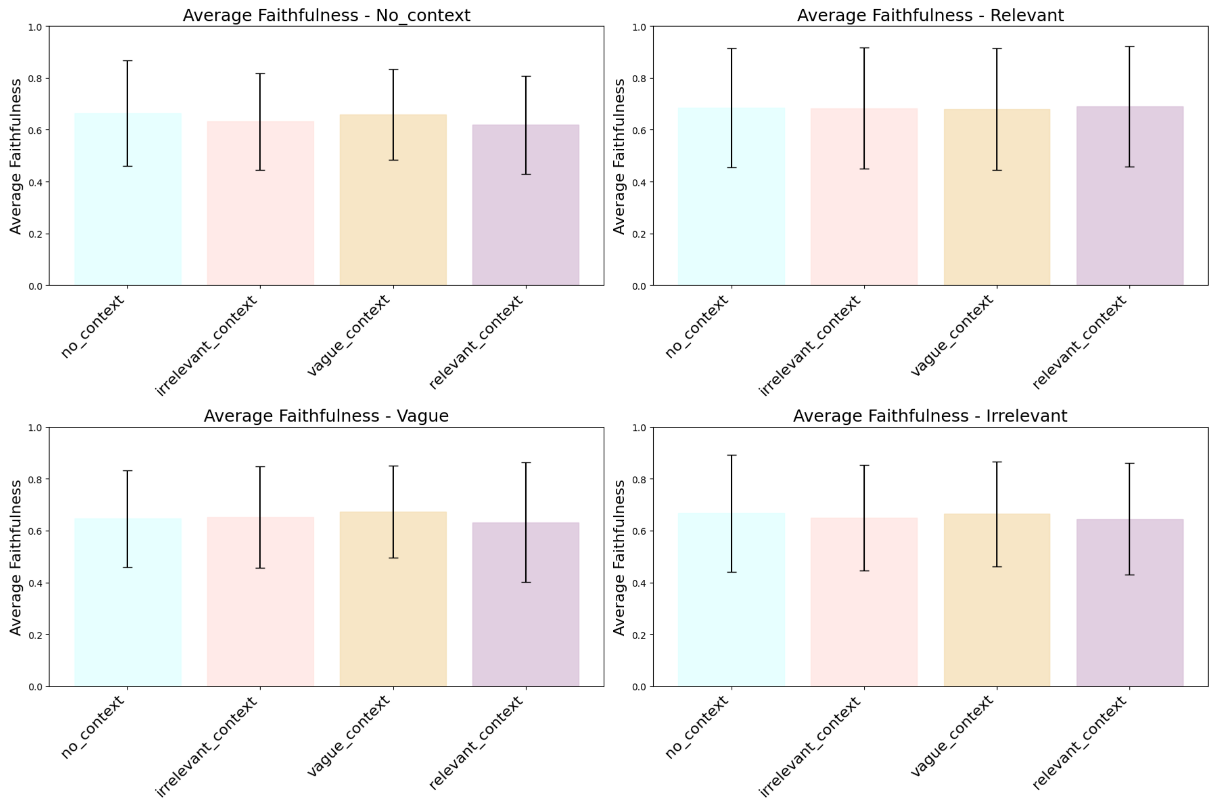

Figure 13: Faithfulness mapping between responses from new prompt and original ones. Each plot shows the mapping of the new prompt with one specific context type against the four different context types from the original response.