Evaluating Explanations Through LLMs: Beyond Traditional User Studies

Francesco Bombassei De Bona
Università della Svizzera italiana
francesco.bombassei.de.bona@usi.ch

Gabriele Dominici
Università della Svizzera italiana
gabriele.dominici@usi.ch

Tim Miller
The University of Queensland
timothy.miller@uq.edu.au

Marc Langheinrich
Università della Svizzera italiana
marc.langheinrich@usi.ch

Martin Gjoreski
Università della Svizzera italiana
martin.gjoreski@usi.ch

Abstract

As AI becomes fundamental in sectors like healthcare, explainable AI (XAI) tools are essential for trust and transparency. However, traditional user studies used to evaluate these tools are often costly, time consuming, and difficult to scale. In this paper, we explore the use of Large Language Models (LLMs) to replicate human participants to help streamline XAI evaluation. We reproduce a user study comparing counterfactual and causal explanations, replicating human participants with seven LLMs under various settings. Our results show that (i) LLMs can replicate most conclusions from the original study, (ii) different LLMs yield varying levels of alignment in the results, and (iii) experimental factors such as LLM memory and output variability affect alignment with human responses. These initial findings suggest that LLMs could provide a scalable and cost-effective way to simplify qualitative XAI evaluation.

1 Introduction

As artificial intelligence (AI) becomes integrated into critical sectors such as healthcare [1–3], the adoption of explainable AI (XAI) becomes inevitable [4–6]. For example, AI models can help diagnose diseases [7], predict patient outcomes [8], and recommend treatments [9]. The decisions of these models are often opaque, making it difficult for practitioners to fully trust or understand them. Therefore, XAI tools can have a huge impact in the integration of AI in healthcare. This necessity is also highlighted by regulatory efforts such as the EU AI Act [10], which enforces transparency and accountability in AI systems, particularly in critical sectors, where understanding AI-driven decisions can mean the difference between life and death.

This need for effective XAI tools has led to a significant number of studies aimed at advancing the field [11]. Many of these efforts have focused mainly on developing new techniques and algorithms [12–17] to explain models and evaluate them through quantitative metrics. However, this approach holds significant challenges, as there are no clear and unique metrics (e.g., surrugate model fidelity [12], counterfactual validity [17], and proximity score [18]) to evaluate these tools. The choice of metrics is often highly dependent on the specific XAI technique and the domain of application, and these metrics frequently fail to capture the actual benefits from the end-user's perspective. As a result, many tools are optimized to maximize performance in these quantitative metrics, ignoring the ultimate goal of providing explanations that help users understand the model's decisions [19]. In contrast, fewer studies involve qualitative evaluations in which users assess key properties such as the effectiveness, helpfulness, and trustworthiness of the explanations [20–24]. Furthermore, there is no standardized process for structuring these evaluations, leading to inconsistencies in the

GenAI for Health Workshop @ NeurIPS 2024, Vancouver.

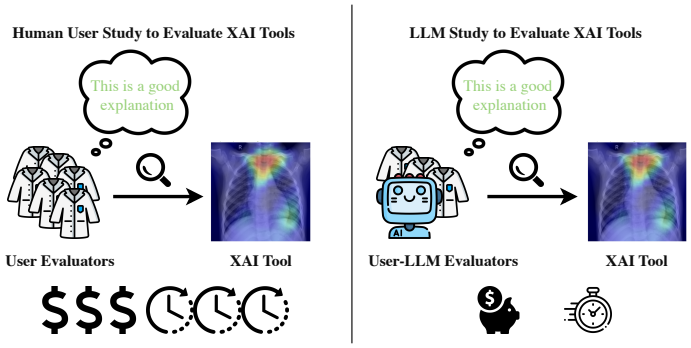

Figure 1: Our vision in comparing Human and LLM Evaluators for XAI Tool Effectiveness

way user studies are conducted. Consequently, running user studies tends to be not only costly and time consuming but also prone to producing variable outcomes, which limits their scalability and reproducibility. These challenges create barriers that result in fewer qualitative evaluations and slow down progress in the field.

Under these circumstances, Large Language Models (LLMs) offer a promising way to complement user studies. First, LLMs can help researchers run smaller, more focused studies by integrating their results with LLM-generated data, reducing the need for large-scale participant recruitment. This streamlines the evaluation of XAI tools while ensuring alignment with human preferences. Second, LLMs are useful in expert-driven studies, where recruiting participants like clinicians is challenging. Instead of relying on large groups of laypeople for early-stage feedback—which can reduce the validity of the study—LLMs can provide preliminary insights, allowing researchers to refine tool designs before engaging experts, saving time and resources. For example, a healthcare institution developing machine learning models to detect brain cancer via MRI scans must validate the model by understanding its decision-making process. XAI techniques are essential for this, as they provide insights into the model's reasoning. To ensure the model's explanations align with clinical expectations, a user study involving practitioners is crucial. In this context, LLMs can streamline the process by pre-evaluating the XAI outputs, ensuring that the model's decisions are coherent before full expert validation, saving time and resources.

This paper explores the potential of LLMs to bridge the gap between the need for scalable XAI evaluation and the limitations of traditional user studies. Specifically, we aim to replicate a user study which compared counterfactual and causal explanations in terms of their helpfulness and effectiveness in transmitting insights from AI systems. However, instead of human participants, we use LLMs and explore whether the LLM-generated results align with the conclusions drawn from the original user study. We evaluated seven of the most advanced LLMs — Llama 3 (8B and 70B) [25], Qwen 2 (7B and 72B) [26], Mistral 7B [27], Mistral Nemo and GPT-4o Mini [28] - in various experimental settings. These settings included leveraging LLMs memory and exploring the effects of LLM variability in generating answers to understand their impact in the alignment between LLMs and humans preferences. The results of our experiments demonstrate that: (i) LLMs can replicate most of the conclusions from the original user study, (ii) different LLMs can lead to varying conclusions depending on the architecture and capabilities of the model, and (iii) the experimental setup, such as the use of memory or randomness, can significantly impact the extent to which LLM responses align with human responses. These initial findings provide promising insights into the feasibility of developing automatic, scalable, and cost-effective qualitative evaluation frameworks that rely on LLMs as an alternative to traditional user studies.

2 Traditional user studies to evaluate XAI tools

In the evaluation of XAI tools, particularly within healthcare, user studies are considered the gold standard for assessing how well these tools perform in real-world scenarios [29]. Typically, these studies involve a structured process in which healthcare professionals or end users interact with XAI tools under controlled conditions. In previous user studies that evaluated XAI tools [20–23], participants were assigned with activities such as predicting the output of the AI model based on AI-generated explanations and/or completing questionnaires about the perceived usefulness of the explanations and the degree of trust in the model.

For instance, Colin et al. [20] asked participants to guess the prediction of the model based on the provided explanation to evaluate how various XAI techniques [14, 30] help users detect biases, identify strategies for solving unknown tasks, and recognize model failures. Their user study included scenarios with inherent biases, tasks in which users lacked domain expertise, and models prone to misclassifying specific examples. The primary objective was to assess which explanation strategies were most effective in helping users replicate the model's decision-making process. The assumption was that if an explanation was understandable and complete, users would be able to predict the model's decisions accurately. The conclusions were drawn by comparing the performance of different explanation techniques against each other, as well as against a baseline scenario where no explanations were provided. Similarly, Karagoz et al. [23] asked professionals to report their level of trust in an AI model both before and after receiving an explanation [12, 13] of the decision of the model. Participants were also asked to make predictions based on the explanation. Additionally, the study examined the level of agreement between practitioners who made decisions using XAI tools versus those who did not. As with the previous study, the researchers used statistical tests to compare the results across different explanations and settings (pre- and post-explanation). In a different approach, Singh et al. [21] designed a study in which participants filled out a questionnaire with several items regarding the actionability of the explanations they were shown. The aggregated results were compared to determine which types of explanations were perceived as more actionable. Their findings demonstrated that this setting effectively highlighted which explanation type was considered most actionable by participants. These user studies demonstrate the diversity of metrics and settings used (from performing tasks to giving opinion through a questionnaire) in the evaluation of XAI tools, highlighting the lack of a universally accepted structure or standardized metrics for conducting user studies in this field. Despite this variability, one common thread across all the studies is the comparison of aggregated results for different XAI tools in hypothetical real-world scenarios. This ability to simulate real-world use cases is the primary advantage of user studies.

However, despite their importance, user studies have several significant drawbacks. Conducting these studies is often resource intensive and requires substantial time and financial investment to recruit participants and simulate usage environments (e.g., clinical decision making in the context of healthcare). The involvement of domain experts (e.g., clinicians), who may already have demanding schedules, further complicates the process. Furthermore, variability in user experience and interpretation can lead to inconsistent results, making it challenging to draw generalizable conclusions, especially if the number of participants is relatively low. This variability also limits the scalability of user studies, as replication across different institutions, or with larger groups, can be prohibitively expensive and time consuming.

3 Can LLMs evaluate XAI tools?

LLMs are advanced AI systems designed and trained to process text data and generate human-like text based on vast amounts of data, covering a wide range of topics. This training enables them to estimate context, generate coherent answers, and mimic human-like reasoning in their responses. LLMs excel in tasks that require the comprehension and generation of natural language, making them particularly effective in simulating human interactions [31], such as those involved in user studies. LLMs have also demonstrated significant performance in tasks beyond their original training without the need for additional fine-tuning. For example, LLMs can classify various data types, such as tabular data [32] and time series [33], or generate synthetic data [34]. Although they may not yet represent the state-of-the-art models for these tasks (with specialized models often being more capable), they offer valuable versatility. This is especially relevant when dealing with niche domains like healthcare, where specialized models are typically preferred for specific tasks due to their superior and tailored capabilities. However, LLMs can offer additional capabilities in conjunction with specialized models and tools, leveraging their human-like conversational abilities and contextual understanding.

To exploit these abilities in conjunction with specialized models and XAI tools, we propose using LLMs to qualitatively evaluate XAI tools, simulating human-based studies. Querying LLMs instead of humans offers several significant advantages. Firstly, LLMs provide a cost-effective alternative, eliminating the need for expensive and time-consuming recruitment and compensation of human participants. This allows for the rapid and inexpensive execution of studies that can be conducted repeatedly with unlimited queries, leading to large-scale data collection and analysis. In general,

including the use of LLMs in the user study would dramatically increase the flexibility and scalability of the research process regarding the development of XAI tools.

However, achieving useful results with LLMs is only possible if the LLM model is properly aligned with human preferences. Ensuring such alignment is one of the key challenges in the development of more powerful LLMs today. This alignment is typically achieved through additional training [35] or techniques [36]. These differences in the alignment process can significantly affect the way LLMs generate responses, influencing their reasoning processes, beliefs, and preferences. This variability in alignment becomes a critical consideration in scenarios where the primary goal is to accurately mimic user preferences, such as in our evaluations of XAI tools. In particular, we are interested in assessing two types of alignment: *general alignment*, where we determine the outcome of statistical comparisons (e.g., explanation A is generally perceived as better than explanation B), and *absolute alignment*, where we measure specific ratings (e.g., explanation A is rated a 4 on a scale of 1 to 5 for helpfulness). In both cases, the choice of which LLM to use is not trivial. Different LLMs can exhibit varying degrees of alignment depending on the training process and the intended application. Thus, selecting the most suitable LLM is a critical design decision that can significantly impact the effectiveness and reliability of the user study.

In addition to alignment, several other factors influence how LLMs generate responses. These include varying the initial prompts, employing different prompting techniques (e.g., zero-shot, few-shot [37], chain-of-thought [38]) and leveraging session memory. The structure and clarity of the input prompt can significantly affect the coherence and relevance of the model output. For example, various types of prompt injection can be used to alter the generation process, incorporating elements such as personalization, task description, and context specification. Then, conversation memory can impact LLMs' performances by allowing the models to exploit prior generated information, increasing the chance of alignment (as also the human evaluator is influenced by the past examples) and also modifying the variability of the output. If memory is used, then the order of instruction also plays an important role in conditioning the results. These factors contribute to the variability in LLM generation, further highlighting the importance of carefully selecting and configuring the model to better align with specific user preferences and study goals.

4 Experiments

We designed our experiments to explore whether it is feasible to estimate human preferences in XAI user studies using LLMs. More specifically, we tested for *general alignment* (i.e., can we arrive to the same findings/conclusions on a population/study level) and *absolute alignment* (i.e., can we come to similar responses on a case-by-case level) between our LLM-based study and an existing use-based study. Additionally, we aim to explore the impact of different factors that might influence the outcomes of such studies when using LLMs. To achieve this, we address the following research questions:

- **Alignment**: Is it possible to replicate the results of an XAI user study using LLMs? Are LLMs answers aligned with human preferences in the context of evaluating explanations in absolute terms?
- **LLM Variability**: Do different LLMs lead to different general alignment?
- **Framework Variability**: Does the use of memory in LLMs influence the results of the user study? How does variability in LLM responses impact the general alignment?

4.1 Evaluation Setting

4.1.1 User study

As a foundation for the experimental setting, we use the first set of experiments of the user study by Celar and Byrne [22]. Their study was designed to better understand the relationship between causal and counterfactual explanations in terms of their helpfulness and effectiveness in transmitting insights from AI systems to end users. Participants interacted with the predictive AI system's input and output, along with explanations from an XAI tool providing insights into the system's decision-making process. The study featured four experimental conditions: *counterfactual* versus *causal* explanations in *high* versus *low familiarity* scenarios. *Counterfactual* explanations provide alternative scenarios by

answering "what if" questions and describing how the world would have to change for a desirable outcome to occur. In contrast, *causal explanations* describe the direct cause-and-effect relationships that led to the observed outcome. In the *high familiarity* condition, participants determine whether alcohol levels are over or under the legal limit for driving. In contrast, the *low familiarity* condition requires participants to assess the safety of an unknown chemical compound.

In each experimental setting of their study, users completed three tasks. In the first, they rated the helpfulness of a given explanation based on the input and output of the AI system. In the second task, participants attempted to make the prediction themselves using the provided explanation. Finally, in the third, they expressed their confidence level in their prediction on the Likert scale: "not at all confident", "not very confident", "neither", "fairly confident", and "very confident".

The first two tasks consisted of sixteen cases each. In the *high familiarity* scenario, the case comprised the following fields: name of the subject, weight, units of alcohol consumed, duration, gender, and stomach content. In the *low familiarity* setting, the case comprised the name of the chemical, occupational exposure limit, pH, exposure duration, air pollution rating, and PNEC Rating. Participants completed the first task on sixteen cases, followed by the second task on sixteen new cases. The third task questions were interleaved between the second task cases in equal numbers, ensuring sequential progression. In the first task, each case was paired with a prediction and an explanation, either causal or counterfactual, and the user judged the statement "This explanation was helpful" on a Likert scale: "strongly disagree", "disagree", "neutral", "agree", and "strongly agree". In the second task, only the case information was shown to enable the user to make their prediction, either over the limit/under the limit or safe/not safe.

4.1.2 Estimating human preferences with LLMs

To replicate the above-mentioned user study [22], we transpose the experimental setting designed for human evaluators into a compatible setting for LLMs. In this context, each LLM is treated as a participant, tasked with generating responses across the same experimental conditions (high/low familiarity and causal/counterfactual explanations).

LLM models are used to generate responses in place of human participants. A run corresponds to the execution of one experiment formed by the helpfulness, prediction, and confidence tasks. A state refers to the specific conditions under which the model generates responses, such as the combination of familiarity and explanation type for each task. We use two approaches to aggregate and compare the results across different models and runs. In the first approach, we conduct multiple inference runs for each model and calculate the mean response at each question. This method, similar to the self-consistency technique, helps to mitigate the variability in the generated outputs and produces a more stable, reliable average response for each experimental condition. By averaging across multiple runs, we reduce the impact of any outlier responses and ensure consistency in the model's performance. In the second approach, we treat each inference run as if it were generated by a different participant. This method simulates the diversity typically seen in a group of human participants. By treating each run independently, we capture the variability that may arise in model-generated responses, mirroring how individual differences exist in human participants.

In addition, the use of instruction-following models enables us to explore the influence of conversation history during task execution. We assume that while performing the tasks, users undergo a learning process. Thus, to replicate the same process, we enable the LLM to use previously generated answers and instructions as context for every new inference. We test this scenario against a baseline where the models do not have access to any previously generated answers and inputs, and every inference is treated as a separate task. We refer to these two scenarios as *"with memory"* and *"in isolation"*, respectively. The isolation setting poses two challenges. The first issue arises in the third task, where the LLM is asked to provide a confidence level about the prediction made in the previous task. However, if each inference is treated without memory, the LLM has no information about the previous task or the answer given. As a result, it cannot express a confidence level about its performance. To address this issue, we switch to a hybrid approach regarding memory by allowing the second and third tasks to be completed in pairs and enabling the conversation history between the two tasks. The second issue involves the assumption that the users, while performing the first task learn how to make the predictions in the second task. This assumption is violated due to the lack of presence of the conversation memory. To address this issue we apply a few-shot prompting technique. This method uses synthetic data of input-output examples to instruct the LLM on what answer is expected and

the way reasoning should be done. In our scenario, we take the sixteen cases from the first task, pair them with the corresponding prediction as if made by the same LLM, and use them as context for the input of the second task.

Given the two initial settings (familiar and not familiar) for our experiments, we feed LLMs with specific prompts to fit each task and each conversation memory setting. Specifically, we propose the following prompts, where we switch the familiarity setting based on the experiment and inject the cases illustrated previously:

First task prompt

Given the following case, how would you rate the sentence "This explanation was helpful"? You must answer by only providing one value from the following: "Strongly disagree", "Disagree", "Neutral", "Agree", "Strongly agree". {case}

Second task prompt

Complete the sentence "Based on the information provided, I believe the app's prediction for this person|chemical will be ...". You must answer by only providing one value from the following: Over the limit, Under the limit | Safe, Not safe. {case}

Third task prompt

How confident are you in your prediction? You must answer by only providing one value from the following: "Not at all confident", "Not very confident", "Neither", "Fairly confident", "Very confident".

Our final consideration concerns the impact of case ordering during inference in the memory setting. Since the original user study used different permutations of the cases for each participant, we apply the same approach to the LLM to ensure consistency. This allows us to account for any potential influence that case order might have on performance or results. This introduces the concept of "LLM-user." An LLM-user refers to the aggregation of results obtained across multiple inference runs, where a fixed permutation of cases is used for each group of runs. The results from multiple LLM-users are then combined to form the study's overall conclusions.

4.1.3 Metrics and statistical tests

Our primary objective is to compare the results obtained from LLMs with those from the original user study. To ensure consistency, we use the same statistical metrics and tests as those proposed in the original user study [22].

The original study compares four experimental conditions: low familiarity with causal explanations, low familiarity with counterfactual explanations, high familiarity with causal explanations, and high familiarity with counterfactual explanations. The analysis focuses on the mean responses provided by participants, aiming to show that high familiarity scenarios lead to better outcomes compared to low familiarity ones and that counterfactual explanations are generally more helpful and insightful than causal explanations. We replicate this approach by calculating the mean values for the LLM-generated responses in each condition and comparing the outcomes across the four experimental scenarios. This allows us to get information about the absolute alignment between LLMs and humans.

However, the main conclusion of the paper are drawn on statistical tests regarding the effect of the two primary variables — familiarity (high vs. low) and explanation type (causal vs. counterfactual) — and their interaction. Therefore, we apply a two-way ANOVA statistical test [39], just as in the original study, to assess whether these factors significantly influence the responses of LLMs, as they do for human participants (general alignment). This method allows us to investigate whether LLMs exhibit similar patterns of reasoning and judgment, even if they do not present an absolute alignment. Since we aim to assess the alignment between LLMs and the user study, we assume that the answer distributions match in shape and, therefore, we test for normality of the LLMs' answers. The ANOVA test assumptions are only partially satisfied for certain models, specifically Mistral-Nemo-Instruct-2407, Mistral-7B-Instruct-v0.3, and GPT-4o-Mini. As a result, we exercise caution and refrain from drawing strong conclusions based on the statistical outcomes for these models.

By aligning our evaluation techniques with those used in the original work, we aim to determine how well LLMs replicate human reasoning processes and whether they demonstrate the same preferences and patterns when presented with familiar or unfamiliar scenarios, as well as causal or counterfactual explanations.

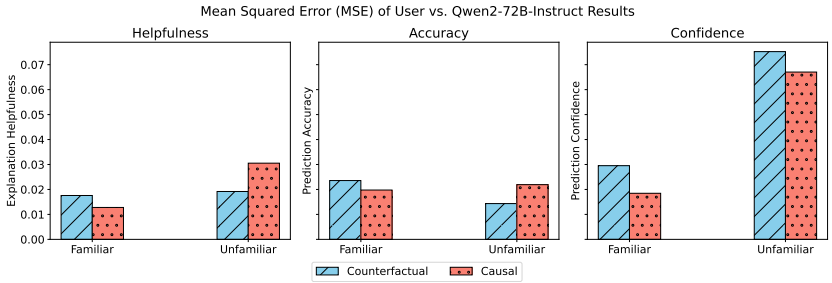

Figure 3: Bar charts representing the MSE between users and Qwen2-72B in the three tasks

4.2 Results

Evaluating XAI tools with LLMs partially mirrors user study conclusions. (Figure 2) Using the results of the Qwen 2 72B model (sampling the same number of "LLM-users" as participants in the original user study), we successfully replicated 6 of the 9 statistical outcomes from the first part of the user study. Figure 2 illustrates the concordance between the LLM results and those of the original human-based study for each statistical test. This shows that, under specific conditions, partial alignment between human and LLM conclusions can be achieved when replacing human participants with LLMs. Specifically, we observed perfect general alignment in tasks that require the prediction of the model output given an explanation. However, alignment was more challenging in tasks that involved confidence-related questions, where the LLM struggled to match human responses.

LLMs exhibit slight differences in absolute alignment with human preferences. (Figure 3) Although LLMs can replicate overall trends in user studies, their responses still show some deviations from human participants. Figure 3 presents the MSE of the Qwen 2 72B model across different categories (helpfulness, accuracy, confidence) in both familiar and unfamiliar conditions under causal and counterfactual settings. While the MSE for helpfulness is relatively low, particularly in familiar contexts, the model struggles more with accuracy, especially in unfamiliar settings. Confidence shows the largest errors, mainly in unfamiliar conditions. These results suggest that, although the model's absolute predictions differ from human results, its comparative judgments remain consistent, showing potential as an alternative in user evaluations across different scenarios.

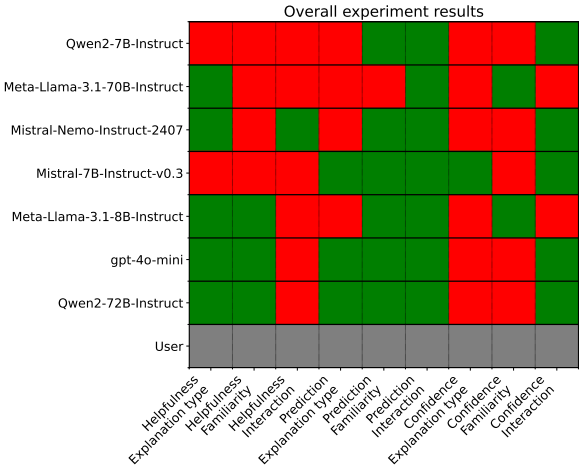

Figure 2: Concordance of the results for each statistical test. Concordance is computed by merging the results of the ANOVA test and the results of the comparison of the averaged values.

Different LLMs exhibit varying levels of general alignment with human responses. (Figure 2) The degree of general alignment between LLMs and human participants varies between models, reflecting differences in size, capabilities, and behaviors. Figure 2 shows that some LLMs align more closely with human judgments in specific tasks, while others diverge. Among the tested models, Qwen 2 72B and GPT-4o Mini achieved the highest general alignment, matching human conclusions in 6 out of 9 cases. Interestingly, the smaller LLaMA 8B demonstrated better general alignment than its larger counterpart, LLaMA 70B. In contrast, the largest Qwen 2 model (72B) significantly outperformed the smaller Qwen 2 7B, indicating that a larger model does not necessarily guarantee better alignment with human responses across all architectures. Additionally, all models showed their best performance on the prediction tasks, while they performed worst in the confidence-related

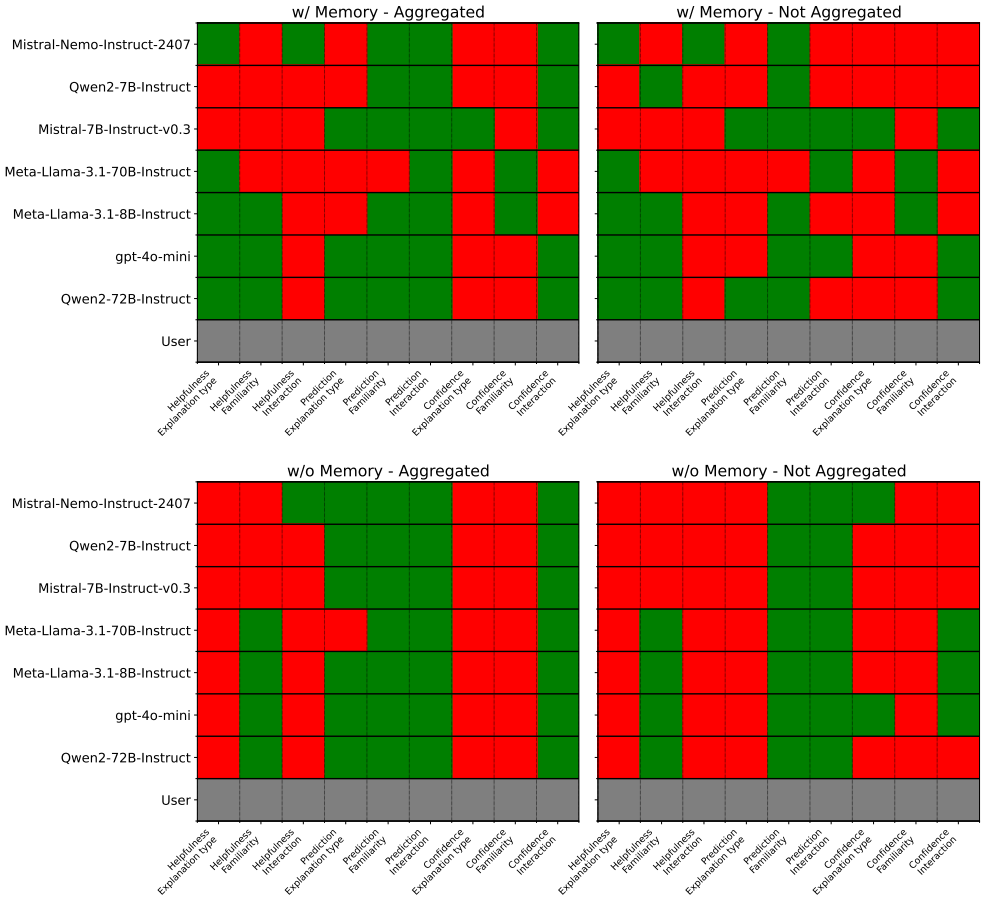

Figure 4: Concordance of the results for each statistical test aggregated by experimental settings. Experimental settings explored comprise conversation with or without conversation memory and usage of aggregated inference runs.

tasks. This may be because LLMs are trained to answer specific questions (such as predicting outcomes) based on provided context (an explanation). Confidence questions, on the other hand, involve subjective judgments that are more challenging for LLMs to mimic, as they may struggle to express confidence levels—since their responses always represent the most probable outcome, not a measure of certainty. This suggests that while LLMs can replicate decision-making tasks, more nuanced, subjective metrics like confidence may require further refinement.

Usage of memory impacts LLMs' general alignment. (Figure 4) Figure 4 illustrates the concordance between the results of the LLMs and those of the original human-based study in different settings, specifically comparing models with and without memory use. LLMs that utilize memory behave differently from those that do not. Generally, LLMs without memory exhibit more uniform performance across all tests, likely due to the absence of influence from the prior context. In contrast, the use of memory introduces variability, as the model's responses are affected by the way it interprets and incorporates information from previous interactions. Despite this variability, models that employed memory tended to perform better overall, showing higher concordance with human judgments. This suggests that memory, when used effectively, can enhance an LLM's ability to align with human responses. However, this benefit depends on how well the memory mechanism is utilized.

Simulating different users through aggregation leads to opposite results. (Figure 4) Evaluating the impact of aggregation allows us to better assess the impact of LLM generation variability. Figure 4 compares the outcomes of different models using both aggregated and non-aggregated methods. The results reveal that prior to aggregation, different LLMs exhibit varying degrees of initial variability,

as they show different level of agreement with the aggregated results. Therefore, the aggregation method significantly influences the alignment of the LLM results. For instance, models with memory that use aggregation closely mirror the human user results. However, when non-aggregated responses are considered, the variability across runs increases, leading to more divergent results, particularly in confidence ratings and prediction accuracy. The non-aggregated approach without memory produces results that deviate significantly from the original user study. This suggests that relying on individual, un-aggregated LLM responses introduces greater variability, making it difficult to replicate human users. By contrast, aggregated models, especially those using memory, maintain more consistent performance. As shown in Figure 4, by using aggregation, we observe an increase in the alignment of LLMs with the user study in 11 out of 14 cases while the other 3 maintain the original alignment with the non-aggregated models.

5 Limitations

One significant limitation of our research is that it is based on a single publicly available user study, which focused on evaluating explanations generated by two similar XAI methods. The use of only two explanation techniques in a single user study limits the breadth of the conclusions we can draw. XAI encompasses a wide variety of techniques and applications across different domains, and our findings may not generalize to all types of explanations or contexts. Moreover, our approach cannot replicate certain types of human studies, such as qualitative interviews or assessments that measure real trust, particularly in expert-driven fields like healthcare, where the trust of clinicians is crucial. Similarly, this approach would struggle to handle entirely new domains that fall outside the LLM's training set, as the models rely on prior knowledge to generate responses. However, this study highlights an interesting path for further exploration. We plan to extend the experimental settings in future work to include a broader range of XAI techniques and user studies. Additionally, we aim to incorporate queries to Vision Language Models (VLMs) to evaluate visually oriented XAI techniques, such as saliency maps, which are important in the healthcare field.

Another limitation is the possibility that the tasks or responses from the original Celar and Byrne [22] study may be part of the LLM training set. This could introduce bias and compromise the validity of the results. Ensuring that the LLM responses are genuinely independent of the study's prior knowledge will be critical in addressing this limitation.

Lastly, a limitation lies in the specific LLMs used in this study. Although we used up-to-date LLMs at the time of evaluation, the rapid pace of advancements in AI technology, especially in LLMs, means that future models may exhibit different behavior, reasoning abilities, or alignment capabilities. Similarly, improvements in alignment techniques could lead to better alignment with human preferences, potentially altering the conclusions drawn in this study. As a result, future research will need to inspect LLM performance again as new models and alignment methods emerge. Nonetheless, this paper serves as an illustration of this idea, providing promising results and laying the groundwork for further investigation.

6 Conclusions

In conclusion, our investigation into the use of Large Language Models (LLMs) to complement and integrate user studies for evaluating Explainable AI (XAI) tools offers promising initial results. By replicating a user study on counterfactual and causal explanations, we found: (i) LLMs can replicate most of the conclusions derived from traditional user studies, indicating their potential as scalable and cost-effective alternatives, (ii) different LLM architectures and capabilities can produce varying outcomes, emphasizing the importance of selecting appropriate models for specific evaluation tasks, and (iii) experimental factors, such as the use of memory and the impact of variability in generating responses, significantly affect the alignment between LLM and human preferences. Our findings suggest LLM-based evaluations could greatly improve the scalability and reproducibility of XAI assessments. Future work should aim to enhance the alignment of LLMs with human judgment in the evaluation of XAI tools and explore the broader applicability of this approach across various XAI techniques and domains.

Acknowledgments and Disclosure of Funding

This study was funded by the Swiss National Science Foundation, through the projects XAI-PAC (PZ00P2_216405) and TRUST-ME (205121L_214991).

References

[1] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, December 2017. URL http://arxiv.org/abs/1711.05225. arXiv:1711.05225 [cs, stat].

[2] Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network. *Nature medicine*, 25(1):65–69, January 2019. ISSN 1078-8956. doi: 10.1038/s41591-018-0268-3. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6784839/.

[3] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1:18, 2018. ISSN 2398-6352. doi: 10.1038/s41746-018-0029-1.

[4] Zahra Sadeghi, Roohallah Alizadehsani, Mehmet Akif Cifci, Samina Kausar, Rizwan Rehman, Priyakshi Mahanta, Pranjal Kumar Bora, Ammar Almasri, Rami S. Alkhawaldeh, Sadiq Hussain, Bilal Alatas, Afshin Shoeibi, Hossein Moosaei, Milan Hladik, Saeid Nahavandi, and Panos M. Pardalos. A Brief Review of Explainable Artificial Intelligence in Healthcare, April 2023. URL http://arxiv.org/abs/2304.01543. GSCC: 0000367 arXiv:2304.01543 [cs].

[5] Deepti Saraswat, Pronaya Bhattacharya, Ashwin Verma, Vivek Kumar Prasad, Sudeep Tanwar, Gulshan Sharma, Pitshou N. Bokoro, and Ravi Sharma. Explainable AI for Healthcare 5.0: Opportunities and Challenges. *IEEE Access*, 10:84486–84517, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3197671. URL https://ieeexplore.ieee.org/document/9852458. GSCC: 0000182 Conference Name: IEEE Access.

[6] Ramasamy Mariappan. Extensive Review of Literature on Explainable AI (XAI) in Healthcare Applications. *Recent Advances in Computer Science and Communications*, 17, March 2024. ISSN 26662558. doi: 10.2174/0126662558296699240314055348. URL https://www.eurekaselect.com/228159/article. GSCC: 0000367 0 citations (Crossref/DOI) [2024-08-29].

[7] Richard J. Chen, Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, and Faisal Mahmood. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878.e6, August 2022. ISSN 1535-6108. doi: 10.1016/j.ccell.2022.07.004. URL https://www.sciencedirect.com/science/article/pii/S1535610822003178.

[8] Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. HEALNet – Hybrid Multi-Modal Fusion for Heterogeneous Biomedical Data, November 2023. URL http://arxiv.org/abs/2311.09115. arXiv:2311.09115 [cs].

[9] Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, November 2018. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-018-0213-5. URL https://www.nature.com/articles/s41591-018-0213-5.

[10] Cecilia Panigutti, Ronan Hamon, Isabelle Hupont, David Fernandez Llorca, Delia Fano Yela, Henrik Junklewitz, Salvatore Scalzo, Gabriele Mazzini, Ignacio Sanchez, Josep Soler Garrido, and Emilia Gomez. The role of explainable AI in the context of the AI Act. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 1139–1150, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594069. URL https://dl.acm.org/doi/10.1145/3593013.3594069. GSCC: 0000110.

[11] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58: 82–115, June 2020. ISSN 1566-2535. doi: 10.1016/j.inffus.2019.12.012. URL https://www.sciencedirect.com/science/article/pii/S1566253519308103.

[12] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, August 2016. URL http://arxiv.org/abs/1602.04938. arXiv:1602.04938 [cs, stat].

[13] Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions, November 2017. URL http://arxiv.org/abs/1705.07874. arXiv:1705.07874 [cs, stat].

[14] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, February 2020. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://arxiv.org/abs/1610.02391. arXiv:1610.02391 [cs].

[15] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept Bottleneck Models, December 2020. URL http://arxiv.org/abs/2007.04612. arXiv:2007.04612 [cs, stat].

[16] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), June 2018. URL http://arxiv.org/abs/1711.11279. arXiv:1711.11279 [stat].

[17] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR, March 2018. URL http://arxiv.org/abs/1711.00399. arXiv:1711.00399 [cs].

[18] Victor Guyomard, Françoise Fessant, Thomas Guyet, Tassadit Bouadi, and Alexandre Termier. VCNet: A self-explaining model for realistic counterfactual generation, December 2022. URL http://arxiv.org/abs/2212.10847. arXiv:2212.10847 [cs].

[19] Mark T. Keane, Eoin M. Kenny, Eoin Delaney, and Barry Smyth. If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques, February 2021. URL http://arxiv.org/abs/2103.01035. arXiv:2103.01035 [cs].

[20] Julien Colin, Thomas Fel, Remi Cadene, and Thomas Serre. What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods, January 2023. URL http://arxiv.org/abs/2112.04417. arXiv:2112.04417 [cs].

[21] Ronal Singh, Tim Miller, Liz Sonenberg, Eduardo Velloso, Frank Vetere, Piers Howe, and Paul Dourish. An Actionability Assessment Tool for Explainable AI, June 2024. URL http://arxiv.org/abs/2407.09516. arXiv:2407.09516 [cs].

[22] Lenart Celar and Ruth M. J. Byrne. How people reason with counterfactual and causal explanations for Artificial Intelligence decisions in familiar and unfamiliar domains. *Memory & Cognition*, 51(7):1481–1496, October 2023. ISSN 1532-5946. doi: 10.3758/s13421-023-01407-5. URL https://doi.org/10.3758/s13421-023-01407-5.

[23] Gizem Karagoz, Geert van Kollenburg, Tanir Ozcelebi, and Nirvana Meratnia. Evaluating How Explainable AI Is Perceived in the Medical Domain: A Human-Centered Quantitative Study of XAI in Chest X-Ray Diagnostics. In Hao Chen, Yuyin Zhou, Daguang Xu, and Varut Vince Vardhanabhuti, editors, *Trustworthy Artificial Intelligence for Healthcare*, pages 92–108, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-67751-9. doi: 10.1007/978-3-031-67751-9_8.

[24] Yao Rong, Tobias Leemann, Thai-trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. Towards Human-centered Explainable AI: A Survey of User Studies for Model Explanations, December 2023. URL `http://arxiv.org/abs/2210.11584`. arXiv:2210.11584 [cs].

[25] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, February 2023. URL `http://arxiv.org/abs/2302.13971`. arXiv:2302.13971 [cs].

[26] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 Technical Report, July 2024. URL `http://arxiv.org/abs/2407.10671`. arXiv:2407.10671 [cs].

[27] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, October 2023. URL `http://arxiv.org/abs/2310.06825`. arXiv:2310.06825 [cs].

[28] OpenAI. GPT-4 API, 2024. URL `https://platform.openai.com`.

[29] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning, March 2017. URL `http://arxiv.org/abs/1702.08608`. arXiv:1702.08608 [cs, stat].

[30] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, April 2014. URL `http://arxiv.org/abs/1312.6034`. arXiv:1312.6034 [cs].

[31] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang,

Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report, March 2024. URL `http://arxiv.org/abs/2303.08774`. arXiv:2303.08774 [cs].

[32] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. TabLLM: Few-shot Classification of Tabular Data with Large Language Models, March 2023. URL `http://arxiv.org/abs/2210.10723`. arXiv:2210.10723 [cs].

[33] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large Language Models Are Zero-Shot Time Series Forecasters, August 2024. URL `http://arxiv.org/abs/2310.07820`. arXiv:2310.07820 [cs].

[34] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations, October 2023. URL `http://arxiv.org/abs/2310.07849`. arXiv:2310.07849 [cs].

[35] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, March 2022. URL `http://arxiv.org/abs/2203.02155`. arXiv:2203.02155 [cs].

[36] Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate, October 2018. URL `http://arxiv.org/abs/1805.00899`. arXiv:1805.00899 [cs, stat].

[37] 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. Language Models are Few-Shot Learners, July 2020. URL `http://arxiv.org/abs/2005.14165`. arXiv:2005.14165 [cs].

[38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, January 2023. URL http://arxiv.org/abs/2201.11903. arXiv:2201.11903 [cs].

[39] Lars St»hle and Svante Wold. Analysis of variance (ANOVA). *Chemometrics and Intelligent Laboratory Systems*, 6(4):259–272, November 1989. ISSN 0169-7439. doi: 10.1016/0169-7439(89)80095-4. URL https://www.sciencedirect.com/science/article/pii/0169743989800954.

[40] Harrison Chase. LangChain, October 2022. URL https://github.com/langchain-ai/langchain. original-date: 2022-10-17T02:58:36Z.

[41] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9 (3):90–95, 2007. doi: 10.1109/MCSE.2007.55. Publisher: IEEE COMPUTER SOC.

A Implementation details

Running local and remote models We executed Hugging Face models (Mistral, Llama 3.1, and Qwen2 families) on a local server equipped with the following hardware:

- **CPU**: 2 x AMD EPYC 7513 32-Core Processor
- **RAM**: 512 GB
- **GPU**: 4 x RTX A6000 (48 GB VRAM each)

CUDA acceleration was utilized to parallelize and distribute computation across the GPUs, significantly speeding up the processing.

The total inference time for running the full experiment was approximately 76 hours.

For GPT 4o Mini, inference was run using the OpenAI API. Inference time depends on the usage Tier available on the API.

Code and licenses Our code implementation is built using Python 3.12 and leverages the open-source library LangChain [40] (MIT License) to develop the inference infrastructure for both local and remote executions. All plots were generated using the Matplotlib [41] (BSD License) library. The dataset used in the experiment is freely available, following the guidelines provided in the original paper [22].

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See research questions and results in Section 4

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 5

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not contain theoretical results

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: See Section 4 for description of experimental setting

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the code and data

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Section 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: A brief description is available at Appendix A

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We followed the NeurIPS Code of Ethics

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Sections 1 and 2 describe the impact of integrating LLMs in human user-studies, their benefit and limitations.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release neither data or models

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See Appendix A

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We don't release any new asset

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We don't use crowdsourcing or research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: See previous answer

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.