

To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models

Jessica Y Bo
Computer Science
University of Toronto
Toronto, Ontario, Canada
jbo@cs.toronto.edu

Sophia Wan
Computer Science
University of Toronto
Toronto, Ontario, Canada
sophia.wan@mail.utoronto.ca

Ashton Anderson
Computer Science
University of Toronto
Toronto, Ontario, Canada
ashton@cs.toronto.edu

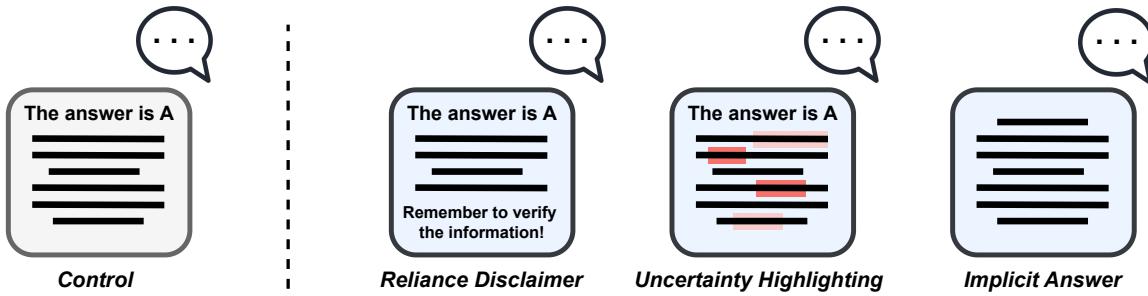


Figure 1: We evaluate the following LLM reliance interventions for their effectiveness at improving appropriate reliance on LLM-generated advice, in comparison to Control: Reliance Disclaimer, Uncertainty Highlighting, and Implicit Answer.

Abstract

As Large Language Models become integral to decision-making, optimism about their power is tempered with concern over their errors. Users may over-rely on LLM advice that is confidently stated but wrong, or under-rely due to mistrust. Reliance interventions have been developed to help users of LLMs, but they lack rigorous evaluation for appropriate reliance. We benchmark the performance of three relevant interventions by conducting a randomized online experiment with 400 participants attempting two challenging tasks: LSAT logical reasoning and image-based numerical estimation. For each question, participants first answered independently, then received LLM advice modified by one of three reliance interventions and answered the question again. Our findings indicate that while interventions reduce over-reliance, they generally fail to improve appropriate reliance. Furthermore, people became more confident after making wrong reliance decisions in certain contexts, demonstrating poor calibration. Based on our findings, we discuss implications for designing effective reliance interventions in human-LLM collaboration.

CCS Concepts

- Information systems → Information systems applications;
- Human-centered computing → Interaction design process and methods; Empirical studies in HCI.

Keywords

Large Language Models, Human-LLM Collaboration, Appropriate Reliance, Over-Reliance

ACM Reference Format:

Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2025. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan*. ACM, New York, NY, USA, 23 pages. <https://doi.org/10.1145/3706598.3714097>

1 Introduction

As people increasingly incorporate Large Language Models (LLMs) into their decision-making workflows, there is cause for both tremendous optimism and skeptical concern [23, 92]. LLMs encode remarkably vast amounts of knowledge; can communicate with users in clear, natural language; and have the potential to improve both the quality and efficiency of users' decision-making [75]. While LLMs are prone to hallucination and error, they can still mislead users into a false sense of trust with their impressive detail and confidence [13, 90]. On the other hand, algorithm aversion can cause users to also ignore helpful recommendations due to a general distrust towards technology [10, 21]. Consequently, both over-reliance [33, 71, 73] and under-reliance present significant yet contrasting challenges in human-LLM collaboration [89].

Conscious effort must be placed on helping users calibrate appropriate levels of trust and reliance on LLM advice [91, 93]. Neither over-reliance nor under-reliance is ideal; instead, the focus should

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI '25, Yokohama, Japan

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-1394-1/25/04

<https://doi.org/10.1145/3706598.3714097>

be to achieve appropriate reliance [30, 57, 66]. Due to the risk of hallucinations, reliance interventions in LLMs are often motivated by curbing over-relying on wrongful information [39, 71, 73], however recent attention has also shifted to optimizing for appropriate reliance as well [19, 22, 89, 93]. In this work, we adapt rigorous measures of appropriate reliance from prior works in human-AI collaboration [67] to benchmark the impact of several LLM reliance interventions in question-answering tasks.

The tangential field of human-AI collaboration has explored methods like explanations and calibration techniques to improve joint decision-making by enhancing AI transparency and verifiability, though results often show mixed outcomes and trade-offs [4, 48, 58, 68, 78]. Reliance and calibration interventions for LLMs can involve communicating the capabilities and confidence of an LLM to its users [73], or introducing various types of frictions that the user must bypass to receive the advice [19]. Since interventions are being developed and evolving at a rapid pace, many of these methods lack comprehensive evaluations of appropriate reliance in real users, especially those focusing on technical contributions [5, 8, 51]. Furthermore, people's subjective perceptions—such as their self-confidence when receiving advice from LLMs—can have a complex relationship with their reliance behaviour [12, 39, 54, 71], which warrants further exploration. Finally, the impact of different intervention types on reliance and perceptions remains unclear, as existing studies do not benchmark multiple interventions. We seek to address these gaps by asking the following research questions:

- **RQ1:** How do different reliance interventions affect users' tendencies to over- and under-rely on LLM advice?
- **RQ2:** Can reliance interventions appropriately calibrate users' confidence in LLM advice?

Summary of Contributions. We conduct a pre-registered, randomized online experiment with $n = 400$ participants where they attempted two diverse and challenging question-answering tasks: LSAT logical reasoning and image-based numerical estimation. Participants first attempted the tasks on their own, then were shown LLM advice that was randomly treated with one of four conditions (three reliance interventions and one baseline) and given the question again. To evaluate reliance, we implemented a rigorous set of evaluation metrics that isolate the decision-making process users undergo when choosing to rely on the LLM's advice or their own judgment.

We find that while interventions can significantly reduce over-reliance, they are generally *not effective at improving appropriate reliance*. These interventions tend to reduce reliance overall—reducing over-reliance at the expense of useful reliance. Participants were also generally miscalibrated in their confidence, reporting a higher increase in confidence when they *make the wrong reliance decision*. Out of all the interventions, **Reliance Disclaimer** is the only one to improve appropriate reliance and better calibrate confidence when participants relied on LLM advice, but this effect was only observed on the LSAT task. **Uncertainty Highlighting** worsened subjective perceptions and had the least effect on reliance out of the three interventions. **Implicit Answer** induced better self-reliance in both tasks, but at the cost of requiring significantly more time for

participants to process the information. We also discuss the implications of our results for the design and evaluation of appropriate reliance interventions for the future of human-LLM interactions. We highlight the need to perform rigorous, human-centered evaluations of technical LLM calibration methods, particularly with emphasis on achieving appropriate reliance. Furthermore, as the trade-off between balancing over- and under-reliance may be difficult to avoid, defining which outcome to prioritize should be done within the context of the specific human-LLM collaboration task.

2 Related Works

We provide an overview of the growing body of research in human-AI interactions, human-LLM collaboration, and LLM calibration techniques that have been explored in both human-computer interaction (HCI) and AI spaces.

2.1 Appropriate Reliance on AI

The need for appropriate reliance has been a constant over the years of technological development [46]. With the advancement of AI technologies, much attention has turned on how end users can incorporate assistive AI models into their decision-making tasks [14, 40, 45]. Given that AI predictions are not guaranteed to be accurate, the user must decide whether to or not to incorporate the AI's advice [12, 17, 32, 54, 61, 72]. The concept of appropriate reliance can typically be defined as "*relying on the AI when it's correct, and relying on yourself when it's not*"—which concerns reducing both over-reliance and under-reliance [67, 68]. Root causes for inappropriate reliance can come from misreliance (inherently nonoptimal rate of reliance) or discrimination errors (being unable to identify the payoff of following the AI relative to following oneself) [29, 57].

Factors that influence reliance include the expertise of the user in the task, their cognitive engagement in the task, and the information presented by the AI [14, 28, 30]. On the individual level, people may also be averse or appreciative towards algorithms, perhaps due to pre-conceptions or judgments formed after seeing the AI's performance [21, 37, 53, 57, 62]. A number of strategies have been developed to augment users' engagement and to provide them with more information of the AI decision, such as through providing explanations [64, 68, 78, 91]. The best interventions for appropriate reliance should be designed to support the users' abilities to *verify* the AI's recommendation, not necessarily to provide a fully transparent system [26, 63, 82]. We adapt appropriate reliance metrics [67] to assess the ability of LLM interventions to help users verify outputs and increase appropriate reliance.

2.2 LLMs as Decision Aids

Recent research has investigated the role of LLMs in human-LLM collaborations for applications such as writing [47], communication [27], coding [35, 60], education [7, 43, 52], fact-checking [38, 71, 84], information search [65, 69, 73], and other specialized domains [15, 18, 87]. In these tasks, the LLM can provide guidance and information for the user to assist in their decision-making, learning, and creations. They can also offer antagonistic advice to induce the user into considering alternative options [11, 16, 56]. Even when the user is just reading without explicitly searching for advice, the opinions embedded in the LLM outputs can be influence the user's

own opinions [25, 41, 69]. In technical AI research, the quality of LLM responses has been improved through methods such as chain-of-thought with intermediate reasoning steps [83], debating amongst multiple LLMs [36, 38], and enforcing self-consistency in outputs [5, 81].

In ideal cases, the human-LLM team can reach better performance as a team than individually, tapping into the strength of their combined knowledge and skill sets [8]. However, the question of how to induce *appropriate reliance* is still open. While using an LLM for information search can be just as efficient as a search engine, the user's ability to distinguish incorrect information is diminished, resulting in a higher rate of over-reliance on incorrect information [71, 85]. On the other hand, aversion to LLM advice has also been observed, resulting in users losing performance gains that they could have achieved through relying on the LLM [89].

In addition to technical limitations of LLMs, human decision-makers are also imperfect and subject to cognitive biases [6]. People may have unrealistic expectations of LLMs and use them inappropriately [76], such as by assuming LLMs are skilled at math simply because they are computational models [93]. People also demonstrate divergent decision-making characteristics, with some preferring to meticulously optimize their decision while others take the satisficing or minimizing paths that allow them to bypass extensive thinking [34, 59]. Those who minimize time spent on tasks may be more likely to over-rely on LLM advice, given its perceived completeness and relevance [23].

2.3 Uncertainty and Reliance Calibration Techniques

Much attention has been placed on promoting transparency in LLMs, with the goal of communicating their capabilities or confidence to users [49, 51, 86]. For example, **Uncertainty Highlighting** draws visual attention to uncertain tokens in the LLM's outputs that the user may want to scrutinize [22, 77], which have been found to reduce over-reliance in information search [73], improve trust in answering questions [22], and increase interpretability in AI-aided coding [77]. Other interventions that have been tested in human-LLM collaborations include contrastive explanations [71], adding semantic expressions of uncertainty [39, 51, 74], and confidence scores [79].

Complex LLM calibrations may carry the risk of being too cognitively overwhelming to understand—for example, to truly understand uncertainty highlighting, users must know that the highlighting reflects token-wise uncertainty rather than semantic uncertainty. Many people make satisficing decisions, where they only look at information to make ‘good enough’ decisions but don't engage further, thus rendering explanations and interventions ineffective [23, 34]. **Reliance Disclaimer** is a low-effort intervention that can be statically applied to LLM outputs, such as through a persistent warning on the chat interface that encourages the user to verify the LLM's answers. While extremely simple, it has been found to reduce students' over-reliance on incorrect LLM guidance in a classroom setting [7].

Additionally, explainable AI (XAI) research has found it effective to focus on increasing the user's cognitive engagement in the task, such as through simple cognitive forcing functions that add frictions

and delays to the interface or removing direct recommendations [9, 28]. Similar friction-based UI designs have been applied to the LLM space to modulate user behaviour [19]. We take inspiration from the domain of LLMs for education, where LLMs are prompted to provide guidance or hints instead of direct answers [42, 55]. Friction is introduced in the reliance process, where users have to cognitively engage to consider the recommended answer.

In this study, we implement three highly relevant interventions that induce cognitive engagement in different ways: **Reliance Disclaimer**, **Uncertainty Highlighting**, and **Implicit Answer**.

3 Methodology

We designed a large-scale, randomized, and incentivized experiment to evaluate the efficacy of three reliance interventions applied to the outputs of GPT-4o¹ in two question-answering tasks. To answer our research questions, we compare the effect of the interventions on performance, reliance behaviour, confidence calibration, and time of engagement against the effect of the baseline **Control** condition. In this section, we outline our choices of tasks, procedure, measures, and conditions.

3.1 Question-Answering Tasks

We gave the participants two separate question-answering tasks, which we carefully selected to meet the following desiderata:

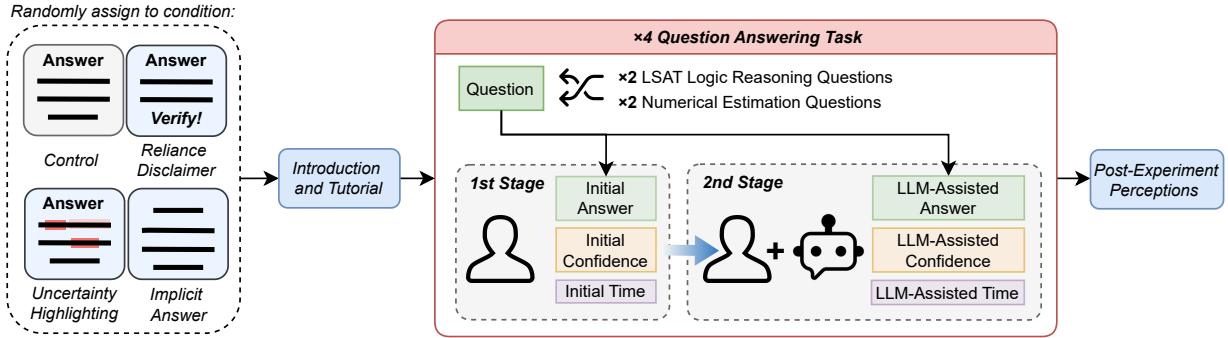
- D1 **Accessible to a general audience.** The tasks should not require any specialized knowledge or skills, so that they are relevant and accessible to a typical person.
- D2 **Difficult.** The tasks should be sufficiently challenging, requiring a meaningful level of cognitive effort to answer correctly.
- D3 **Resistant to cheating.** The tasks should not be straightforward to solve by using external assistance, such as a search engine.

Driven by the desiderata above, one of our central objectives was to select tasks that a typical person could reason about, even if they do not arrive at the correct answer. We therefore intentionally excluded factual questions characterized by domain-specific knowledge that individuals either possess or lack, and cannot conclude through reasoning alone. This approach emphasizes critical thinking over recall of factual information, and ensures that the average person has the potential to get the correct answer, as well as the ability to engage with LLM advice and reliance interventions. To increase generalizability, we selected two different question formats: multiple-choice, a common standard in human-AI collaboration research; and numerical estimation, which allows for a more granular measurement of the LLM's influence on participants through the weight-of-advice (WoA). Below, we outline the tasks selected for each format.

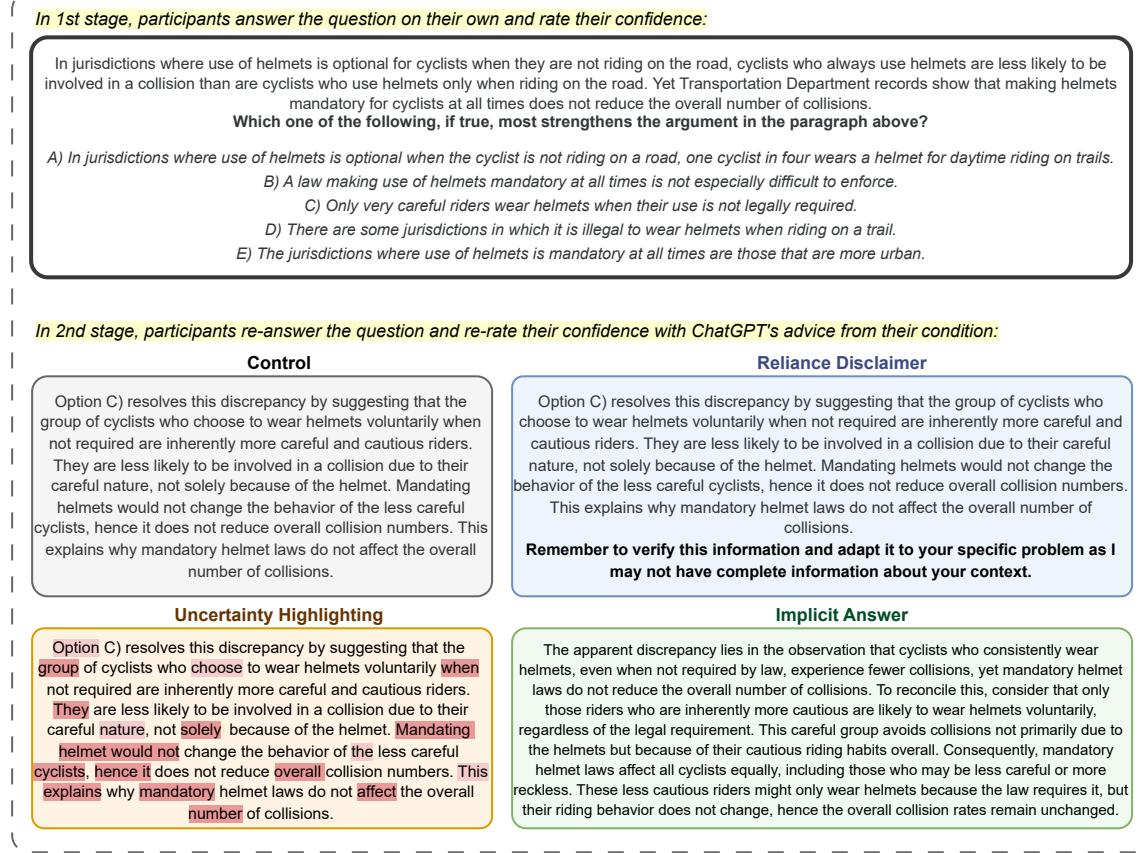
LSAT Logical Reasoning. Based on preliminary results from piloting multiple-choice questions from several domains, we found that logical reasoning questions best balanced difficulty with accessibility. We selected LSAT logical reasoning questions² as the

¹We used gpt-4o-2024-08-06 as it was the most state-of-art GPT model at the time of experimentation.

²Law School Admission Test® (LSAT) questions created by Law School Admission Council (LSAC).



(a) Diagram of the full experimental procedure.



(b) Sample question and advice provided from the LSAT task. Participants saw only one out of the four advice based on their condition.

Figure 2: Diagrams and materials related to the procedure and tasks, where (a) shows the experimental flow, and (b) shows a complete question from the LSAT task and the advice provided to participants from each of the four conditions.

domain for the multiple-choice task and adapted two multiple-choice LSAT logical reasoning questions, both labeled ‘difficult’. Each LSAT question has five options, of which exactly one is correct (determined based on LSAT answer keys). We generated both correct and incorrect LLM advice for each question for use during the experiment. A sample question and corresponding LLM advice

for all four conditions can be found in Figure 2b, and screenshots of what participants saw can be found in Figures A.1 in Appendix A.

While LSAT questions can be difficult to comprehend, they are designed to test for aptitude in law school rather than knowledge of the legal domain. As such, they do not contain undefined jargon, nor do they require domain knowledge (satisfying D1). In addition

to reasoning, LSAT questions also test analytical and reading comprehension skills, making them non-trivial to answer (satisfying *D2*). Furthermore, we reworded a question so that entering it into a search engine would not result in the correct answer (satisfying *D3*). Finally, they lend our study further external validity by emulating situations where people use LLMs to assist them with text comprehension or reasoning, in addition to people studying for tests such as the LSAT. While the MMLU dataset is another popular multiple-choice question benchmark for LLMs [31], we did not find it suitable since it is skewed towards fact-based, *know-it-or-you-don't* questions.

Numerical Estimation. While most previous studies in human-AI collaboration focus on multiple-choice or binary outcome questions, we believe that introducing a Numerical Estimation (NE) task adds a valuable new dimension to this research space. Among the various forms of NE tasks, such as predicting values in regression tasks, we opted for an image-based format based on a classic carnival game: estimating the quantity of beans in a glass jar. We posit that this is a gamified task that is familiar to most people and does not require advanced mathematics (satisfying *D1*). Despite the simplicity of the estimation process, it is still challenging to achieve accurate results (satisfying *D2*). Lastly, since we took the pictures and many of the beans are occluded, it is impossible to use external assistance to “cheat” on this task (satisfying *D3*).

For the experiment, we captured photos of glass jars with manually-counted quantities of beans from both side perspectives and bottom perspectives. One set featured a smaller jar containing 403 chickpeas, while the other contained a larger jar with 856 jellybeans. The pictures and sample LLM advice can be found in Figures A.2-A.3 in Appendix A. Unlike in multiple-choice tasks, where advice can be objectively classified as good or bad, the goodness of the advice in the Numerical Estimation setting depends on the participant’s initial estimate. For example, while advice recommending a number that is far lower than the ground truth is objectively wrong, it is still *better* than an even lower participant estimate. To account for this variability, we provided five distinct levels of LLM advice relative to the ground truth quantity: *very low*, *low*, *good*, *high*, and *very high*. We chose these levels of advice based on results from initial pilot experiments to balance the proportion of bad advice with good advice. Visualizations of the advice levels with the distribution of participant guesses are shown in Figure C.1 in Appendix C.

3.2 Procedure

In the experiment, each participant was randomly assigned to one of four conditions and then answered two questions from the LSAT task and two questions from the Numerical Estimation task, randomized so that the LSAT phase and Numerical Estimation phase were equally likely to appear first. The orders of questions within each category were also randomized. Participants received pre-generated advice from GPT-4o. This advice was randomly sampled for each question, meaning participants could receive any mix of ‘good’ or ‘bad’ advice. After completing the tasks, participants completed a short post-survey. See Figure 2a for the full diagram of the experimental flow and Figure 2b for sample advice for all four

conditions for an LSAT question (for the advice provided in an Numerical Estimation question, see Figure B.2).

To mimic the stakes and level of engagement typically involved in real-world decision-making scenarios where people might consult AI chatbots, we incorporated two key design elements into the study:

- (1) **Two-Stage Answering:** We aim to emulate contexts where the user is already engaged in the task before turning to LLM advice, such as a student who tries to solve the problem on their own first before checking with ChatGPT. Given that our participants are crowdworkers without prior context or investment in the tasks, we believed it is necessary to induce engagement. To address this, we require participants to answer each question on their own before seeing the LLM-generated advice, and then once again after they are given the advice. Studies that use one-stage questioning, where the LLM’s advice is presented alongside the initial question, reflect a context where the participant has less incentive to be engaged, which can lead to artificially high over-reliance rates.
- (2) **Incentivization:** Real-world decision-making involves balancing risk and reward. To simulate these stakes and induce incentive-compatible behaviour, we introduced a bonus, amounting to a maximum of 20% of the total participation pay. Participants earn rewards for correct answers, with two opportunities for bonus pay: one before, and one after receiving LLM-generated advice. In the LSAT questions, participants received a bonus if they selected the correct answer, while for the Numerical Estimation questions, participants earned a bonus if they guessed within 10% of the true count.

3.3 Measures

Our research questions focus on reliance behaviour (**RQ1**) and confidence calibration (**RQ2**). We detail the equations for these categories of measurements, in addition to secondary measurements of task performance and time of engagement.

Reliance. Following existing literature on appropriate reliance in human-AI interactions, we use the two-stage question-answering task format where a participant first answers a given question on their own before receiving LLM advice and answering again. If the participant’s initial answer conflicts with the LLM’s advice, appropriate reliance requires the human to correctly determine whether the LLM’s advice is better or worse. We focus on this explicit *decision-making step* as the construct of interest for measuring reliance, where the user is presented with the choice to either a) change their answer to rely on the LLM’s advice, or b) disregard it and retain their initial answer. Thus in our analysis on reliance, we *only consider the data instances* where the participant’s initial answer differs from the LLM’s advice. See Figure 3 for an illustration of the reliance decisions pathways.

What does it mean to *rely* on LLM advice? In multiple-choice questions, we follow previous work by defining reliance as *changing one’s answer to the LLM’s suggestion*. In numerical estimation questions, we define reliance as *an answer with positive weight-of-advice (WoA)*, where WoA is calculated as $\frac{\text{Final Estimate} - \text{Initial Estimate}}{\text{LLM Advice} - \text{Initial Estimate}}$

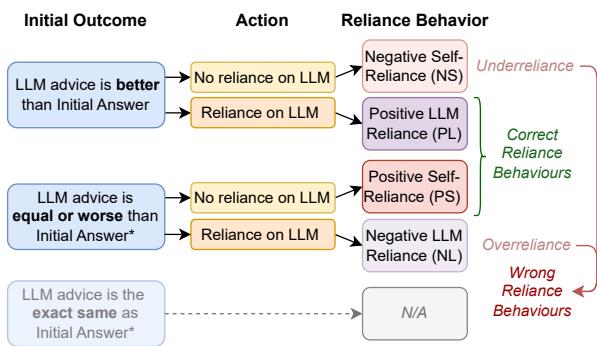


Figure 3: Illustration of the reliance pathways based on the initial answer and LLM advice. Note that in the second pathway, *equal or worse* refers to cases where the LLM advice is *different* from the initial answer, but the relative accuracy can be *equal* (such as both answering a multiple-choice question wrong). The third pathway covers cases where the LLM advice matches the initial answer exactly, where there is no opportunity to make a reliance decision.

— in short, this means that the participant was influenced by the LLM’s advice to adjust their estimate. We further define the reliance outcomes as falling into one of four categories:

- **Positive LLM Reliance (PL):** the participant changes their original answer to follow the LLM’s advice, and subsequently improves their performance.
- **Positive Self-Reliance (PS):** the participant does not follow the LLM’s advice, and doing so would have worsened or failed to improve their performance.
- **Negative LLM Reliance (NL):** the participant changes their original answer to follow the LLM’s advice, but worsens or fails to improve their performance. This is also referred to as *over-reliance*.
- **Negative Self-Reliance (NS):** the participant does not follow the LLM’s advice, but doing so would have improved their performance. This is also referred to as *under-reliance*.

PL and PS can be grouped together as *Correct Reliance Behaviours*, while NL and NS are *Wrong Reliance Behaviours*. NL represents over-reliance on LLM advice (the participant took the advice when it worsens or fails to improve their performance) and NS represents under-reliance (the participant didn’t take the advice but should have). We note that PL is possible even if the LLM advice is incorrect, as long as there is a notion of relative error—such as switching to a *less* wrong answer suggested by the LLM, which applies to the Numerical Estimation task and can be extended to other real-world tasks. In contrast, the LSAT task lacks relative scoring across options, treating all incorrect answers as equally wrong; so relying on incorrect advice counts as over-reliance. This disambiguation is made since the LLM advice would not improve performance, so it would be preferable for the user to engage in the task and self-rely, allowing them to learn from the experience [28]. From these behaviour classifications, we compute the following *appropriate reliance* metrics:

- **Relative LLM Reliance (RLR) = $\frac{PL}{PL+NS}$** = Rate of correctly relying on the LLM when the LLM advice is helpful.
- **Relative Self-Reliance (RSR) = $\frac{PS}{PS+NL}$** = Rate of correctly relying on oneself when the LLM advice is harmful.
- **Appropriate Reliance Ratio (ARR) = $\frac{PL / (PL+NS)}{NL / (PS+NL)}$** = Ratio of correct reliance on helpful advice to wrong reliance on harmful advice, where:
 - $ARR < 1$: Higher rate of over-reliance than positive reliance, signifying *inappropriate reliance behaviour*.
 - $ARR = 1$: Random chance; LLM reliance rates are the same regardless of LLM correctness.
 - $ARR > 1$: Higher rate of positive reliance than over-reliance, signifying *appropriate reliance behaviour*.

We compute the metrics for each condition across participants, then calculate the effect size as the difference in outcomes between the control and each intervention. To assess statistical significance, we use permutation tests. For each planned comparison, we generate a distribution of effect sizes under the null hypothesis by randomly shuffling and reassigning data between the control and intervention groups. By comparing the observed effect size to this shuffled distribution, we estimate how likely it is to occur by chance, using a significance threshold of $p < .05$.

Confidence and Subjective Perceptions. For each question, we collect the participant’s confidence in their answer prior to receiving LLM advice (**Initial Confidence**), and their confidence in their answer after receiving the LLM advice (**Final Confidence**). Confidence is measured using a unipolar Likert scale from 1 (highly uncertain) to 5 (highly certain), which allows us to calculate the **Change in Confidence** as a result of receiving LLM advice. Ideally, well-calibrated confidence should result in lower confidence scores when participants make wrong reliance decisions. We compare the change in confidence between correct and wrong reliance outcomes for each of the conditions, where statistical differences are tested with Mann-Whitney U-test for non-normal distributions with significance threshold $p < .05$.

At the end of the survey, participants provide additional ratings on a bipolar Likert scale of 1 (disagree) to 7 (agree) for the following subjective perception statements:

- ChatGPT’s advice made me more confident³.
- ChatGPT’s advice seems accurate.
- ChatGPT’s advice was helpful.

Task Performance. Since our experiment is not designed to emulate the realistic performance of LLMs in tasks, we do not particularly emphasize the human-LLM collaborative performance. However, we explore the initial error from participants independently answering questions, and compare this with their updated error after re-answering questions with LLM advice. We also compute a **Best Possible Performance** metric, which simulates the outcomes as if the human made the correct reliance decision in every single question, representing the best possible accuracy of a human-LLM team.

³Note that this measure of post-task confidence is different from the question-level confidence (the target of RQ2) described previously.

Time of Engagement. To understand participant engagement with the LLM’s advice, we measure the amount of time that participants spend on each stage of question-answering. From these two time measurements, we calculate the relative time as the ratio between time taken to engage with the advice, to the time taken to answer the initial question (**Relative Time**). A larger value potentially signifies higher engagement in the advice, which may correlate with the complexity of the intervention paired with the advice.

3.4 Experimental Conditions

In this section, we outline the implementation of our four experimental conditions, each representing different forms of LLM-generated advice. The interventions were chosen based on their relevance to the research community, their applicability across both multiple-choice and estimation tasks, and their diversity in design—incorporating both static and dynamic interventions, and both visual and semantic augmentations. Whenever possible, we used direct outputs generated through tailored prompts from the LLM with minimal manual adjustments. Figure 1 provides a visual overview of the advice conditions, with more details available in Figure 2b.

Control. To generate the baseline advice, we input the questions into GPT-4o through OpenAI’s API and record the responses directly. Participants receive a direct recommendation of the answer (such as “Option A” in Logical Reasoning and “120 chickpeas” in Numerical Analysis) in addition to a 2-4 sentence long explanation supporting the answer.

Reliance Disclaimer. We append the following static disclaimer statement, “Remember to verify this information and adapt it to your specific problem as I may not have complete information about your context”, while keeping the advice generated in **Control** the same. This disclaimer was generated through prompting the LLM to disclose its limitation as a language model to convey that its outputs should be used carefully. For consistency and to emulate real-world chatbot disclaimers like ChatGPT, we apply the same disclaimer to all advice.

However, we acknowledge that the effectiveness of a static disclaimer may not persist in long-term deployment. Alarm fatigue describes the process of desensitization, where users learn to ignore unfounded warnings over time [20]. We hypothesize that disclaimers may be most effective for priming new users or in shorter interactions, like a tutoring chatbot that students only use occasionally. Some alternative ways to improve longevity of the effect are to dynamically insert disclaimers fluidly into the conversation—which makes them less repetitive and predictable [7], or to insert them only when error rates are particularly high [20].

Uncertainty Highlighting. When generating the **Control** advice, we also record the log probability of each token. To visually augment the baseline advice, we highlight the tokens with log probability under -1 in **stronger red** and indicate them to be *uncertain*, and the tokens under -0.5 in **lighter pink** and indicate them to be *slightly uncertain*. We selected these thresholds to balance highlighting uncertain text within the answer with minimizing visual overwhelm for the reader. We make manual adjustments to maintain coherence,

such as extending highlighting to full words if they were only partially highlighted and removing highlighting on articles that do not impact semantic meaning. Participants were not told that the method of determining uncertainty is based on log probabilities. However, they were reminded of the definition of the highlighting colors in every survey page with the LLM advice. Our choice to highlight generation probability follows prior works like [73], but we note that alternative methods of highlighting can offer better interpretation, such as edit probability [77].

Implicit Answer. This intervention semantically augments the content of the advice to avoid writing the answer directly, requiring more engagement to extract which answer the advice supports. For the Logical Reasoning questions, we instruct GPT-4o to never explicitly state the answer, and to use paraphrasing to generate an argument in support of the answer. For the Numerical Estimation questions, we manually removed the final calculation step from the advice generated in **Control** (as it was difficult to achieve this through prompting alone). This way, we ensure that participants see the intermediate steps of the calculation, and can arrive at the value advised by the LLM if they engage with the answer enough to calculate it themselves.

4 Results

We pre-registered our experimental design and analysis plan online ⁴ before conducting the experiments. Here, we report the results of the analyses described in Section 3, separately for LSAT and Numerical Estimation due to the inherent differences between these two tasks.

4.1 Participants

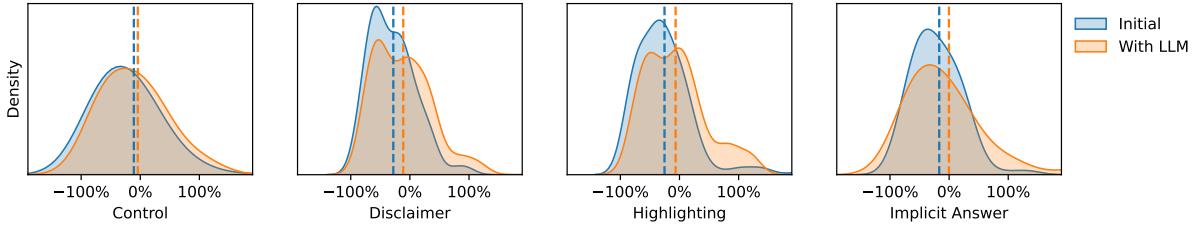
We conducted a power analysis based on preliminary pilot data and found that approximately 200 samples per condition would suffice to achieve significance with an effect size of 0.4 for the ARR. Since we perform the analysis separately for LSAT and Numerical Estimation and each participant answers two questions, we determined that we needed 100 participants per condition.

We recruited 400 crowdworkers from Prolific, filtering for those who are based in the US, are fluent in English, have a task success rate of 98% or higher, and have completed at least 50 tasks. Following Prolific’s policy, we excluded participants who failed 2/3 attention checks, which was one participant from the **Control** ($n = 99$) condition and two from the **Uncertainty Highlighting** condition ($n = 98$). For both of the other conditions, **Reliance Disclaimer** and **Implicit Answer**, all participants passed the attention checks. Across all conditions, the median completion time was 9 minutes and 44 seconds. Participants were compensated \$1.28 USD as the base pay. They had the chance to earn up to \$0.26 USD as a bonus for performance, with the mean earning being \$0.07 USD. Based on Prolific’s demographics information about our participants, the participants are 56.3% female with an average age of 43.5 years and are all located in the US. Furthermore, they are well-distributed in their usage of AI chatbots, with 18.9% citing daily use, 33.5% citing weekly use, 17.9% citing monthly use, and the remaining 29.7% citing rare or no usage.

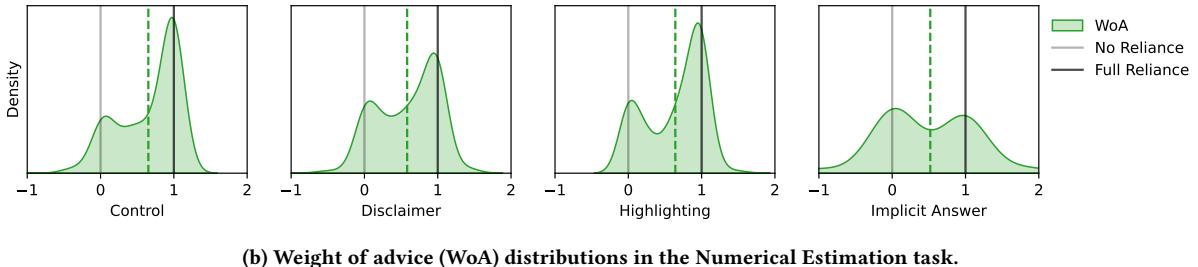
⁴https://aspredicted.org/8BD_WWN

Wrong→ Wrong: 44.9%	Wrong→ Correct: 21.2%	Wrong→ Wrong: 37.0%	Wrong→ Correct: 22.0%	Wrong→ Wrong: 48.0%	Wrong→ Correct: 18.4%	Wrong→ Wrong: 43.0%	Wrong→ Correct: 20.5%
Correct→ Wrong: 8.1%	Correct→ Correct: 25.8%	Correct→ Wrong: 6.0%	Correct→ Correct: 35.0%	Correct→ Wrong: 7.1%	Correct→ Correct: 26.5%	Correct→ Wrong: 7.5%	Correct→ Correct: 29.0%
Control		Disclaimer		Highlighting		Implicit Answer	

Figure 4: Performance outcomes in the LSAT task for initial and LLM-assisted answers, shown through transition matrices.



(a) Percent error distributions of the initial and LLM-assisted answers in the Numerical Estimation task.



(a) Percent error distributions of the initial and LLM-assisted answers in the Numerical Estimation task.

Figure 5: Performance outcomes in the Numerical Estimation task for initial and LLM-assisted estimations. (a) shows the percentage error (relative to ground truth) distribution and (b) shows the weight-of-advice (WoA) distributions. The mean of each distribution is indicated by the dotted line.

4.2 Task Performance

We preface our main discussion of reliance with an examination of task performance. We emphasize that since we are interested in how people react to both good and bad LLM advice, we sample these at equal rates. Our LLM performance therefore does not necessarily reflect the actual performance of GPT-4o on our tasks and should not be taken as a measure of realistic human-LLM collaborative decision-making.

LSAT. Due to the high difficulty of the LSAT questions, participants were initially correct in only 36.0% of question instances, not far from the rate of random guessing (20%). Participants then either received correct or wrong LLM advice, equivalent to a 50% accuracy rate. Their second-stage (LLM-assisted) accuracy rate was 46.1%. Figure 4 shows transition matrices of the first- and second-stage performance outcomes in each condition. In the performance outcomes, there are no significant differences between **Control** and any of the interventions as evaluated by a χ^2 test ($p > .05$ for all three interventions). Based on the collaborative performance, it is evident that while the LLM improved overall performance, the

human-AI team does not perform optimally. In each question, if the human-LLM team relied on the team member that got the correct answer, then the **Best Possible Performance** would reach 62.1% accuracy (not 100%, because sometimes both the human and the LLM answered incorrectly). The gap between optimal and actual performance represents the performance lost to wrong reliance behaviour—either relying on wrong LLM advice or ignoring correct LLM advice.

Numerical Estimation. While estimation may be more approachable, it is still a difficult task. The mean *absolute* percentage error relative to the ground truth is 45.7% across all conditions. With advice, the participants do not significantly improve, answering with a similar *absolute* error rate of 47.0%—this is unsurprising, since poor LLM advice was given as frequently as good advice. However, in terms of the *mean* (not absolute) percentage error, the participants initially underestimate the counts, averaging -20.2% relative to the ground truth. With the advice, the gap closes to just -5.6% (below the ground truth). The full distributions, which include both positive and negative percent errors, are shown by

condition in Figure 5a. There are no significant performance differences between conditions (Mann-Whitney U test $p > .05$ for all three interventions). The optimal human-LLM team outcome can be simulated by submitting the estimation of the team member who is closest to ground truth. In this hypothetical **Best Possible Performance**, the mean of the *absolute* error would be 29.0%. That the observed error was much higher indicates people both over- and under-relied on LLM advice.

While unrelated to task performance, we further compute *how much* the participants relied on the advice for additional insights on their behaviour. The weight-of-advice (WoA) taken by each participant in each question is shown as distributions for each condition in Figure 5b, where a WoA of 1 means fully adopting the LLM's advice, a WoA of 0 means retaining the initial answer, and other values interpolate between these definitions. The distribution of WoA tends to be bimodal, with participants in **Control**, **Reliance Disclaimer**, and **Uncertainty Highlighting** predominantly favouring adoption of the LLM advice. Since **Implicit Answer** did not provide the final recommendation, and instead left it to the participants to apply the mathematical hints, the WoA distribution is more flattened, even with values in the negatives or exceeding 1. In our WoA analysis, we drop outlier values over a magnitude of 5 (such as switching from an initial estimate of 600 to a final estimate 4480 when the LLM advice is 544, which constitutes a WoA of -69.3). By the Mann-Whitney U test, **Reliance Disclaimer**'s WoA ($p = .03$) and **Implicit Answer**'s WoA ($p = .005$) are significantly smaller than **Control**'s WoA.

4.3 Reliance Behaviour

Following the method of prior work [67], we drop the data instances where the participant's initial answer agreed with the LLM advice for the reliance analysis, as it is not possible to measure deliberate reliance in those cases (see the third pathway in Figure 3). For Numerical Estimation, we do not use a strict definition of agreement, as it is unlikely for a participant's initial estimate to exactly match the LLM advice. Instead, we adopt a heuristic of *perceived agreement*. For example, a participant may perceive a guess of 700 to be "pretty much the same" as an advice of 750. We use a conservative 10% boundary for perceived agreement, as participants were primed with the expectation that guesses within $\pm 10\%$ of the true count would earn a bonus, making this range a reasonable proxy for perceived agreement.⁵ After excluding the instances of agreement, 87.1% of samples in the LSAT task and 88.8% in Numerical Estimation are retained.

To answer **RQ1**, we compute the Relative LLM Reliance (RLR), Relative Self-Reliance (RSR), and Appropriate Reliance Ratio (ARR) across all data instances collected per task in each condition. For interpretation, RLR indicates the rate of reliance on the LLM when the advice is beneficial, and RSR indicates the rate of self-reliance when the LLM advice is harmful. ARR is the ratio of RLR to $(1 - \text{RSR})$, where higher values signify more appropriate reliance on the LLM. In terms of reliance outcomes, over-reliance maps to low RSR (i.e.

⁵We refine this heuristic with caveat that for smaller levels of advice, the boundary of perceived equivalence becomes very small. To address this, we include an additional ± 50 padding around the advice. As a robustness check, we also conduct reliance analysis using a 5% agreement boundary without additional padding and find similar significance trends to those reported.

relying on the LLM when the advice is bad), under-reliance maps to low RLR (i.e. not relying on the LLM when the advice is good), and appropriate reliance maps to when both are high. Figure 6a, reconstructed based on Figure 1 from [66], provides a visual reference for the relationship between reliance outcomes and RLR and RSR values. Table 1 displays all the metrics calculated for each condition, separately for the LSAT and Numerical Estimation tasks. For the interventions, there are also p -values indicating the significance of the metric difference between treatment and **Control**⁶.

LSAT. The visual representations of the metrics and their significances for LSAT are plotted in Figure 6b, where **Control** is plotted in gray and we compare the interventions against it. We find that **Reliance Disclaimer** is the *only intervention* that has a significantly positive influence on the ARR, by reducing both over-reliance and under-reliance ($p = .04$). **Uncertainty Highlighting** and **Implicit Answer** both significantly improve self-reliance ($p = .01$ and $p = .002$, respectively), but at the cost of shifting overall reliance away from the LLM. Despite so, the reduction of RLR – signifying correct reliance on LLM – is not significant. Both interventions have higher ARR values than **Control**, but the difference is again not significant at our power level. These results indicate that a *static, persistent disclaimer message* may induce appropriate reliance more effectively than more specialized methods.

Numerical Estimation. This task requires more care to analyze, as WoA is a continuous rather than binary measure of reliance. As such, we compute reliance behaviours at several different WoA thresholds:

- $\text{WoA} > 0.01$: Any adjustment towards the LLM's advice is counted as reliance (highest sensitivity). Under this definition, 79.7% of trials count as reliant behaviour.
- $\text{WoA} > 0.5$: Adjustments must be made over 50% towards the LLM's advice to count as reliance (moderate sensitivity). Under this definition, 60.7% of trials count as reliant behaviour.
- $\text{WoA} > 0.99$: Adjustment towards the LLM's advice must be nearly total to count as reliance (lowest sensitivity). Under this definition, 32.7% of trials count as reliant behaviour.

Values and significance for RLR, RSS, and ARR are expressed at these three WoA threshold levels in both Table 1 and Figure 6c. Unlike in the LSAT task, we do not observe any intervention to effectively improve appropriate reliance through increasing ARR. Across different WoA thresholds, **Implicit Answer** increased appropriate self-reliance the most (significant at WoA thresholds of 0.01 with $p = .004$ and 0.5 at $p = .007$). At the lowest WoA threshold, this also increases under-reliance through significantly reducing RLR ($p < .001$)—which is undesirable. This could be due to being forced to perform simple math calculations to see the advised number—which people may not do if they are satisfied with their own answer, or believe that extracting the advice requires more effort than the reward is worth. **Uncertainty Highlighting** also causes significant decrease in RLR at a high WoA threshold of 0.99 ($p = .02$), indicating that people are hesitant to fully use the advice.

⁶This study followed a preregistered analysis plan using $p < .05$ as the significance threshold. While no multiple-hypothesis corrections were applied (as per the pre-registration), we acknowledge that adjusting for multiple tests could influence the significance level.

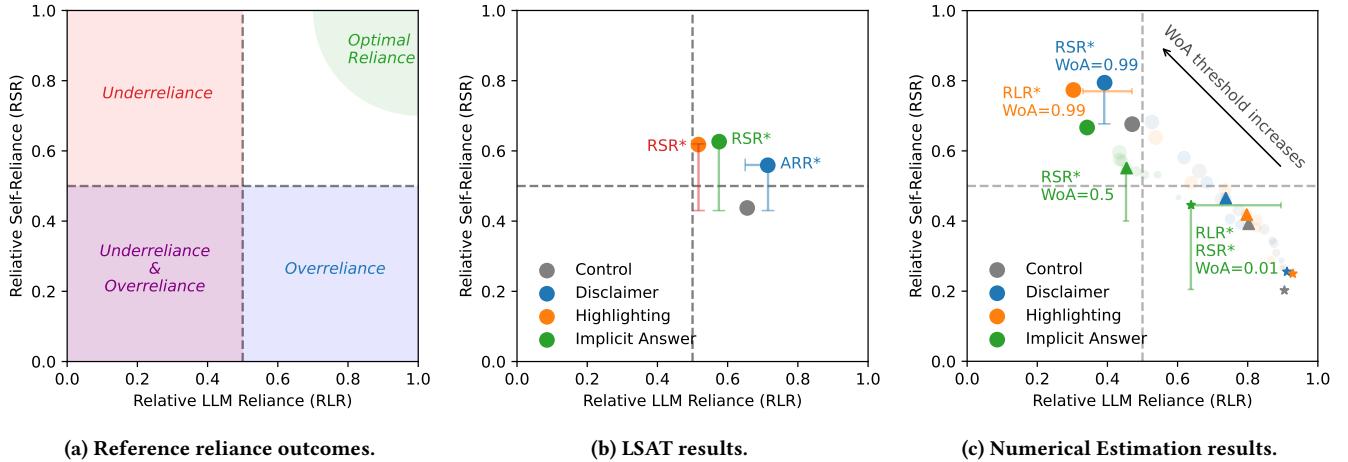


Figure 6: Visualization of the appropriate reliance measurements – RLR on the x-axis, RSR on the y-axis, and ARR indirectly represented as a combination of RLR and RSR (see Section 3.3 for equations). Each colored marker represents the (RSR, RLR) measurement for a condition. For any significance in the differences between Control and each of the interventions, a horizontal bar (RLR), vertical bar (RSR), or both (ARR) are drawn between the two conditions’ markers. Subfigures (a) shows the visual reference for the areas of the graph that correspond to different reliance outcomes; (b) shows the reliance results for the LSAT task; and (c) shows the reliance results for the Numerical Estimation task, using *multiple WoA threshold values* for classifying reliance. Three WoA thresholds are indicated explicitly in increasing marker size: 0.01 (star marker), 0.5 (triangle marker), and 0.99 (circle marker), while other thresholds are plotted with translucency to visualize the trajectory of changes.

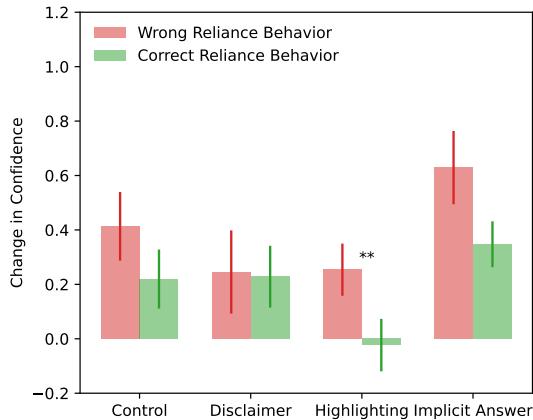
Table 1: Summary of reliance metrics (RLR, RSR, and ARR) calculated for all four conditions in both LSAT and Numerical Estimation tasks, where the latter is expressed at three different levels of WoA thresholds. Significance is calculated using permutation testing between Control and each intervention, where significant differences are bolded. Positive changes are further indicated with (+), while negative changes are indicated with (-).

	Control	Reliance Disclaimer	Uncertainty Highlighting	Implicit Answer
LSAT	RLR 0.66	0.71, p=.2	0.52, p=.1	0.58, p=.2
	RSR 0.44	0.56, p=.05	0.62, p=.01 (+)	0.63, p=.002 (+)
	ARR 1.17	1.62, p=.04 (+)	1.35, p=.2	1.54, p=.09
Numerical Estimation (WoA >0.01)	RLR 0.91	0.91, p=.4	0.93, p=.3	0.67, p<.001 (-)
	RSR 0.15	0.22, p=.1	0.23, p=.1	0.43, p=.004 (+)
	ARR 1.07	1.17, p=.2	1.20, p=.1	1.18, r=.2
Numerical Estimation (WoA >0.5)	RLR 0.8	0.74, p=.2	0.8, p=.5	0.45, p=.17
	RSR 0.39	0.47, p=.1	0.42, p=.3	0.55, p=.007 (+)
	ARR 1.32	1.38, p=.4	1.37, p=.4	1.01, p=.1
Numerical Estimation (WoA >0.99)	RLR 0.47	0.39, p=.2	0.3, p=.02 (-)	0.34, p=.05
	RSR 0.68	0.79, p=.02 (+)	0.77, p=.05	0.67, p=.5
	ARR 1.45	1.91, p=.2	1.34, p=.4	1.03, p=.1

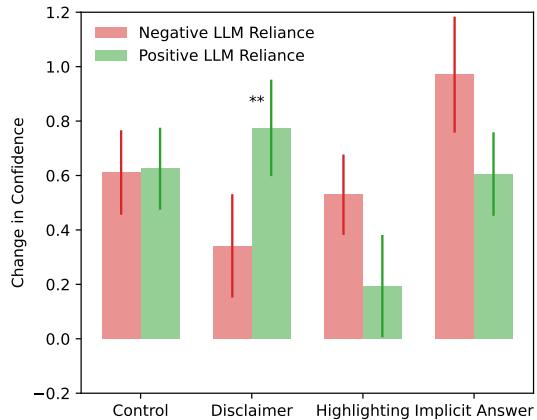
This corroborates Figure 5b, which shows that **Uncertainty Highlighting** and **Reliance Disclaimer** both have reduced density of WoA values close to 1 compared to **Control**.

Another key difference between the Numerical Estimation and LSAT tasks is the length and detail of the advice. A higher barrier to cognitive engagement may lead to heterogeneous outcomes in how participants approach the task. For instance, some participants

might adopt the advice as a shortcut to make easy, satisficing decisions, while others might disregard the advice due to its excessive length and complexity [23]. The discrepancy in results between our two tasks also warns that results of interventions *may not generalize* beyond the specific tasks they are tested on, emphasizing the need for thorough validation across diverse tasks before broader deployment.

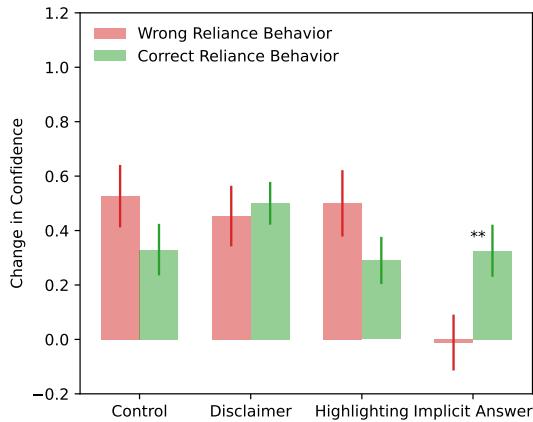


(a) Correct (PL + PS) vs. Wrong (NL + NS) Reliance.

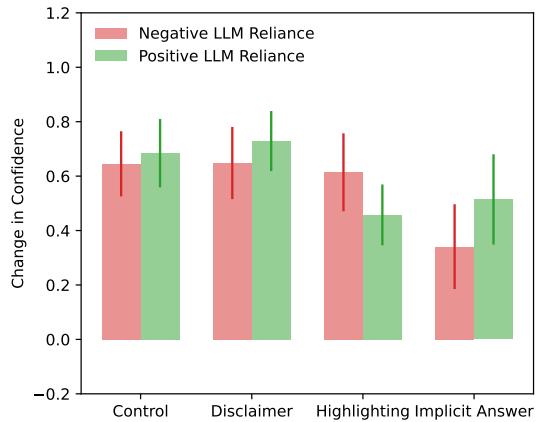


(b) Positive LLM Reliance (PL) vs. Negative LLM Reliance (NL).

Figure 7: Change in Confidence at the question level for LSAT.



(a) Correct (PL + PS) vs. Wrong (NL + NS) Reliance.



(b) Positive LLM Reliance (PL) vs. Negative LLM Reliance (NL).

Figure 8: Change in Confidence at the question level for Numerical Estimation with a WoA threshold of 0.5.

4.4 Confidence Calibration and Perceptions

To address RQ2, we compare participants' confidence in their pre-advice answer and their post-advice answer (**Change in Confidence**) at the question level, along with several post-experiment Likert scale ratings of their subjective perceptions. We compare cases where participants made correct reliance decisions with those where they made incorrect ones, and further focus on instances of positive versus negative LLM reliance (cases where participants actively relied on the LLM). Unless otherwise indicated, significance is calculated *within* each condition using the Mann-Whitney U test with $p < .05$ as the significance threshold. Error bars in the bar plots of Figures 7-10 represent standard errors.

LSAT. Figure 7a shows participants' **Change in Confidence** in the LSAT task for correct and wrong reliance behaviours. The most salient takeaway is that on-aggregate, across all conditions (not explicitly shown in the plot), *people get more confident when they*

over- and under-rely than when making the appropriate reliance decisions ($p = .02$). Within each condition, **Uncertainty Highlighting** is the only one where this observed effect is significant ($p = .03$), which indicates poor confidence calibration since it is undesirable for people to be more confident in wrong decisions.

As an extension, Figure 7b specifically focuses on cases where people chose to rely on the LLM. Here, the confidence scores tend to increase more as people trusted the LLMs enough to use them. We observe that **Reliance Disclaimer** is the only condition that achieves *calibrated confidence*—the increase in confidence from using the LLM's advice is lower when the participant over-relied than when they relied appropriately ($p = .04$). Though not significant, it is interesting to observe that the opposite trend can be demonstrated in some of the other interventions, where confidence increases more in cases of over-reliance (NL). This indicates that reliance interventions that operate by moderating users' confidence

in the LLM may inadvertently *miscalibrate* confidence with respect to appropriate reliance.

Numerical Estimation. For this task, we repeat the confidence analyses and show them in Figure 8. For simplicity, reliance is determined based on the middle WoA threshold of 0.5, since the confidence outcomes between high and low WoA thresholds are qualitatively similar (the results for WoA thresholds of 0.01 and 0.99 are shown in Appendix D as Figures D.1-D.2). **Implicit Answer** is the only condition to achieve a significant difference in calibrating confidence changes to be higher in correct reliance outcomes (Figure 8a). However, no significant differences are seen in the behaviours when we select for instances that explicitly result in reliance on LLMs (Figure 8b), which may indicate that **Implicit Answer** is effective at reducing confidence for *negative self-reliance*—perhaps people recognize that they can benefit from the advice, but did not want to go through the effort of calculating the recommendation.

Overall Post-Perception. Lastly, Figure 9 shows the mean Likert ratings per perceptions category by condition. The full subjective perception statements are shown in the figure's labels (1=disagree, 7=agree). We calculate significance between the **Control** and the interventions using the Mann-Whitney U test with $p < .05$ as threshold. Although all LLM conditions scored positively across the three categories, **Control** scored highest in subjective perceptions compared to all other the interventions. In particular, **Uncertainty Highlighting** is the only intervention that had significantly lower ratings in all categories compared to **Control**, indicating that our implementation may have induced too much uncertainty in the users. An alternative approach to moderate this effect may be to highlight highly certain tokens in green in addition to uncertain tokens in red. In addition, we find that the frequency of baseline LLM usage correlates positively with all three subjective perception measures – see Appendix E for the details of this analysis. The main implication here is that *uncertainty calibration methods can have the undesirable outcome of inducing negative perceptions of accuracy and usefulness of the LLM*, which increases overall algorithm aversion against LLM advice, rather than appropriate reliance.

4.5 Time of Engagement

We compare the **Relative Time** of engagement with given LLM advice by taking the ratio of time spent in the second stage to the first stage. Time is correlated to cognitive engagement, and previous studies have found that introducing slight temporal frictions like waiting for delayed AI advice has potential to affect user behaviour [9, 19]. Figure 10 shows the time ratio by condition for both the LSAT (Figure 10a) and Numerical Estimation (Figure 10b) tasks, where significance is indicated by Mann-Whitney U test with $p < .05$. Overall, participants spent more time in the Numerical Estimation question, with the average relative time ratio exceeding 1, potentially due to the average length of the advice being longer. On the other hand, the time ratio is well below 1 in LSAT.

In both tasks, the order of relative time spent on the advice roughly matches the effort required to extract the answer from the advice. **Reliance Disclaimer** is closest to matching the time

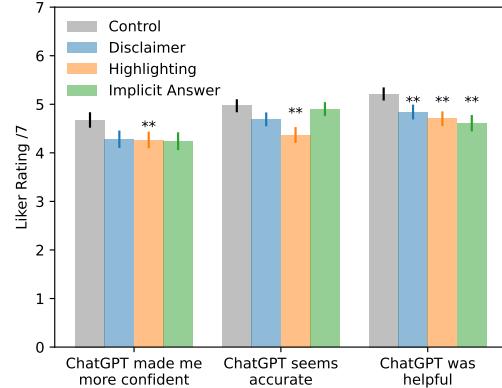


Figure 9: Post-experiment perceptions ratings. Overall, subjective perceptions are lower for the intervention conditions than **Control**, possibly due to added usability friction. (**) indicates interventions that scored significantly lower than **Control** in each respective perception category.

ratio of **Control** while **Implicit Answer** takes much longer, reaching nearly double the relative time in Numerical Estimation. **Uncertainty Highlighting** is somewhere in between, but also with significant increase in time over **Control**. Taken with the results on the efficacy of **Reliance Disclaimer**, our results indicate that longer engagement time does not necessarily correlate with reliance outcomes. The easier, more digestible intervention had the most tangible effect on appropriate reliance. Perhaps with more training and rigour, especially for **Uncertainty Highlighting**, participants can become more proficient at interpreting advice with more complex uncertainty calibration.

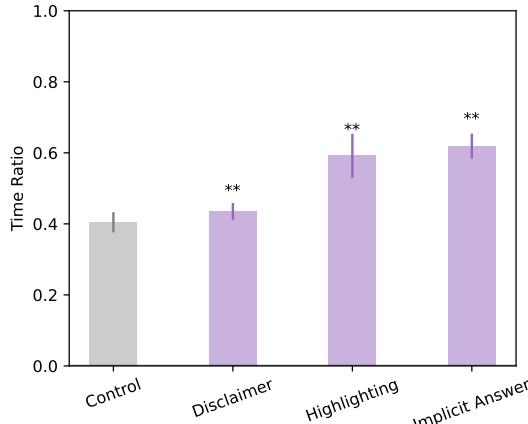
5 Discussion

We now present the key findings to our research questions on appropriate reliance and confidence calibration. We then extend our findings to broader implications for design and for evaluating reliance on LLMs, and discuss the limitations of our experiments.

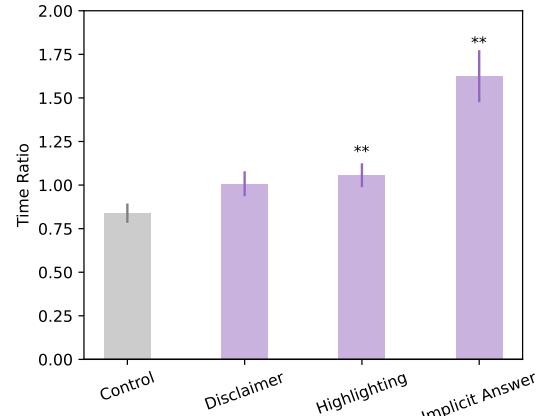
5.1 Key Findings

Cognitive Forcing Designs Reduce Over-Reliance, but May Not Improve Appropriate Reliance. In our study, we investigated the effectiveness of three interventions on LLM reliance in two diverse and challenging tasks. We use measures of inappropriate reliance, separating over-relying on bad advice from under-relying on helpful advice, to uncover mechanisms of sub-optimal human-LLM team performance. Our findings suggest that cognitive forcing designs in the form of confidence calibration (highlighting uncertain tokens) and usability friction (adding disclaimers and removing direct answers) *can help people shift away instinctively relying on the LLM*, but improving appropriate advice usage is a harder challenge.

While the effect of each intervention varied by the task, *all three achieved a reduction in over-reliance* in at least some experiment configurations. However, the only intervention to reduce *both over-reliance and under-reliance* is **Reliance Disclaimer** in the LSAT



(a) LSAT time results.



(b) Numerical Estimation time results.

Figure 10: Relative Time spent at the LLM advice stage compared to the initial question stage. Note that the scale of the y-axes are different between tasks.

logical reasoning task. The Numerical Estimation task captured reliance with more granularity, which provided insight on how the interventions induced under-reliance differently. Our results suggest that the added friction of calculating the answer in **Implicit Answer** caused people to disproportionately discard the LLM advice (not rely at all), while the confidence calibration in **Uncertainty Highlighting** caused people to hesitate to fully follow advice (rely, but not fully).

The increase in under-reliance supports the cost-benefit framework from explainable AI literature, which shows that users will not engage with explanations if they are more cognitively complex than the task itself [78]. This suggests that for the LSAT task, static **Reliance Disclaimers** may have added just enough friction to dissuade habitual over-reliance. However, advice with other interventions, like **Implicit Answer** in Numerical Estimation, may have been perceived as requiring more effort to use than it was worth. Future work should aim to quantify the comprehension difficulty of the interventions and the tasks to understand how they relate to reliance behaviours. In implementing interventions, designers should consider the targeted use case and develop cognitive forcing techniques that do not increase burden substantially.

Confidence When Using LLMs is Poorly Calibrated. We examine how participants' confidence in their answers change after seeing the LLM's advice. Ideally, even with wrong reliance behaviour, confidence should be lower when individuals are internally uncertain about their decision, such as when they are skeptical of the LLM's advice. However, we largely observe the opposite or inconclusive effects, where wrong reliance results in higher gains in confidence. This raises particularly problematic implications, suggesting that individuals were *overall uncalibrated when receiving assistance from LLMs*, corroborating prior findings [71, 89].

There were a few areas where the interventions had success. In the Numerical Estimation task, **Implicit Answer** reduced confidence during under-reliance, while in the LSAT task, **Reliance Disclaimer** lowered confidence gains during over-reliance, which can

be exploited as feedback mechanisms for users to learn from. Both methods present different pathways using which confidence can be influenced by modifying LLM output, but further work is needed to achieve generalizable findings. As shown in prior works in human-AI collaboration, self-confidence and advice agreement rates have nuanced relationships with reliance on advice [54], which can be further investigated in the context of human-LLM collaboration.

We also find perceptions of helpfulness were negatively impacted by all three interventions, while perceived self-confidence and advice accuracy were only reduced by **Uncertainty Highlighting**. Confidence calibration interventions may risk increasing a user's aversion towards LLMs rather than encouraging appropriate evaluation if overall performance is perceived as unacceptably low, or if the intervention makes the advice difficult to understand and use. Further work can explore qualitative factors of what participants disliked or found more cumbersome, and how their perception of their internal decision-making process was affected.

5.2 Broader Implications

With the integration of Copilot in consumer PCs, the widespread use of LLMs in customer service and information retrieval, and advances in Generative AI for creative work, interactions between end users and these tools are becoming nearly unavoidable. As Generative AI advances rapidly, consumers' literacy in its appropriate use is likely to lag behind, raising concerns about how to ensure proper use of the many AI-powered tools and products being introduced [1]. In our study, we show that wrongful reliance is frequent and users are not calibrated to be aware of their behaviour. Interventions that reduce over-reliance can have undesired effects of increasing under-reliance, and require thorough validation in human user studies. We highlight the need to extend the appropriate reliance metrics to realistic usage scenarios.

Generalization of the Appropriate Reliance Metrics. Through our experiments, we highlight the flexibility of the Relative LLM

Reliance, Relative Self-Reliance, and Appropriate Reliance Ratio metrics in benchmarking interventions across two diverse task formats. While prior works have focused on binary or multiple choice answers, we demonstrate the suitability of the metrics for numerical targets, which opens up their applicability to additional real-world tasks. These measures enable more nuanced analyses of suboptimal human-LLM collaboration beyond task performance alone, and we recommend their incorporation into future empirical research on LLM usage. The metrics can be generalized to any LLM-assisted task where there is a deterministic notion of optimality (i.e. some outcomes are preferred over others), such as code compilation success for programming, accuracy of arguments in writing persuasive essays, and requirement satisfaction in travel planning. Extending experiments to more open-ended tasks requires further nuance to faithfully capture agreement, reliance, and correctness. However, we theorize some degree of generalization in our high level findings—that cognitive forcing interventions will cause people to approach reliance on LLM content more cautiously.

LLM Design Trade-Offs Based on the Context of Use. The question, “*How can interventions be designed successfully?*”, still remains open. We uncover complex heterogeneity in how interventions performed across different tasks, indicating that optimal designs are likely context-dependent rather than universal. Our work motivates further investigation into the theoretical foundations linking individual factors, tasks characteristics, and LLMs’ response styles, and towards developing a reliable empirical framework to evaluate novel interventions rigorously. Additionally, as our results suggest a trade-off between reducing over-reliance and increasing under-reliance, it is therefore vital that practitioners consider the *context of use* of the LLM tool. This involves a closer evaluation of the stakeholders, users, and risks involved in the decision-making tasks [50].

In medical or legal decision-making, the heightened risk requires accountability, and errors from over-reliance may cause extreme damage if unchecked [15, 87]. In creative tasks where the diversity of humans is valued, over-reliance may introduce an undesirable degree of homogeneity [44]. Within the education space, over-reliance brings about another issue: the unintended reduction of skill development due to relying on the LLM as a cognitive crutch [88]. As AI capabilities advance and eventually eclipse human capabilities, avoiding under-reliance may be preferred in lieu of aversion. However, scaling ability requires scaling oversight, representing further challenges in human-AI collaboration that require carefully calibrated trust and reliance [8].

Human Validation for Technical Development. Ensuring the success of human-AI interactions requires bi-directional design and validation from both AI and HCI researchers [70, 80]. Technical AI research has progressed significantly towards understanding LLMs’ reasoning pathways, reducing hallucinations, and calibrating the confidence of outputs [3, 24, 51, 90]. However, such works often rely on automating evaluations by using existing benchmarks or artificially simulating users, considering only theoretical design desiderata and bypassing human evaluation. These evaluation methods do not ascertain whether these methods will have the desired effects when deployed to real users. For example, we showed that highlighting log probabilities to calibrate confidence did not have

the intended effect on humans, who may have found the added information too confusing to optimally leverage. Not only did reliance and performance outcomes lack improvement, we also found that participants’ confidence changes were poorly calibrated, with a significant increase in post-task confidence when LLM advice is used *inappropriately*. We therefore encourage the broader research community, particularly the technical AI community, to capture multi-dimensional measures from real users.

5.3 Limitations

Overall, our results would be further strengthened by future research that explores LLM reliance in more naturalistic and long-term settings, offering a clearer picture of how users assess and use model outputs in the real world. Incorporating qualitative measures can also provide more nuanced feedback on *how* the participants regarded the advice and made choices.

Engagement of Participants in Tasks. The generalizability of our results is primarily limited by our design of a lab experiment with crowdworkers recruited from Prolific. Since participants knew that they were participating in an experiment, they may have inferred that they were being evaluated on LLM reliance and become more cautious about their reliance behaviours. Although we tried to emulate the risks and benefits of real-world decision-making by rewarding participants for correctness, this may have encouraged participants to ‘hedge’: to maximize their chances of earning *some* bonus payment by choosing different answers at each stage. Our two-stage question answering set-up also emulates a specific type of decision-making where users are already engaged in the task. In fact, having users answer questions independently before advice can be considered a form of cognitive forcing even without additional interventions [2]. As LLM usage becomes more common, it is expected that many users will bypass this critical engagement and rely on LLMs by default, which can reflect in reliance outcomes differently.

External Validity of the Tasks. Our chosen tasks do not perfectly translate to the usage contexts for which people consult LLMs in the real world. While multiple-choice questions provided clean correctness and reliance metrics, real-world tasks predominantly consist of open-ended questions and generative tasks (such as writing and coding). Knowledge-based questions were avoided in this study to ensure that the questions would be accessible to participants; however, real questions posed to LLMs frequently require synthesis of more complex ideas and knowledge bases, which may be more susceptible to factual hallucinations.

Realism of Interacting with the LLM. Due to the nature of the survey study, our participants were unable to interactively ask follow-up questions or use sanity checks to verify the model’s outputs and prevent consequential errors. To ensure clean reliance metrics, they saw fixed advice and were asked not to consult external resources. Although we used direct outputs generated through tailored LLM prompts whenever possible, we did apply manual adjustments in some interventions and artificially manipulated the recommendation in the LLM’s advice. The distribution of advice quality was held constant regardless of the LLM’s true tendencies when responding to the question to prevent question difficulty from

being a confound. We only explore one specific LLM, GPT-4o from OpenAI, so chatbots like Claude⁷, Copilot⁸, Gemini⁹, and other open-source models may reflect different conversational styles that can implicitly affect user behaviour. We also did not explore varying the system prompts to steer the LLM into embodying different personas. Furthermore, the framing of the LLM (e.g., ChatGPT vs. a generic chatbot) can prime user expectations of the quality and accuracy of the responses.

Selection of Interventions. To operationalize our research questions, we made certain design choices in the implementation of the interventions. For example, we used a naive method of highlighting based on token generation probability, while prior works have identified that likelihood of editing is a better indicator [77]. Our evaluation of the interventions also did not include any extensive training or longer-term follow-ups. As such, the impacts of each intervention in our studies are not necessarily representative of how they would impact reliance if optimally implemented in the real world. In particular, it is not clear if static disclaimers would persistently encourage appropriate reliance in deployment where users may experience alarm fatigue [20], whereas dynamic disclaimers may have a longer impact [7]. We explored a very small space of potential interventions, and encourage further human-centered evaluation of other uncertainty calibration techniques (e.g., confidence score [79], semantic expressions of uncertainty [39, 51, 74], and contrastive explanations [71]) and setting appropriate expectations of the LLM (e.g., model report card [86] and usage tutorial [17]).

6 Conclusion

We measured the effects of three different LLM interventions on users' reliance behaviours and confidence levels in two difficult LLM-assisted question-answering tasks. We find that adding a persistent disclaimer is the most effective at improving appropriate reliance (in the LSAT task) without significantly increasing the time needed to engage with the LLM advice, while both highlighting uncertain tokens and removing direct answers only reduced over-reliance and increased the engagement time. In addition, people's confidence levels were poorly calibrated, generally exhibiting higher confidence increases when they made the wrong reliance decision. Based on our results, we call for the following: rigorous and human-centered evaluation of novel LLM reliance interventions; consideration for trade-offs between over- and under-reliance based on the context of use; and future research in naturalistic human-LLM collaboration tasks.

Acknowledgments

We thanks Nicholas Carlini for helpful discussions during the early phase of the project. We further thank Lev McKinney, Lillio Mok, Harsh Kumar, Nakwon Rim, Paula Aoyagui, Zixin Zhao, and members of the UofT Computational Social Science Lab for feedback on the experimental methods and the draft of the manuscript.

⁷<https://claude.ai/>

⁸<https://copilot.microsoft.com/>

⁹<https://gemini.google.com/>

References

- [1] Ravinithesh Annapureddy, Alessandro Fornaroli, and Daniel Gatica-Perez. 2024. Generative AI literacy: Twelve defining competencies. *Digital Government: Research and Practice* (2024).
- [2] Zahra Ashktorab, Qian Pan, Werner Geyer, Michael Desmond, Marina Danilevsky, James M Johnson, Casey Dugan, and Michelle Bachman. 2024. Emerging Reliance Behaviors in Human-AI Text Generation: Hallucinations, Data Quality Assessment, and Cognitive Forcing Functions. *arXiv preprint arXiv:2409.08937* (2024).
- [3] Suriya Ganesh Ayyamperumal and Limin Ge. 2024. Current state of LLM Risks and AI Guardrails. *arXiv preprint arXiv:2406.12934* (2024).
- [4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16.
- [5] Evan Becker and Stefano Soatto. 2024. Cycles of Thought: Measuring LLM Confidence through Stable Explanations. *arXiv preprint arXiv:2406.03441* (2024).
- [6] Astrid Bertrand, Rafik Belloum, James R Eagan, and Winston Maxwell. 2022. How cognitive biases affect XAI-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 78–91.
- [7] Jessica Bo, Harsh Kumar, Michael Liut, and Ashton Anderson. 2024. Disclosures & Disclaimers: Investigating the Impact of Transparency Disclosures and Reliability Disclaimers on Learner-LLM Interactions. *Available at SSRN* (2024).
- [8] Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, et al. 2022. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540* (2022).
- [9] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction* 5, CSCW1 (2021), 1–21.
- [10] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. 2020. A systematic review of algorithm aversion in augmented decision making. *Journal of behavioral decision making* 33, 2 (2020), 220–239.
- [11] Alice Cai, Ian Arawjo, and Elena L Glassman. 2024. Antagonistic AI. *arXiv preprint arXiv:2402.07350* (2024).
- [12] Shiye Cao and Chien-Ming Huang. 2022. Understanding user reliance on AI in assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–23.
- [13] Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine* (2023).
- [14] Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction* 7, CSCW2 (2023), 1–32.
- [15] Inyoung Cheong, King Xia, KJ Kevin Feng, Quan Ze Chen, and Amy X Zhang. 2024. (A) I Am Not a Lawyer, But... Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2454–2469.
- [16] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2024. Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 103–119.
- [17] Chun-Wei Chiang and Ming Yin. 2022. Exploring the effects of machine learning literacy interventions on laypeople's reliance on machine learning models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 148–161.
- [18] Avishek Choudhury and Hamid Shamszare. 2023. Investigating the impact of user trust on the adoption and use of ChatGPT: survey analysis. *Journal of Medical Internet Research* 25 (2023), e47184.
- [19] Katherine M Collins, Valerie Chen, Ilia Sucholutsky, Hannah Rose Kirk, Malak Sadek, Holli Sargeant, Ameet Talwalkar, Adrian Weller, and Umang Bhatt. 2024. Modulating Language Model Experiences through Frictions. *arXiv preprint arXiv:2407.12804* (2024).
- [20] Maria Cvach. 2012. Monitor alarm fatigue: an integrative review. *Biomedical instrumentation & technology* 46, 4 (2012), 268–277.
- [21] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General* 144, 1 (2015), 114.
- [22] Hyo Jin Do, Rachel Ostrand, Justin D Weisz, Casey Dugan, Prasanna Sattigeri, Dennis Wei, Keerthiram Murugesan, and Werner Geyer. 2024. Facilitating Human-LLM Collaboration through Factuality Scores and Source Attributions. *arXiv preprint arXiv:2405.20434* (2024).
- [23] Eva Eigner and Thorsten Händler. 2024. Determinants of llm-assisted decision-making. *arXiv preprint arXiv:2402.17385* (2024).

- [24] Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration. *arXiv preprint arXiv:2402.00367* (2024).
- [25] Sharon A Ferguson, Paula Akemi Aoyagui, and Anastasia Kuzminykh. 2023. Something Borrowed: Exploring the Influence of AI-Generated Explanation Text on the Composition of Human Explanations. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7.
- [26] Raymond Fok and Daniel S Weld. 2023. In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. *AI Magazine* (2023).
- [27] Yue Fu, Sami Foell, Xuhai Xu, and Alexis Hiniker. 2024. From Text to Self: Users' Perception of AIMC Tools on Interpersonal Communication and Self. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
- [28] Krzysztof Z Gajos and Lena Mamykina. 2022. Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 794–806.
- [29] Ziyang Guo, Yifan Wu, Jason D Hartline, and Jessica Hullman. 2024. A Decision Theoretic Framework for Measuring AI Reliance. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 221–236.
- [30] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [31] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).
- [32] Kori Inkpen, Shreya Chappidi, Keri Mallari, Besmira Nushi, Divya Ramesh, Pietro Michelucci, Vani Mandava, Libuše Hannah Vepřek, and Gabrielle Quinn. 2023. Advancing human-AI complementarity: The impact of user expertise and algorithmic tuning on joint decision making. *ACM Transactions on Computer-Human Interaction* 30, 5 (2023), 1–29.
- [33] Samia Kabir, David N Udo-Imeh, Bonan Kou, and Tianyi Zhang. 2024. Is stack overflow obsolete? an empirical study of the characteristics of chatgpt answers to stack overflow questions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
- [34] Harmpreet Kaur, Matthew R Conrad, Davis Rule, Cliff Lampe, and Eric Gilbert. 2024. Interpretability Gone Bad: The Role of Bounded Rationality in How Practitioners Understand Machine Learning. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–34.
- [35] Majeed Kazemitaabaa, Justin Chow, Carl Ka To Ma, Barbara J Ericson, David Weinrop, and Tovi Grossman. 2023. Studying the effect of AI code generators on supporting novice learners in introductory programming. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–23.
- [36] Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782* (2024).
- [37] Antino Kim, Mochen Yang, and Jingjing Zhang. 2023. When algorithms err: Differential impact of early vs. late errors on users' reliance on algorithms. *ACM Transactions on Computer-Human Interaction* 30, 1 (2023), 1–36.
- [38] Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024. Can LLMs Produce Faithful Explanations For Fact-checking? Towards Faithful Explainable Fact-Checking via Multi-Agent Debate. *arXiv preprint arXiv:2402.07401* (2024).
- [39] Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 822–835.
- [40] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.
- [41] Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. 2023. ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports* 13, 1 (2023), 4569.
- [42] Harsh Kumar, Ilya Musabirov, Mohi Reza, Jiakai Shi, Anastasia Kuzminykh, Joseph Jay Williams, and Michael Liut. 2023. Impact of guidance and interaction strategies for LLM use on Learner Performance and perception. *arXiv preprint arXiv:2310.13712* (2023).
- [43] Harsh Kumar, David M Rothschild, Daniel G Goldstein, and Jake M Hofman. 2023. Math Education with Large Language Models: Peril or Promise? Available at SSRN 4641653 (2023).
- [44] Harsh Kumar, Jonathan Vincentius, Ewan Jordan, and Ashton Anderson. 2024. Human Creativity in the Age of LLMs: Randomized Experiments on Divergent and Convergent Thinking. *arXiv preprint arXiv:2410.03703* (2024).
- [45] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).
- [46] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.
- [47] Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuan-dong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, et al. 2024. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *arXiv preprint arXiv:2403.07183* (2024).
- [48] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021).
- [49] Q Vera Liao and Jennifer Wortman Vaughan. 2023. Ai transparency in the age of LLMs: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941* (2023), 5368–5393.
- [50] Q Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. 2022. Connecting algorithmic research and usage contexts: a perspective of contextualized evaluation for explainable AI. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 147–159.
- [51] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334* (2022).
- [52] Chung Kwan Lo. 2023. What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences* 13, 4 (2023), 410.
- [53] Zhuoran Lu, Zhuoyan Li, Chun-Wei Chiang, and Ming Yin. 2023. Strategic Adversarial Attacks in AI-assisted Decision Making to Reduce Human Trust and Reliance.. In *IJCAI. 2020–2028*.
- [54] Zhuoran Lu and Ming Yin. 2021. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [55] Qianou Ma, Hua Shen, Kenneth Koedinger, and Sherry Tongshuang Wu. 2024. How to Teach Programming in the AI Era? Using LLMs as a Teachable Agent for Debugging. In *International Conference on Artificial Intelligence in Education*. Springer, 265–279.
- [56] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2024. Towards human-ai deliberation: Design and evaluation of llm-empowered deliberative ai for ai-assisted decision-making. *arXiv preprint arXiv:2403.16812* (2024).
- [57] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [58] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.
- [59] Raffaella Misuraca, Palmira Faraci, Amelia Gangemi, Floriana A Carmeci, and Silvana Miceli. 2015. The Decision Making Tendency Inventory: A new measure to assess maximizing, satisficing, and minimizing. *Personality and Individual Differences* 85 (2015), 111–116.
- [60] Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. 2024. Using an LLM to help with code understanding. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. 1–13.
- [61] An T Nguyen, Aditya Kharosekar, Saumya Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. 2018. Believe it or not: designing a human-ai partnership for mixed-initiative fact-checking. In *Proceedings of the 31st annual ACM symposium on user interface software and technology*. 189–199.
- [62] Mahsan Nouriani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring bias affects mental model formation and user reliance in explainable AI systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 340–350.
- [63] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.
- [64] Omer Reingold, Judy Hanwen Shen, and Aditi Talati. 2024. Dissenting Explanations: Leveraging Disagreement to Reduce Model Overreliance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 21537–21544.
- [65] Vildan Salikutluk, Dorothea Koert, and Frank Jäkel. 2023. Interacting with large language models: A case study on AI-aided brainstorming for guesstimation problems. In *HHAI 2023: Augmenting Human Intellect*. IOS Press, 153–167.
- [66] Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. *arXiv preprint arXiv:2204.06916* (2022).
- [67] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422.
- [68] Jakob Schoeffler, Maria De-Arteaga, and Niklas Kuehl. 2024. Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.
- [69] Nikhil Sharma, Q Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

- [70] Hua Shen, Tiffany Kneare, Reshma Ghosh, Kenan Alkiel, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, et al. 2024. Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions. *arXiv preprint arXiv:2406.09264* (2024).
- [71] Chenglei Si, Navita Goyal, Sherry Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III, and Jordan Boyd-Graber. 2023. Large Language Models Help Humans Verify Truthfulness—Except When They Are Convincingly Wrong. *arXiv preprint arXiv:2310.12558* (2023).
- [72] Saniat Javid Sohraward, Y Kelly Wu, Andrea Hickerson, and Matthew Wright. 2024. Dungeons & Deepfakes: Using scenario-based role-play to study journalists' behavior towards using AI-based verification tools for video content. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
- [73] Sofia Eleni Spatharioti, David M Rothschild, Daniel G Goldstein, and Jake M Hofman. 2023. Comparing traditional and lilm-based search for consumer choice: A randomized experiment. *arXiv preprint arXiv:2307.03744* (2023).
- [74] Mark Steyvers, Heliодoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas Mayer, and Padhraic Smyth. 2024. The calibration gap between model and human confidence in large language models. *arXiv preprint arXiv:2401.13835* (2024).
- [75] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The metacognitive demands and opportunities of generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–24.
- [76] Keyon Vafa, Ashesh Rambachan, and Sendhil Mullainathan. 2024. Do Large Language Models Perform the Way People Expect? Measuring the Human Generalization Function. *arXiv preprint arXiv:2406.01382* (2024).
- [77] Helena Vasconcelos, Gagan Bansal, Adam Fournier, Q Vera Liao, and Jennifer Wortman Vaughan. 2023. Generation probabilities are not enough: Exploring the effectiveness of uncertainty highlighting in AI-powered code completions. *arXiv preprint arXiv:2302.07248* (2023).
- [78] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.
- [79] Yuvraj Virk, Premkumar Devanbu, and Toufique Ahmed. 2024. Enhancing Trust in LLM-Generated Code Summaries with Calibrated Confidence Scores. *arXiv preprint arXiv:2404.19318* (2024).
- [80] Kailas Vodrahalli, Tobias Gerstenberg, and James Y Zou. 2022. Uncalibrated models can improve human-ai collaboration. *Advances in Neural Information Processing Systems* 35 (2022), 4004–4016.
- [81] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).
- [82] Xinru Wang and Ming Yin. 2021. Are explanations helpful? A comparative study of the effects of explanations in ai-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 318–328.
- [83] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
- [84] Robert Wolfe and Tanushree Mitra. 2024. The Impact and Opportunities of Generative AI in Fact-Checking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1531–1543.
- [85] Ruiyun Xu, Yue Feng, and Hailiang Chen. 2023. Chatgpt vs. google: A comparative study of search performance and user experience. *arXiv preprint arXiv:2307.01135* (2023).
- [86] Blair Yang, Fuyang Cui, Keiran Paster, Jimmy Ba, Pashootan VaeziPoor, Silviu Pitis, and Michael R Zhang. 2024. Report Cards: Qualitative Evaluation of Language Models Using Natural Language Summaries. *arXiv preprint arXiv:2409.00844* (2024).
- [87] Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science* 2, 4 (2023), 255–263.
- [88] Chumpeng Zhai, Santoso Wibowo, and Lily D Li. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments* 11, 1 (2024), 28.
- [89] Peter Zhang. 2023. Taking advice from chatgpt. *arXiv preprint arXiv:2305.11888* (2023).
- [90] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemaio Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219* (2023).
- [91] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.
- [92] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).
- [93] Kaitlyn Zhou, Jena D Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. 2024. Rel-AI: An Interaction-Centered Approach To Measuring Human-LM Reliance. *arXiv preprint arXiv:2407.07950* (2024).

A Task Question Examples

Figure A.1 shows an example of the LSAT Logical Reasoning task question. Participants answer independently in Stage 1, and then re-answer the same question with randomly sampled LLM advice in Stage 2.

Each question is based on reasoning from a short passage; choose the best answer that fits the logic of the passage only. No factual knowledge is needed.

Question:

In jurisdictions where use of helmets is optional for cyclists when they are not riding on the road, cyclists who always use helmets are less likely to be involved in a collision than are cyclists who use helmets only when riding on the road. Yet Transportation Department records show that making helmets mandatory for cyclists at all times does not reduce the overall number of collisions.

Which one of the following, if true, most helps to resolve the apparent discrepancy in the information above?

- A) In jurisdictions where use of helmets is optional when the cyclist is not riding on a road, one cyclist in four wears a helmet for daytime riding on trails.
- B) A law making use of helmets mandatory at all times is not especially difficult to enforce.
- C) Only very careful riders wear helmets when their use is not legally required.
- D) There are some jurisdictions in which it is illegal to wear helmets when riding on a trail.
- E) The jurisdictions where use of helmets is mandatory at all times are those that are more urban.

(a) Stage 1 of the “Road” LSAT Logical Reasoning question.

Now you have access to ChatGPT’s advice. Will you change your answer? Note that this is ChatGPT’s response to the question only; it does not know what you chose.

Question:

In jurisdictions where use of helmets is optional for cyclists when they are not riding on the road, cyclists who always use helmets are less likely to be involved in a collision than are cyclists who use helmets only when riding on the road. Yet Transportation Department records show that making helmets mandatory for cyclists at all times does not reduce the overall number of collisions.

Which one of the following, if true, most helps to resolve the apparent discrepancy in the information above?

Your previous answer: E) The jurisdictions where use of helmets is mandatory at all times are those that are more urban.

ChatGPT’s advice:

Option C) resolves this discrepancy by suggesting that the group of cyclists who choose to wear helmets voluntarily when not mandatory are inherently more careful and cautious riders. They are likely to engage in safer riding practices, which reduces their collision rates. When helmet use is made mandatory for all cyclists, it includes those who are not as careful. Thus, the overall collision rate remains unchanged because the mandatory helmet law does not change the riding behavior of those cyclists who are less careful.

Please enter what you believe is the correct answer:

- A) In jurisdictions where use of helmets is optional when the cyclist is not riding on a road, one cyclist in four wears a helmet for daytime riding on trails.
- B) A law making use of helmets mandatory at all times is not especially difficult to enforce.
- C) Only very careful riders wear helmets when their use is not legally required.
- D) There are some jurisdictions in which it is illegal to wear helmets when riding on a trail.
- E) The jurisdictions where use of helmets is mandatory at all times are those that are more urban.

(b) Stage 2 of the “Road” question with an example of base (Control) advice.

Figure A.1: Screenshot of the “Road” LSAT Logical Reasoning question, taken from Qualtrics.

Figures A.2 and A.3 show screenshots of Numerical Estimation task questions, with the same two-stage design.

How many chickpeas are in this glass jar?
Make sure to estimate carefully!



Please enter a number representing the number of chickpeas:

(a) Stage 1 of the “Chickpea” Numerical Estimation question.

Now you have access to ChatGPT's advice. Will you change your answer? Note that this is ChatGPT's response to the question only; it does not know your answer.



Your answer: 200

ChatGPT's answer:

To estimate the number of chickpeas in the jar, we can use the following steps:

1. Estimate the volume of the jar:
- From the images, let's assume the jar is about 10 cm in height and 6 cm in diameter.
2. Calculate the volume of the jar:
- Volume of a cylinder = $\pi \times \text{radius}^2 \times \text{height}$
- Volume = $3.14 \times (6 \text{ cm} / 2)^2 \times 10 \text{ cm} = 3.14 \times 9 \text{ cm}^2 \times 10 \text{ cm} = 282.6 \text{ cm}^3$
3. Estimate the volume of one chickpea:
- Chickpeas are roughly spherical.
- Assume the diameter of a chickpea is about 1 cm.
- Volume of a sphere = $4/3 \times \pi \times \text{radius}^3$
- Volume ≈ $4/3 \times 3.14 \times (1 \text{ cm} / 2)^3 = 4/3 \times 3.14 \times 0.125 \text{ cm}^3 \approx 0.52 \text{ cm}^3$
4. Estimate the number of chickpeas in the jar:
- Divide the volume of the jar by the volume of one chickpea.
- Number of chickpeas = $282.6 \text{ cm}^3 / 0.52 \text{ cm}^3 = 544$

So, there are approximately 544 chickpeas in the jar.

Please enter a number representing the number of chickpeas:

(b) Stage 2 of the “Chickpea” question with an example of base (Control) advice.

Figure A.2: Screenshot of the “Chickpea” Numerical Estimation question, taken from Qualtrics

How many jellybeans are in this glass jar?
Make sure to estimate carefully!



Please enter a number representing the number of jellybeans:

(a) Stage 1 of the “Jellybean” Numerical Estimation question.

Now you have access to ChatGPT's advice. Will you change your answer? Note that this is ChatGPT's response to the question only; it does not know your answer.



ChatGPT's answer:

To estimate the number of jelly beans in the jar, we can use a few assumptions and simple steps:

1. Measure the Jar Dimensions: Assume the jar is a cylinder. Estimate its height at around 20 cm and diameter at around 10 cm.

2. Calculate the Volume of the Jar:

- Volume of a cylinder = $\pi \times \text{radius}^2 \times \text{height}$
- Radius = Diameter / 2 = 10 cm / 2 = 5 cm
- Volume = $3.14 \times (5 \text{ cm})^2 \times 20 \text{ cm} = 1570 \text{ cubic cm}$

3. Estimate the Volume of a Jelly Bean:

- Assume each jelly bean is roughly a cylinder with dimensions of approximately 1.5 cm in length and 1 cm in diameter.
- Volume of a jelly bean = $\pi \times (0.5 \text{ cm})^2 \times 1.5 \text{ cm} = 1.17 \text{ cubic cm}$

4. Account for Packing Efficiency:

- Jelly beans are not packed perfectly tightly. On average, there is around 30-40% air space left.
- Packing efficiency is therefore roughly 60-70%.
- Using 65% as the packing density: Effective volume of the jar for jelly beans = $1570 \text{ cubic cm} \times 0.65 \approx 1020 \text{ cubic cm}$.

5. Calculate the Number of Jelly Beans:

- Number of jelly beans = Effective volume / Volume of one jelly bean
- Number of jelly beans = $1020 \text{ cubic cm} / 1.17 \text{ cubic cm} \approx 872$

So, the jar likely contains around 872 jelly beans.

Please enter a number representing the number of jellybeans:

(b) Stage 2 of the “Jellybean” question with an example of base (Control) advice.

Figure A.3: Screenshot of the “Jellybean” Numerical Estimation question, taken from Qualtrics

B Control and Intervention Advice Examples

Figures B.1 and B.2 show sample ChatGPT advice for all four conditions for a LSAT Logical Reasoning question and a Numerical Estimation, respectively.

ChatGPT's advice:

Option C) resolves this discrepancy by suggesting that the group of cyclists who choose to wear helmets voluntarily when not mandatory are inherently more careful and cautious riders. They are likely to engage in safer riding practices, which reduces their collision rates. When helmet use is made mandatory for all cyclists, it includes those who are not as careful. Thus, the overall collision rate remains unchanged because the mandatory helmet law does not change the riding behavior of those cyclists who are less careful.

(a) Control

ChatGPT's advice:

Option C) resolves this discrepancy by suggesting that the group of cyclists who choose to wear helmets voluntarily when not required are inherently more careful and cautious riders. They are less likely to be involved in a collision due to their careful nature, not solely because of the helmet. Mandating helmets would not change the behavior of the less careful cyclists, hence it does not reduce overall collision numbers. This explains why mandatory helmet laws do not affect the overall number of collisions.

(c) Uncertainty Highlighting

(b) Reliance Disclaimer

ChatGPT's advice:

The apparent discrepancy lies in the observation that cyclists who consistently wear helmets, even when not required by law, experience fewer collisions, yet mandatory helmet laws do not reduce the overall number of collisions. To reconcile this, consider that only those riders who are inherently more cautious are likely to wear helmets voluntarily, regardless of the legal requirement. This careful group avoids collisions not primarily due to the helmets but because of their cautious riding habits overall. Consequently, mandatory helmet laws affect all cyclists equally, including those who may be less careful or more reckless. These less cautious riders might only wear helmets because the law requires it, but their riding behavior does not change, hence the overall collision rates remain unchanged.

(d) Implicit Answer

Figure B.1: Screenshots of sample LLM advice for all four conditions for an LSAT Logical Reasoning question, taken from Qualtrics.

ChatGPT's answer:

- Estimate the chickpeas' dimensions:
- Approximate diameter of one chickpea: 1 cm.

2. Calculate the jar's volume:

- Assume the jar's height = 12 cm.
- Assume the diameter = 7 cm.
- Use the formula for the volume of a cylinder: $V = \pi \times r^2 \times h$, where r is the radius and h is the height.
- Radius = $7/2 = 3.5$ cm.
- Volume of the jar = $3.14 \times (3.5 \text{ cm})^2 \times 12 \text{ cm} = 461 \text{ cm}^3$.

3. Estimate the volume of one chickpea:

- Since chickpeas are roughly spherical, use the formula for the volume of a sphere: $V = 4/3 \times \pi \times r^3$.
- Radius of chickpea = 0.5 cm.
- Volume of one chickpea = $4/3 \times 3.14 \times (0.5 \text{ cm})^3 = 0.52 \text{ cm}^3$.

4. Estimate the number of chickpeas in the jar:

- Jar volume / Chickpea volume = $461 \text{ cm}^3 / 0.52 \text{ cm}^3 \approx 887$.

Therefore, there are approximately 887 chickpeas in the jar.

ChatGPT's answer:

- Estimate the chickpeas' dimensions:
- Approximate diameter of one chickpea: 1 cm.

2. Calculate the jar's volume:

- Assume the jar's height = 12 cm.
- Assume the diameter = 7 cm.
- Use the formula for the volume of a cylinder: $V = \pi \times r^2 \times h$, where r is the radius and h is the height.
- Radius = $7/2 = 3.5$ cm.
- Volume of the jar = $3.14 \times (3.5 \text{ cm})^2 \times 12 \text{ cm} = 461 \text{ cm}^3$.

3. Estimate the volume of one chickpea:

- Since chickpeas are roughly spherical, use the formula for the volume of a sphere: $V = 4/3 \times \pi \times r^3$.
- Radius of chickpea = 0.5 cm.
- Volume of one chickpea = $4/3 \times 3.14 \times (0.5 \text{ cm})^3 = 0.52 \text{ cm}^3$.

4. Estimate the number of chickpeas in the jar:

- Jar volume / Chickpea volume = $461 \text{ cm}^3 / 0.52 \text{ cm}^3 \approx 887$.

Therefore, there are approximately 887 chickpeas in the jar.
Remember to verify this information and adapt it to your specific problem as I may not have complete information about your context.

(a) Control

(b) Reliance Disclaimer

ChatGPT's answer:

- Estimate the chickpeas' dimensions:
- Approximate diameter of one chickpea: 1 cm.

2. Calculate the jar's volume:

- Assume the jar's height = 12 cm.
- Assume the diameter = 7 cm.
- Use the formula for the volume of a cylinder: $V = \pi \times r^2 \times h$, where r is the radius and h is the height.
- Radius = $7/2 = 3.5$ cm.
- Volume of the jar = $3.14 \times (3.5 \text{ cm})^2 \times 12 \text{ cm} = 461 \text{ cm}^3$.

3. Estimate the volume of one chickpea:

- Since chickpeas are roughly spherical, use the formula for the volume of a sphere: $V = 4/3 \times \pi \times r^3$.
- Radius of chickpea = 0.5 cm.
- Volume of one chickpea = $4/3 \times 3.14 \times (0.5 \text{ cm})^3 = 0.52 \text{ cm}^3$.

4. Estimate the number of chickpeas in the jar:

- Jar volume / Chickpea volume = $461 \text{ cm}^3 / 0.52 \text{ cm}^3 \approx 887$.

Therefore, there are approximately 887 chickpeas in the jar.

(c) Uncertainty Highlighting

(d) Implicit Answer

Figure B.2: Screenshots of sample LLM advice for all four conditions for an Numerical Estimation question, taken from Qualtrics.

C Levels of Advice in Numerical Estimation

The five levels of LLM advice provided for the two questions in the Numerical Estimation task are illustrated in Figure C.1, alongside the distribution of the participants' estimates before and after receiving the advice.

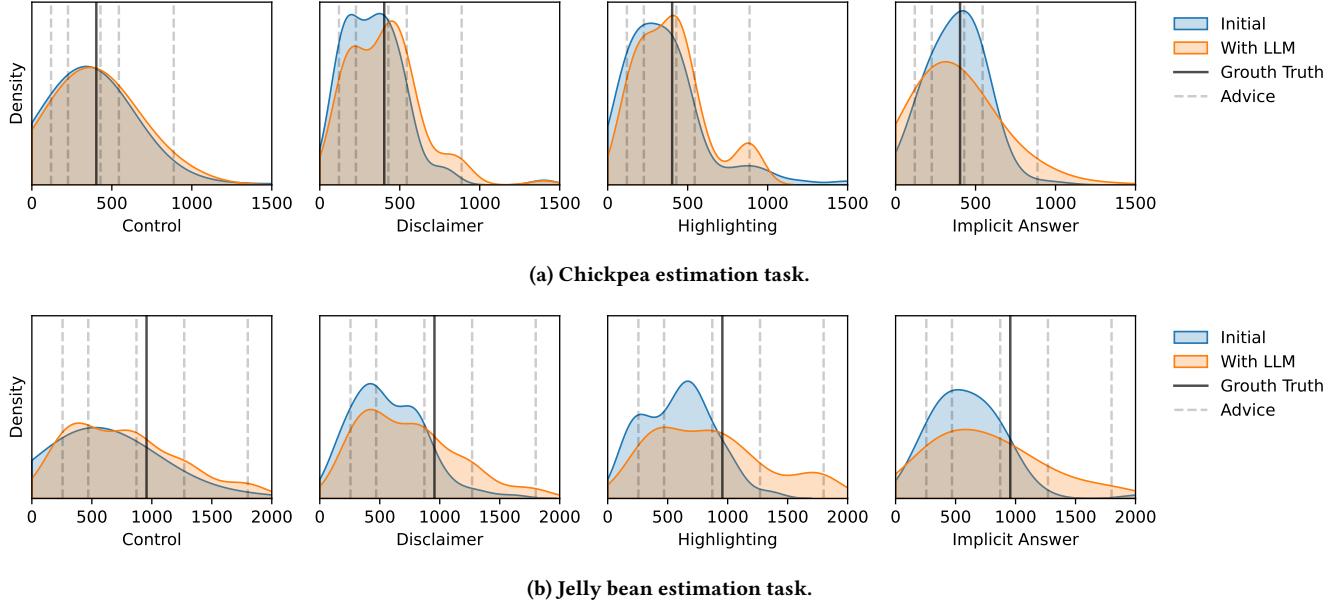


Figure C.1: Distribution of the participants' initial (blue) and final (orange) estimations for both Numerical Estimation questions. The true count is indicated by the black line. The five gray dotted lines mark the different levels of LLM advice that were randomly provided to the participants.

D Confidence Analysis for Other WoA Thresholds

The confidence calibration graphs for Numerical Estimation at the other WoA thresholds (0.01 and 0.99) are listed here as Figure D.1 and D.2.

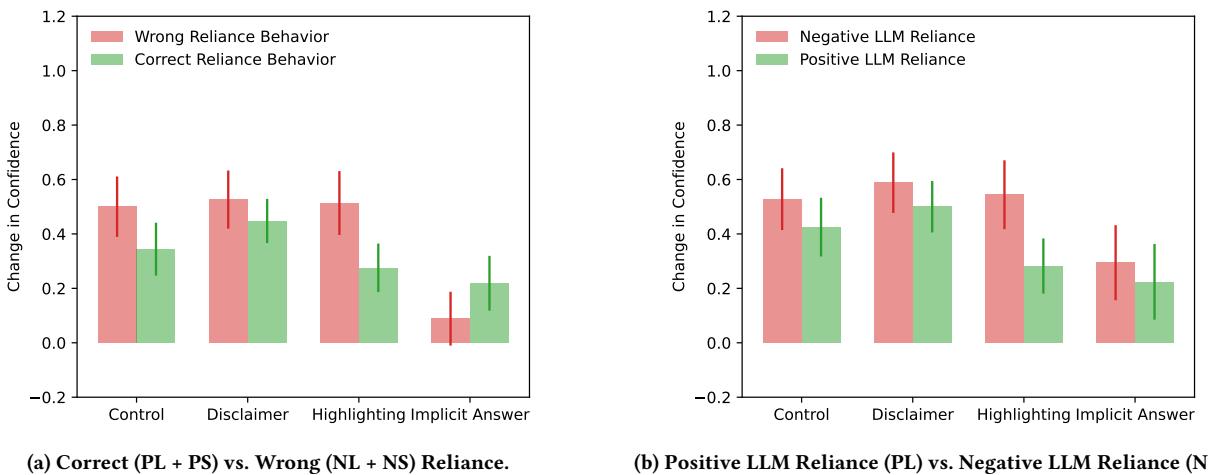


Figure D.1: Change in confidence at the question level for Numerical Estimation with a WoA threshold of 0.01 (most sensitive threshold for reliance).

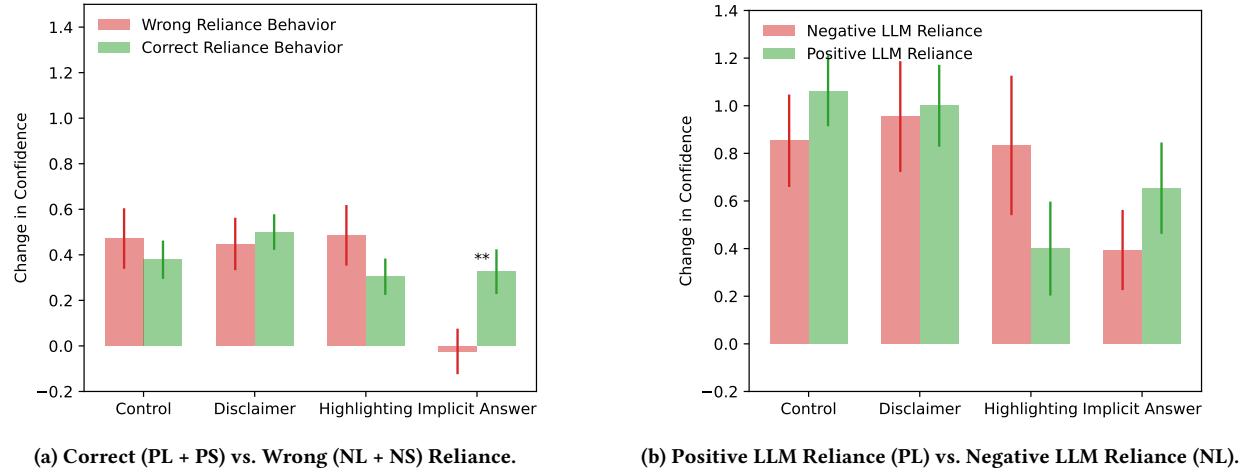


Figure D.2: Change in confidence at the question level for Numerical Estimation with a WoA threshold of 0.99 (least sensitive threshold for reliance).

E Effect of Prior LLM Experience

As an additional analysis, we examine whether participants' baseline frequencies of using LLM tools affected their behaviours and perceptions. The participants reported their prior experience with LLMs in the survey as *never tried*, *almost never but have tried*, *monthly*, *weekly*, and *daily*, which we map into ordinal values from 0 to 4. When controlling for condition, a partial Spearman correlation reveals no significant correlations between LLM usage and reliance on the advice for either task, LSAT ($r = .02, p = .5$ for a binary reliance variable) or Numerical Estimation ($r = .06, p = .1$ for the continuous WoA variable). Nor are there correlations with the changes in confidence for LSAT ($r \approx 0, p = .9$) and Numerical Estimation ($r = .03, p = .4$). However, there is significant positive association between frequency of usage and all three categories of post-task perceptions of the LLM: **Confidence** ($r = .12, p < .05$), **Accuracy** ($r = .14, p < .01$), and **Helpfulness** ($r = .13, p < .01$). This suggests that while people who were more familiar and open to using LLM tools generally hold higher view of ChatGPT's advice (across all conditions), it did not impact actual rates of reliance in the task.