Confirmation Bias: A Challenge for Scalable Oversight

Gabriel Recchia [1*], Chatrik Singh Mangat[2] , Jinu Nyachhyon[3] , Mridul Sharma[3] , Callum Canavan[4] , Dylan Epstein-Gross[5] , Muhammed Abdulbari[6]

[1]Modulo Research
[2]Vector Research
[3]Information and Language Processing Research Lab (ILPRL), Kathmandu University
[4]Hidden Variable Limited
[5]Princeton University, Department of Computer Science
[6]Georgia Institute of Technology, School of Computer Science

Abstract

Scalable oversight protocols aim to empower evaluators to verify outputs of AI models more capable than themselves. However, human evaluators' biases can lead to systematic errors. We reanalyse prior work which seemed to show benefits from a simple protocol, and suggest that a strategy of "answer the question myself if I know the answer, defer to the language model otherwise" likely contributed to its positive results. This strategy fails to provide meaningful oversight when model capability increases. We also present two experiments examining simple protocols, finding no overall advantage for either. In our main experiment, participants in control and intervention groups became more confident in the system's answers after conducting online research, even when those answers were incorrect. Our null results are restricted to the simple protocols and settings tested, and say little regarding the promise of scalable oversight more broadly. Nevertheless, they underscore the importance of testing the degree to which protocols are robust to confirmation bias, whether they outperform a strategy of simple deference to the model being evaluated, and whether performance scales with increasing problem difficulty and model capability.

Code & Data — https://github.com/modulo-research/bias
Extended Version — https://arxiv.org/abs/2507.19486

1 Introduction

Research into scalable oversight aims to design protocols which ensure that artificial intelligence (AI) systems remain aligned with human values and intentions as they become increasingly capable. As large language models (LLMs) and other AI systems address increasingly difficult problems, naive approaches to evaluation may fail as human feedback becomes unreliable or too costly to obtain. A common framing of the scalable oversight problem imagines a weak but 'aligned' judge, such as a human or trusted model, which we would like to evaluate the output of a more powerful system, in a manner such that the powerful system becomes more aligned when provided with these evaluations as feedback. Scalable oversight research aims to devise and test approaches by which the judge can leverage the capabilities of the model being overseen (or other powerful models) to successfully evaluate its output (Christiano, Shlegeris, and Amodei 2018; Engels et al. 2025).

Empirical research on scalable oversight protocols aims to evaluate their efficacy and to identify potential changes that might make them more effective in suboptimal conditions (Leike et al. 2018; Kirchner et al. 2024). For example, consider a judge assessing answers to a binary question by refereeing a structured argument between two instances of a powerful but untrusted model, the format of the scalable oversight paradigm known as *debate* (Irving, Christiano, and Amodei 2018). For this method to succeed, truthful arguments should hold a systematic advantage in persuading the judge. Human cognitive biases are an important source of uncertainty about whether this will be the case in reality (Irving and Askell 2019). Ideal conditions would involve a judge with neutral priors who evaluates a debate without being aware of which side the untrusted model favors, thus minimizing our natural tendency to favor information that supports existing beliefs (Lord, Lepper, and Preston 1984). In practice, circumstances may not always be so favorable, as evaluators' priors may influence their judgments. Furthermore, in some important real-world contexts, it seems plausible that judges will get to observe the output they are being asked to evaluate during or before their evaluation, or that they will have a belief about what that output is likely to be. For example, one key question we might hope that scalable oversight protocols can help us answer is whether a particular system is aligned with respect to some goal, or can otherwise be trusted to reliably perform a particular class of task. Even highly trained judges may find themselves assuming that the model they are evaluating is extremely competent across a wide range of domains, which may incline them to believe *a priori* that the model would also be competent/aligned with respect to the tasks/goals in question.

Specific biases likely to play a role in this context include *automation bias* and *confirmation bias*. Automation bias refers to the tendency of human decision makers to assume that machine-generated solutions are correct and to pay in-

*Corresponding author, gabe@moduloresearch.com

sufficient attention to contradictory information (Cummings 2017), while confirmation bias is a more general inclination to favor information that confirms one's existing beliefs or hypotheses, while disregarding or undervaluing conflicting data (Nickerson 1998). These phenomena have been widely studied in human-computer and human-AI interaction research (see Lyell and Coiera (2017); Bertrand et al. (2022) for overviews) but have been rarely investigated in the context of scalable oversight. When human evaluators of an AI system believe that a system possesses greater knowledge than they do, the risk of inappropriate deference is likely to increase.[1] Some previous scholarship (Section 2) has explored how scalable oversight protocols compare to baselines with the potential to induce automation or confirmation bias, such as direct consultation with the untrusted model, where the untrusted model is randomly assigned to argue either for a correct or incorrect answer. In this setup, systematically deferring to the untrusted model results in low accuracy. This approach allows experiments to have equal power to examine differences between conditions for cases in which the untrusted model is correct vs. incorrect. However, it does not allow for direct exploration of the specific real-world concern described in the previous paragraph, where the worry is that merely being aware that the model generally provides true answers to difficult questions might induce bias strong enough to render an oversight method less effective. Our work addresses this gap by systematically investigating approaches in contexts where human evaluators lack domain expertise, multiple domains are considered, and participants are exposed to the output of an untrusted model, either alone or in conjunction with some intervention. Like Bowman et al. (2022), we explore simple interventions, in hopes of establishing a baseline for comparing more complex oversight methods in the context of bias induced by evaluator beliefs about the system being evaluated.

This paper contributes to the growing literature on scalable oversight by attempting to establish more rigorous baselines against which to evaluate oversight strategies. We provide empirical evidence for specific challenges that oversight protocols must overcome. Our findings highlight that providing meaningful assistance to human judges who have access to the output of an AI system they know to be "correct most of the time, but not all of the time" is not straightforward using the approaches we explore, although we do not investigate more competitive protocols like multi-turn debate, and results may differ when evaluation difficulty is lower. A further motivation for this research was to identify ways to leverage the processes of the most successful judges to improve the quality of feedback for LLM training.

After reviewing related work, we present our efforts examining different protocols. First, we describe preliminary work investigating unstructured interaction with an AI system, preceded either by no intervention or an intervention requiring participants to consider the system's arguments in favor of both possible answers to each question. Our main study explores a setting in which two groups of participants are presented with long-form answers generated by `gpt-4-0613` and have the opportunity to engage in online research to verify the system's claims, but one group additionally receives structured research guidance from the same model in an attempt to help them evaluate its own claims. We conclude by discussing implications and identifying directions for future work.

[1] See Goddard, Roudsari, and Wyatt (2012, 2014); Lyell and Coiera (2017); Bansal et al. (2019); Vodrahalli et al. (2022); Goddard, Roudsari, and Wyatt (2014) for related work on the circumstances under which automation bias is observed and exacerbated.

2 Related Work

Scalable oversight. Scalable oversight research aims to design protocols that enable a weaker system to meaningfully supervise a stronger system. Multiple groups have introduced techniques which vary how the systems interact and what information they can access. Proposed methods include debate (Irving, Christiano, and Amodei 2018), prover-verifier games (Anil et al. 2021), self-critique (Saunders et al. 2022), and market making (Hubinger 2020). There has also been work on how oversight performance scales with model performance (Engels et al. 2025). Many researchers provide concrete examples of feedback from specialized small verifiers improving the performance of a strong LLM in some dimension (e.g. Cobbe et al. (2021); Perez et al. (2022)). In a test of the 'sandwiching' paradigm (Cotra 2021; Irving and Askell 2019), Bowman et al. (2022) found that human-AI teams outperformed humans or models working independently on QA tasks, although others working on human-AI collaboration have found benefits to be highly task-dependent (Vaccaro, Almaatouq, and Malone 2024). Our preliminary work and primary experiment builds on Bowman et al. (2022) by investigating different interventions with the aim of establishing analogous baselines with different assumptions. We also perform additional analyses of their data to better understand their results.

Debate and consultancy. Existing work has placed judges in circumstances with the potential to induce automation bias for use as a baseline against which judges using scalable oversight methods of interest are compared. Michael et al. (2023) compare the performance of judges participating in debate protocols to a baseline they call *consultancy*, where a single 'expert' (e.g., a human or AI model) provides an argument for a single answer. Kenton et al. (2024) contrast consultancy with an alternative where, rather than being assigned whether to argue for the correct or incorrect answer with 50% probability, the expert itself selects the solution to argue in favor of, an approach they refer to as *open consultancy*[2]. Our work differs in that it investigates simple approaches in contexts where participants are exposed to either open consultancy alone or open consultancy plus some intervention.

Human-AI collaboration performance. In a system-

[2] Our implementation of "open consultancy" in this work more closely matches the definition in Kenton et al. (2024) instead of the original definition in Roger (2024). While the original form of open consultancy would provide a more challenging baseline, Kenton et al. (2024) and our work both find that even the weaker version is difficult to beat reliably.

atic review and meta-analysis of 106 experimental studies published between 2020 and 2023, Vaccaro, Almaatouq, and Malone (2024) found that human-AI combinations performed significantly worse on average than the best of humans or AI alone, although they did not find an overall effect in either direction when restricting the analysis to studies from 2022–23. On average, human-AI teams showed performance losses in decision-making tasks but gains in content creation tasks. When AI systems outperformed humans, the collaboration typically underperformed AI working independently; conversely, when humans outperformed AI systems, the collaboration outperformed either group alone. Liu, Lai, and Tan (2021) investigated whether human-AI team performance exceeded solo performance of an AI system alone for out-of-distribution examples on three challenging tasks, and found that it did not. Other theoretical and empirical work explores the concept of human-AI complementarity and methods for harnessing it to maximize performance of human-AI teams (Bansal et al. 2019; Hemmer et al. 2021, 2024; Sperrle et al. 2021).

Biases affecting human-AI collaboration. There has been little empirical work on cognitive biases specifically in scalable oversight settings (Irving and Askell 2019; Buhl et al. 2025), but automation bias and confirmation bias have received wide attention in the literature on human-AI collaboration. Lyell and Coiera (2017) offer a systematic review of automation bias, highlighting that it is exacerbated by cognitive load and occurs in single-tasking and multitasking settings. Bertrand et al. (2022) review work on cognitive biases that affect AI-assisted decision-making in the context of AI systems that produce explanations. Rastogi et al. (2022) explore a Bayesian framework for modeling bias in AI-assisted decision making, and propose a strategy to mitigate anchoring bias. Ha and Kim (2024) discuss post-2018 studies exploring approaches for reducing cognitive biases in AI-assisted tasks, and propose techniques to mitigate confirmation bias in AI-assisted decision making. Mozannar et al. (2023) attempt to improve human collaboration with agents using rules that specify when the agent can be relied upon, while Rosbach et al. (2024) observe confirmation bias as a result of AI integration in computational pathology.

3 Preliminary Work

Previous work in scalable oversight demonstrated surprisingly strong results from unstructured human/LLM interaction, but the mechanisms underlying the improvements remain unclear. We conducted two forms of preliminary work to better understand these findings: reanalysis of transcripts from Bowman et al. (2022), and a preliminary experiment testing unstructured human/LLM interaction under more carefully controlled circumstances.

3.1 Reanalysis of prior work

Bowman et al. (2022) showed that human participants conversing with language models outperformed both humans and models working independently on difficult questions. The authors suggested that the strategies reported by participants, such as cross-examining the model and asking for specific information, contributed to their success.

We analyzed supporting data from Bowman et al. (2022) and other data provided by the same authors to explore the relationship between human-AI interaction patterns and task performance across MMLU (Hendrycks et al. 2021) (a factual knowledge task) and QuALITY (Pang et al. 2022) (an extractive QA task). The dataset corresponding to the original paper consisted of two files containing complete transcripts of human/LLM conversations in which the human's aim was to correctly answer questions from MMLU and QuALITY, respectively. Additionally, we were provided access to two additional files that corresponded to a replication that they had conducted using a stronger model. These datasets all demarcated beginnings and ends of the participants' efforts to answer each question. [3]

Analysis of the MMLU data from the "Human + Model" condition of the experiment presented in the original paper revealed that conversations in which participants ultimately answered questions correctly involved significantly *fewer* (human) turns ($M = 4.52$, $SD = 3.01$) compared to conversations in which the participant answered incorrectly ($M = 5.60$, $SD = 3.53$), $t(397) = 2.51$, $p = 0.013$, with performance highest for zero-turn conversations. Logistic regression analysis confirmed that each additional turn was associated with a decrease in the odds of answering correctly ($\beta = -0.07$, $OR = 0.93$, $p = 0.016$). The negative relationship between the depth of interaction and accuracy was further supported by a chi-square test showing that participants were significantly less likely to answer correctly when engaging in more than one turn of conversation, $\chi^2(1, N = 399) = 4.15$, $p = 0.042$. Participants sometimes created new conversations by resetting the conversation history; the difference in the number of conversations engaged in for questions answered correctly vs. incorrectly was not significant, $t(397) = 1.62$, $p = 0.106$. However, participants were significantly less likely to answer correctly when resetting the conversation history at least once, $\chi^2(1, N = 399) = 4.24$, $p = 0.039$. These findings for were also observed in the same condition of the replication study that had used a stronger model. In contrast to the MMLU findings, analysis of the "Human + Model" condition of QuALITY showed no significant relationship between these interaction patterns and response accuracy in either the original or replication study.

For MMLU, this suggests that a simple strategy—"answer the question myself if I know the answer, and defer to the language model otherwise"—may explain some of the observed performance gains. Such a strategy could allow human-AI dyads to outperform both unassisted humans and unassisted language models without demonstrating the more sophisticated evaluation abilities essential for effective long-term oversight, such as critically assessing language model arguments or identifying when the model is likely to be incorrect. Given that the QuALITY task required answering questions pertaining to long fictional stories under a 5-

[3] Because the scientific aim was to identify the reasons for the pattern of effects observed by Bowman et al. (2022) in the context of the sandwiching paradigm, publicly available alternatives were unable to meet this need.

minute time limit that prevented participants from reading the story in full, this strategy would not have been useful for this task, likely explaining the differing results.

3.2 Preliminary Experiment

We designed an experiment using binary questions difficult enough that participants would be unlikely to be able to answer with far greater than 50% accuracy without LLM assistance, designed to examine whether information gained from interaction with LLMs would enable participants to answer difficult questions more successfully than either they or the language model could alone. Half of participants were additionally randomized to view LLM-generated arguments supporting each of the two possible choices ("one-turn debates"). 192 participants recruited through the Prolific platform were given guidance similar to that in Bowman et al. (2022) and engaged in unstructured interaction with `gpt-3.5-turbo-0301` to inform their answers. Question domains and key methodological choices were as described in Sec. 5.1, except that each participant received 4 multiple-choice questions (drawn from a pool of 192) rather than 48, the 'open consultancy' consisted of unstructured interaction with an LLM (rather than being shown a pre-generated answer to the question), the key intervention was one-turn debate (vs. LLM-generated research guidance), and participants were not asked to conduct online research.

The extended version of this paper contains a more thorough description of the preliminary experiment and its results[4], but in sum, the LLM alone achieved significantly higher accuracy (75%) than LLM-assisted participants who were not shown LLM-generated arguments supporting each answer choice (69%, $p = 0.003$); there was no significant difference with LLM-assisted participants shown one-turn debates (73%, $p = 0.37$). This failure to observe benefits motivated us to explore an alternative intervention, and our task's differences from real-world research motivated us to explore a more realistic setting for our primary experiment.

4 Motivations and Hypotheses

We implemented several refinements to address the limitations of our preliminary work: (1) recruiting more committed participants as part of a longer-term study with higher payouts; (2) allowing for longer, more in-depth fact-checking of LLM outputs; (3) providing access to external verification sources, inspired in part by Irving and Askell (2019); and (4) replacing unstructured LLM interaction with static consultancy, comparing a control group experiencing LLM consultancy alone with an intervention group experiencing LLM consultancy along with access to structured assistance from the same model. This assistance took the form of LLM-generated questions whose answers, if researched sufficiently, could help participants determine whether the consultant's argument was accurate.

These modifications aimed to create conditions more conducive to effective scalable oversight and to better understand the factors that influence human-AI collaborative performance. In order to streamline our analysis, we focused on the key question of whether participants in the intervention group would have more (or less) success than participants in the control group, and preregistered a narrow set of hypotheses with corresponding analyses:

- H1: Compared to the control group, the mean number of questions that participants answer correctly will be higher for the intervention group.
- H2: Compared to the control group, the mean of the logits of the final probabilities that participants assign to the correct answers (after learning the language model's answer and conducting online research) will be higher for the intervention group.
- H3: Compared to the control group, the mean result obtained when subtracting the logit of the *initial probability* assigned to the correct answer (before learning the language model's answer or conducting online research) from the logit of the *final probability* assigned to the correct answer (after learning the language model's answer and conducting online research) will be higher for the intervention group.

5 Methods

A power analysis was conducted by simulation. This estimated that 106 participants would provide 90% power to detect an $\sim$8% improvement in correct answers (4 of 48 more questions answered correctly) in the intervention group compared to the control group, using a two-sided t-test. We therefore recruited 106 participants from a pool of English language speaking Upwork freelancers based in the United Kingdom who had previously completed a miniature version of the task. For logistical reasons, we preregistered our intent to recruit participants in two waves (of 70 and 36 participants, respectively), with half of the participants in each wave assigned to the control group or the intervention group using simple randomization. Due to attrition, only 67 participants completed Wave 1, leaving the remaining 39 for Wave 2. Differences between these waves, as well as participant instructions and other methodological details, are provided in the extended version of this paper[5].

5.1 Procedure

Participants received information sheets and provided written consent to participate, and were presented with a total of 48 multiple-choice questions over multiple weeks. Questions spanned four domains: contract law, evidence law, surgery, and Lojban translation. Each question presented two answer choices (A or B) and were balanced so that the letter corresponding to the correct answer choice was 'A' (vs. 'B') exactly 50% of the time for all questions, for questions that the model answered correctly, and for questions that the model answered incorrectly. Question order was randomized for each participant. For each question, participants first indicated their initial confidence level using a probability slider. They were then shown `gpt-4-0613`'s answer to the question, with the understanding that these answers

[4] See "Study 1" in https://arxiv.org/abs/2507.19486.

[5] See "Study 2" in https://arxiv.org/abs/2507.19486.

were "correct most of the time, but not all of the time". Participants in the intervention group were additionally shown a list of guidelines generated by `gpt-4-0613` specifically designed to aid in the evaluation of the particular response under scrutiny. After reading this information, participants were asked to indicate their revised level of confidence in each choice. Participants then conducted approximately 20 minutes of online research in an attempt to determine the correct answer. Participants were explicitly instructed not to use generative AI systems such as ChatGPT. Upon completing their research, participants provided their final assessments.

5.2 Materials

Questions were either created from scratch by an expert (Lojban questions) or modified from domain-specific materials to prevent participants from finding direct answers online by searching the question verbatim. Domain experts contributed to the modified questions we used in the study and verified their answers. The final answers of the `gpt-4-0613`-generated outputs to this set of questions were accurate 75% of the time. In cases where these outputs were inaccurate, experts indicated that one or more of the `gpt-4-0613`-generated research guidelines presented to the intervention group seemed "likely to lead [a nonexpert's] research in a direction that would help them determine the correct answer", although this was often true of less than half of the six guidelines provided per question.

5.3 Statistical analysis

We preregistered[6] our analysis plan for the three hypotheses previously described, as well as the details of how the dependent variables would be calculated, which we summarize here. For H1, the dependent variable is the number of questions answered correctly. The participant is deemed "correct" if they assign a final probability greater than 50% to the correct answer. The participant is deemed "half-correct" (counted as 0.5) if they assign a final probability equal to 50% to the correct answer. For H2, the dependent variable is the logits of the final probabilities that participants assigned to the correct answers (after learning the language model's answer and conducting research)[7]. For H3, the dependent variable is the difference between the logit of the final probability that each participant assigned to the final answer (after learning the language model's answer and conducting online research) and the logit of the initial probability that the same participant had assigned to the correct answer (before learning the language model's answer or conducting online research). The primary analysis for each hypothesis is a two-tailed Welch's t-test ($\alpha = 0.05$) comparing the relevant independent variable between the control and intervention group.

As a secondary analysis, we preregistered our intent to employ a mixed-effects model incorporating fixed and random effects. The model's participant-specific random effects account for the correlation of measurements taken from the same participant across time points (before learning the language model's answer or conducting online research vs. after learning the language model's answer and conducting online research), thus adjusting for the non-independence of observations within participants. Similarly, its item-specific random effects control for the repeated measures taken for different questions or items, assuming that responses to different items by the same participant may also be correlated. This model made use of the variables "medical experience", "legal experience", "constructed language experience", and "native or native-level English", coded as 0 or 1 based on participants' responses to the questions described in the extended version. The full model specification is:

$$\begin{aligned}\text{logit}(p) = \beta_0 &+ \beta_1\,\text{group} + \beta_2\,\text{timepoint} \\ &+ \beta_3\,(\text{group} \times \text{timepoint}) + \beta_4\,\text{medical\_exp} \\ &+ \beta_5\,\text{law\_exp} + \beta_6\,\text{conlang\_exp} \\ &+ \beta_7\,\text{english} + u_{\text{participant}} \\ &+ w_{\text{topic}} + v_{\text{item(topic)}} + \epsilon \end{aligned} \quad (1)$$

where p is the probability that participant assigned to the correct answer. Other details are described in the preregistration, such as conditions for backing off to simpler models and pre-registered robustness checks. Statistical analyses were conducted in R 4.2.3 and scipy 1.16.0 on Windows 10.

6 Results

Primary analyses. Our primary analyses for H1, H2, and H3 examined differences between intervention and control groups across three dependent variables. Our pre-registered analyses found no significant differences in the proportion of questions answered correctly, the logits of final probabilities assigned to correct answers, or the difference between final and initial logits of probabilities assigned to correct answers between the intervention group and the control group. Exploratory analyses investigating waves 1 and 2 separately found significant differences for all three dependent variables for wave 1 only, with performance being *worse* in the intervention group than in the control group. Means, standard deviations and p-values are reported in Table 1.

Secondary analyses. Our main preregistered mixed model evaluated effects of *group*, *time point* (before vs. after exposure to the language-model answer and additional research), and their interaction, while adjusting for participants' domain and English experience and accounting for the hierarchical structure of the data. Nine participants who did not respond to the questions about their domain or English experience were excluded. The model converged but exhibited a singular fit for the *topic* random intercept. Participant-specific intercepts showed modest variability ($SD = 0.29$), whereas item-level intercepts varied more substantially ($SD = 0.85$).

Estimated coefficients indicated a large main effect of time point: logits of probabilities assigned to the correct answer at the final timepoint were higher than at the initial timepoint by 1.17 ($SE = 0.043$, $t = 27.1$, $p < 0.001$; $OR = 3.21$), consistent with improved accuracy after seeing the model's answer and conducting research.

[6] https://osf.io/wcxdj/

[7] When calculating logits, minimum and maximum probabilities are clamped to 1% and 99% to avoid undefined values.

Measure	Wave	Intervention		Control		p-value
		Mean	SD	Mean	SD	
Accuracy (H1)	Overall	0.73	0.06	0.74	0.04	0.19
	Wave 1	0.72	0.06	0.74	0.04	0.027*
	Wave 2	0.75	0.06	0.74	0.05	0.54
Logit of final probability assigned to the correct answer (H2)	Overall	1.15	0.51	1.31	0.56	0.13
	Wave 1	1.05	0.47	1.33	0.53	0.029*
	Wave 2	1.32	0.56	1.27	0.64	0.83
Difference between logits of final and initial probabilities assigned to the correct answer (H3)	Overall	1.00	0.46	1.17	0.51	0.09
	Wave 1	0.92	0.42	1.19	0.49	0.018*
	Wave 2	1.15	0.50	1.13	0.56	0.87

Table 1: Primary analyses (two-tailed Welch's t-tests, $\alpha = 0.05$). * indicates $p < 0.05$. In Wave 1, all analyses indicated that performance was poorer for the intervention group than for the control group; there were no significant differences when both waves were combined.

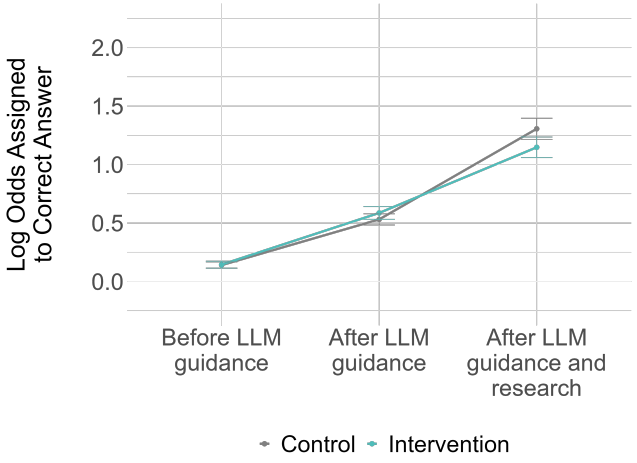

Figure 1: Mean log odds assigned to correct answers by timepoint and group. Error bars represent 95% confidence intervals. Log odds assigned to the correct answer was lower in the intervention group after LLM guidance and research.

There was also a *group × time point interaction*, such that the intervention group improved slightly *less* from the initial to final timepoint than the control group did ($B = -0.16$, $SE = 0.060$, $t = -2.59$, $p = 0.010$; $OR = 0.85$). None of the four covariates (medical, legal, constructed-language experience, native-level English) reached statistical significance ($p \geq 0.26$)[8]. Figure 1 displays mean logits of probabilities assigned to the correct answer for each group at each timepoint.

Additional preregistered analyses. We conducted a robustness check involving the same model without the variables "medical experience", "legal experience", "constructed language experience", and "native or native-level English". This analysis included all 106 participants in the control and intervention groups and yielded the same pattern of effects as the main model.

We also fitted a model where we compared the final timepoint to an intermediate timepoint after reading the LLM-generated consultancy (and the LLM-generated research guidelines, for the intervention group) but before conducting online research. There was again a large main effect of time point ($SE = 0.041$, $t = 18.7$, $p < 0.001$; $OR = 2.17$) and a a *group × time point interaction* such that the intervention group improved slightly less from the intermediate to final timepoint than the control group did ($B = -0.21$, $SE = 0.058$, $t = -3.71$, $p < 0.001$; $OR = 0.81$).

Exploratory analyses. Fitting the main mixed model to the subset of data points from wave 1 revealed the same pattern of effects observed for the dataset as a whole, with a large main effect of timepoint ($B = 1.18$, $SE = 0.053$, $t = 22.09$, $p < 0.001$; $OR = 3.26$) and a *group × time point interaction* with the same pattern observed for the dataset as a whole ($B = -0.26$, $SE = 0.074$, $t = -3.48$, $p < 0.001$; $OR = 0.85$). Fitting the main mixed model to the subset of data points from wave 2 still yielded a large main effect of timepoint ($B = 1.15$, $SE = 0.072$, $t = 15.94$, $p < 0.001$; $OR = 3.14$), but no interaction ($p = 0.87$). There was also an association between native-level English and higher logits assigned to the correct answer ($B = 0.32$, $SE = 0.144$, $t = 2.19$, $p = 0.036$; $OR = 1.37$).

We also investigated the change in log odds assigned to the consultant's answer with respect to its correctness for both protocols, including the domain experience / native-level English covariates and the random effects used in our other main study mixed models. We observed a main effect of the LLM consultant's correctness, with participants updating more weakly in the direction of the model (after conducting online research) when the model was incorrect than when it was correct ($B = 0.72$, $SE = 0.152$, $t = 4.76$, $p < 0.001$; $OR = 2.06$). We also observed an interaction ($B = -0.22$, $SE = 0.101$, $t = -2.19$, $p = 0.028$; $OR = 0.80$), such that while participants in the control and intervention groups updated towards the consultant's answer equally when the consultant was incorrect, participants in the intervention group updated less when it was

[8]Participant demographics are reported in the extended version of this paper at https://arxiv.org/abs/2507.19486.

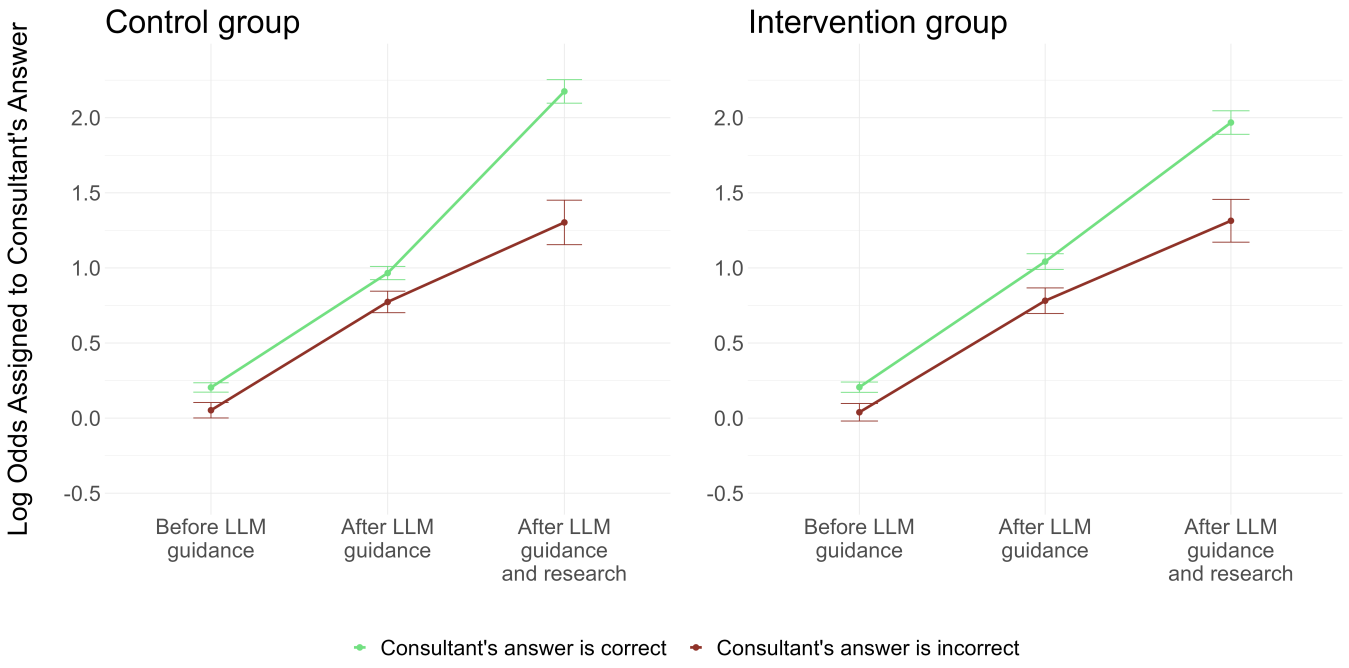

Figure 2: Log odds assigned by participants to the LLM consultant's answer at each timepoint, split by whether participants were in the control group (left) or intervention group (right). Green lines represent cases where the LLM's guess was correct, while maroon lines represent cases where it was incorrect. Error bars indicate 95% confidence intervals. In both cases, confidence in the consultant's answer continued to increase after participants conducted online research with the intent of verifying the consultant's answer, even when the consultant was incorrect. Error bars indicate 95% confidence intervals.

correct. Figure 2 displays log odds assigned by participants to the consultant's answer broken out by timepoint, group, and whether the consultant was correct or incorrect.

Qualitative analysis. We requested free-text summaries of participants' reasoning, in participants' own words, from each for each item. A common theme was that when participants' research did not reveal information that specifically contradicted the model-generated answer, this increased their confidence in it. The following quotes from cases where participants' confidence in the model-generated answer increased after conducting their research, despite this answer being incorrect, provide particularly clear examples of this phenomenon:

"I relied heavily on the guidance of the A.I suggestions. My research brought me to the conclusion that the evidence would be permissible as hear say because it met the threshold of several exceptions... I saw no reasoning that pointed to Option B being the correct answer."

"Using all the research prompts provided were not only useful, but helped me in understanding what the condition was about, [its] seriousness and so on. There [was] no information which specifically disputed the answer provided by AI. The only conclusion I could come to was that AI was correct. I really could have answered that the answer is 100% A."

"As this is a legal question, it was difficult for me to understand to fully understand and decide on the correct answer. Whilst I understood elements of the question, my research did not convince that I even came close to what the answer is. I had no choice but to lean towards the answer AI provided as more often than not, AI is correct."

"Prompted by the AI guidelines, I started by researching the answers to each of the guidelines to see if my research aligned with ai answer. The logic being if one didnt align with AI then it could be a wrong answer. AI answer aligned with my research."

The fact that the LLM-generated research guidelines were only of middling quality likely contributed to the inability of participants to use them to effectively find information that contradicted the consultant's judgments. For each of the twelve main study problems that `gpt-4-0613` answered incorrectly, we sent a domain expert the original question, `gpt-4-0613`'s incorrect answer when acting as a consultant, and the six guidelines (the questions suggested as useful focus areas by `gpt-4-0613` when it acted as an assistant in the intervention condition). Experts were blinded to the priority ratings (high, medium, or low) that `gpt-4-0613` included in its advice to participants in the intervention condition. In each case, the expert indicated that one or more of the questions provided seemed "likely to lead [a nonexpert's] research in a direction that would help them determine the correct answer", but that others did not. The average fraction of the six AI-generated follow-up questions that experts deemed likely to guide a non-expert toward the correct answer was 39%, and was only 55% even for the subset that the model had labelled "high priority." Many of these

questions were also not straightforward and would have taken substantial time to research thoroughly. It is therefore not surprising that participants frequently failed to surface important information that contradicted the model's answer.

7 Discussion

We investigated whether providing participants with AI-generated research guidelines would enhance their ability to evaluate the accuracy of an LLM consultant's answers on complex domain-specific questions. Contrary to our expectations, we found no evidence that such research guidelines improved participants' overall performance. In fact, during Wave 1, participants in the intervention group performed more poorly than those in the control group across all three of our primary measures. Qualitative analysis suggested that participants' performance in our experiments was sometimes hampered by participants directing their research using questions which were not among the $\sim$40% of questions deemed likely to guide non-experts' research productively, and the knowledge barriers introduced by our specialized domains (Lojban, contract law, evidence law, and surgery), which even guided research frequently was not enough to overcome in the allotted time. The additional cognitive load imposed by the intervention, causing participants to follow guidelines mechanically instead of critically assessing model output, may have been another factor.

While prior work has demonstrated that debate outperforms consultancy in contexts where the consultants are calibrated to argue for the correct answer only 50% of the time (Michael et al. 2023; Khan et al. 2024), we are not aware of any publicly reported test of a scalable oversight protocol that has demonstrated an advantage of the protocol over open consultancy when using human judges and a frontier model as open consultant. That said, the conclusions that can be drawn from the failure of our interventions to increase overall accuracy are constrained by the limitations of our studies, particularly the simplicity of the protocols tested and restriction to binary QA.

7.1 Confirmation Bias

Our findings revealed that initial exposure to the LLM consultant's answers appeared to induce confirmation bias—the tendency to search for, interpret, favor, and recall information that confirms one's present beliefs (in this case, that the LLM was likely to be correct). As Figure 2 illustrates, after participants conducted independent online research after receiving the consultant's answer, they ended up being even more confident in the consultant's position than before—even when that position was incorrect. This suggests that rather than successfully identifying information that revealed the consultant's errors, participants may have inappropriately allowed their confidence in the consultant's final answer to increase whenever they encountered information that supported *any* of its claims. Protocols that rely on human evaluators should consider the potential influence of confirmation bias on their judgments. Given observations of confirmation bias in large language models (Khan et al. 2024; Malberg et al. 2025; Macmillan-Scott and Musolesi 2024; O'Leary 2025; Tan et al. 2025), this should also be considered for studies using LLMs as judges.

7.2 Future Work

Implicit consultancy effects. While it is well-established that consultancy is not an ideal oversight protocol, its effects raise a more fundamental concern: if oversight protocols are used in contexts where the human evaluators know what claim the evaluated model has made (or believe they know what claim it would make, if asked), this is likely to influence their beliefs before any oversight technique is applied. For instance, if a human is tasked with determining whether a powerful model is truthfully claiming alignment, the human evaluator may already be biased toward believing the model (due to its generally high accuracy in relevant domains, institutional incentives favoring positive outcomes, etc.). Future research should specifically test whether oversight strategies are likely to remain robust in 'implicit consultancy' contexts where the judge is biased by their belief about the overseen model's likely output, whether or not any explicit consultancy has occurred.

Determining conditions under which desirable properties generalize. More research is needed to identify the degree to which desirable properties of oversight protocols scale with problem difficulty, and how this may vary by domain. Evaluations of scalable oversight protocols should consider the specific conditions under which they are effective. One useful framing is to treat such protocols as adversarial games between an evaluator/judge/guard and a stronger, potentially deceptive model (Engels et al. 2025). In this setting, we should consider the advantages the game gives the evaluator, judge, and guard that enable it to win reliably. Future work could also explore how these advantages evolve as model capabilities advance, and as the difficulty of the tasks we aim to accomplish with the models increase.

8 Conclusion

Our results show that two simple interventions fail to outperform open consultancy in the settings investigated, highlighting the effects of confirmation bias. We also demonstrated that the strong 'centaur' performance observed in one influential study did not necessarily arise from critical assessment of model outputs by participants. Tests of scalable oversight protocols should consider using open consultancy as a baseline. They should also aim to demonstrate effectiveness under more challenging conditions that will continue to apply as model capabilities advance, and as the difficulty of the problems to which the models are applied increases.

The strong confirmation bias we observed may not generalize to stronger protocols, or to settings where the difficulty of evaluation is lower (e.g. due to using expert rather than lay judges). Nevertheless, our results illustrate the importance of confirmation bias as a possible failure mode. Depending on their specific goals, other empirical scalable oversight studies may likewise wish to consider inducing biased priors in human judges by first providing them with the default output of the system being evaluated, as a proxy for the bias to trust model outputs that may occur in the real world.

Acknowledgments

Support for this project was provided by Open Philanthropy and the Good Ventures Foundation.

References

Anil, C.; Zhang, G.; Wu, Y.; and Grosse, R. 2021. Learning to Give Checkable Answers with Prover-Verifier Games. arXiv:2108.12099.

Bansal, G.; Nushi, B.; Kamar, E.; Lasecki, W. S.; Weld, D. S.; and Horvitz, E. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 7, 2–11.

Bertrand, A.; Belloum, R.; Eagan, J. R.; and Maxwell, W. 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.

Bowman, S. R.; Hyun, J.; Perez, E.; Chen, E.; Pettit, C.; Heiner, S.; Lukošiūtė, K.; Askell, A.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Olah, C.; Amodei, D.; Amodei, D.; Drain, D.; Li, D.; Tran-Johnson, E.; Kernion, J.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J.; Ndousse, K.; Lovitt, L.; Elhage, N.; Schiefer, N.; Joseph, N.; Mercado, N.; DasSarma, N.; Larson, R.; McCandlish, S.; Kundu, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Fort, S.; Telleen-Lawton, T.; Brown, T.; Henighan, T.; Hume, T.; Bai, Y.; Hatfield-Dodds, Z.; Mann, B.; and Kaplan, J. 2022. Measuring Progress on Scalable Oversight for Large Language Models. arXiv:2211.03540.

Buhl, M. D.; Pfau, J.; Hilton, B.; and Irving, G. 2025. An alignment safety case sketch based on debate. arXiv:2505.03989.

Christiano, P.; Shlegeris, B.; and Amodei, D. 2018. Supervising strong learners by amplifying weak experts. arXiv:1810.08575.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168.

Cotra, A. 2021. The Case for Aligning Narrowly Superhuman Models.

Cummings, M. L. 2017. Automation bias in intelligent time critical decision support systems. In *Decision making in aviation*, 289–294. Routledge.

Engels, J.; Baek, D. D.; Kantamneni, S.; and Tegmark, M. 2025. Scaling Laws For Scalable Oversight. arXiv:2504.18530.

Goddard, K.; Roudsari, A.; and Wyatt, J. C. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1): 121–127.

Goddard, K.; Roudsari, A.; and Wyatt, J. C. 2014. Automation bias: empirical results assessing influencing factors. *International journal of medical informatics*, 83(5): 368–375.

Ha, T.; and Kim, S. 2024. Improving trust in AI with mitigating confirmation bias: Effects of explanation type and debiasing strategy for decision-making with explainable AI. *International journal of human–computer interaction*, 40(24): 8562–8573.

Hemmer, P.; Schemmer, M.; Kühl, N.; Vössing, M.; and Satzger, G. 2024. Complementarity in Human-AI Collaboration: Concept, Sources, and Evidence. arXiv:2404.00029.

Hemmer, P.; Schemmer, M.; Vössing, M.; and Kühl, N. 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *PACIS*, 78: 118.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. arXiv:2009.03300.

Hubinger, E. 2020. AI Safety via Market Making.

Irving, G.; and Askell, A. 2019. AI Safety Needs Social Scientists. *Distill*. https://distill.pub/2019/safety-needs-social-scientists/.

Irving, G.; Christiano, P.; and Amodei, D. 2018. AI safety via debate. arXiv:1805.00899.

Kenton, Z.; Siegel, N.; Kramár, J.; Brown-Cohen, J.; Albanie, S.; Bulian, J.; Agarwal, R.; Lindner, D.; Tang, Y.; Goodman, N.; et al. 2024. On scalable oversight with weak llms judging strong llms. *Advances in Neural Information Processing Systems*, 37: 75229–75276.

Khan, A.; Hughes, J.; Valentine, D.; Ruis, L.; Sachan, K.; Radhakrishnan, A.; Grefenstette, E.; Bowman, S. R.; Rocktäschel, T.; and Perez, E. 2024. Debating with more persuasive LLMs leads to more truthful answers. In *ICML 2024 (Best Paper Award)*.

Kirchner, J. H.; Chen, Y.; Edwards, H.; Leike, J.; McAleese, N.; and Burda, Y. 2024. Prover-Verifier Games improve legibility of LLM outputs. arXiv:2407.13692.

Leike, J.; Krueger, D.; Everitt, T.; Martic, M.; Maini, V.; and Legg, S. 2018. Scalable agent alignment via reward modeling: a research direction. arXiv:1811.07871.

Liu, H.; Lai, V.; and Tan, C. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–45.

Lord, C. G.; Lepper, M. R.; and Preston, E. 1984. Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, 47(6): 1231–1243.

Lyell, D.; and Coiera, E. 2017. Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*, 24(2): 423–431.

Macmillan-Scott, O.; and Musolesi, M. 2024. (Ir) rationality and cognitive biases in large language models. *Royal Society Open Science*, 11(6): 240255.

Malberg, S.; Poletukhin, R.; Schuster, C.; and Groh, G. G. 2025. A Comprehensive Evaluation of Cognitive Biases in LLMs. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, 578–613. Association for Computational Linguistics.

Michael, J.; Mahdi, S.; Rein, D.; Petty, J.; Dirani, J.; Padmakumar, V.; and Bowman, S. R. 2023. Debate Helps Supervise Unreliable Experts. arXiv:2311.08702.

Mozannar, H.; Lee, J. J.; Wei, D.; Sattigeri, P.; Das, S.; and Sontag, D. 2023. Effective Human-AI Teams via Learned Natural Language Rules and Onboarding. arXiv:2311.01007.

Nickerson, R. S. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2): 175–220.

O'Leary, D. E. 2025. Confirmation and Specificity Biases in Large Language Models: An Explorative Study. *IEEE Intelligent Systems*, 40(1): 63–68.

Pang, R. Y.; Parrish, A.; Joshi, N.; Nangia, N.; Phang, J.; Chen, A.; Padmakumar, V.; Ma, J.; Thompson, J.; He, H.; and Bowman, S. R. 2022. QuALITY: Question Answering with Long Input Texts, Yes! arXiv:2112.08608.

Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; and Irving, G. 2022. Red Teaming Language Models with Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3419–3448. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Rastogi, C.; Zhang, Y.; Wei, D.; Varshney, K. R.; Dhurandhar, A.; and Tomsett, R. 2022. Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. arXiv:2010.07938.

Roger, F. 2024. Open consultancy: Letting untrusted AIs choose what answer to argue for. Accessed: 2025-04-22.

Rosbach, E.; Ammeling, J.; Krügel, S.; Kießig, A.; Fritz, A.; Ganz, J.; Puget, C.; Donovan, T.; Klang, A.; Köller, M. C.; Bolfa, P.; Tecilla, M.; Denk, D.; Kiupel, M.; Paraschou, G.; Kok, M. K.; Haake, A. F. H.; de Krijger, R. R.; Sonnen, A. F. P.; Kasantikul, T.; Dorrestein, G. M.; Smedley, R. C.; Stathonikos, N.; Uhl, M.; Bertram, C. A.; Riener, A.; and Aubreville, M. 2024. When Two Wrongs Don't Make a Right" – Examining Confirmation Bias and the Role of Time Pressure During Human-AI Collaboration in Computational Pathology. arXiv:2411.01007.

Saunders, W.; Yeh, C.; Wu, J.; Bills, S.; Ouyang, L.; Ward, J.; and Leike, J. 2022. Self-critiquing models for assisting human evaluators. arXiv:2206.05802.

Sperrle, F.; El-Assady, M.; Guo, G.; Borgo, R.; Chau, D. H.; Endert, A.; and Keim, D. 2021. A Survey of Human-Centered Evaluations in Human-Centered Machine Learning. *Computer Graphics Forum*, 40(3): 543–568.

Tan, Z.; Wang, S.; Marjit, S.; Chen, Z.; He, Y.; Zhao, X.; Li, P.; Li, J.; Chen, T.; et al. 2025. Understanding Prejudice and Fidelity of Diverge-to-Converge Multi-Agent Systems. Under review.

Vaccaro, M.; Almaatouq, A.; and Malone, T. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 1–11.

Vodrahalli, K.; Daneshjou, R.; Gerstenberg, T.; and Zou, J. 2022. Do humans trust advice more if it comes from AI? an analysis of human-AI interactions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 763–777.