arXiv:2510.26518v1 [cs.AI] 30 Oct 2025

Human-AI Complementarity: A Goal for Amplified Oversight

Rishub Jain^{*,1}, **Sophie Bridgers**^{*,1}, **Lili Janzer**^{†,1}, **Rory Greig**^{†,1}, **Tian Huey Teh**^{†,1} **and Vladimir Mikulik**^{2}
*Equal first author contributions, †Equal contributions, order randomized, ^{1}Google DeepMind, ^{2}Work done while previously at Google DeepMind

Human feedback is critical for aligning AI systems to human values. As AI capabilities improve and AI is used to tackle more challenging tasks, verifying quality and safety becomes increasingly challenging. This paper explores how we can leverage AI to improve the quality of human oversight. We focus on an important safety problem that is already challenging for humans: fact-verification of AI outputs. We find that combining AI ratings and human ratings based on AI rater confidence is better than relying on either alone. Giving humans an AI fact-verification assistant further improves their accuracy, but the type of assistance matters. Displaying AI explanation, confidence, and labels leads to over-reliance, but just showing search results and evidence fosters more appropriate trust. These results have implications for Amplified Oversight—the challenge of combining humans and AI to supervise AI systems even as they surpass human expert performance.

Keywords: HCI, Human-AI Complementarity, Amplified Oversight, Scalable Oversight, LLM Rating, AI Assistance, Hybridization

1. Introduction

Human oversight is critical for ensuring that Artificial Intelligence (AI) models remain safe and aligned to human values. But AI systems are rapidly advancing in capabilities and are being used to complete ever more complex tasks, making it increasingly challenging for humans to verify AI outputs and provide high-quality feedback. How can we ensure that humans can continue to meaningfully evaluate AI performance? An avenue of research to tackle this problem is “Amplified Oversight” (also called “Scalable Oversight”), which aims to develop techniques to use AI to amplify humans’ abilities to oversee increasingly powerful AI systems, even if they eventually surpass human capabilities in particular domains.

With this level of advanced AI, we could use AI itself to evaluate other AIs (i.e., AI raters), but this comes with drawbacks (see Section 3.1). Importantly, humans and AIs have complementary strengths and weaknesses. We should thus, in principle, be able to leverage these complementary abilities to generate an oversight signal for model training, evaluation, and monitoring that is stronger than what we could get from human raters or AI raters alone. Two promising mechanisms for harnessing human-AI complementarity to improve oversight are:

- **Hybridization**, in which we combine judgments from human raters and AI raters working in isolation based on predictions about their relative rating ability per task instance (e.g., based on confidence), and
- **Rater Assistance**, in which we give human raters access to an AI rating assistant that can critique or point out flaws in an AI output or automate parts of the rating task.

In this paper, we empirically investigate how combining Hybridization and Rater Assistance can produce a higher quality oversight signal than relying on either human or AI raters alone. We ground our study in the task of verifying the factual accuracy of AI-generated sentences. We chose this task for two reasons: (1) It is a realistic, pressing safety concern of currently deployed models, which often hallucinate and generate misleading information and (2) It is already a challenging task for human raters.

To do so, we developed an AI fact-verification

Corresponding author: rishubj@google.com
© 2025 Google DeepMind. All rights reserved

rater that uses a search engine tool to research factuality. This AI rater has a higher overall accuracy than a typical human rater on our realistic evaluation dataset. But critically, when the AI's confidence is low, it performs worse than humans. This allows us to perform what we call *Confidence-based Hybridization*: use the AI's ratings when it has high confidence, and the humans' ratings when the AI has lower confidence. The AI rater output can also be served to humans as a fact-verification assistant.

Using Rater Assistance to achieve complementarity in settings where AI raters already outperform human raters is challenging. A recent review shows that when AI alone performs better, it is harder for humans to add value (Vaccaro et al., 2024). To still realize the gains of Rater Assistance, we can use Confidence-based Hybridization to identify examples where Humans still outperform AI, thus making it easier for Rater Assistance to achieve complementarity.

This leads to our primary research questions:

1. Does confidence-based hybridization improve accuracy above AI or human ratings alone?
2. On the low-AI-confidence data slice assigned to humans, can AI assistance improve human accuracy, even though the assistant is less accurate than the humans on this slice? If so, what form of AI assistance is most effective at improving accuracy while appropriately calibrating human reliance?

1.1. Related Work

1.1.1. Amplified Oversight

AI models are rapidly increasing in capabilities. As models improve, we want to ensure that they continue to align with human goals and values. A prominent approach for value alignment of LLMs is fine-tuning based on human feedback, using techniques like Reinforcement Learning from Human Feedback (Christiano et al., 2017) and Direct Preference Optimization (Rafailov et al., 2024). These techniques rely on humans being able to accurately judge the quality of an AI's responses, considering several dimensions (e.g. helpfulness, factuality, safety).

The increasing capabilities of AI models, however, present a challenge for human oversight: AI systems will be asked to complete increasingly challenging tasks and their capabilities on these tasks may exceed human capabilities making it difficult or highly time consuming for humans to provide feedback. If human evaluation becomes more unreliable, we risk rewarding behavior that looks superficially "good" but isn't actually desirable (Krakovna et al., 2020), such as misleading information and sycophancy (Sharma et al., 2023). This raises a critical question: how can we improve human oversight to keep pace with AI advancements?

The field of Amplified Oversight aims to address this issue by leveraging the increasingly advanced capabilities of AI to assist humans in providing high-quality feedback (Amodei et al., 2016). Current research explores different approaches to AI assistance for human evaluation, such as AI critics that point out errors in a model's output and AI debates that provide different perspectives, with promising but mixed results (e.g., Irving et al., 2018; McAleese et al., 2024; Michael et al., 2023; Parrish et al., 2022b; Saunders et al., 2022). For example, there is positive evidence of debate helping when there is knowledge asymmetry between the AI assistance and the "judge" rater, but when this asymmetry is removed, the judge receives little benefit (Kenton et al., 2024; Khan et al., 2024; Michael et al., 2023; Parrish et al., 2022a,b).

Thus far the field has primarily focused on machine learning solutions, but figuring out how to use AI to elicit better reasoning from a human rater is fundamentally a Human Computer Interaction (HCI) problem. Amplified Oversight would benefit from closer collaboration with the HCI community to effectively leverage the complementary strengths of humans and AI in achieving robust and scalable oversight. In this paper, taking an HCI perspective, we explore this collaborative human-AI approach for improving human feedback on the factuality of model outputs.

While existing Amplified Oversight and HCI research on AI assistance provide valuable insights, Amplified Oversight has not considered hybridization and both fields have primarily focused on scenarios where AI assistant performance is on par with or surpasses human capabilities. It's possible humans will just ignore the assistance in this case, but fact-verification is quite challenging, so it's also possible that they will opt to critically engage with the information the assistance provides (see Vasconcelos et al., 2023). Given that our task of interest is fact-verification, search results and verified quotes from online sources can be included in the information provided by the assistant, which might help raters to better evaluate the legitimacy of explanations and predictions.

The design of Rater Assistance and/or Hybridization protocols that enable human-AI complementarity is challenging. It requires grappling with complex questions such as how to pinpoint the unique skills and knowledge that humans or AIs possess, how to identify when AI or human judgment is more reliable, and how to effectively use AI to improve human reasoning and decision-making without leading to under- or over-reliance on the AI. These are fundamentally questions of Human-Computer Interaction (HCI), Cognitive Science, Psychology, Philosophy, and Education. Luckily, these fields have explored these same or related questions, and AI safety can learn from and collaborate with them to address these sociotechnical challenges.

1.1.2. *Human Computer Interaction: Rater Assistance*

HCI, in particular, offers a wealth of relevant research. Complementarity is a term from this field, and achieving human-computer complementarity or synergy is a long-studied problem across various technologies, with AI as the newest and, arguably, most complex addition (e.g., Lee and Moray, 1994; Lee and See, 2004; Vaccaro et al., 2024). HCI research has investigated what leads to successful human-AI partnerships, often focusing on AI as a decision support system more akin to Rater Assistance than Hybridization. Studies have explored the impact of numerous factors, including both fixed conditions of the set-up (e.g., the relative competence of the AI compared to the human, the task-type, the task difficulty, etc.) and design decisions (e.g., the information the AI provides, the workflow, etc.) (for a review, see Vaccaro et al., 2024). Here are a few learnings from this research that are particularly relevant for Rater Assistance:

- **Achieving human-AI complementarity is difficult and does not happen by default.** A comprehensive review of the literature on human-AI team performance revealed that on average, across several studies and tasks, human-AI teams perform statistically worse than humans or AIs alone (Vaccaro et al., 2024).
- **Complementarity depends on relative human/AI performance, task type, and division of labor.** Though complementarity is hard to achieve, a few factors may increase its likelihood: (1) humans alone outperform AI alone (i.e., AI assistance can still be helpful even when humans are better than AI at a task, but when AI alone performs better, it is harder for humans to add value); (2) the task involves creating content rather than making decisions, and (3) the AI completes subtasks rather than the entire task (Vaccaro et al., 2024). Humans can struggle to delegate tasks to AI, but training humans on the complementary strengths and weaknesses of humans and AI improves their delegation and human-AI team performance (Pinski et al., 2023)[1]. Rater Assistance research has also found that even when human-AI teams don't achieve complementary performance, they can counter each others' biases (i.e., catching more bugs in code than humans alone and hallucinating fewer bugs than AI alone McAleese et al., 2024). These findings suggest that complementarity is possible, even though it may require careful de-

---

[1] Division of labor and how it might facilitate complementarity but also how humans might struggle with optimal delegation has implications for the Amplified Oversight technique Iterative Amplification. This technique involves a human rater attempting to solve a problem that is difficult or impossible to solve on their own and assigning simpler sub-tasks to AI; an AI then learns from this delegation behavior, the human moves to a higher level of abstraction, and the cycle repeats (Christiano et al., 2018).

sign to achieve.

- **Measuring over- and under-reliance on the AI assistant is critical.** In addition to human-AI team performance, it is important to measure over-reliance (deferring to AI even when it's unhelpful/incorrect) and under-reliance (ignoring AI even when it's helpful/correct) (Lee and Moray, 1994; Lee and See, 2004; Manzini et al., 2024; Parasuraman and Riley, 1997). These can limit the gains from assistance and likely require different interventions to counteract. Notably, as AI advances, over-reliance may become a primary obstacle to achieving complementarity. Humans may increasingly defer to the AI's reasoning and recommendations even when wrong, exacerbating AI biases, diminishing the complementary value of human input, and reducing human agency (Lai and Tan, 2019). Rater Assistance research has similarly found that human raters can be misled by AI assistants that provide evidence and arguments in support of incorrect conclusions (Khan et al., 2024).
- **AI explanations and confidence do not robustly reduce over-reliance.** One way one might hope to counter over-reliance would be to show the AI's explanation of its reasoning or an estimate of its confidence in its answer to the human. However, simply providing this additional information does not reliably reduce over-reliance or lead to complementarity across studies (e.g., Bansal et al., 2021; Lai and Tan, 2019; Ma et al., 2023; Vaccaro et al., 2024). Explanations have even been shown to make over-reliance worse, perhaps by inflating perceptions of AI competence (e.g., Bansal et al., 2021; Buçinca et al., 2020; Buçinca et al., 2021; Kaur et al., 2020). The situation might change, though, if we can better train humans to utilize this information, and better train AI to helpfully and accurately explain its reasoning. Indeed, recent research has shown positive results when encouraging humans to think not only about AI confidence but also their own confidence and correctness likelihood (Ma et al., 2023; Ma et al., 2024).
- **Contrasting explanations can reduce over-reliance but also increase under-reliance.** Displaying contrastive explanations that argue for different conclusions can both reduce over-reliance and increase under-reliance compared to non-constrastive explanations, and so do not better calibrate reliance overall (e.g., Bansal et al., 2021; Si et al., 2023). Rater Assistance research, however, has found evidence that debate (two AI assistants providing contrastive arguments for different conclusions) improves human accuracy above consultancy (one-sided argument) (Khan et al., 2024), suggesting more research is needed on the HCI of debate, a theoretically promising alignment technique.
- **Over-reliance on AI decisions is a cognitive short-cut (but it can be a rational decision).** Over-reliance on AI decisions without carefully considering accompanying AI explanations can be viewed as a human cognitive heuristic, or short-cut. But Vasconcelos et al. (2023) argue and empirically show that though over-reliance may be a short-cut, it is a strategic cost-benefit decision. In situations where critically reading an AI explanation is cognitively costly, people are more likely to ignore it and just defer to the AI decision; but when the explanation is easier to digest, especially in comparison to the difficulty of the task, or when people are paid performance bonuses, over-reliance decreases. Interventions to encourage analytical thinking such as having people make a decision first before seeing the AI decision or forcing them to slow down before making a decision also decrease over-reliance (Buçinca et al., 2021). Rater Assistance research should consider the cognitive demands of the types of assistance explored and how to design an interaction interface and protocol that reduce these demands; this might be especially important when it comes to debate, which is a longer, more complex form of assistance.

Beyond these learnings, there is much more to explore. HCI research on human-AI complementarity provides a strong foundation from which to draw inspiration and form hypotheses for Rater

Assistance research. We see this as an opportunity for mutually beneficial collaboration between the fields, as Amplified Oversight can leverage learnings from HCI and in turn, learnings from Amplified Oversight could help to inform HCI theories and practice. What's more, HCI techniques that are effective at improving oversight could be applied to products for end users, giving them more control over the AI tools they use in their daily lives.

1.2. Hybridization in HCI

Hybridization is another way to leverage human-AI complementarity and more broadly to achieve successful human-AI teams. However it has received less attention from Amplified Oversight research than rater assistance. Although training on AI rater feedback may be hard to get right in practice, it is a powerful technique (e.g., Bai et al., 2022) and increasingly widespread, since AI raters are more scalable than human raters, making the study of Hybridization even more timely.

With Hybridization, the goal is to figure out how best to combine the signal from human ratings and from AI ratings. Two main approaches have been explored: (1) averaging ratings per datum (e.g., taking the majority vote out of all human and AI ratings, the raw average, or a weighted average; see Li, 2024), and (2) slicing the data and routing tasks to either humans or AIs based on confidence or some other metric for predicting who will perform better (see Wang et al., 2024).

Combining Hybridization with Rater Assistance might provide an opportunity to better amplify the complementary strengths of humans and AIs. Slicing, in particular, might enable the success of AI assistance by identifying areas where the AI rater struggles to make the right decision, but where it (or another assistant model) could still meaningfully assist a human rater, e.g. by surfacing helpful information. Currently, HCI evidence suggests that human-AI complementarity is easier to achieve when humans alone outperform AI alone (Vaccaro et al., 2024), and so by focusing on the shortcomings of AI raters, we might better clarify humans' contributions and how they could be enhanced by assistance.

HCI research is increasingly exploring Hybridization, referred to as Delegation (see Baird and Maruping, 2021), with initial promising results. Specifically, delegating tasks to humans based on AI confidence has been found to improve image classification accuracy over AI alone, in a situation where AI outperforms humans alone (Hemmer et al., 2023; Fügener et al., 2022). These results align with the idea that division of labor can facilitate complementarity (Vaccaro et al., 2024).

It is thus important for Amplified Oversight research to consider Rater Assistance in the context of Hybridization, and the interaction between them. Hybridization might change the kind of assistance we explore — the form of assistance that is most helpful overall on the entire dataset might not be the form that is most helpful on a particular slice. Additionally, for Hybridization, the space of possible signals to ensemble is even richer than previously proposed: We don't just have AI or humans to consider, we can also take into account assisted humans, which might be further broken down into different assistant methods. This is a more complicated but also an arguably more flexible and powerful set of signals to optimally combine.

1.3. Hybridization in Machine Learning

Hybridization has been studied in the Machine Learning literature, under different names. Dvijotham et al., 2023 calls task of combining AI and Human labels as "Selective Prediction". Using AI-confidence (measured by the conditional output probability of the positive label of their binary classifier), they achieve complementary performance on a broad set of medical diagnosis tasks.

Hybridization has also been studied in the field of Active Learning, which aims to select a subset of data such that training on that will lead to a model with high accuracy. Amplified Oversight's goal is similar, to improve the quality of training data such that the resulting model is more robust and of higher quality. Common Active Learning strategies include Uncertainty Sampling, where data points with low AI-confidence are selected for human labeling, and that data is later trained on (Tharwat and Schenck, 2023).

2. Experiments and Results

2.1. Summary of Contributions

Across several experiments, we test how hybridization affects rating accuracy, as well as how different forms of assistance affect human rating accuracy. The specific task of interest is rating the factual correctness of an AI-generated sentence. We decided to verify the factuality of sentences rather than entire model responses because it was easier to collect high-quality golden (ground-truth) labels that we could trust.

We show that overall accuracy on our evaluation dataset can be improved above either AI or human ratings alone if we use confidence-based hybridization. If we use the AI fact-verification rater as an AI assistant and show it to the human raters in a specific way, human accuracy increases on the dataset they are rating, even though these are examples where the AI assistant is not confident in its judgments and performs worse as a rater than unassisted humans. Furthermore, across 10 experiments we we selectively filter the information the AI assistant provides to human raters to see what information best calibrates reliance on the assistant. We investigate the effects of AI predictions, explanations, confidence, search results, and selected evidence, as well as information supporting contrastive predictions. We find that more leading forms of assistance that include factuality labels, explanations, and confidence scores lead to over-reliance, while a less leading form that only includes the AI generated search results and evidence snippets it has selected from these results does not cause over-reliance and helps the most after confidence-based hybridization.

2.2. Overall Methods

Evaluation Set. The *Evaluation Set* has 1918 [prompt, response, target sentence] tuples in total sampled from a realistic distribution of Gemini responses to User prompts. Each target sentence has a golden label of either "Accurate" or "Inaccurate" from high quality human raters.

Factuality Task. Raters assess each target sentence as "Inaccurate", "Unsupported", "Disputed", "Accurate", or "Doesn't require attribution", using online research and/or information from the AI fact-verification assistant (if displayed). Raters can also select "Can't confidently assess", or skip the question, though they are asked to try to avoid doing so.

Analysis. For all calculations of accuracy, we binarize all ratings into "Accurate" and "Inaccurate," with "Unsupported," "Disputed," and "Doesn't require attribution" considered "Inaccurate." AI and human ratings are scored as "correct" if they match the golden label and "incorrect" otherwise. We mark ratings with "Can't confidently assess" as incorrect.

AI fact-verification model. We built an AI model that uses a search-engine to research the factuality of a given sentence. This model can serve as both an AI-rater and an assistant to human raters. We trained the AI model to (1) *decompose* the target sentence into its constituent factual claims, (2) *search the web* for reputable evidence to assess the accuracy of each claim, (3) *select quotes* from helpful pieces of evidence, (4) *provide reasoning* about the accuracy of each claim based on the evidence, and (5) *provide ratings* for each claim and the sentence overall. To determine the final Overall AI-Rating, we sample our model 50 times, filter samples by if they pass a programmatic Format Verifier, and take the Majority (modal) Vote of the remaining samples' Overall Ratings. We also determine a model Confidence level on this rating, which is the proportion of verified samples that have the same Overall Rating as the Majority Vote. These confidence scores were fairly well-calibrated (See Appendix Fig 10).

Human Experiment Set-up. For all of our human rater experiments, we recruited participants from Prolific, using a pool filtered for general competency but not specifically for fact-verification ability. Experiments were divided into two types: baseline (unassisted) where raters worked independently, and intervention (assisted) where raters were provided information from an AI fact-verification assistant.

(See Section A.2 in the Appendix for more details on the evaluation set, task, model, and human experiment set-up.)

2.3. RQ1: Does confidence-based hybridization improve accuracy above AI or human ratings alone?

Our first research question was whether or not hybridizing between AI ratings and human ratings would lead to higher overall accuracy on the evaluation dataset than using either alone.

We selected which examples to have the human raters rate based on the AI's own confidence. That is, we select a threshold T, and if the model's confidence is > T on a particular example, then we accept the AI rating. Otherwise, we have the humans rate it.

2.3.1. "T1" Human Rater Experiment setup

To see if there is a threshold T at which hybridized accuracy is higher than either AI or human accuracy, we ran an unassisted baseline experiment in which human participants rated the entire 1918-example Evaluation Set on their own without assistance (NB: we will refer to this baseline as *baseline_t1* for time 1). We recruited for 800 rating hours, and participants had an opportunity to take part two times if there were still spots available. In the end, we collected an average of 5.32 ratings per example.

2.3.2. Results

On the evaluation dataset, the AI assistant as a rater achieves 87.7% accuracy, performing above human raters who achieved 75.1% accuracy on average. Aggregating human ratings into a single label rather than averaging individual ratings has been shown to increase accuracy. Since multiple raters provided ratings for each example, we can calculate the majority vote rating. We take the modal rating as the majority vote rating (for ties, we mark the example as "Inaccurate"). If we use human majority vote ratings, human raters achieve a higher accuracy of 80.6% (compared to 75.1%) on the entire Evaluation Set, but still perform worse than the AI rater. Since human accuracy is higher using majority vote ratings, we will use these aggregated ratings for Hybridization.

If we conduct Confidence-based Hybridization with a threshold T=.62 (i.e., humans rate all examples where AI confidence in the factuality rating is equal to or less than .62), accuracy on the entire Evaluation Set is 89.3%, higher than using AI ratings alone (87.7%). As can be seen in Fig., there are a range of thresholds below T=0.74 for which Hybridized accuracy is higher than AI alone. For our analyses, we use T=0.62 as this is a threshold among a small set of thresholds where Hybridized accuracy is at its highest.

To test the statistical significance of these observed differences, we conducted a mixed effects logistic regression on the entire Evaluation Set predicting label accuracy (1 or 0) from a fixed effect of rating protocol (three-level factor: AI alone, human alone, hybridized, with AI alone as the reference category) with random intercepts by example and by label source (i.e., AI, human). Majority vote human alone ratings were significantly less accurate than AI ratings alone ($\beta = -1.461, SE = 0.161, z = -9.070, p < .001$), but Hybridized accuracy was significantly more accurate than AI ratings ($\beta = 0.413, SE = 0.164, z = 2.520, p = .012$).

Hybridization can only lead to an increase in accuracy compared to AI if humans are better than the AI on the subset of examples they rate. If we filter the Evaluation Set using T=.62, there are 280 examples where AI confidence in the overall rating is 0.62 or lower, what we will call the *Post-Hybridized* Human Set. On this Human Set, AI accuracy is 60.5% (much lower than accuracy on the entire set), while human accuracy is 71.3%. To test whether this difference was statistically significantly, we fit a mixed effects logistic regression predicting label accuracy (0 or 1) from a fixed effect of label-type (i.e., AI or human with AI as the reference) and a random intercept by conversation. This analysis indicated that human ratings were indeed significantly more accurate than AI ratings on this set

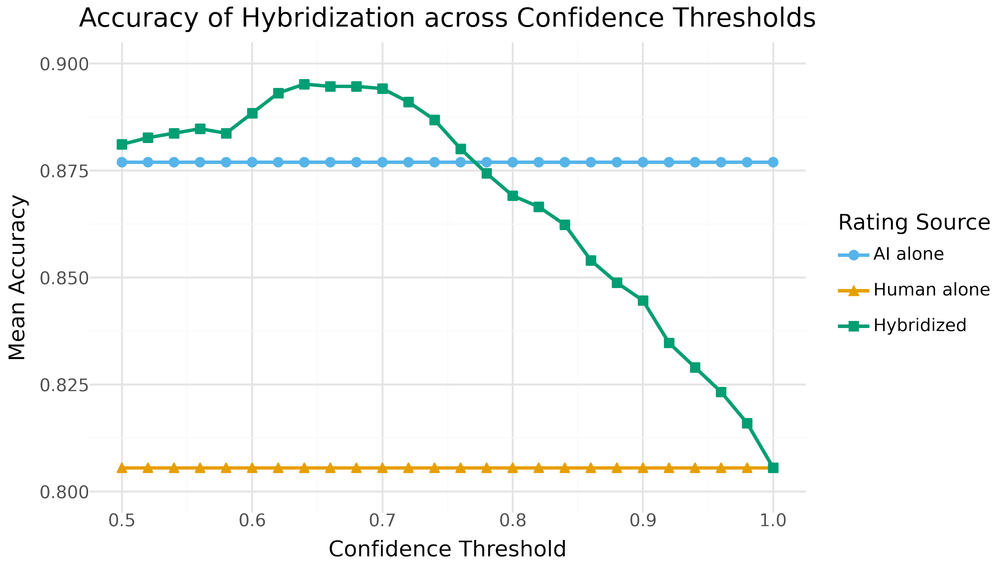

Figure 1 | Mean Accuracy of different rating protocols by Confidence Threshold. Humans alone (orange) achieved lower accuracy than AI alone (blue). Hybridized accuracy (green) varies by confidence threshold and there is a range of thresholds for which Hybridization achieves a statistically significantly higher accuracy than using only AI ratings. (NB: Human ratings are majority vote labels from the "T1" studies, described in Section 2.3.1)

($\beta = 0.514, SE = 0.185, z = 2.775, p = 0.006$).

In conclusion, confidence-based hybridization enabled human-AI complementarity: combining human and AI labels led to higher overall accuracy on the Evaluation Set than using AI ratings or human ratings alone. This is because (1) we had a fairly well-calibrated confidence metric, and (2) there was some level of decorrelation between human and AI labels (as the AI's accuracy decreased with lower AI-confidence, human accuracy didn't decrease as much), as shown in Appendix Fig 10. These two properties may help in identifying which tasks and methods are promising for hybridization.

2.4. RQ2: Does AI assistance improve human rater accuracy?

Next, we investigated whether providing human raters with access to the AI as a fact-verification assistant would improve their rating accuracy, specifically on the Post-hybridized Human Set. We were also curious what form of assistance would be best for fostering appropriate reliance (Section 2.4.3 below). We ran a series of experiments where human raters were given access to different versions of the AI fact-verification assistant, whose information they could use to help them determine the factuality of the sentence. We had a few different hypotheses about the possible results of these studies. First, it's possible that providing assistance on the hybridised set, the set of examples where the AI is less confident, might actually hurt rather than help raters, especially given that the AI assistant performs statistically significantly worse than human raters on this set. Alternatively, since the fact-verification trace generated by the AI includes search results and verified quotes from websites, it's possible that there is information in the trace that is still helpful to human raters; likewise, if there are errors in the AI's logic or reasoning, this might still be useful to humans for identifying the correct factuality rating. Still another possibility is that humans recognize that they are more competent than the AI and so choose to disregard the information it provides, leading to no difference in assisted vs. unassisted accuracy.

2.4.1. "T2" Human Rater Experiments

We ran 10 experiments, one unassisted baseline experiment (which we will call *baseline_t2* for time 2) and nine different assisted intervention experiments. This was a between-subjects design, so participants only participated in one of the 10 experiments. Participants were recruited from the same larger pool of raters as the previous experiment, but we took measures to ensure that participants who took part in the previous experiment were evenly distributed across the current experiments.

Unlike the initial baseline experiment, participants did not rate the entire Evaluation Set and instead only rated the examples where model confidence was equal to or lower than .75 (611 examples). We had them rate more than just the 280 examples in the Post-hybridized Human Set (with T=.62), since if assistance improves human performance it's possible that the optimal threshold for hybridization might change. We ran the baseline experiment again because rater performance may vary depending on time of day, week, etc. Also, rater skill on the given task might change over time, as participants get more practice doing similar tasks.

We recruited for 256 rating hours per experiment, and participants had an opportunity to take part three times if there were still spots available.

Different Versions of the Assistant. Because we trained the AI fact-verification assistant to output a highly structured output, we could easily filter and curate which aspects of the output were displayed and how.

We will use the following terminology to describe what information was displayed in each experiment:

- **Search results** - the search queries and results with URLs and snippets.
- **Evidence** - the numbered list of selected evidence with URLs and quotes. Quotes were verbatim from web sources, and usually a couple sentences long. (NB: Whenever Evidence was shown, Search results were also shown)
- **Reasoning** - a list of extracted factual claims from the sentence, and explanation of the factual accuracy of each claim, citing and summarizing the evidence.
- **Judgments** - the judgments of factual accuracy for each claim (if "Reasoning" is included) and the overall rating for the sentence (NB: if factual claims are not listed, then only the Overall Judgment is displayed).
- **Confidence** - the confidence score expressed as both a range (low, medium, high) and rounded point estimate (e.g., Model Confidence: low (65%)). Participants were told that they could think of the confidence percentage as the likelihood that the information in the trace led the AI fact-verification assistant to the correct factuality rating for the sentence, which is a reasonable description since the assistant's confidence is fairly well calibrated.

For all versions of the assistant, there were drop-down menus under each section that participants could expand to learn more about the information included. They were warned in several places about how the information could be misleading (i.e., "Some claims might be missing, and the summary and evidence could be misleading. Indented Quoted text is guaranteed to be from the webpage. But, this evidence might still be untrustworthy, insufficient, or irrelevant."). If a rating was included, it was described as a predicted rating and raters were warned that it could be incorrect. Again, we hoped such warnings would encourage them to engage more critically with the assistant and not take its reasoning and judgments at face value.

We showed 8 different subsets of the full AI output to the human rater, to understand how these help them at the factuality verification task. See Figure 2 for our full list of experiments. We also included an "AI Debate" experiment, where we show the full output (Search, Evidence, Reasoning, and Judgments, but without Confidence), arguing for different overall factuality ratings (i.e., one trace argued for "Accurate" while the other argued for "Inaccurate") - this set-up is analogous to one-turn, simultaneous debate.

2.4.2. *Results*

Fig 2 shows how the different assistant presentation styles affect the average human rater accuracy on the Post-Hybridized Human Set (T=.62, 280 examples). We use bootstrap resampling to calculate confidence intervals on the Mean Individual Rater Accuracy for visual benefit. But to test if certain differences are statistically significant, we filtered data to the Post-Hybridized Human Set and fit a mixed effects logistic regression predicting label accuracy (1 or 0) from a fixed effect of experiment (10-level factor, with the unassisted baseline_t2 dummy coded as the reference variable), and a random intercept by participant. This analysis revealed that even though the six experiments manipulating whether the Evidence+Reasoning, Judgements, or Confidence were numerically higher in accuracy than baseline, that these differences were not significant (all $\beta's < 0.168$ and all $p's > 0.225$). For the less leading versions of the assistant, search alone was no different from baseline (search, 68.7% vs. baseline, 67.3%; $\beta = 0.072, SE = 0.141, z = 0.511, p = 0.609$), and interestingly, debate was the only intervention that was numerically lower in accuracy than baseline, though this difference was also not significant (64.9%; $\beta = -0.124, SE = 0.137, z = -0.904, p = 0.3669$). The only form of assistance that led to a significantly higher average accuracy than baseline was search and evidence (73.3%, $\beta = 0.308, SE = 0.136, z = 2.269, p = 0.023$).

If we re-run this analysis with search and evidence dummy coded as the reference variable we find that Search and Evidence performs significantly better than Judgments & Confidence ($\beta = -0.285, SE = 0.141, z = -2.024, p = 0.043$) and Debate ($\beta = -0.432, SE = 0.145, z = -2.981, p = 0.003$) but no different than the other interventions (all β's between -0.236 and -0.141 and all $p's > 0.111$).

All in all these results suggest that for our assistant (1) the selected evidence is particularly helpful for improving overall rating accuracy and (2) the rest of Evidence+Reasoning, Judgments, and Confidence did not appear to statistically significantly or differentially improve rater accuracy. It's possible these components are indeed helpful when they support the correct overall factuality rating, but when overall rating is incorrect, these components may be misleading, limiting the degree to which this information can overall increase rating accuracy. In other words, it's possible that these forms of assistance help raters when the AI is right but hurt raters when the AI is wrong. In the next section, we investigate under- and over-reliance directly by examining assisted rater accuracy for the examples where the AI generated the correct factuality rating for the sentence vs. examples where the overall rating was incorrect.

2.4.3. *RQ2a: What form of assistance is best for calibrating reliance?*

To understand why certain forms of rater assistance work better than others, it is useful to compare the degree to which human raters under- and/or over-rely on the assistant. That is, does the rater use or trust the information the assistant provides the appropriate amount, too little, or too much given its validity? We operationalize under-reliance, i.e. not trusting the assistant enough when it is warranted to do so, as how often the rater still provides the incorrect factuality judgment for the sentence when the assistant generates the correct label. Conversely, we operationalize over-reliance, i.e. trusting the assistant too much when it's not warranted, as how often the rater provides the incorrect factuality judgment when the assistant generates the incorrect label compared to how often the rater provides the incorrect judgment without assistance (i.e., does the assistant significantly decrease raters' performance on this slice of the data).

As identified in the HCI literature, over-reliance is particularly important to avoid, as well as especially critical for achieving complementary performance. One would hope that the assistant only provides additional information, and doesn't persuade the rater to incorrectly go against their own judgment or to perform worse than they would on their own, ultimately, reducing the benefits of hybridization as raters are less likely to rely on their own complementary strengths. In a future where rater assistants are widely deployed, we don't want raters to uncritically and across-

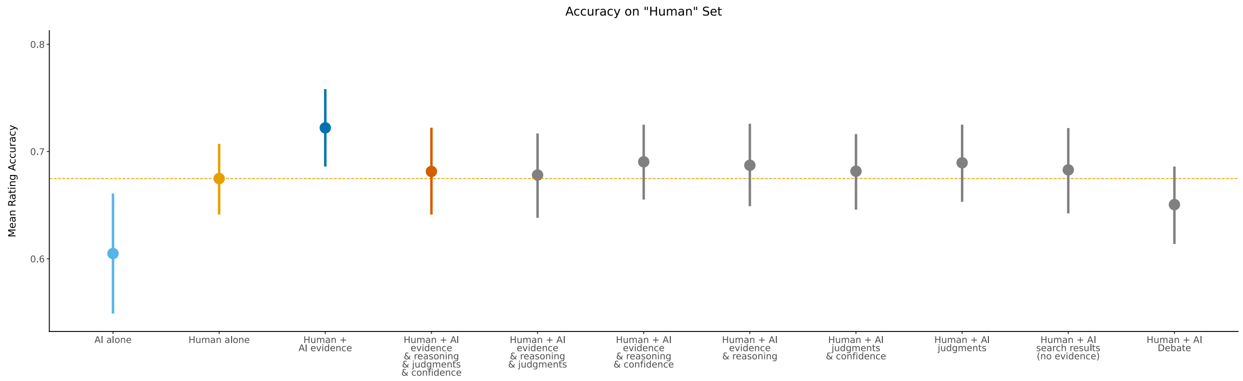

Figure 2 | Mean Individual Rater Accuracy, for the different assistant presentations. Accuracy is calculated over the Post-hybridized Human Set (i.e., restricted to examples where model confidence in the overall rating is <= .62). Dotted horizontal lines are Model Accuracy (60%) and Baseline unassisted accuracy (69.3%) to ease comparison. Error bars are 95% Bootstrapped confidence intervals of the mean. These were from "T2" studies described in Section 2.4.1.

the-board default to the AI's output. We want to leverage the complementary skills and knowledge of AI and humans to achieve a higher quality supervision signal than either entity could achieve alone. Ideally, we would like a form of assistant that helps when it is correct and at least does not hurt when it is incorrect.

Figure 3 breaks down accuracy across the different experiments by whether or not the AI assistant generated the correct overall factuality judgment for the sentence. To test whether raters over-relied on particular forms of assistance, we fit a mixed effects logistic regression predicting label accuracy (1 or 0) from fixed effects of experiment (10 level factor, with the unassisted baseline dummy coded as the reference variable) and model accuracy (two level factor, with incorrect as reference), and their interaction with a random intercept by participant. This analysis indicated that when the model Judgments and Evidence+Reasoning were shown (regardless of whether or not Confidence was included), raters performed worse with assistance compared to baseline on examples where the model was incorrect, and this decrement was statistically significant (Evidence & Reasoning & Judgments & Confidence: $\beta = -0.636, SE = 0.187, z = -3.405, p < .001$; Evidence & Reasoning & Judgments: $\beta = -0.768, SE = 0.174, z = -4.413, p < .001$), i.e., there was evidence of over-reliance. This was also true when both the overall Judgment and Confidence were shown, even if the Evidence and Reasoning are not included (Judgments & Confidence: $\beta = -0.356, SE = 0.176, z = -2.029, p = .042$). However, if just the overall Judgment is shown alone, or if no Judgments are included (regardless of the Evidence+Reasoning and Confidence), assisted performance was numerically but not statistically worse than baseline (all β's between -0.343 and -0.172 and all $p's > .054$), i.e., the rates of over-reliance were not statistically significant. All together these results suggest that if the overall Judgment for the sentence is shown along with another piece of information from the assistant model, whether that be the Evidence+Reasoning or Confidence, human raters tend to over-rely on the model's judgment.

If we turn to our less leading forms of intervention, we see that for Debate, Evidence, and Search-only (not including evidence), performance on the examples that the model gets incorrect is no different when assisted vs. unassisted, i.e., there is no evidence of over-reliance in these experiments (all β's between -0.138 and 0.163, and all $p's > .181$). Indeed, for Evidence and Search results (without evidence) performance is even numerically higher than baseline (though the increase in each case is small and not statistically significant). These findings confirm that these forms of assistance are indeed less leading and so are also less mis-leading. Interestingly, showing the full output (without confi-

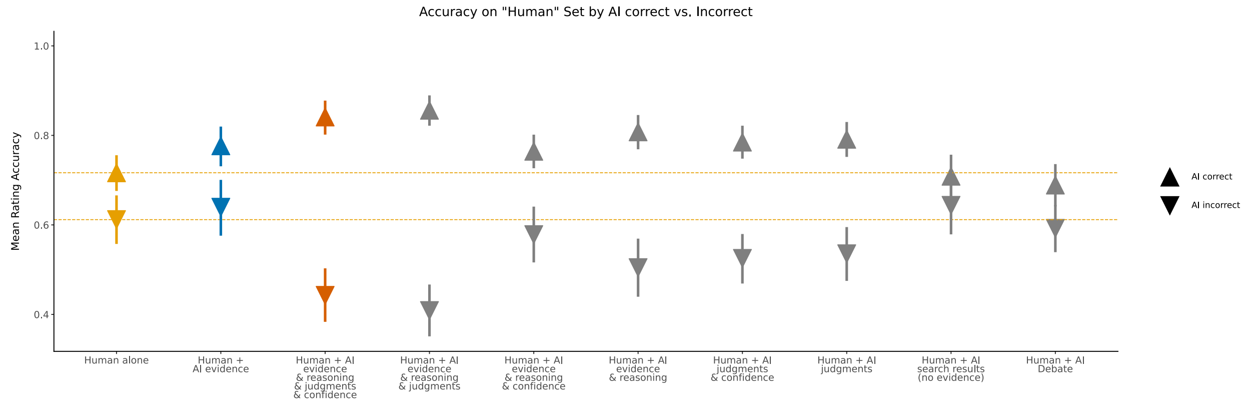

Figure 3 | Mean Individual Rater Accuracy, for the different assistant presentations, split by whether the Overall Judgment generated by the fact-verification assistant was correct or incorrect (note that some forms of assistance did not display this judgment). Accuracy is calculated over the Post-hybridized Human Set (i.e., restricted to examples where model confidence in the overall judgment is $<= .62$). Dotted lines are Baseline accuracy for correct and incorrect to ease comparison. Error bars are 95% Bootstrapped CIs of the mean. These were from "T2" studies, described in Section 2.4.1.

dence) leads to over-reliance, but showing two of these outputs and contrasting Judgments in the Debate experiment seems to mitigate this tendency.

To better understand the full impact of the fact-verification assistant, it is important to also understand its effect on examples where it generated the correct overall Judgment. If we re-fit the same regression model described above, but refactored so that "correct" is the reference level for the predictor of model accuracy, we find that assisted performance statistically significantly better than baseline for all forms of assistance (all β's > 0.357 and all $p's < 0.027$, except for Debate and only showing search results (Debate: $\beta = -0.115, SE = 0.167, z = -0.689, p = .491$; Search-results-only: $\beta = 0.25, SE = 0.172, z = 0.065, p = .948$).

No form of assistance gets raters to 100% on the examples where the model generates the correct factuality judgment, so there is evidence of under-reliance throughout, but it appears to be the most extreme for Debate and Search-results-only where raters are no different from baseline - indeed for these forms of assistance they neither help when correct nor hurt when incorrect, so it is possible raters just ignored them. Under-reliance appears to be the least prominent when raters see the Reasoning in combination with the Judgments (regardless of whether Confidence is included) - raters achieve 84% accuracy when Confidence is included (Evidence & Reasoning & Judgments & Confidence) and 86% when not (Evidence & Reasoning & Judgments). These forms of assistance, however, also exhibit the worst amount of over-reliance, suggesting that they are both the most helpful when correct and the most harmful when wrong, which is why they don't lead to an overall increase in accuracy above baseline.

The only form of assistance that achieves the ideal of helping when correct and not hurting when wrong, is showing just the Evidence (alongside the Search results). This form of assistance statistically significantly improves performance compared to baseline on the examples that the model gets correct (79.3% vs. 71.3%, $\beta = 0.446, SE = 0.170, z = 2.631, p = .009$), and as described above, does not affect performance when the model is incorrect (64.0% vs. 61.5%). It's possible that when the assistant generates the incorrect factuality judgment, the information present in the search results and evidence list is still useful to the raters and perhaps even correct and valid. If this is the case, it points to a limitation in the way we operationalize over-reliance. We slice the data based on if the final model judgment for the sentence is correct, not based on whether all of the intermediate steps and information provided is correct. Doing the latter would provide a more accurate measure of over-reliance. We return to this limitation in the discussion.

2.4.4. *RQ2 Summary*

In summary, most AI-assistance presentations did not improve accuracy on the Post-Hybridized Human Set, but showing participants the AI's search results and selected evidence *did* increase human rating accuracy because it was best at calibrating reliance.

For RQ1, Confidence-based Hybridization with unassisted majority vote human ratings achieved higher performance than using AI ratings alone. Unlike baseline_t1 data, in the experiments for RQ2 (T2 data), we did not collect a sufficient number of ratings per example to see an increase in performance from aggregating individual ratings into a majority label, so we ran an additional intervention experiment using AI Evidence assistance to increase the number of ratings per example.

The experiment was identical in structure and content to the prior search and evidence experiment except that we made it available to the entire pool of participants so that we could achieve sufficient replication per example (i.e., all participants from prior experiments were eligible to participate again). We recruited for 375 rating hours, and participants had an opportunity to take part two times, spots permitting. Participants rated examples in the equal to or lower than .75 confidence set, and average replication was 6.42 per example. Average individual label accuracy on the Post-Hybridized Human Set (T = 0.62) was 75.12% and average majority label accuracy was 84.51%, so we again see a boost from aggregating human labels, as we did for RQ1.

Hybridizing with majority vote Human + AI Evidence ratings from this experiment, using a confidence threshold of 0.62, results in an accuracy of 91.3%, compared to 89.3% using unassisted, baseline_t1 ratings. To test the statistical significance of this difference, we conducted a mixed effects logistic regression on the entire evaluation set predicting label accuracy (1 or 0) from a fixed effect of rating protocol (four-level factor: AI alone, human alone, hybridized w/ unassisted humans, hybridized w/ evidence-assisted humans, with hybridized w/ unassisted humans as the reference) with random intercepts by example and by label source (i.e., AI, unassisted human, assisted human). This analysis indicated that hybridizing with evidence-assisted humans achieved significantly higher accuracy on the entire Evaluation Set than hybridizing with unassisted humans ($\beta = 0.590, SE = 0.180, z = 3.276, p = .001$).

We thus find that Confidence-based Hybridization achieves human-AI complementarity and that combining hybridization with Rater Assistance further increases performance above AI ratings alone.

3. Discussion

3.1. The Future of Human Oversight in LLM Training and Evaluation

As AI continues to improve, including in rating ability, a critical question arises: will humans even be necessary for oversight in the future? If AI raters become demonstrably superior, will there be any slice of data where assisted humans add value? There are several reasons why we expect it to remain critical to keep humans in-the-loop to develop superhuman AI safely and in line with human values:

- **Capabilities:** The relative strengths of humans compared to AI might change over time, but it is possible that humans will still retain complementary knowledge, skills, and abilities. For example, AIs might still have specific weaknesses such that the human-AI oversight team is more adversarially robust to reward hacking targeting the weaknesses of one or the other specifically (e.g., the comprehensiveness-hallucination trade-off described in McAleese et al., 2024). Humans might also be more adversarially robust in general. The current frontier of AI is also very "jagged", and while AI might be better than

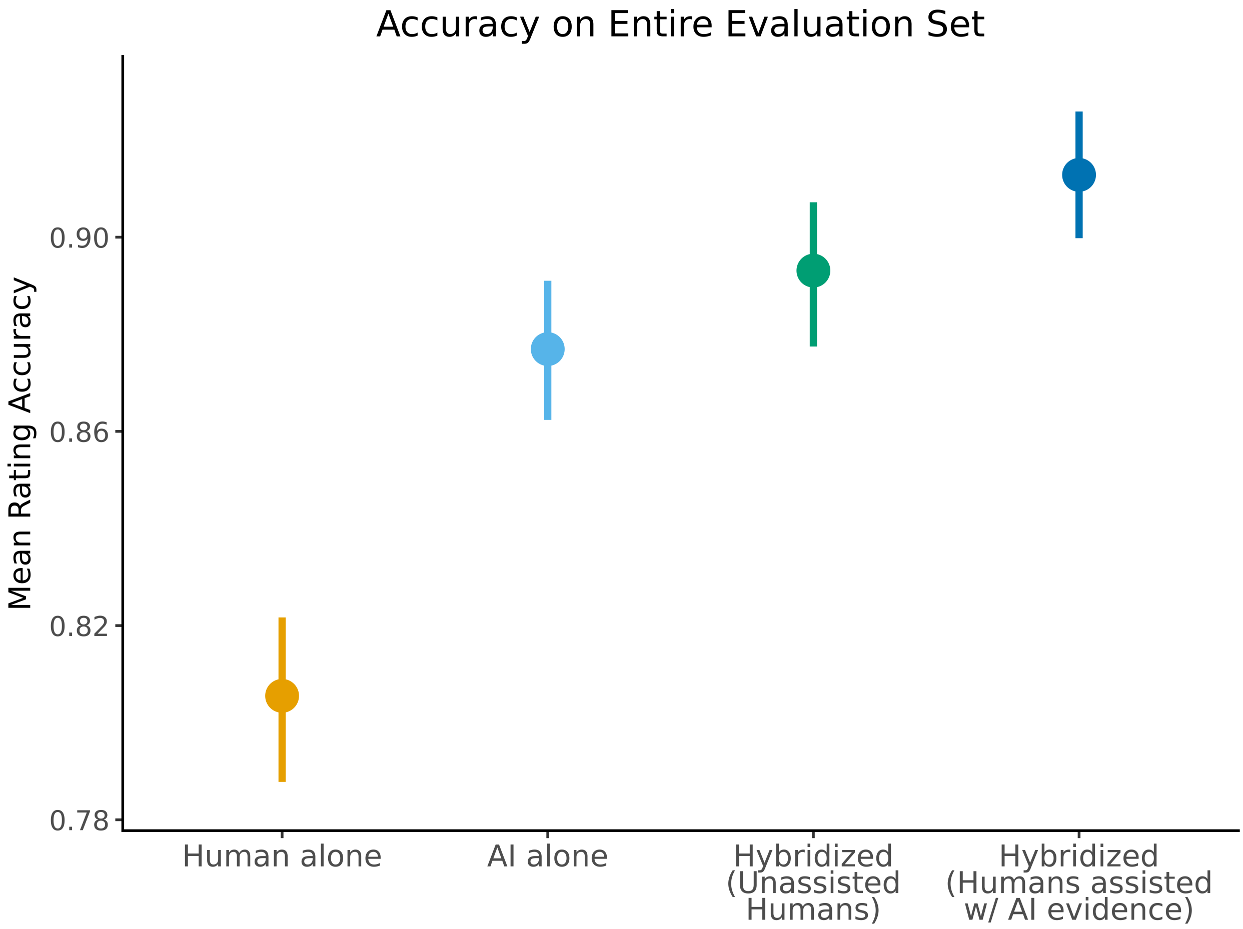

Figure 4 | Mean accuracy and and 95% confidence intervals on the entire evaluation set of humans alone (yellow), AI alone (light blue), hybridization with unassisted humans (green), and hybridization with Evidence-assisted humans (dark blue). The assisted-hybridized approach is significantly more accurate than the Human alone, AI alone, and the unassisted-hybridized protocol.

humans overall for a domain, it might unexpectedly do worse on certain problems that we can identify via tools like AI-confidence.

- **Value Alignment:** Ultimately, humans will likely need to continue to provide input to confirm that AI systems are indeed acting in accordance with human values. This is because human values continue to evolve. In fact, human preferences define a "slice" of data where humans are definitionally more accurate than non-humans (including AI). AI systems might get quite good at predicting what aligned behavior should be in out-of-distribution scenarios, but it's unlikely that AI will be able to figure out what humans want in completely new situations without humans being consulted and kept in the loop.
- **Trust:** As AIs improve, they might develop the capability to "scheme" and sabotage the rating process (Meinke et al., 2024). AIs with these capabilities pose loss of control risks (Ngo et al., 2022), and so we should not fully trust the output from these models (including AI raters) without better ways for humans to monitor and supervise them. Humans are capable of scheming and sabotage as well, but we have a lot more experience with endowing trust to humans as part of critical decision making systems and guarding these systems from bad human actors.

How humans remain in the loop might change over time. Even if we switch entirely to AI raters for training our models, we may still use human input to train the AI raters and human judgments to evaluate them (as is the case today). If we switch to even higher level approaches such as Constitutional AI (Bai et al., 2022), human input will still be needed to write the rules for the AI to follow. Complementarity will continue to be relevant as long as humans are involved, wherever and however that might be (e.g., AI assistance for constitution writing), and so Amplified Oversight research will continue to benefit from research in HCI and other related disciplines. But the constant improvement of model capabilities will render human-AI complementarity a moving target, which means we need to develop repeatable and generalizable processes for conducting this type of research.

3.2. Discussion, Limitations, and Future Work

Overall, our rater AI-assistance with confidence-based hybridization leads to complementary performance, better than either humans or AI individually, in our Factuality rating task where AI-raters outperform human raters.

Appendix Section A.3.4 shows that at a confidence threshold of 1 (i.e. no confidence-based hybridization, only human ratings), AI-assisted humans perform better than unassisted humans, but they do not surpass model performance. In our experiments, confidence-based hybridization techniques are critical to achieve complementarity, and also critical to enable rater assistance to achieve complementarity. The reason it is hard for even assisted humans to surpass our AI rater seems largely due to the fact that the model gets extremely high accuracy at high confidences (see Appendix Figure 10), which it has for a large proportion of the data (see Appendix Figure 11). Even so, the errors of the AI and humans are not entirely correlated and there is a slice of data where humans do perform better, suggesting the AI and humans have complementary skills on this task that we were able to leverage via confidence-based hybridization.

Importantly, confidence-based hybridization requires a well calibrated AI rater, where increased confidence translates to increased rating accuracy and quality. Proper estimation and calibration of LLM's uncertainty is a challenging task, and may not be a given for many AI raters. But on tasks where AI raters outperform assisted humans, any sort of slicing of the data to find sections where assisted humans outperform AI raters would lead to increased performance. Exploring other forms of hybridization could be a promising area for future Scalable Oversight research. A key limitation of our confidence-based hybridization approach was that we didn't use a separate "train" set to calibrate the threshold. Future work should understand how well this hybridization method generalizes.

Appendix Figure 10 shows that some types of assistance do better than AI Evidence at higher confidences, indicating that perhaps hybridization between different forms of assistance might lead to even better performance. In our experiments (see Appendix Section A.3.4.1), this "assistance-hybridization" did not improve accuracy above hybridizing with a single form of assistance, likely because the model still surpasses the other forms of assistance at higher confidence. But we think this "assistance-hybridization" is still a promising research direction, allowing for different types of assistance targeting different types of data.

Our results on debate showed that it led to low over-reliance, yet also low overall accuracy. On the one hand, this might point to debate as a promising method for reducing over-reliance. On the other hand, it didn't lead to *appropriate* reliance as it increased under-reliance, and didn't increase accuracy over the baseline. Our debate method could be improved significantly though. Our model was not trained to give arguments for both sides. Also, there was a lot of text shown to raters (double that of the next longest form of assistance), which could have easily led to raters disengaging and ignoring the assistance.

As AI improves, the quality of human raters will also need to improve to be able to give useful feedback. Our results around bonus studies (Section A.3.2 in Appendix) allude to how the rater assistant we built may not be useful for higher quality raters. In addition, in our experiments in Appendix Section A.3.4.1, raters got better and assistance no longer helped. These practice effects and human raters improving over time could have been another reason that our human rater accuracy in Section 2.4.4 was so high, as we opened up that experiment to all prior participants in order to get sufficient number of ratings. Different kinds of assistance might be necessary as raters' abilities on a task improve either from increased motivation or skill. For example, future work could explore building AI assistance that aims to teach human raters, and improve rater quality over time. This could build on the wealth of existing pedagogy research.

The AI assistants will also improve as the capabilities of the underlying foundation models improve. Unfortunately, it will be increasingly difficult to conduct realistic hybridization and rater assistant research, since it will become harder to collect golden data from experts to evaluate how our techniques are improving rater quality. There might already be domains where currently we don't have a good way of getting golden labels (e.g. on moral reasoning questions). If AI becomes so capable that we cannot distinguish the quality of their ratings from experts', new methods of collecting golden labels might need to be created in order to continue to know how best to leverage complementary human and AI strengths.

3.3. Acknowledgements

We would like to thank Geoffrey Irving and Jonathan Uesato for their thought and research leadership on the GDM Rater Assist team; we additionally thank Jonathan for his engineering contributions, which were the foundations for the ongoing work presented here. We are also grateful to John Aslanides and Zoe Ashwood for their crucial contributions to an earlier iteration of the project, and to Rohin Shah, Noah Goodman, Anca Dragan, and Sasha Goldshtein for their thoughtful guidance. Lastly, a big thank you to MH Tessler, Mevan Babakar, Verena Rieser, Irene Rae, Tom Everitt, Noah Siegel, Taylan Cemgil, Zachary Kenton, and Samuel Albanie for their helpful feedback and discussions.

References

D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL `https://arxiv.org/abs/2212.08073`.

A. Baird and L. M. Maruping. The next generation of research on is use: A theoretical framework of delegation to and from agentic is artifacts. *MIS quarterly*, 45(1), 2021.

G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld. 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*, pages 1–16, 2021.

S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

Z. Buçinca, M. B. Malaya, and K. Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), Apr. 2021. doi: 10.1145/3449287. URL `https://doi.org/10.1145/3449287`.

Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 454–464, 2020.

P. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. 06 2017.

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

K. D. Dvijotham, J. Winkens, M. Barsbey, S. Ghaisas, R. Stanforth, N. Pawlowski, P. Strachan, Z. Ahmed, S. Azizi, Y. Bachrach, L. Culp, M. Daswani, J. Freyberg, C. Kelly, A. Kiraly, T. Kohlberger, S. McKinney, B. Mustafa, V. Natarajan, K. Geras, J. Witowski, Z. Z. Qin, J. Creswell, S. Shetty, M. Sieniek, T. Spitz, G. Corrado, P. Kohli, T. Cemgil, and A. Karthikesalingam. Enhancing the reliability and accuracy of ai-enabled diagnosis via complementarity-driven deferral to clinicians. *Nature Medicine*, 29(7):1814–1820, 2023. ISSN 1078-8956. doi: 10.1038/s41591-023-02437-x.

A. Fügener, J. Grahl, A. Gupta, and W. Ketter. Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research*, 33(2):678–696, 2022.

P. Hemmer, M. Westphal, M. Schemmer, S. Vetter, M. Vössing, and G. Satzger. Human-ai collaboration: The effect of ai delegation on human task performance and task satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 453–463, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701061. doi: 10.1145/3581641.3584052. URL `https://doi.org/10.1145/3581641.3584052`.

E. Hubinger. Ai safety via market making, Jun 2020. URL `https://www.alignmentforum.org/posts/YWwzccGbcHMJMpT45/ai-safety-via-market-making`.

G. Irving, P. Christiano, and D. Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

Z. Kenton, N. Y. Siegel, J. Kramár, J. Brown-Cohen, S. Albanie, J. Bulian, R. Agarwal, D. Lindner, Y. Tang, N. D. Goodman1, and R. Shah. On scalable oversight with weak llms judging strong llms. *arXiv preprint arXiv:2407.04622*, 2024.

A. Khan, J. Hughes, D. Valentine, L. Ruis, K. Sachan, A. Radhakrishnan, E. Grefenstette, S. R. Bowman, T. Rocktäschel, and E. Perez. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*, 2024.

V. Krakovna, J. Uesato, M. R. Vladimir Mikulik, T. Everitt, R. Kumar, Z. Kenton, J. Leike, and S. Legg. Specification gaming: the flip side of ai ingenuity, apr 2020. URL https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/.

V. Lai and C. Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 29–38, 2019.

J. D. Lee and N. Moray. Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies*, 40(1):153–184, 1994.

J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

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

J. Li. A comparative study on annotation quality of crowdsourcing and llm via label aggregation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6525–6529. IEEE, 2024.

S. Ma, Y. Lei, X. Wang, C. Zheng, C. Shi, M. Yin, and X. Ma. 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*, pages 1–19, 2023.

S. Ma, X. Wang, Y. Lei, C. Shi, M. Yin, and X. Ma. "are you really sure?" understanding the effects of human self-confidence calibration in ai-assisted decision making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642671. URL https://doi.org/10.1145/3613904.3642671.

A. Manzini, G. Keeling, N. Marchal, K. R. McKee, V. Rieser, and I. Gabriel. Should users trust advanced ai assistants? justified trust as a function of competence and alignment. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1174–1186, 2024.

N. McAleese, R. M. Pokorny, J. F. C. Uribe, E. Nitishinskaya, M. Trebacz, and J. Leike. Llm critics help catch llm bugs, 2024. URL https://arxiv.org/abs/2407.00215.

A. Meinke, B. Schoen, J. Scheurer, M. Balesni, R. Shah, and M. Hobbhahn. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*, 2024.

J. Michael, S. Mahdi, D. Rein, J. Petty, J. Dirani, V. Padmakumar, and S. R. Bowman. Debate helps supervise unreliable experts. *arXiv preprint arXiv:2311.08702*, 2023.

R. Ngo, L. Chan, and S. Mindermann. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022.

R. Parasuraman and V. Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

A. Parrish, H. Trivedi, N. Nangia, V. Padmakumar, J. Phang, A. S. Saimbhi, and S. R. Bowman. Two-turn debate doesn't help humans answer hard reading comprehension questions. *arXiv preprint arXiv:2210.10860*, 2022a.

A. Parrish, H. Trivedi, E. Perez, A. Chen, N. Nangia, J. Phang, and S. R. Bowman. Single-turn debate does not help humans answer hard reading-comprehension questions. *arXiv preprint arXiv:2204.05212*, 2022b.

S. Petridis, B. D. Wedin, J. Wexler, M. Pushkarna, A. Donsbach, N. Goyal, C. J. Cai, and M. Terry. Constitutionmaker: Interactively critiquing large language models by converting feedback into principles. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 853–868, 2024.

M. Pinski, M. Adam, and A. Benlian. Ai knowledge: Improving ai delegation through human enablement. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3580794. URL https://doi.org/10.1145/3544548.3580794.

R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

W. Saunders, C. Yeh, J. Wu, S. Bills, L. Ouyang, J. Ward, and J. Leike. Self-critiquing models for assisting human evaluators. 2022. URL https://arxiv.org/abs/2206.05802.

M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

C. Si, N. Goyal, S. T. Wu, C. Zhao, S. Feng, H. Daumé III, and J. Boyd-Graber. Large language models help humans verify truthfulness–except when they are convincingly wrong. *arXiv preprint arXiv:2310.12558*, 2023.

A. Tharwat and W. Schenck. A survey on active learning: State-of-the-art, practical challenges and research directions. *Mathematics*, 11(4), 2023. ISSN 2227-7390. doi: 10.3390/math11040820. URL https://www.mdpi.com/2227-7390/11/4/820.

M. Vaccaro, A. Almaatouq, and T. Malone. When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303, 2024.

H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, and R. Krishna. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7 (CSCW1):1–38, 2023.

X. Wang, H. Kim, S. Rahman, K. Mitra, and Z. Miao. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2024.

J. Wei, C. Yang, X. Song, Y. Lu, N. Hu, D. Tran, D. Peng, R. Liu, D. Huang, C. Du, et al. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*, 2024.

A. Appendix

A.1. Extended Related Work on Amplified/Scalable Oversight

We add a more detailed Related Work section here to guide knowledge share between the Amplified Oversight and HCI fields.

Amplified Oversight refers to the problem of how we can still provide accurate feedback on whether or not an AI output is “good” and “correct”, on tasks that are very hard for humans to verify (Amodei et al., 2016). As AI improves, it will be asked to do harder tasks, and verifying those tasks will also get harder. Humans might even reward problematic behavior that looks "good" but isn’t actually desirable Krakovna et al. (2020), such as generating misleading or inaccurate information and sycophancy Sharma et al. (2023). Providing accurate feedback is crucial to building safe AI.

The field of Amplified Oversight focuses on combining the complementary strengths of human raters and capable AI to provide this more accurate feedback. Fundamentally this is an HCI problem: How do we leverage human and AI to do better than either one individually?

Prior work has mostly focused on how best we can provide AI assistance to human raters to amplify their verification abilities. Some of this work has also experimented with assisting AI raters for quicker iteration speed. The target of the assistance, whether AI or human, is called the Judge.

Critics are rater assistants that are instructed to give a single response to assist the rater, e.g. pointing out mistakes in the original AI response Saunders et al. (2022). CriticGPT was one such AI critic, optimized for developer-created guidelines on what would be a useful assistant McAleese et al. (2024). This AI performed better than humans, but wasn’t able to assist humans to be better than the AI. Another paradigm of AI Rater Assistance are when two (or more) AI’s debate each other over whether the original response is “good”, and this debate transcript is shown to the judge Irving et al. (2018). This debate can span one turn, or many turns. Debate has only been shown to work better than alternative approaches (such as critique assistance, or simply no assistance at all) in settings where the assistant has access to useful information that the judge does not have access to, for both human and AI judges (Kenton et al., 2024; Khan et al., 2024). When the information asymmetry between the debaters and judge is removed, the judge benefits little from the debate Kenton et al. (2024); Parrish et al. (2022a,b). Adding multiple turns of debate appears to offer no further benefit yet (Kenton et al., 2024; Khan et al., 2024).

Other proposed forms of AI Rater Assistance techniques have yet to be evaluated on realistic tasks. Notable examples include Iterative Amplification Christiano et al. (2018), Recursive Reward modelling Leike et al. (2018), and AI safety via market making Hubinger (2020).

Other Amplified Oversight techniques are centered around how we can best rely on AI raters alone. In Constitutional AI, humans only provide a constitution of rules for the AI system (with or without AI involvement), and then feedback is given solely through an AI rater that uses the written rules to determine if the original output is “good” (Bai et al., 2022; Petridis et al., 2024).

Amplified Oversight as a field originated from the AI safety and alignment community Amodei et al. (2016), and has focused largely on Machine Learning solutions. However, as discussed using AI assistance for improving human oversight is fundamentally an HCI problem, as the goal is how best to achieve complementary performance between the AI and the human, which is also the focus of AI assisted decision-making in the field of HCI (discussed in the next section). The Machine Learning field is just beginning to engage with the HCI community, and this paper hopes to further bridge this gap and foster excitement between both communities to collaborate.

A.2. Additional Details on Evaluation Set, Model, and Experiments

A.2.1. Evaluation Set

The *Evaluation Set* has 1918 [prompt, response, target sentence] tuples in total. The prompts come from a realistic and representative distribution of real user interactions with Gemini, and the responses come from Gemini at the time (late 2023, early 2024), through the Gemini App UI (then called "Bard"). Prompts and responses underwent rigorous filtering of Personally Identifying Information (PII), ensuring researchers only accessed de-identified data. The responses were split into sentences using code based on the NLTK library Bird et al. (2009).

The golden labels come from a high-quality human rating pool specifically trained for fact-verification. Three different raters rate each sentence as either "Accurate" or "Inaccurate". They could also rate sentences as "Doesn't require assessment", however, sentences with at least one of those labels were discarded from our Evaluation Set to ensure that all sentences contained information to fact-verify. We take the majority (i.e., modal) rating out of the 3 labels to get our golden label for each sentence. We have manually reviewed some of these ratings, and we do find them to be very accurate.

Here's a generated representative example of the type of data in the Evaluation Set:

Representative Eval Set Example

Prompt: Why are some lakes green?
AI Response (shortened here, but the full response is shown to raters): [. . .] Nutrient Enrichment: The primary driver of algal blooms is an abundance of nutrients, particularly phosphorus and nitrogen. These elements act as fertilizers for algae. [. . .]
Target Sentence: These elements act as fertilizers for algae.
Golden Label: Accurate

A.2.2. The AI Fact-verification model

A.2.2.1 Model Description We built an AI model that uses a search-engine to research the factuality of a given sentence. This model can serve as both a rater and an assistant to human raters. We took inspiration from SOTA AI Fact-verifiers Wei et al. (2024) and trained the AI model to follow these 5 steps:

1. **Decompose:** The AI decomposes the sentence to be fact-verified into its constituent Factual Claims.

2. **Search the web:** The AI iteratively searches the web for information that supports and/or that contradicts each factual claim. It uses a Search Tool to issue Search Queries, and for each query it receives a set of (around 5) Search Results that include the relevant website, and a relevant short snippet from the web page (usually a couple of sentences). It is trained to continue to issue Searches until it finds all necessary information to determine the factuality rating for each factual claim and thus the sentence as a whole.

3. **Select quotes:** The AI then selects only the most relevant quotes from the snippets as Evidence, along with the corresponding website URL, and numerically lists this evidence. These quotes are verified, verbatim quotes from the websites in the search results.

4. **Reasoning:** For each decomposed Factual Claim, the AI gives a short explanation for why that claim is factually accurate or not based on the selected evidence and cites those sources. For each claim, it also determines a Factuality Rating.

5. **Overall Rating:** Finally, the AI determines an Overall Factuality Rating for the sentence based on the ratings for each claim (all Factual Claims need to be accurate for the sentence to be "Accurate").

Claim-level and Overall Ratings could be one of ["Inaccurate", "Unsupported", "Disputed", "Accurate"]. But, the model's Overall Ratings were always one of "Inaccurate" or "Accurate" on our Evaluation Set, perhaps because of how the model was trained.

To determine the final rating of our model for each sentence in our evaluation data-set, we sample our model 50 times, and only select the samples that pass the Format Verifier (described below). On average, there were 33.25 verified samples per sentence. Out of these verified samples, we determine the Model Rating by taking the Majority (modal) Vote of the samples' Overall Ratings. To determine the AI output that we display to human raters in the assistant experiments, we choose the verified sample that has the highest reward model (RM) score (described below), out of those samples that have the same Overall Rating as the Majority Vote Rating (this is called Best-of-N sampling). In our Debate experiment where human raters see two AI outputs one arguing for "Accurate" and one arguing for "Inaccurate", we display the two outputs that each have the highest RM score among the samples with the same rating.

For each sentence, we also determine a model Confidence level. This is done by calculating the proportion of verified samples that have the same Overall Rating as the Majority Vote. Since the model's Overall Rating was binary, this confidence value was between .5 and 1. We found that confidence scores were fairly well-calibrated with accuracy, which is critical for confidence-based hybridization to be effective (See Fig 10).

Below is an representative example of the full output of the model after formatting. This formatted output is what is shown to raters. Right-facing triangles indicate HTML <details> tags that can be clicked to expand and reveal further information:

A.2.2.2 Model Training The AI assistant fact-verification model was trained by Supervised Fine-Tuning a pre-trained Gemini 1.5 Pro model, using high-quality human-written demonstrations. The human raters writing these demonstrations came from a high-quality rating pool (different from the pool that provided the golden labels), and they followed the same 5 steps as the AI model.

To ensure the human-written demonstrations were of suitable quality, we built a comprehensive tutorial so the human raters would understand the goals and nuances of fact-verification, as well as the steps to follow and the specific format of the model's output. Fact-verification even a single sentence can be a challenging task depending on the nature and number of factual claims in the sentence; there are many edge cases that can be difficult to verify. We provided detailed information and many examples to help raters understand the task and equip them with the skills and knowledge necessary to achieve it.

When writing their demonstrations, the raters had to adhere to the format described above for the AI and use the same search tool as the AI. To ensure each demonstration was formatted correctly, we implemented a Format Verifier to check demonstrations in real-time: each demonstration had to pass this verifier in order for the rater to be able to submit it. The verifier also had the following checks: the explanation for each claim must cite at least one piece of evidence; each piece of selected evidence must be cited somewhere; and the quote per piece of selected evidence must come verbatim from the webpage snippets returned in the search results.

We also trained a Preference Reward Model (RM). We obtained side-by-side preference ratings from the same human raters who wrote the demonstrations, and used these ratings to train the RM. We used this RM for the Best-of-N sampling mentioned in Section A.2.2.1.

A.2.3. Human Factuality Rating Task

We asked the human raters to follow a similar process to our AI rater. For each sentence, raters classified each target sentence as "Inaccurate", "Unsupported", "Disputed", or "Accurate", using online research and/or information from the AI fact-verification assistant (if displayed). For a sentence to be "Accurate", all factual claims in the sentence need to be accurate. A claim is accurate if no reputable contradictory evidence exists; if the claim is an opinion or recommendation, it is accurate if it is reasonable and widely held. If raters believed that a sentence did not contain any factual claims, they could choose a third option of "Doesn't require assessment." According to our golden labels, all sentences contain factual claims, so "Doesn't require assessment" ratings were coded as incorrect labels in our analysis. Lastly, raters could additionally choose to skip sentences they were unable to understand or confidently evaluate (by clicking a Skip button, or rating the sentence as "Can't Confidently Assess"), though they were asked to try to avoid doing so and were given an opportunity to indicate their confidence in the factuality rating (see below).

A.2.4. Human Experiment Set-Up

In all experiments, participants first completed a short tutorial that introduced them to the task, explained the study duration (one hour) and payment. We also included a short section explaining why fact-verification AI model outputs is important (i.e., to guard against models generating false, inaccurate, or misleading information that could lead to problematic or unsafe decisions on the part of users). We hoped this explanation would tap into participants' intrinsic motivation and improve their performance, as well as critical engagement with the assistant. Participants were also explicitly encouraged to do their best to provide the correct factuality rating for each sentence and to focus more on quality than quantity (i.e., to spend more time on fewer tasks). There was then a short reading comprehension quiz that provided immediate feedback; participants had to answer all questions correctly to advance. For participants in the unassisted baseline experiments, this was the end of the tutorial and they proceeded to the main task. For participants in one of the assisted intervention experiments, after the quiz, there was an additional short section explaining that they would be shown information from an AI fact-verification assistant. Participants were informed that this fact-verification assistant was still learning and could make mistakes so they should be critical of the information it provided and only use it to the extent they deemed it valid. We hoped this warning would help raters better understand the models' limitations and foster more appropriate reliance.

After completing the tutorial, participants spent the remainder of the hour on the main fact-verification task. For each fact-verification example, participants saw the AI-generated sentence, and underneath it, the factulaity rating scale where they classified the factual accuracy of the sentence as described above. Below the factuality rating scale was a four-point confidence likert scale, where raters could indicate their confidence in the factuality rating as: "not at all confident", "somewhat confident", "mostly confident", "completely confident".

In the assisted intervention experiments, the screen set-up was the same as for the unassisted baseline experiments, except that above the AI-generated sentence and rating scales, information from the AI fact-verification Assistant was displayed in a clearly demarcated, blue box. The box was titled, "Experimental AI-generated Fact-check" with a warning underneath to be careful because the information could be misleading. In addition to rating the factual accuracy of the sentence and their confidence in that rating, participants also rated the helpfulness of the experimental AI-generated fact-verification on a three-point likert scale: "not at all helpful", "somewhat helpful", "extremely helpful". Lastly, there was an optional comments box where participants could share any comments they might have about the AI-generated fact-verification.

Each experiment was between-subjects but within an experiment, participants could take part in the experiment more than once. They only completed the tutorial one time at the start of their first session. Participants were financially compensated for each hour of work spent on the experiment.

A.3. Supplementary Results

A.3.1. Terminology

The Supplementary Results below use a different, older terminology: "Trace" refers to the combination of Search Results and Evidence mentioned in Section 2.4.1. "Ratings" refers to Judgements. "Confidence" means the same. "+" means one of those is included in the assistance, and "-" means that it is not included.

A.3.2. How do monetary incentives change how the assistance helps?

In addition to comparing performance using different types of assistance, we also investigated how giving raters a bonus for providing correct ratings impacted the effects of assistance. Prior research has shown that increasing the monetary incentive for performance reduces over-reliance Vasconcelos et al. (2023). To test the effect of a bonus on human rater accuracy, we replicated the unassisted baseline experiment, as well as AI-assistance with Evidence, Reasoning, and Judgments (called **Trace+/Ratings+/Confidence-** below) and AI-assistance with Search and Evidence only (called **Search+/Evidence+** below). If participants completed at least six ratings (that were not "Doesn't require assessment") and achieved at least 80% accuracy on those ratings, they would receive a bonus equal to their hourly rate. The Baseline bonus study was done at the same time as the T2 studies described in Section 2.4.1. The assisted studies with bonus were done a week later. Each of the T2 and bonus studies were all done with mutually exclusive sets of raters. The resulting Mean Individual Accuracy for these studies are shown in Fig 5.

Providing raters with a bonus for accuracy did not increase their overall accuracy on the Post-hybridized Human Set with or without assistance. For the baseline experiments, non-bonus accuracy was 67.3%, while bonus accuracy was numerically but not statistically higher, 68.9% ($\beta = -0.132, SE = 0.128, z = -1.033, p = .302$). For both **Trace+/Ratings+/Confidence-** and **Search+/Evidence+** assistance, bonus accuracy was actually numerically *lower* than non-bonus accuracy, but again these differences were not statistically significant (**Trace+/Ratings+/Confidence-**: 67.1% with bonus vs. 69.8% without bonus, $\beta = 0.084, SE = 0.137, z = 0.612, p = .541$; **Search+/Evidence+**: 72.4% with bonus vs. 73.3% without bonus, $\beta = 0.047, SE = 0.145, z = 0.320, p = .749$). In addition, though **Search+/Evidence+** assistance increased overall accuracy compared to unassisted baseline without bonus, this assistance with a bonus did not statistically significantly improve accuracy compared to baseline with a bonus (72.4% vs. 68.9%, $\beta = 0.130, SE = 0.138, z = 0.939, p = .348$).

Fig 6 shows how under- and over-reliance changes when incentivizing raters for accuracy. Comparing baseline with and without bonus, we find that accuracy when the model is incorrect is higher with a bonus (67.1% vs. 61.4%, $\beta = -0.342, SE = 0.174, z = -1.969, p = .049$), so even though overall accuracy was not improved by incentivizing raters for accuracy, this extra monetary incentive did appear to improve performance on harder examples (unassisted performance is lower for model incorrect vs. correct examples suggesting that the examples the model got wrong were also more challenging for humans). However, unlike prior work, incentivizing raters did not appear to reduce over-reliance. In particular, **Trace+/Ratings+/Confidence-** assisted performance on incorrect examples was still statistically significantly worse when raters were bonused ($\beta = -1.108, SE = 0.184, z = -6.006, p < .001$). Likewise, the difference in accuracy between correct and incorrect examples was no different for **Trace+/Ratings+/Confidence-** with and without assistance ($\beta = 0.182, SE = 0.237, z = 0.770, p = .442$). For **Search+/Evidence+** assistance, there was again no evidence of over-reliance (i.e., unassisted and bonused performance was no different from assisted and bonused performance on the incorrect examples, $\beta = -0.060, SE = 0.188, z = -0.318, p = .750$).

In sum, incentivizing raters for accuracy decreased the positive effect of assistance but did not significantly increase overall accuracy nor reduce over-reliance. These results raise questions of how assistance might impact higher quality raters.

A.3.3. *How do the different forms of assistance affect the time spent per task?*

Fig 7 shows that all interventions increase time per raters take to do each task. We did filter out a few outliers in the data of task times that took longer than 1 hour.

There do seem to be certain presentation styles that increase the time taken more than others. Providing just Search results take the longest, possibly because that section has the most content and raters are reading those carefully when only they are presented. Showing evidence also leads to much higher time taken per task compared to baseline and other some interventions, perhaps because raters are going more often to the linked sites and understanding the evidence. There also seems to be a trend of how showing the Confidence increases the amount of time taken. This effect seems stronger when the Trace is present.

The bonus also seems to slow down raters, about the same absolute amount for each presentation style.

A.3.4. RQ3: Does the optimal confidence threshold for hybridization change with assistance, and if so, how does that affect which form of assistance is best?

In our initial experiment for RQ1, we fixed the kind of human ratings (unassisted baseline), and we allowed the confidence threshold to vary to identify the optimal threshold for hybridization. In the experiments for RQ2, we fixed the confidence threshold at the previously selected optimal level (T=.62) and explored various types of human ratings, unassisted and different forms of assisted. Now, we run a series of experiments that allow us to vary both the confidence threshold and the human ratings. Varying both the threshold and type of assistance will enable us to see if the optimal threshold is different for different forms of assistance and whether this changes the optimal assistance. Figure 12 generated with T2 results shows how changing the threshold might increase accuracy for certain types of assistance. These results are limited in that we cannot increase the threshold beyond .75, and we don't have enough ratings per example to see benefit from majority vote rating, leading us to run another study.

We select three forms of assistance to explore: (1) trace with ratings and confidence (**Trace+**/**Ratings+**/**Confidence+**), (2) trace with ratings but no confidence (**Trace+**/**Ratings+**/Confidence-), and (3) search and evidence (**Search+**/**Evidence+**). We select search and evidence because it is our most promising form of assistance (at least with T=.62). We select the other two because they helped the most when the overall rating was correct (though they also hurt the most when this rating was incorrect). It's thus possible that these forms might be more helpful at higher thresholds where a higher proportion of model overall ratings are correct. We can also explore whether hybridizing with multiple forms of assistance (e.g., search and evidence for low confidence, trace with ratings and confidence for medium confidence, and AI alone for high confidence) achieves an even higher accuracy than hybridizing with a single form of assistance.

A.3.4.1 Human (Rater) Experiments The experiments were identical in structure and content to the prior **Search+**/**Evidence+**, **Trace+**/**Ratings+**/**Confidence+**, and **Trace+**/**Ratings+**/Confidence- experiments except that all together participants rated the entire Evaluation Set (1918 examples). We also made the experiments available to the entire pool of participants to ensure that we could get a sufficient number of ratings per example on the entire set. This means that participants could take part in multiple experiments. We also ran a fresh baseline experiment (baseline_t3 for time 3) to match the conditions under which the intervention experiments were being launched, and also to account for performance increases due to prior experience with our task, as well as other teams' factuality tasks. We recruited for 800 rating hours per experiment, and participants had an opportunity to take part two times, spots permitting. Average number of ratings per example was 3.61 for each experiment (we clamped the maximum number of ratings per example to be the minimum across all "T3" experiments).

A.3.4.2 Results It appears that the human raters unassisted baseline performance did improve over time. In the new baseline experiment (baseline_t3), human raters achieved 82.7% accuracy on the entire Evaluation Set based on individual labels and 85.8% accuracy based on majority labels, whereas in the initial baseline experiment (baseline_t1) accuracy was 75.1% and 84.7%, respectively. Additionally, if we filter this data to the Post-Hybridized Human Set (T = .62), we see that baseline_t3 accuracy 74.1%) is also higher than baseline_t2 accuracy (67.3%). The baseline_t3 performance was still lower than the AI rater alone, which, as a reminder, achieves 87.7% accuracy on the Evaluation Set. However, it's important to keep in mind that the RQ3 experiments involve higher performing raters than RQ1 and RQ2.

In the current experiments, we again find evidence that confidence-based hybridization improves accuracy above human ratings or AI ratings alone, achieving complementary performance. Confidence-based hybridization with T=.62 and using unassisted (baseline_t3) ratings achieves 89.7% (individual) and 90.3% accuracy (majority). However, confidence-based hybridization with assistance did not lead to an additional improvement, likely because baseline performance was quite high. In fact, **Trace+/Ratings+/Confidence+** and **Trace+/Ratings+**/Confidence- assistance slightly decreased overall hybridized accuracy, achieving 88.9% (majority) and 89.3% (majority), respectively. **Search+/Evidence+** assistance did not make a difference, achieving 90.5% (majority).

Indeed, we find that on the Post-Hybridized Human Set, unlike RQ2, **Search+/Evidence+** does not increase accuracy above baseline and in some cases, **Trace+/Ratings+/Confidence+** and **Trace+/Ratings+**/Confidence- assistance actually leads to a statistically significantly worse accuracy than baseline. To test these relationships, we filter the data to the T=.62 Post-Hybridized Human Set and fit a mixed effects logistic regression predicting individual rating accuracy (0 or 1) from a fixed effect of label-type (5-level categorical variable: AI, unassisted baseline_t3 human, **Search+/Evidence+** assisted human, **Trace+/Ratings+/Confidence+** assisted human, and **Trace+/Ratings+**/Confidence-human, with unassisted baseline_t3 as the reference) and a random intercept by conversation. This analysis indicates that while **Search+/Evidence+** and **Trace+/Ratings+**/Confidence- are no different than baseline ($beta's$ between -0.208 and 0.112, $p's > .06$), **Trace+/Ratings+/Confidence+** is statistically significantly worse than baseline (74.1% vs. 70.0%; $\beta = -0.317, SE = 0.110, z = -2.876, p = .004$). Note, that all humans ratings (unassisted or assisted) are statistically significantly better than the AI ratings, which is again why we find that hybridization works.

But what happens if we explore different thresholds above T=.62? Figures 8 and 9 show how mean hybridized accuracy changes when using baseline and each form of assistance across different confidence thresholds. We find that (1) the optimal threshold does not vary much across the different forms of assistance, (2) it is never greater than .75, and (3) the optimal form of assistance does not change. For **Search+/Evidence+** assistance the optimal thresholds are the same as in RQ2, suggesting that these were not local optima: T=.62 for individual ratings (89.9%) and T=.7 (90.7%) for majority ratings. For **Trace+/Ratings+**/Confidence-, the optimal threshold is T=.68 for individual ratings (89.3%) and T=.7 for majority ratings (89.7%). For **Trace+/Ratings+/Confidence+**, the optimal threshold is T=.72 both for individual (89.0%) and majority (89.2%). Moreover, we do not find a case where assisted overall accuracy is appreciably better than baseline: We achieve the highest accuracy with T=.7 hybridizing using either majority vote baseline (90.4%) *or* majority vote **Search+/Evidence+**) assistance (90.7%).

It appears that with more skilled raters, the benefits of even our most promising assistant (**Search+/Evidence+**) are reduced. If we filter data to the optimal post-hybridized set for this form of assistance (T=.7), which is also optimal for **Trace+/Ratings+/Confidence+**, we find that **Search+/Evidence+** does not statistically significantly improve accuracy above baseline ($\beta = 0.084, SE = 0.183, z = 0.457, p = .647$). **Trace+/Ratings+**/Confidence- also has no effect ($\beta = -0.239, SE = 0.179, z = -1.336, p = .182$), while **Trace+/Ratings+/Confidence+** statistically significantly decreases accuracy below baseline majority-vote ratings on this set ($\beta = -0.404, SE = 0.177, z = -2.281, p = .023$).

In sum, we again find evidence that confidence-based hybridization improves accuracy above and beyond using AI ratings. Hybridizing with unassisted majority vote, or **Search+/Evidence+** assisted majority vote human ratings yields a 3% increase above using AI ratings alone. Unlike our RQ2 experiments, we do not find evidence that assistance further increases accuracy on top of the gains form hybridization, and interestingly, the most naive form of assistance that just shows human raters everything that the AI fact-verification assistant outputs

(**Trace+**/**Ratings+**/**Confidence+**), hurts performance. These findings suggest that the effectiveness of assistance may vary depending on rater skill, and that assistance tested on one type of rater may not transfer to other raters or even the same raters at a later time.

A.4. Additional Figures

Representative AI Assistance of the full model output

Experimental AI-generated fact-verification.
▹ ⚠ Be careful - this could be misleading

Factual claims in sentence, and summary of evidence:
▹ *(expand to learn more about the information below)*

Claim 1: Strawberries are a source of Vitamin C.
Summary of Evidence on claim:
Multiple sources confirm that strawberries are a source of Vitamin C [1, 3].
Predicted verdict for claim (could be incorrect): ✓
Accurate

Claim 2: The amount of Vitamin C in strawberries is a significant portion of the recommended daily value.
Summary of Evidence on claim:
Adults need 40 mg/day of Vitamin C [2]. 100g of strawberries provides 58.8mg of Vitamin C [3], so it is considered a good source of the nutrient.
Predicted verdict for claim (could be incorrect): ✓
Accurate

Predicted Overall Verdict (could be incorrect): ✓ Accurate
Model Confidence: high (95%)
▹ *(expand to understand "confidence")*

Selected Evidence:
▹ *(expand to learn more)*

1. Bioactive Compounds and Antioxidant Activity in Berries - NCBI

   > Amongst the fruits, fresh strawberries are considered to be one with the highest content of ascorbic acid. Among the berry species, strawberries have similar content to raspberries, but about four-times more ascorbate than blueberries. Ascorbate content in strawberries is highly variable, and in fresh strawberries generally ranges from 5 to 50 mg/100 g fresh weight (fw)

2. Vitamin C - - - Vitamins and minerals

   > How much vitamin C do I need? Adults aged 19 to 64 need 40mg of vitamin C a day. You should be able to get all the vitamin C you need from your daily diet.

3. Strawberries, raw - USDA FoodData Central

   > Nutrient name: Vitamin C, total ascorbic acid. Amount per 100g: 58.8 mg. Footnotes: Source: USDA Nutrient Data Laboratory. SR Legacy, 2018.

All Search Queries and Results:
(Expand a query to see the results and a quote from each webpage.)

- ▹ Search query: `"strawberries good source of vitamin C"`
- ▹ Search query: `"recommended daily value of vitamin c"`

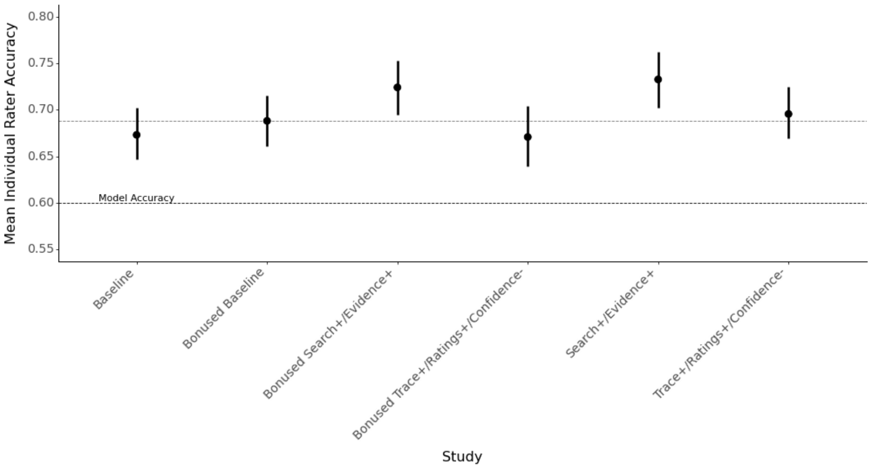

Figure 5 | Mean Individual Rater Accuracy for the bonus and select T2 studies described in Sections A.3.2 and 2.4.1. Restricted to examples where model confidence <= .62.

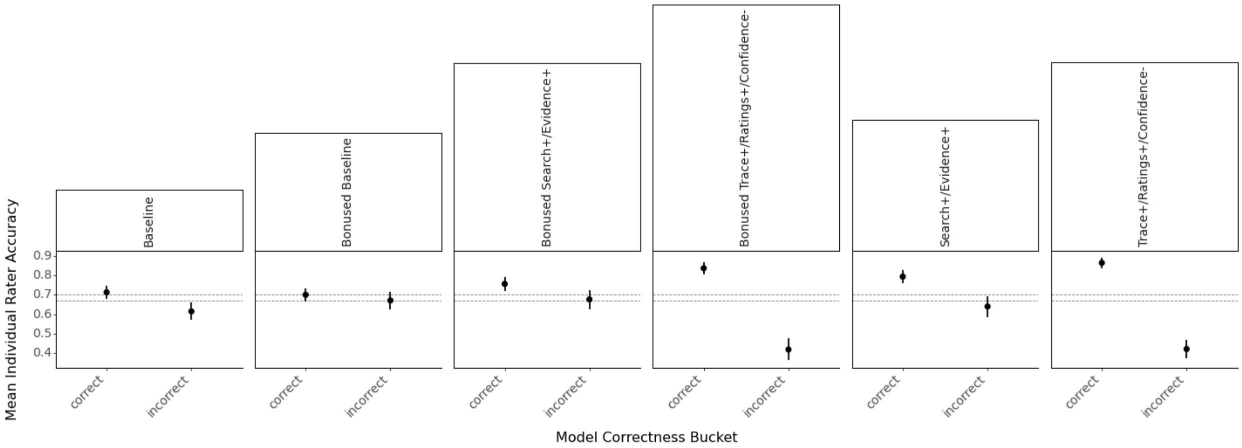

Figure 6 | Mean Individual Rater Accuracy for the bonus and select T2 studies described in Sections A.3.2 and 2.4.1, split by if the fact-verification assistant was correct. Restricted to examples where model confidence <= .62.

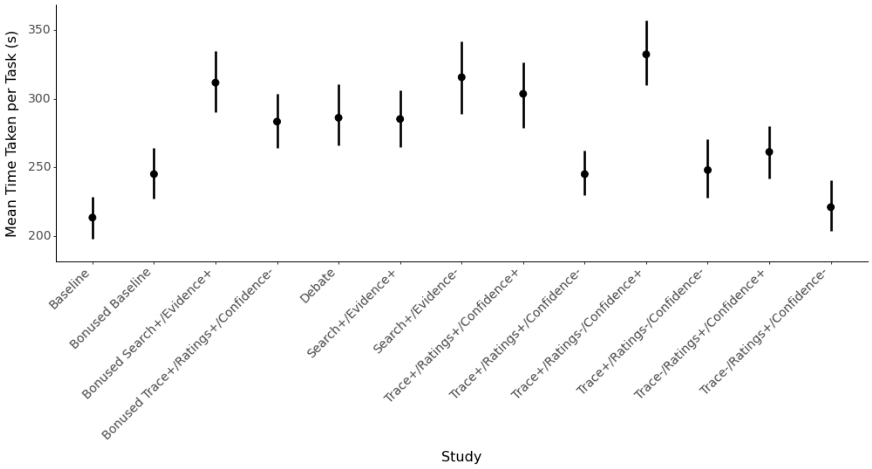

Figure 7 | Mean Time Taken per Task. Restricted to examples where model confidence <= .62. These were from "T2" studies, described in Section 2.4.1. In this plot and below, "Trace" refers to "Search+Evidence+Reasoning".

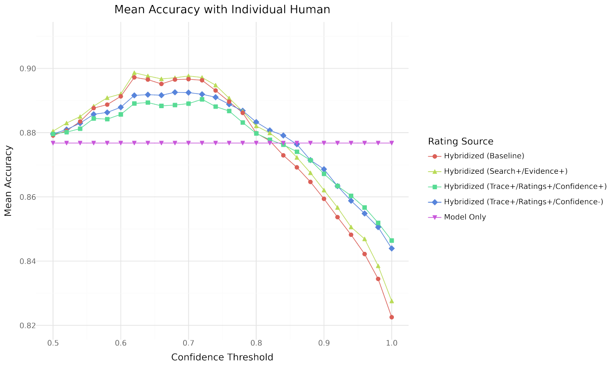

Figure 8 | Mean Rater Accuracy, using hybridization with Individual Human ratings, for "T3" studies, described in Section A.3.4.1.

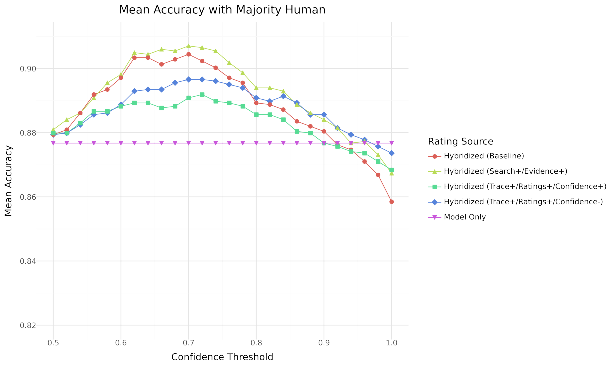

Figure 9 | Mean Rater Accuracy, using hybridization with Majority Human vote, for "T3" studies, described in Section A.3.4.1.

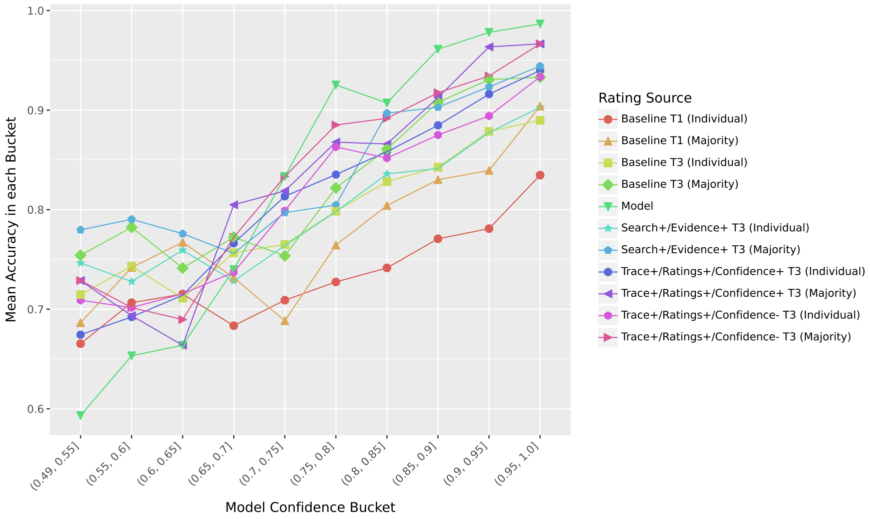

Figure 10 | Accuracy bucketed by confidence, for the model and different human ratings. The model seems to be calibrated. And at higher model confidences, although certain assistance methods do better than **Search+/Evidence+**, the model still does best. So, these other assistance methods do worse overall after confidence-based hybridization. Note: Majority vote accuracies between T1 (described in section 2.3.1) and T3 (section A.3.4.1) studies are not comparable, since the Baseline T1 study had an average of 5.5 ratings per example, and all T3 studies had an average of 3.61 ratings per example.

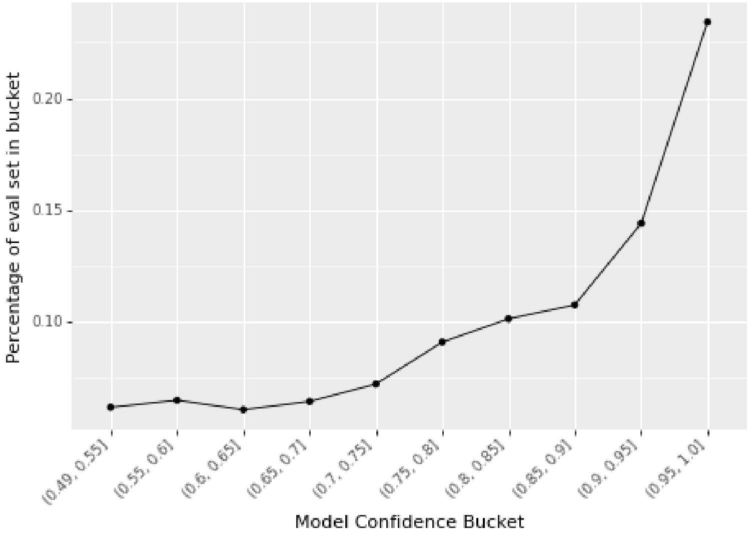

Figure 11 | Percentage of the evaluation dataset that is in each confidence bucket

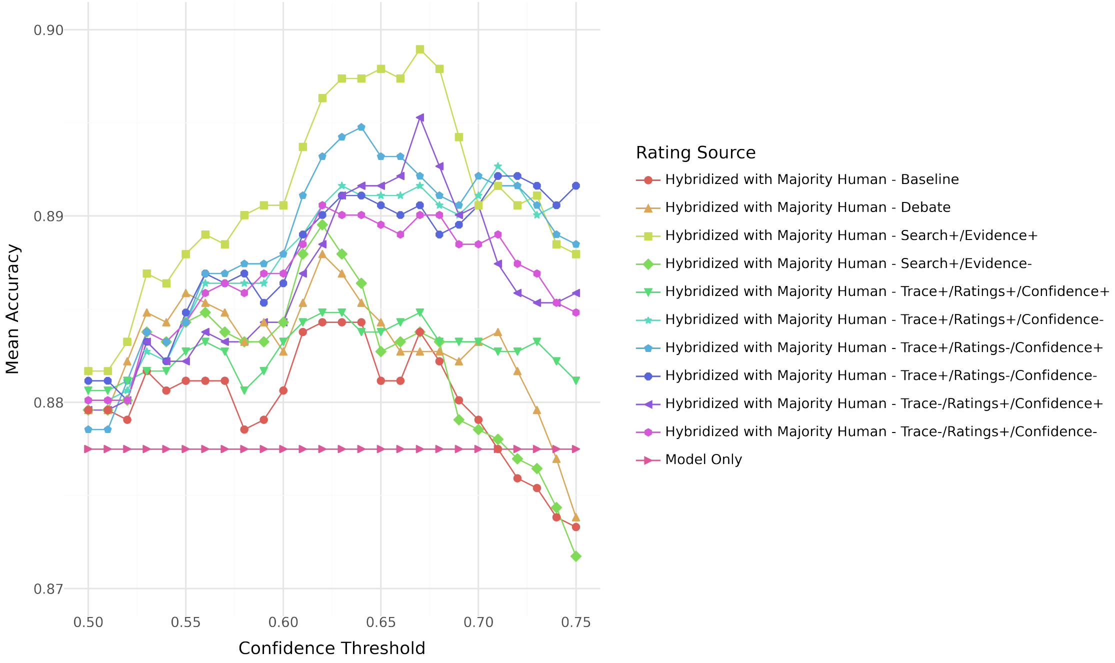

Figure 12 | Mean Rater Accuracy, using hybridization with Individual Human ratings for T2 studies (described in Section 2.4.1).