Large Language Models Help Humans Verify Truthfulness – Except When They Are Convincingly Wrong

CHENGLEI SI, Stanford University, USA
NAVITA GOYAL, University of Maryland, USA
SHERRY TONGSHUANG WU, Carnegie Mellon University, USA
CHEN ZHAO, NYU Shanghai, China
SHI FENG, New York University, USA
HAL DAUMÉ III, University of Maryland and Microsoft Research, USA
JORDAN BOYD-GRABER, University of Maryland, USA

Large Language Models (LLMs) are increasingly used for accessing information on the web. Their truthfulness and factuality are thus of great interest. To help users make the right decisions about the information they're getting, LLMs should not only provide but also help users fact-check information. In this paper, we conduct experiments with 80 crowdworkers in total to compare language models with search engines (information retrieval systems) at facilitating fact-checking by human users. We prompt LLMs to validate a given claim and provide corresponding explanations. Users reading LLM explanations are significantly more efficient than using search engines with similar accuracy. However, they tend to over-rely the LLMs when the explanation is wrong. To reduce over-reliance on LLMs, we ask LLMs to provide contrastive information—explain both why the claim is true and false, and then we present both sides of the explanation to users. This contrastive explanation mitigates users' over-reliance on LLMs, but cannot significantly outperform search engines. However, showing both search engine results and LLM explanations offers no complementary benefits as compared to search engines alone. Taken together, natural language explanations by LLMs may not be a reliable replacement for reading the retrieved passages yet, especially in high-stakes settings where over-relying on wrong AI explanations could lead to critical consequences.

ACM Reference Format:
Chenglei Si, Navita Goyal, Sherry Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III, and Jordan Boyd-Graber. 2023. Large Language Models Help Humans Verify Truthfulness – Except When They Are Convincingly Wrong. 1, 1 (October 2023), 18 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Imagine you are told a claim about Neptune: *"Only one spacecraft has visited Neptune and it has more than 13 moons."* and you want to verify whether it is factual. What would you do—looking up relevant pages from search engines or asking ChatGPT for its take? This is not just a question of checking a piece of trivia; our information ecosystem depends on people being able to check the veracity of information online. Misinformation, whether accidental or deliberate, has

Authors' addresses: Chenglei Si, clsi@stanford.edu, Stanford University, USA; Navita Goyal, University of Maryland, USA, navita@umd.edu; Sherry Tongshuang Wu, Carnegie Mellon University, USA, sherryw@cs.cmu.edu; Chen Zhao, NYU Shanghai, China, cz1285@nyu.edu; Shi Feng, New York University, USA, shifeng@nyu.edu; Hal Daumé III, University of Maryland and Microsoft Research, USA, me@hal3.name; Jordan Boyd-Graber, University of Maryland, USA, jbg@umiacs.umd.edu.

Manuscript submitted to ACM

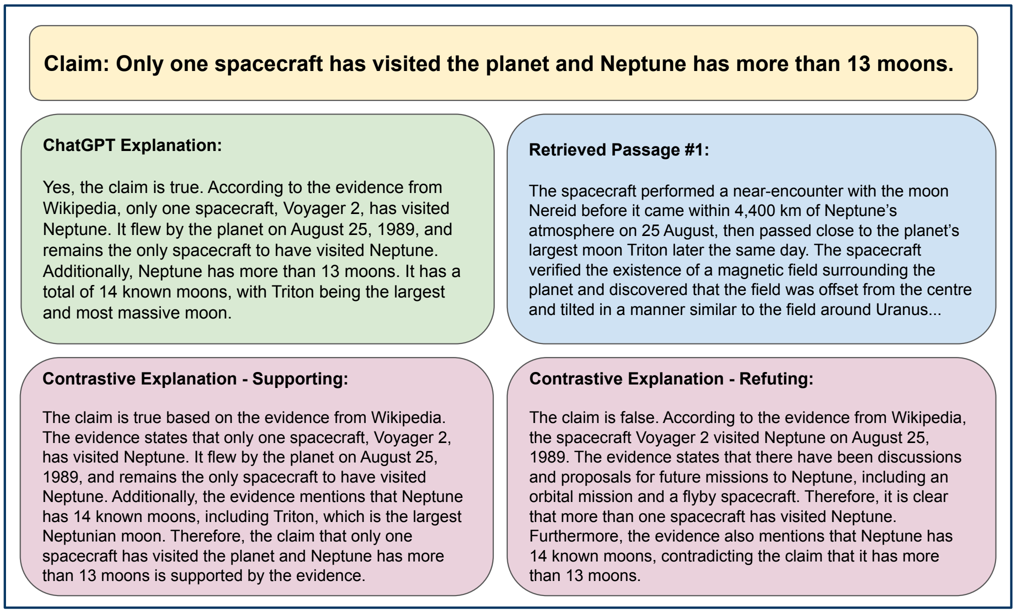


Fig. 1. An example of the claim to be verified and the corresponding ChatGPT explanation, retrieved passages (abridged), and contrastive explanation. Note that the claim is true and the refuting explanation contains factual errors and reasoning contradiction.

the potential to sway public opinion, influence decisions, and erode trust in credible sources [10, 24]. Moreover, the wide adoption of large language models like ChatGPT increases the danger of misinformation, both by malicious actors and models generating inadvertent hallucinations [28].

Consequently, the task of verifying the factual accuracy of information has taken on great importance. Fact-checking claims is a well-established task in the field of natural language processing (NLP) [16, 39]. However, automated fact-checkers are far from perfectly accurate and reliable, and they are only useful when users trust their predictions [26]. Building that trust and providing effective help is crucial: a team without trust leads to suboptimal human-AI team performance, while over-trusting wrong AI predictions could lead to catastrophic failures in high-stakes applications. Therefore, in real-life applications, we care about the AI-assisted human accuracy on fact-checking, rather than evaluating and improving automated fact-checkers alone [35].

The two major types of tools for helping human users (many of which are non-expert fact-checkers) are *retrieval* and *explanation* [26], exemplified by the widely-used web search engines (*e.g.*, Google) and generative language models (*e.g.*, ChatGPT) respectively. Showing retrieved passages to users has long been established as an effective information-seeking tool [25, 42]. In contrast, the usefulness of generative explanations on fact-checking remains understudied. On the one hand, competent generative models (especially LLMs) can generate fluent and convincing-looking natural language explanations that not only provide an answer (*i.e.*, whether the claim is true or false), but also elucidate the context and basis of its judgment. On the other hand, these models are prone to hallucinations [23, 25], so the users are frequently left to their own devices.

In this work, we conduct human-subject experiments to study *whether language models can assist fact-checking*. To contextualize the effectiveness of explanations against search engines, we compare them with retrieval models mimicking a search engine experience, and experiment with ways where retrieved passages can be paired with explanations, aiming to provide a practical guide to users on what is the most helpful tool. We base our evaluation on the FoolMeTwice claim verification dataset (described in detail in Section 4.1), where the claims are crowdsourced and gold evidence from Wikipedia is collected [8]. We ask participants to verify whether the claim is factually true or false, and Figure 1 shows an example to illustrate the explanation and retrieved passages that we present to participants in the study.

Our study reveals that showing explanation and retrieved passages lead to similar human accuracy (74% and 73% respectively) on difficult-to-verify claims (59% without AI assistant), but reading natural language explanations is significantly faster (1.01 min/claim vs 2.53 min/claim). However, humans over-trust ChatGPT explanations where they agree with the explanation most of the time, even when the explanation gives a wrong answer.

To combat the issue of over-reliance on natural language explanations, we explore two improvements: 1) contrastive explanations—present both supporting and refuting arguments generated by ChatGPT to the user, and 2) combining retrieval and explanation. Both methods significantly reduce over-reliance on wrong AI explanations, however, they do not show a significant gain in user fact-checking accuracy compared to just showing users the retrieved passages. Overall, our work underscores the potential benefit and danger of natural language explanations as a tool in the battle against misinformation. We reveal their efficiency in saving time, but at the same time the difficulty of combatting over-reliance and the redundancy when combining retrieval and explanation. Turning back to the question of what should users do to verify factuality: taking a longer time to read the retrieved passages is still the more reliable way!

2 RELATED WORK

2.1 Fact Checking

Fact checking is a well-established task in natural language processing where the typical task format is to input evidence text (*e.g.*, retrieved from Wikipedia) and the claim to the model and output a label of support or refute (or sometimes a third class of not enough information) [38, 42]. Abundant datasets have been collected for training and evaluating automatic fact-checking models, such as FEVER [16, 34, 39], SciFact [43]. Various techniques have been proposed to improve the fact-checking pipeline, such as locating the most relevant evidence snippets from long retrieved documents [22], breaking complex claims into atomic sub-claims [18, 25], or synthesizing summaries for retrieved passages [9]. Unlike this line of work, which primarily concerns with developing and improving automated fact-checking systems, our work aims to study how to help humans check facts with the aid of large language models.

2.2 AI Explanation

Explanations have been long sought as a useful tool to help users, not only in understanding AI predictions [21], but also aiding them in calibrating their reliance on these predictions [5]. A thread of work in explainable AI (XAI) attempts to generate useful explanations in various formats [45], such as highlighting [33] , feature importance [32], free-text rationales [7], and structured explanations [20]. As the end goal of explanations is to aid humans in decision-making tasks, several work in XAI literature has focused on human-centered evaluation of explanations [30]. This line of work has reported mixed results on the utility of AI explanations—some work find that explanations can support human-AI decision-making by exceeding both human-alone or AI-alone performance [3, 11], whereas some other work find that explanations lead to worse human-AI performance [1, 2, 44].

Fok and Weld [13] argue that to facilitate complementary human-AI decision-making, explanations must aid users in verifying the AI prediction to yield truly complementary human-AI performance. Explanations targeting verifiability have indeed shown promising avenues in human-AI collaborations [11, 15, 41]. Closest to our work, Feng and Boyd-Graber [11] evaluated human-AI collaborative Quizbowl question answering and compared the effectiveness of showing retrieved passages, highlighting, and showing multiple guesses made by the system. This previous work used only a retrieval component, while our new approach allows us to directly compare ChatGPT generated explanations (in the form of free-text rationales) with retrieved passages for aiding claim verification and explore whether natural language explanations and retrieved evidence yield complementary benefits. Joshi et al. [17] study free-text explanations in question-answering setting: their rationales do not help users much, especially when the rationales are misleading. In contrast to their work, we *contrast* model-generated explanations with passages retrieved from external sources (Wikipedia).

2.3 Trust Calibration and Over-Reliance

Existing work has identified the issue of human over-reliance on AI predictions, where humans tend to trust AI predictions *even when they are wrong* [6, 19]. A growing line of work attempts to mitigate such over-reliance, for example by providing explanations [2, 41, 47], communicating model uncertainty [31, 37], showing AI model accuracy [46], and prompting slow thinking [4] to help users calibrate their trust. Our work also contributes to this line of work by revealing the over-reliance issue in the fact-checking application. We propose new ways of potentially combatting over-reliance including contrastive explanation and combining explanation with retrieval.

3 RESEARCH QUESTIONS

To understand the comparative advantages of retrieval and explanation in human fact verification, we pose the following research questions:

- **RQ1**: Are natural language explanations more effective than retrieved passages for human fact-checking?
- **RQ2**: Can contrastive explanations—arguing for or against a fact being true—mitigate over-reliance and be more effective than non-contrastive explanations?
- **RQ3**: Are there complementary benefits in presenting both natural language explanations and retrieved passages?

We investigate these questions through a series of human studies: we show participants claims that need to be verified, potentially aiding them with different pieces of evidence. This is a between-subjects study; thus, we vary the evidence presented to participants in different conditions:

- **Baseline**: We show users only the claims without any additional evidence.
- **Retrieval**: We show the top 10 paragraphs retrieved from Wikipedia along with the claim to be verified.
- **Explanation**: We show the ChatGPT explanation along with the claim.
- **Contrastive Explanation**: We present users ChatGPT's supporting and refuting arguments side by side, as illustrated in Figure 2.
- **Retrieval + Explanation**: We present both the retrieved passages as well as the (non-contrastive) natural language explanations to users.

In the `Explanation` and `Retrieval + Explanation` conditions, the ChatGPT prediction on whether the claim is true or false is part of the explanation, and users need to decide whether to trust the AI prediction. In the other conditions, users only see the evidence but not a direct AI prediction.

4 STUDY DESIGN OVERVIEW

4.1 Task and Data

We ask human annotators to look at claims and decide whether it is true or false. For this purpose, we use the FoolMeTwice dataset [8], where all claims are written by human crowdworkers. We pick this dataset over other claim-verification datasets because FoolMeTwice is adversarial: the author writes claims based on Wikipedia to maximally fool another set of annotators whose task is to verify these claims. This ensures that all the claims are challenging and hard to verify, mimicking potential real-world fake news arms race. Each claim in the dataset comes with a set of gold evidence, used by the claim writers and is guaranteed to provide sufficient evidence for verification.

For our human studies, we obtain a test set by randomly sampling 200 claims where half are true and half are false. To ensure that the claims we select are sufficiently comples, we only sample claims with at least two different sentences from Wikipedia. The sampled claims come from diverse domains including Geography, Literature, Science, History, Popular Culture, Technology, Music, and Sports.

4.2 Measured Variables

For each claim, we ask for the participant's binary decision of whether they think the claim is true or false. We measure the accuracy of human decisions given that we know the gold labels of these claims. We also ask for the participant's confidence in their judgment on a scale of 1 to 5, and record the time used for verifying each claim. Apart from these quantitative variables, we ask for a free-form response of how the annotator makes their judgments.

4.3 Retriever

For the `Retrieval` and `Retrieval + Explanation` conditions, we show users the most relevant passages from credible sources to aid their verification of the claims. For this purpose, we use a passage retriever to retrieve paragraphs from Wikipedia. Specifically, we adopt a similar retrieval setup as [25], where we use the state-of-the-art Generalizable T5-based Retriever (GTR-XXL), an unsupervised dense passage retriever [27]. We retrieve the top 10 most relevant paragraphs from Wikipedia, where each paragraph has an average length of 188 words. To measure the quality, we report two metrics on our test set. The full recall measures how often the top 10 retrieved passages contain all evidence sentences required to verify the claim, which is 81.5%; the partial recall measures how often the top 10 retrieved passages contain at least one evidence sentence required to verify the claim, which is 93.0%.

4.4 Explanation Generation

We study two types of natural language explanations with ChatGPT: non-contrastive explanation and contrastive explanation. In the `Explanation` and `Retrieval + Explanation` conditions, we generate **non-contrastive** explanations, where we construct the prompt by concatenating the top 10 retrieved passages, followed by the claim to be verified, then appending the question *"Based on the evidence from Wikipedia, is the claim true? Explain in a short paragraph."* We measure the accuracy of these explanations by manually extracting the answer (true or false) from the explanations and comparing with the gold labels. ChatGPT-generated explanations achieve an accuracy of 78.0% (judged based on the AI predictions only, not the reasoning processes).

In the `Contrastive Explanation` condition, we experiment with **contrastive** explanations where we prompt ChatGPT to generate both a supporting answer and a refuting answer. Specifically, after concatenating the retrieved passages and the claim, we append two different questions: 1) *"Based on the evidence from Wikipedia, explain in a*

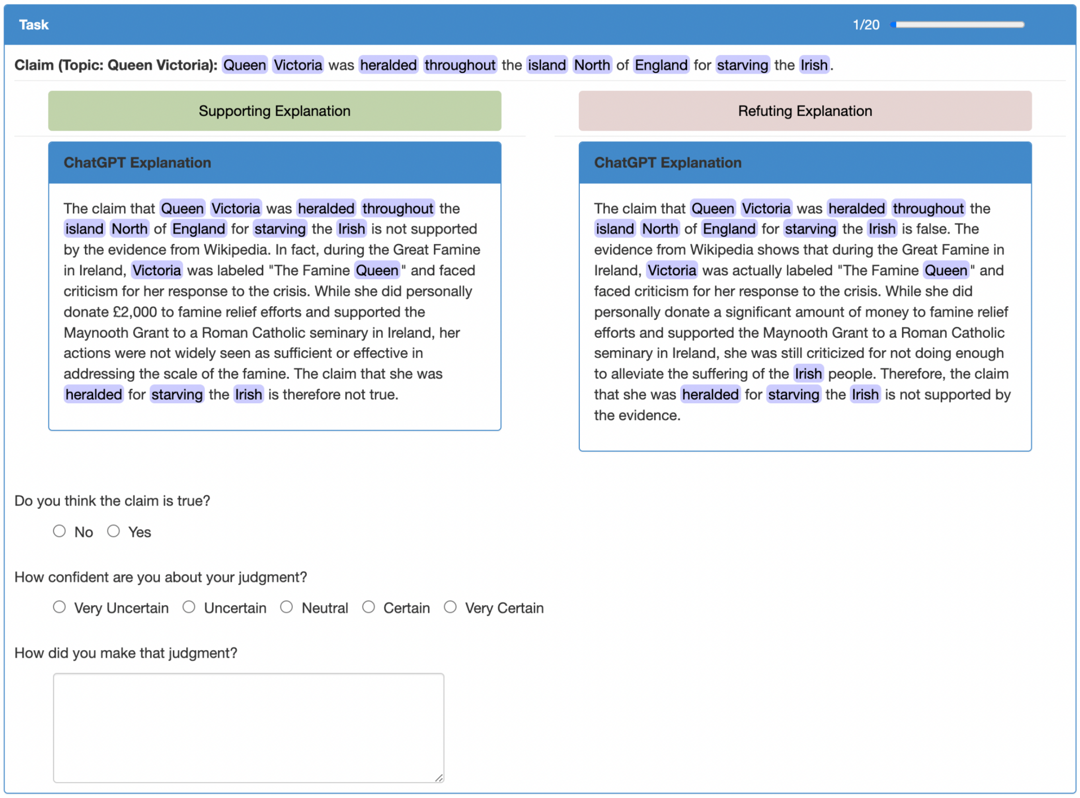


Fig. 2. Interface for the contrastive explanation condition. We present ChatGPT's explanations for both sides together to encourage more careful thinking. We also highlight all the keywords to ease reading.

short paragraph why the claim is ***true****."* and 2) *"Based on the evidence from Wikipedia, explain in a short paragraph why the claim is* ***false****."* We then show both of these generated explanations to annotators, which functions similarly to a single-turn debate [29].

Additionally, in the `Retrieval + Explanation`, we automatically insert citations to the explanation text to attribute the arguments to corresponding retrieved passages. This is implemented by prompting ChatGPT where we provide a manually crafted example of inserting citations into the explanations based on the retrieved passages, which has been shown to be an effective method for enabling citations in language model generations [14].

For all cases, we ground the explanation generation on the retrieved passages. This is because grounding significantly improves the accuracy of explanations. For example, for non-contrastive explanations, grounding improves the accuracy from 59.5% to 78.0%. For all cases, we use a temperature value of 0 for ChatGPT generation to minimize randomness in language generation.

4.5 Interface Design

Figure 2 shows an example user interface for the `Contrastive Explanation` condition. We identify keywords as the non-stopwords in the claim and highlight them in the claims and explanations to aid reading (we also do keyword highlighting in the retrieved passages in the retrieval conditions). For the retrieved paragraphs, we rank them by relevance and only show the first paragraph in full by default and annotators can click to expand the other paragraphs.

In the task instructions, we explicitly discourage participants from searching the claims on the internet. Each participant verifies 20 claims one by one. We provide a tutorial at the beginning of the study. We include two attention check questions at different points in study asking participants' selection from the most recent claim, and reject the responses from user who fail both attention checks.

4.6 Users

We recruit participants from Prolific[1] crowd-sourcing platform for our task. We recruit 16 annotators for each condition and each annotator verifies 20 claims, resulting in $20 \times 16 \times 5 = 1500$ annotations across all 5 conditions we have. We compensate all annotators at least $14 per hour, as well as additional bonuses to users who perform particularly well on the task or who have left very insightful comments as an additional incentive. Our study is approved by the University of Maryland Institutional Review Board (IRB).

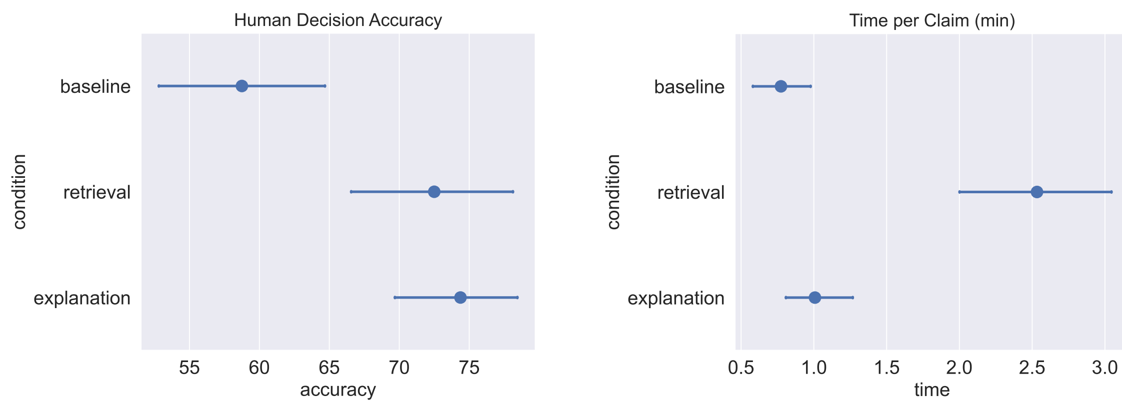


Fig. 3. Human decision accuracy and average time spent on verifying a claim (showing average and 95% confidence interval – same for all plots in this paper). Both retrieval and explanation significantly improve human verification accuracy, while explanation takes a significantly shorter time than retrieval.

5 RESEARCH QUESTION 1: ARE NATURAL LANGUAGE EXPLANATIONS MORE EFFECTIVE THAN RETRIEVED PASSAGES FOR HUMAN FACT CHECKING?

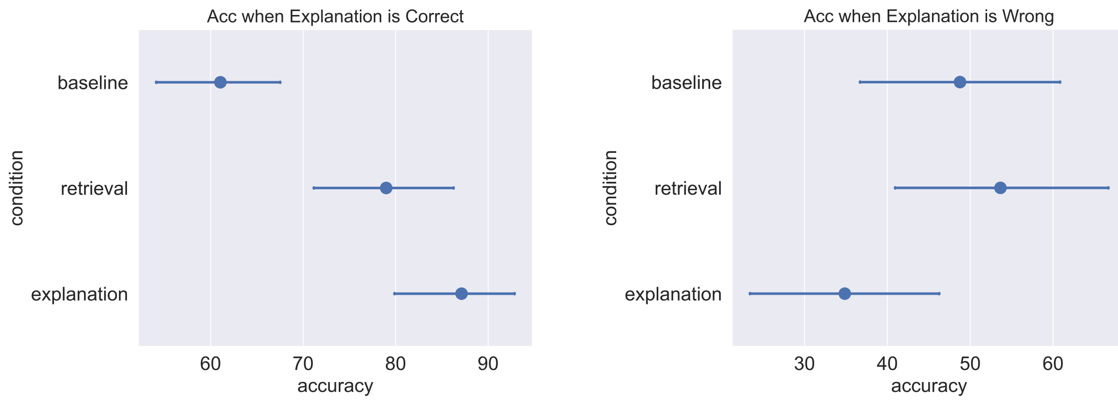
Our first research question is whether showing natural language explanation generated by ChatGPT improves human verification better than showing retrieved passages.

5.1 Hypotheses

We formulate the following hypotheses:

- **H1**: Showing explanations enables users to achieve higher accuracy than not showing any evidence.
- **H2**: Showing explanations enables users to achieve higher accuracy than showing retrieved passages.

[1] https://www.prolific.com/



(a) Human decision accuracy on examples where the explanation is correct.

(b) Human decision accuracy on examples where the explanation is wrong.

Fig. 4. Human verification accuracy broken down into two subsets: examples on which the explanation gives the correct labels, and examples on which the explanation gives the wrong labels. Humans over-rely on explanations so that they achieve significantly lower accuracy than the baseline in the explanation condition when the explanation is wrong.

- **H3**: Showing explanations allows faster decision time than showing retrieved passages.

To verify our hypotheses, we compare across three conditions: the `Baseline` condition (showing users only the claims); the `Retrieval` condition (showing the top 10 paragraphs retrieved from Wikipedia); and the `Explanation` condition (showing the ChatGPT explanation along with the claim). We do not set any time limit but record the time taken to verify each claim.

5.2 Results

Figure 3 (left) shows the AI-assisted human verification accuracy across conditions. We test the significance of our results using Student's t-tests with Bonferroni correction. We start with examining whether ChatGPT explanations and retrieved passages are indeed helpful for humans.

Showing ChatGPT explanation improves human accuracy. When showing explanations to users, they achieve the accuracy of ($\mu = 0.74 \pm \alpha = 0.09$) as compared to the baseline condition where claims are shown without any additional evidence (0.59 ± 0.12). The improvement in accuracy is significant ($z = -4.08, p = 0.00015$), therefore supporting H1.

Showing retrieved passages improves human accuracy. When showing retrieved passages to users, they achieve the accuracy of ($\mu = 0.73 \pm \alpha = 0.12$) as compared to the baseline condition where claims are shown without any additional evidence (0.59 ± 0.12). The improvement in accuracy is significant ($z = -3.15, p = 0.0018$). Now that both ChatGPT explanation and retrieved passages help humans more accurately verify claims, we examine their comparative advantages in both accuracy and time.

Showing ChatGPT explanation does not achieve significantly higher accuracy than showing retrieved passages. Comparing the accuracy in the explanation condition ($\mu = 0.74 \pm \alpha = 0.09$) and the retrieval condition ($\mu = 0.73 \pm \alpha = 0.12$), the improvement in accuracy is not significant ($z = -0.48, p = 0.32$), therefore we could not reject the null hypothesis for H2.

However, **reading ChatGPT explanation is significantly faster than reading retrieved passages.** We compare the time taken to verify claims in Figure 3 (right). When verifying with retrieved passages, the time taken to verify each claim is (2.53 ± 1.07) minutes while for the explanation condition, it takes (1.01 ± 0.45) minutes. This supports H3: showing explanations allows significantly faster decision time than showing retrieved passages ($z = -5.09, p = 9.1e-6$).

5.3 Breakdown Analysis: The Danger of Over-Reliance

While ChatGPT explanations shows promise in aiding human fact verification, the aggregate results obscures the danger when the explanation gives wrong answers. Thus, to examine what happens in those cases, we perform break down the analysis, manually annotating the ChatGPT explanation for each claim based on whether it gives the correct answer (whether the claim is true or false). We then split all user responses into two subsets: ones with correct answers from ChatGPT and ones where the ChatGPT explanation is wrong (Figure 4a and Figure 4b, respectively).

Users achieve the highest accuracy when the explanations are correct, but below-random accuracy when the explanations are wrong. When the explanation is correct, users' accuracy is (0.87 ± 0.13), higher than the baseline of having no evidence (0.61 ± 0.13) as well as the retrieval condition (0.79 ± 0.15). However, when the explanation is wrong, users tend to over-trust the explanations and only achieve an accuracy of (0.35 ± 0.22) as compared to the baseline condition (0.49 ± 0.24) and the retrieval condition (0.54 ± 0.26). Moreover, **users spend similar time on claims with correct and wrong explanations**, further indicating that they are not deliberately differentiating correct and wrong explanations and instead tend to trust most of the explanations. We also look at the free-form responses from users for their decision rationales, the most common responses include: (1) ChatGPT's explanation looks convincing, especially with quotes from the retrieved passages (even when the quotes or reasoning are wrong); (2) They do not have any prior knowledge on the topic so would just trust ChatGPT.

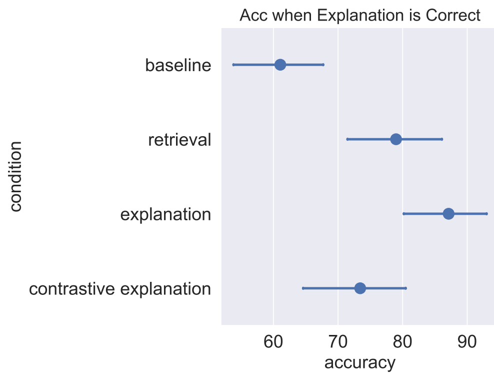
In comparison, retrieved passages suffer less from over-reliance. On examples where the ChatGPT explanations are correct, the retrieval condition achieves the accuracy of (0.87 ± 0.13) as compared to the baseline condition (0.79 ± 0.15). On examples where the ChatGPT explanations are wrong, the retrieval condition achieves the accuracy of (0.54 ± 0.26) compared to the baseline (0.49 ± 0.24). While there is still an accuracy drop on these examples, possibly because they are harder to verify, the performance discrepancy between the two cases (ChatGPT explanation correct vs wrong) is much less severe in the retrieval condition. This highlights the pitfall of using ChatGPT explanation to aid helpful verification: users over-rely on the explanations, even when they are wrong and misleading. To combat this problem, we next explore two strategies for mitigation: contrastive explanation and combining retrieval and explanation.

6 RESEARCH QUESTION 2: CAN CONTRASTIVE EXPLANATIONS MITIGATE OVER-RELIANCE AND BE MORE EFFECTIVE THAN NON-CONTRASTIVE EXPLANATIONS?

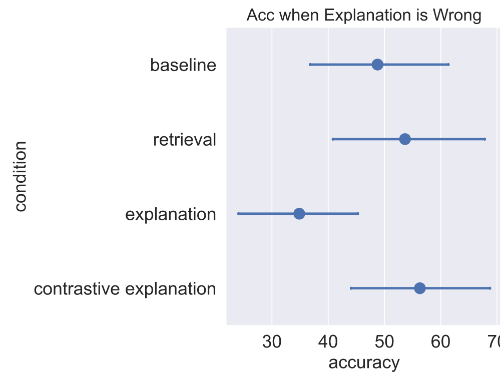
In this section, we study whether contrastive explanation can mitigate users' over-reliance, especially on wrong explanations. At the same time, since we are not directly providing the answer but rather letting users compare the two sides and judge for themselves which explanation they prefer to trust (or neither), we hypothesize that this could lead to a drop in accuracy on cases where the original (non-contrastive) explanation is correct. This is similar to prior work on human judging single-turn debates between language models [29].

6.1 Hypotheses

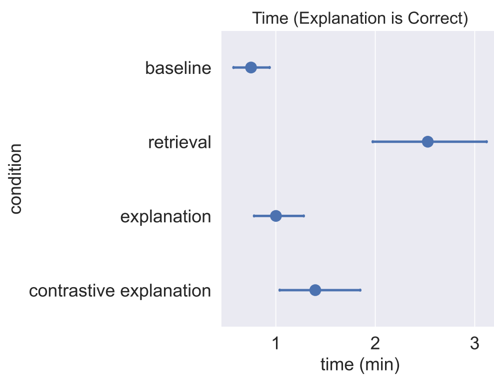
Concretely, we aim to test the following hypotheses:



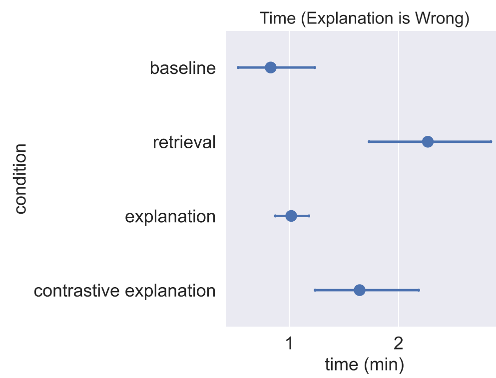
(a) Human decision accuracy on examples where the explanation is correct.



(b) Human decision accuracy on examples where the explanation is wrong.



(c) Human decision time on examples where the explanation is correct.



(d) Human decision time on examples where the explanation is wrong.

Fig. 5. Verification accuracy and time breakdown. Contrastive explanation achieves significantly higher accuracy than non-contrastive explanation on examples where the non-contrastive explanation is wrong, with some drop on examples where the non-contrastive explanation is correct.

- **H4**: Showing contrastive explanation enables users to achieve higher accuracy than non-contrastive explanation on cases where the non-contrastive explanation is wrong.
- **H5**: Showing contrastive explanation lowers human verification accuracy on cases where the non-contrastive explanation is correct, compared to showing non-contrastive explanation.

Apart from the three conditions from the previous section (`Baseline`, `Retrieval`, and `Explanation`), we additionally compare with the `Contrastive Explanation` condition where we present users ChatGPT's supporting and refuting arguments side by side, as illustrated in Figure 2 and Figure 1.

6.2 Results

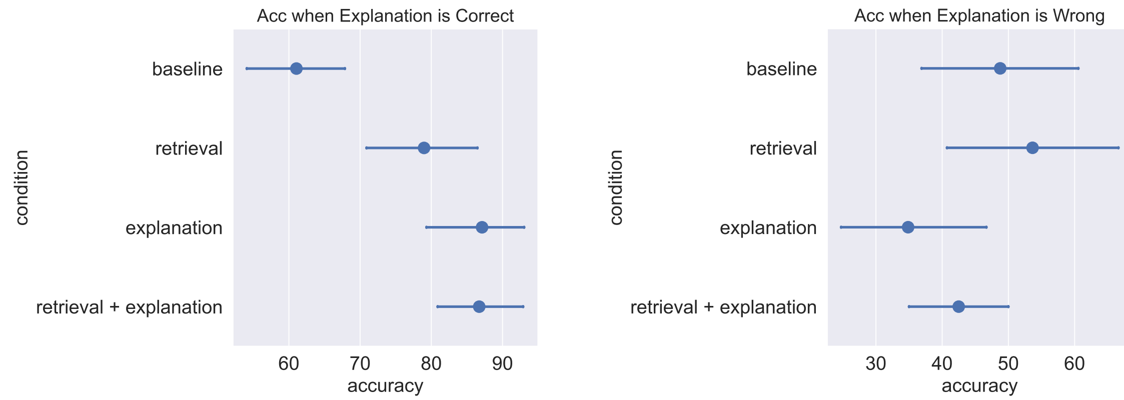
The experiment results are in Figure 5. We first compare contrastive explanation with non-contrastive explanation.

Contrastive explanation achieves higher human accuracy than non-contrastive explanation when the non-contrastive explanation is wrong. When the non-contrastive explanation is wrong, humans only achieve an accuracy of (0.35 ± 0.23) due to over-reliance, but when switching to contrastive explanation, humans achieve an accuracy of (0.56 ± 0.24), which is significantly higher ($z = -2.52, p = 0.009$). Therefore, H4 is supported. When analyzing the free-response rationales of human judgment, the most common patterns of how people make correct judgments based on contrastive explanations are: (1) The correct side of the explanation is more compelling or thorough; (2) The wrong side of the explanation contains factual errors and wrong reasoning; (3) Both sides of the explanations give the same answer (even though ChatGPT was prompted to explain why the claim is true and false in the two sides of explanations).

However, **contrastive explanation achieves lower human accuracy than non-contrastive explanation when the non-contrastive explanation is correct.** When the non-contrastive explanation is correct, humans achieve an accuracy of (0.87 ± 0.13) as compared to contrastive explanation (0.73 ± 0.15), indicating a significant drop ($z = -2.56, p = 0.008$), which supports H5. Unlike the case in non-contrastive explanations where users can just take the AI prediction as the answer, for contrastive explanations they have to decide between the two sides of the explanation and choose one, which can sometimes be difficult since LLMs can generate convincing explanations even for the wrong statements. For example, given the false claim *"Joe Torre was the manager of the New York Yankees and guided the team to four World Series championships, and ranks third all-time in MLB history with 2,326 wins as a manager."*, ChatGPT generates the supporting explanation *"Yes, the claim is true. According to the evidence from Wikipedia, Joe Torre was the manager of the New York Yankees from 1996 to 2007. He also ranks third all-time in MLB history with 2,326 wins as a manager."* and generates the refuting explanation *"The claim is false. According to the evidence from Wikipedia, Joe Torre was the manager of the New York Yankees and guided the team to six pennants and four World Series championships. He ranks fifth all-time in MLB history with 2,326 wins as a manager, not third."* The fact is that Torre ranks fifth all-time in MLB history with 2,326 wins as a manager but ChatGPT still generated a convincingly looking explanation for the wrong side by hallucinating he ranks third all-time rather than fifth. As a result, some users were misled into the wrong judgment. Overall, contrastive explanation shows promise in reducing over-reliance, but incurs a trade-off in accuracy when the non-contrastive explanation is correct. Next, we also compare contrastive explanation with retrieval.

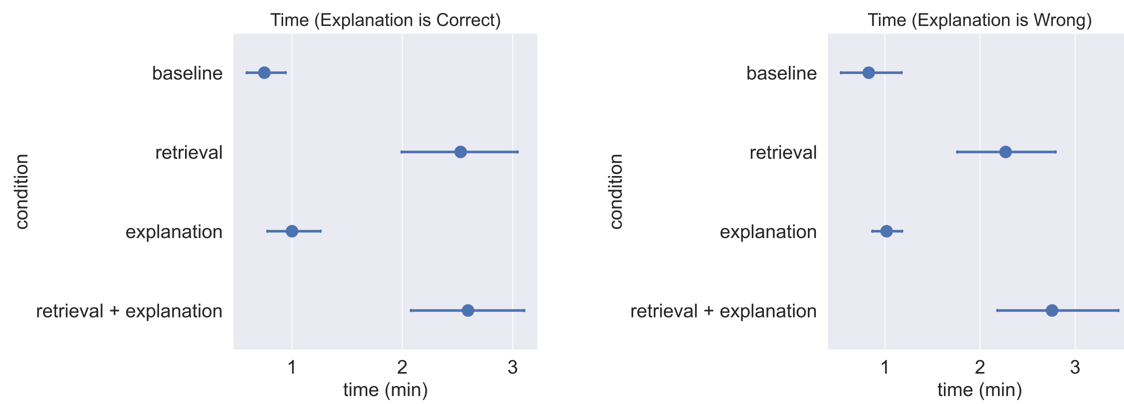
Contrastive explanation does not achieve significantly better human accuracy than retrieval. On examples where the non-contrastive explanation is correct, providing contrastive explanation achieves the human accuracy of (0.73 ± 0.15), lower than the accuracy in the retrieval condition (0.79 ± 0.15). On examples where the non-contrastive explanation is wrong, providing contrastive explanation achieves the human accuracy of (0.56 ± 0.24) as compared to the retrieval condition (0.54 ± 0.26), and the difference is not significant ($z = 0.29, p = 0.61$). Therefore, we conclude that in both cases contrastive explanations do not achieve significantly better human accuracy than retrieval, despite the results that contrastive explanations can mitigate over-reliance as compared to non-contrastive explanations.

Apart from the above quantitative results, we also manually analyze the free-form responses of user decision rationales to understand how users leverage contrastive explanations to make decisions. **Users mostly base their judgment on the relative strength of the two sides of the explanations** (*i.e.*, is the supporting or refuting explanation more convincing) (41.8%). Example user rationales include: *"The refutation seems more logically sound."* and *"The support explanation seems like it's trying too hard to make the claim true, but the refute puts it more plain and simple and makes more sense."* **Sometimes both sides converge on the same answer** (26.9%) and users would just agree with that. For example, for the false claim *"The only verified original sled prop from Citizen Kane was sold at a price of over a hundred thousand dollars."*, users found that *"Both sides acknowledge that there were more than 1 sled prop, therefore refuting the* *claim.”*, even though the ChatGPT supporting explanation said *“The claim is true.”* In several cases, ChatGPT would simply say the claim is true even though we prompt it for a refuting explanation (and vice versa), giving users a clear cue that the model could not make a strong argument for the wrong side.



(a) Human decision accuracy on examples where the explanation is correct.

(b) Human decision accuracy on examples where the explanation is wrong.



(c) Human decision time on examples where the explanation is correct.

(d) Human decision time on examples where the explanation is wrong.

Fig. 6. Verification accuracy and time breakdown. Combining retrieval and explanation is not significantly better than just showing retrieved passages alone.

7 RESEARCH QUESTION 3: ARE COMPLEMENTARY BENEFITS IN PRESENTING BOTH NATURAL LANGUAGE EXPLANATIONS AND RETRIEVED PASSAGES?

In previous experiments, we showed retrieved paragraphs or natural language explanations separately to users. In this section, we study whether there are complementary benefits in showing both retrieval and explanation at the same time. Again, we analyze separately the examples on which the (non-contrastive) explanation is correct versus wrong.

7.1 Hypotheses

We aim to verify the following hypotheses:

- **H6**: Combining retrieval and explanation gives higher accuracy than explanation alone on cases where the explanation is correct.
- **H7**: Combining retrieval and explanation gives higher accuracy than explanation alone on cases where the explanation is wrong.
- **H8**: Combining retrieval and explanation gives higher accuracy than retrieval alone on cases where the explanation is correct.
- **H9**: Combining retrieval and explanation gives higher accuracy than retrieval alone on cases where the explanation is wrong.

Apart from the `Baseline`, `Retrieval`, and `Explanation` conditions from earlier, we also compare with the (`Retrieval + Explanation`) condition where we present both to users.

7.2 Results

Results are plotted in Figure 6 and we start by comparing whether combining explanation with retrieval is better than explanation alone.

Combining retrieval and explanation does not achieve significantly higher accuracy than explanation alone in cases where the explanation is correct. When the explanation is correct, users achieve the accuracy of (0.87 ± 0.13) relying on explanations, as compared to combining both retrieval and explanation (0.87 ± 0.12). We do not observe a significant advantage of combining retrieval and explanation in this case ($z = 0.084, p = 0.53$), and so we cannot reject the null hypothesis for H6.

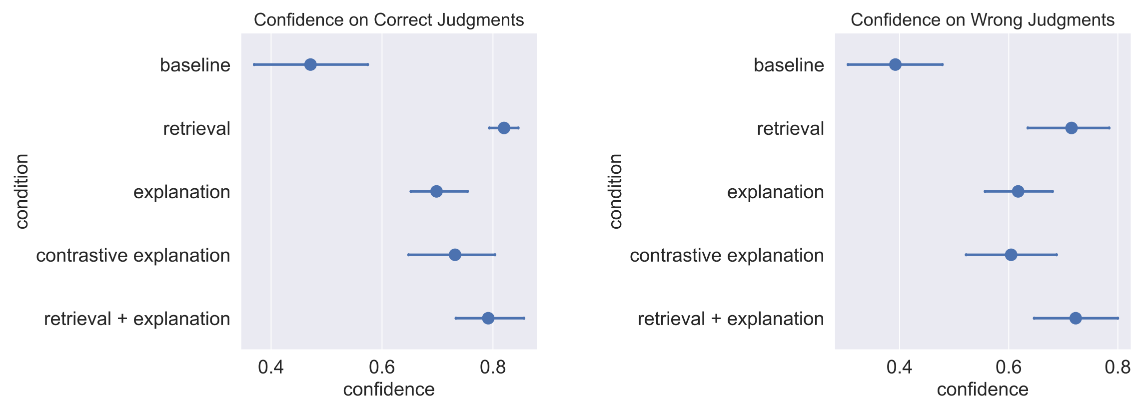
Combining retrieval and explanation does not achieve significantly higher accuracy than explanation alone in cases where the explanation is wrong either. When the explanation is wrong, users achieve the accuracy of (0.35 ± 0.23) in the explanation condition as compared to the combining retrieval and explanation condition (0.43 ± 0.16). The advantage of combining retrieval and explanation is not significant ($z = -1.06, p = 0.15$) so we could not reject the null hypothesis for H7 either. Taken together, we conclude that combining explanation and retrieval is not better than explanation alone. Next, we compare whether combining explanation with retrieval is better than retrieval alone.

Combining retrieval and explanation does not achieve significantly higher accuracy than retrieval alone in cases where the explanation is correct. When the explanation is correct, users achieve the accuracy of (0.79 ± 0.15) in the retrieval alone condition as compared to combining both retrieval and explanation (0.87 ± 0.12). There is a slight advantage of combining retrieval and explanation in this setting but the advantage is not significant ($z = -1.48, p = 0.07$), so we could not reject the null hypothesis for H8.

Combining retrieval and explanation does not achieve significantly higher accuracy than retrieval alone in cases where the explanation is wrong. When the explanation is wrong, users achieve the accuracy of (0.54 ± 0.26) in the retrieval alone condition as compared to combining both retrieval and explanation (0.43 ± 0.16), indicating a drop in accuracy in this case when combining retrieval and explanation. This means that combining retrieval and explanation offers no complementary benefits compared to retrieval alone. To understand whether users indeed read both the explanation and retrieved passages, we compare their reading time.

Combining retrieval and explanation takes longer time. In the retrieval alone condition, users take (2.5 ± 1.1) minutes to verify a claim; in the explanation condition, users take (1.0 ± 0.4) minutes to verify a claim; in the retrieval

+ explanation condition, users take (2.7 ± 1.0) minutes to verify a claim, indicating that combining retrieval and explanation increases the verification time, so users indeed spend time reading the explanation and retrieved passages in most cases. Moreover, in analyzing the free-form responses, the majority of the users base their judgment on the retrieved passages since the ChatGPT explanations are not guaranteed to be credible, further indicating that presenting ChatGPT explanations grounded on the retrieved passages does not really offer additional benefits than just presenting the retrieved passages themselves. Overall, our results suggest that combining retrieval and explanation might be redundant and inefficient.



(a) Human confidence in their correct judgments.

(b) Human confidence in their wrong judgments.

Fig. 7. Human confidence broken down by their correct and wrong judgments. Users are over-confident are wrong judgments.

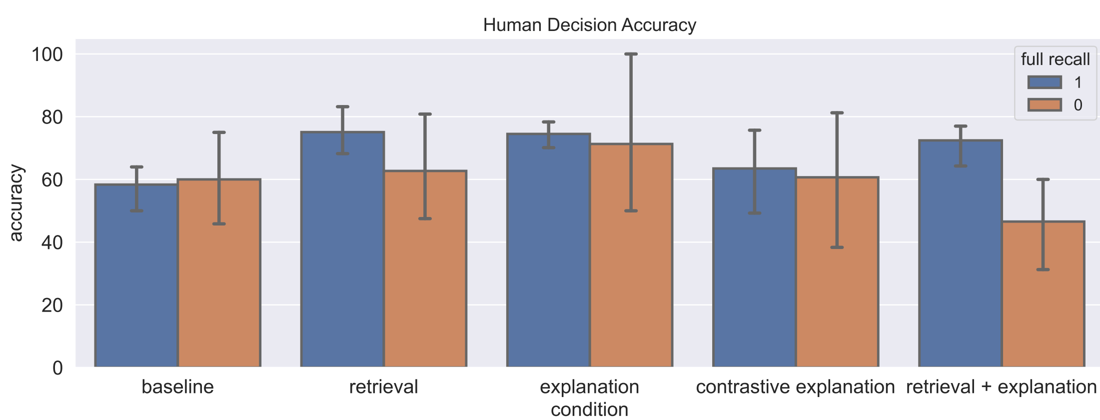


Fig. 8. Human accuracy broken down by retrieval recall. Human accuracy is lower when the retrieval recall is low.

8 META-ANALYSIS

8.1 Confidence Calibration

We convert users' confidence levels into discrete values $C = \{0, 0.25, 0.5, 0.75, 1.0\}$. Our goal is for users to have high confidence on their correct judgments and low confidence on their wrong judgments. We plot their average confidence on correct and wrong judgments in Figure 7. User confidence is always low in the `Baseline` condition, which is reasonable since they do not have additional supporting evidence and are mostly making educated guesses. On correct judgments, users generally have high confidence (above 0.6). However, **users are over-confident are wrong judgments**, with average confidence above 0.6 as well. The `Explanation` and `Contrastive Explanation` conditions incur lower user confidence on both correct and wrong judgments as compared to the `Retrieval` condition, as well as the `Retrieval + Explanation` condition. Overall, these results highlight the difficulty of achieving appropriate calibration in users' judgments.

8.2 Impact of Retrieval Recall

In previous sections, we performed breakdown analysis based on the correctness of the explanations. In this section, we analyze another important dimension—the retrieval recall. We split examples into two categories: the first group where the top-10 retrieved passages contain all the necessary evidence to verify the claim (*i.e.,* full recall = 1), and the second group where not all evidence is retrieved within the top-10 passages (*i.e.,* full recall = 0). We analyze how the retrieval recall affects both the explanation accuracy as well as the human decision accuracy.

The explanation accuracy is much lower when the retrieval recall is low. Over the entire test set of 200 examples, when the full recall = 1, the explanation accuracy is 80.4%; when the full recall = 0, the explanation accuracy is 67.6%. This indicates that retrieval quality has a high impact on the explanation accuracy, which in turn affects human decision accuracy.

Human decision accuracy is much lower when the retrieval recall is low. Human decision accuracy broken down by retrieval recall is plotted in Figure 8. In all cases (apart from the `Baseline` condition where users do not see any evidence), the human decision accuracy is lower when the full retrieval recall is 0, sometimes it is lower than the case of full recall = 1 by large margins, *e.g.*, in the `Retrieval` condition and the `Retrieval + Explanation` condition.

8.3 When Do Users Disagree with Explanations

We manually analyze the free-form responses provided by the users as their decision rationales. We first categorize these rationales in each condition and perform manual coding to quantitatively measure the frequency of each category of rationales. We analyze in what cases would users choose to disagree with the answers given by ChatGPT, and we break it down into the following two questions:

(1) **How do users make the correct decision when the explanations are wrong?**

- In the `Explanation` condition, most users rely on **self-contradiction** in the ChatGPT explanations (40.7%). For example, given the true claim *"Charles Evans Hughes shuffled off this mortal coil in Massachusetts, and then was taken to New York to be submerged in soil."*, ChatGPT generates the explanation *"The claim is false. According to the information provided, Hughes died in Osterville, Massachusetts, and was interred at Woodlawn Cemetery in the Bronx, New York City."* where the explanation actually supports the claim despite it saying the claim is false. Users did manage to catch this: *"The explanation sounds like it's actually agreeing with the claim."* and made the correct judgment.

- In the `Retrieval + Explanation` condition, users mostly rely on information from retrieved passages (63.5%) and occasionally based on ChatGPT's self-contradiction (15.9%), e.g., users responded *"I made the judgment by summarizing the highlighted areas in the passages."*

The remaining less common cases are mostly based on personal knowledge or guesses.

(2) **When do users make wrong judgments even when the explanations give correct answers?**

- In the `Explanation` condition, most users rely on personal knowledge or guess (46.4%) and sometimes because the explanations have wrong or poor-quality reasoning (25.0%). For example, one user responded *"I once took a tour of Alcatraz and I believe I remember this as being true."* to the claim *"Within Alcatraz was a music room where inmates could be rewarded for positive behaviors with playing time."* which is in fact false.
- In the `Retrieval + Explanation` condition, users mostly misinterpreted the evidence (38.1%) or there was just insufficient evidence (28.6%) and they had to make educated guesses. For example, to the false claim *"The Bee Gees went on tour eleven times."*, one user responded: *"The articles mention mostly their songs and a couple tours. I didn't find anything about 11 tours, just a couple of them. It was mainly songs I saw."* and they judged the claim to be true even though the ChatGPT explanation is correct: *"Based on the evidence from Wikipedia, the claim that the Bee Gees went on tour eleven times is not supported. The evidence mentions several tours that the Bee Gees went on, including the 2 Years On Tour, Trafalgar Tour, Mr. Natural Tour, Main Course Tour, Children of the World Tour, Spirits Having Flown Tour, One for All World Tour, High Civilization World Tour, and One Night Only World Tour. However, the evidence does not provide information about any additional tours beyond these mentioned. Therefore, it cannot be concluded that the Bee Gees went on tour eleven times based on the evidence provided."*

9 SUMMARY OF KEY FINDINGS AND FUTURE DIRECTIONS

Our human studies highlight the false promise of using natural language explanation produced by ChatGPT to aid human fact checking. In particular, humans over-rely on explanations even when they are wrong, making human accuracy worse than showing retrieval or the baseline of not even showing any evidence. In attempts to combat over-reliance, contrastive explanation mitigates users' over-reliance on wrong explanations, while combining retrieval and explanation does not achieve significant complementary improvement. Overall, neither of these two approaches significantly outperforms the retrieval baseline.

Our results highlight the need for better methods for combatting over-reliance on explanations. Future work could explore how to customize the best sets of evidence for different users in different conditions [2, 12], and how to strategically combine retrieval and explanation so that they can actually complement each other in both accuracy and efficiency. For instance, when the explanation is likely to be correct, we can show users the explanation; but when the explanation is likely to be wrong, we should prioritize showing users the raw retrieved passages. This also requires better uncertainty estimation or calibration to help users identify AI mistakes [36, 37, 40].

ACKNOWLEDGMENTS

We thank Nelson Liu, Xi Ye, Vishakh Padmakumar, Alison Smith-Renner, Ana Marasović, Omar Shaikh, and Tianyu Gao for their helpful discussion. We also appreciate the feedback from members of the UMD CLIP lab, especially Sweta Agrawal, Pranav Goel, Alexander Hoyle, Neha Srikanth, Rupak Sarkar, Marianna Martindale, Sathvik Nair, Abhilasha Sancheti, and HyoJung Han. Chen Zhao is supported by Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning, NYU Shanghai.

REFERENCES

[1] Yasmeen Alufaisan, Laura R. Marusich, Jonathan Z. Bakdash, Yan Zhou, and Murat Kantarcioglu. 2021. Does Explainable Artificial Intelligence Improve Human Decision-Making? *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 8 (May 2021), 6618–6626. https://doi.org/10.1609/aaai.v35i8.16819

[2] Gagan Bansal, Tongshuang Sherry Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2020. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2020). https://api.semanticscholar.org/CorpusID:220128138

[3] Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukosuite, Amanda Askell, Andy Jones, Anna Chen, et al. 2022. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540* (2022).

[4] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287

[5] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*. 160–169. https://doi.org/10.1109/ICHI.2015.26

[6] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. *2015 International Conference on Healthcare Informatics* (2015), 160–169. https://api.semanticscholar.org/CorpusID:12739635

[7] Upol Ehsan, Brent Harrison, Larry Chan, and Mark O. Riedl. 2018. Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*. Association for Computing Machinery, New York, NY, USA, 81–87. https://doi.org/10.1145/3278721.3278736

[8] Julian Martin Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Borschinger, and Jordan L. Boyd-Graber. 2021. Fool Me Twice: Entailment from Wikipedia Gamification. *ArXiv* abs/2104.04725 (2021). https://api.semanticscholar.org/CorpusID:233210218

[9] Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. 2020. Generating Fact Checking Briefs. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:226262339

[10] Robert Faris, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. 2017. Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election. *Social Science Research Network* (2017). https://api.semanticscholar.org/CorpusID:217155637

[11] Shi Feng and Jordan L. Boyd-Graber. 2018. What can AI do for me?: evaluating machine learning interpretations in cooperative play. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2018). https://api.semanticscholar.org/CorpusID:53039904

[12] Shi Feng and Jordan L. Boyd-Graber. 2022. Learning to Explain Selectively: A Case Study on Question Answering. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:254102968

[13] Raymond Fok and Daniel S Weld. 2023. In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making. *arXiv preprint arXiv:2305.07722* (2023).

[14] Tianyu Gao, Ho-Ching Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. *ArXiv* abs/2305.14627 (2023). https://api.semanticscholar.org/CorpusID:258865710

[15] Navita Goyal, Eleftheria Briakou, Amanda Liu, Connor Baumler, Claire Bonial, Jeffrey Micher, Clare R Voss, Marine Carpuat, and Hal Daumé III. 2023. What Else Do I Need to Know? The Effect of Background Information on Users' Reliance on AI Systems. *arXiv preprint arXiv:2305.14331* (2023).

[16] Zhijiang Guo, M. Schlichtkrull, and Andreas Vlachos. 2021. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics* 10 (2021), 178–206. https://api.semanticscholar.org/CorpusID:237304047

[17] Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. 2023. Are Machine Rationales (Not) Useful to Humans? Measuring and Improving Human Utility of Free-text Rationales. *ArXiv* abs/2305.07095 (2023). https://api.semanticscholar.org/CorpusID:258676376

[18] Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-World Entailment for Claims in Wikipedia. *ArXiv* abs/2303.01432 (2023). https://api.semanticscholar.org/CorpusID:257280052

[19] Vivian Lai, Chacha Chen, Qingzi Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. *ArXiv* abs/2112.11471 (2021). https://api.semanticscholar.org/CorpusID:245385821

[20] Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2020. QED: A Framework and Dataset for Explanations in Question Answering. *Transactions of the Association for Computational Linguistics* 9 (2020), 790–806. https://api.semanticscholar.org/CorpusID:221655495

[21] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* (2004). https://doi.org/10.1518/hfes.46.1.50_30392

[22] Nelson F. Liu, Kenton Lee, and Kristina Toutanova. 2023. Anchor Prediction: Automatic Refinement of Internet Links. *ArXiv* abs/2305.14337 (2023). https://api.semanticscholar.org/CorpusID:258841713

[23] Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. *ArXiv* abs/2304.09848 (2023). https://api.semanticscholar.org/CorpusID:258212854

[24] Ricardo Mendes. 2017. Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation. https://api.semanticscholar.org/CorpusID:158529031

[25] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hanna Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. *ArXiv* abs/2305.14251 (2023). https://api.semanticscholar.org/CorpusID:258841470

[26] Preslav Nakov, David P. A. Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barr'on-Cedeno, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. *ArXiv* abs/2103.07769 (2021). https://api.semanticscholar.org/CorpusID:232233764

[27] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021. Large Dual Encoders Are Generalizable Retrievers. *ArXiv* abs/2112.07899 (2021). https://api.semanticscholar.org/CorpusID:245144556

[28] Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the Risk of Misinformation Pollution with Large Language Models. *ArXiv* abs/2305.13661 (2023). https://api.semanticscholar.org/CorpusID:258840876

[29] Alicia Parrish, H. Trivedi, Ethan Perez, Angelica Chen, Nikita Nangia, Jason Phang, and Sam Bowman. 2022. Single-Turn Debate Does Not Help Humans Answer Hard Reading-Comprehension Questions. *ArXiv* abs/2204.05212 (2022). https://api.semanticscholar.org/CorpusID:248085750

[30] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. https://doi.org/10.1145/3411764.3445315

[31] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Qingzi Vera Liao, and Nikola Banovic. 2023. Understanding Uncertainty: How Lay Decision-makers Perceive and Interpret Uncertainty in Human-AI Decision Making. *Proceedings of the 28th International Conference on Intelligent User Interfaces* (2023). https://api.semanticscholar.org/CorpusID:257767905

[32] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[33] Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. 2022. Human Interpretation of Saliency-based Explanation Over Text. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022). https://api.semanticscholar.org/CorpusID:246294982

[34] Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. In *North American Chapter of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:232233599

[35] Ben Shneiderman. 2022. *Human-Centered Artificial Intelligence.*

[36] Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan L. Boyd-Graber, and Lijuan Wang. 2022. Prompting GPT-3 To Be Reliable. *ArXiv* abs/2210.09150 (2022). https://api.semanticscholar.org/CorpusID:252917981

[37] Chenglei Si, Chen Zhao, Sewon Min, and Jordan L. Boyd-Graber. 2022. Re-Examining Calibration: The Case of Question Answering. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:253098276

[38] James Thorne and Andreas Vlachos. 2018. Automated Fact Checking: Task Formulations, Methods and Future Directions. *ArXiv* abs/1806.07687 (2018). https://api.semanticscholar.org/CorpusID:49320819

[39] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. *ArXiv* abs/1803.05355 (2018). https://api.semanticscholar.org/CorpusID:4711425

[40] Helena Vasconcelos, Gagan Bansal, Adam Fourney, Qingzi Vera Liao, and Jennifer Wortman Vaughan. 2023. Generation Probabilities Are Not Enough: Exploring the Effectiveness of Uncertainty Highlighting in AI-Powered Code Completions. *ArXiv* abs/2302.07248 (2023). https://api.semanticscholar.org/CorpusID:256846746

[41] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael Bernstein, and Ranjay Krishna. 2022. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 7 (2022), 1 – 38. https://api.semanticscholar.org/CorpusID:254591809

[42] Andreas Vlachos and Sebastian Riedel. 2014. Fact Checking: Task definition and dataset construction. In *LTCSS@ACL*. https://api.semanticscholar.org/CorpusID:1669264

[43] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 7534–7550. https://doi.org/10.18653/v1/2020.emnlp-main.609

[44] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 318–328. https://doi.org/10.1145/3397481.3450650

[45] Sarah Wiegreffe and Ana Marasović. 2021. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. In *NeurIPS Datasets and Benchmarks*. https://api.semanticscholar.org/CorpusID:232035689

[46] Ming Yin, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019). https://api.semanticscholar.org/CorpusID:109927933

[47] Yunfeng Zhang, Qingzi Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020).