# Show me the evidence: Evaluating the role of evidence and natural language explanations in AI-supported fact-checking

GRETA WARREN, Department of Computer Science, University of Copenhagen, Denmark
JINGYI SUN, Department of Computer Science, University of Copenhagen, Denmark
IRINA SHKLOVSKI, University of Copenhagen, Denmark & Linköping University, Sweden
ISABELLE AUGENSTEIN, Department of Computer Science, University of Copenhagen, Denmark

Although much research has focused on AI explanations to support decisions in complex information-seeking tasks such as fact-checking, the role of evidence is surprisingly under-researched. In our study, we systematically varied explanation type, AI prediction certainty, and correctness of AI system advice for non-expert participants, who evaluated the veracity of claims and AI system predictions. Participants were provided the option of easily inspecting the underlying evidence. We found that participants consistently relied on evidence to validate AI claims across all experimental conditions. When participants were presented with natural language explanations, evidence was used less frequently although they relied on it when these explanations seemed insufficient or flawed. Qualitative data suggests that participants attempted to infer evidence source reliability, despite source identities being deliberately omitted. Our results demonstrate that evidence is a key ingredient in how people evaluate the reliability of information presented by an AI system and, in combination with natural language explanations, offers valuable support for decision-making. Further research is urgently needed to understand how evidence ought to be presented and how people engage with it in practice.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; *Empirical studies in collaborative and social computing*; • **Computing methodologies** → *Natural language processing*.

Additional Key Words and Phrases: explainable AI, fact-checking, explanation, natural language processing

**ACM Reference Format:**
Greta Warren, Jingyi Sun, Irina Shklovski, and Isabelle Augenstein. 2026. Show me the evidence: Evaluating the role of evidence and natural language explanations in AI-supported fact-checking. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 36 pages. https://doi.org/10.1145/3805689.3812358

## 1 Introduction

Large Language Models (LLMs) are increasingly popular tools for decision support in high-stakes tasks that require retrieving and reasoning over complex information, such as fact-checking [14, 21, 59]. However, these artificial intelligence (AI) models suffer from issues with factuality [4, 12, 30] and bias [22, 46] in their outputs, which means that relying on LLMs for decision-making is risky. At the same time, the linguistic fluency and persuasive qualities of LLM-generated text [54] can convince people to erroneously accept their outputs [6]. As a

Authors' Contact Information: Greta Warren, grwa@di.ku.dk, Department of Computer Science, University of Copenhagen, Copenhagen, Denmark; Jingyi Sun, jisu@di.ku.dk, Department of Computer Science, University of Copenhagen, Copenhagen, Denmark; Irina Shklovski, ias@di.ku.dk, University of Copenhagen, Copenhagen, Denmark & and Linköping University, Linköping, Sweden; Isabelle Augenstein, augenstein@di.ku.dk, Department of Computer Science, University of Copenhagen, Copenhagen, Denmark.


*FAccT '26, Montreal, QC, Canada*

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812358

result, overreliance, that is, accepting or following incorrect AI recommendations, is an increasingly recognised problem for human-AI interaction [5], particularly in information-seeking and decision-support tasks [10, 35].

Common mitigation approaches for overreliance include encouraging people to evaluate AI input before coming to a final decision [8, 50] and providing explanations for AI decisions [62]. A recent critique of explainable AI efforts highlights a lack of consideration for the importance of evidence in how people can integrate AI support into practice [50]. As commercial search systems integrate LLM output, these services increasingly include both explanations and links to retrieved evidence to help people decide whether to rely on AI-generated search summaries or investigate further [35, 44]. Although source- or evidence-seeking behaviour is considered important to understanding and calibrating reliance on AI advice, surprisingly little research has addressed this issue directly [29]. Some LLM-assisted information-seeking studies have shown that including explanations and links to evidence sources can help mitigate AI overreliance, although few people actively inspect the underlying evidence provided by the sources in the course of completing their tasks [35, 36]. Moreover, there is scant research on what may be more or less effective ways of presenting evidence for different types of information search and decision support activities. In particular, little prior work explores providing evidence as part of the experimental setup. Making the evidence readily accessible alongside the AI system output decreases friction inherent in clicking through links to external information, and may enable people to more easily inspect the validity of the system's predictions, allowing for greater oversight and reliance calibration.

A second stream of overreliance mitigation approaches is the provision of explanations of AI system decisions, with the goal of supporting critical engagement with AI output, especially in decision-support tasks [49, 62, 73]. However, research on the relationship between explanations and reliance on AI system advice has produced conflicting findings. In some cases, explanations appear to enable more efficient use of AI support [73], although different types of explanations may also increase overreliance [5]. Communicating model uncertainty in AI outputs has been found to help people identify when advice may be unreliable [11, 43], induce more deliberative thinking [58], and decrease overreliance [34]. While most have used numerical scores [37, 53, 75], or verbal hedges (e.g., "I'm not sure but..." [34]), to indicate model uncertainty, recent work suggests that people may find natural language explanations of uncertainty helpful and compelling [25, 69, 70]. Explanations of uncertainty may be an effective way to highlight inconsistencies in AI output, which can help mitigate overreliance [35, 67].

Building on this work, we sought to explore the role of evidence, that is, the information provided to the model as input (e.g., evidence documents in the form of retrieved news articles) in an AI-assisted information-seeking and decision-making task. Given the impacts of explanations and uncertainty expressions on reliance and decision-making, we also sought to test the effects of two natural language explanation types during the task: traditional verdict-focused explanations, which focus on the relationship between the evidence and the model's fact-checking *verdict* (e.g., true or false), versus uncertainty explanations, which focus on the relationship between the evidence and the model's *uncertainty* in the verdict. We examined to what extent people used AI explanations and evidence underlying the AI system decision when making decisions in a fact-checking context, to address the following research questions:

**RQ1:** What is the relationship between behavioural and subjective reliance in AI-supported fact-checking?
**RQ2:** How do people utilise evidence and AI output when deciding whether to rely on AI advice?
**RQ3:** How do natural language explanations impact people's reliance on AI advice in a fact-checking task?

We conducted a mixed methods 3x2x2 experiment (N=208) with crowdworkers to examine the effects of providing explanations and evidence in a fact-checking scenario and collected qualitative responses about how participants made their decisions. We compared two types of natural language explanations: standard verdict-focused explanations and explanations of AI uncertainty, along with a control condition displaying only numerical uncertainty. Participants had the option to access the evidence documents used by the AI system across all conditions. We measured which sources of information (e.g., evidence, explanations, own knowledge) participants

found useful, their agreement with the AI system's advice, decision accuracy, subjective measures of confidence, usefulness, and reliance and qualitative evaluations of the information available.

We found that participants overwhelmingly identified evidence as the most useful information source, regardless of the correctness of AI advice, level of AI certainty, and whether they saw an explanation or not. When participants were provided with explanations, these were also viewed as helpful, particularly when the AI system's advice was correct or highly confident. We also discovered an association between participants' self-reported reliance on the AI system and the information they used: those who reported higher subjective reliance on the AI system tended to use the AI explanations more, while those with lower subjective reliance scores were more likely to use the underlying evidence. Together, our results highlight that making the underlying evidence for AI system outputs accessible can have significant implications for effective decision-support in information-seeking tasks.

## 2 Related Work

### 2.1 Mitigating overreliance in decision-support systems

Overreliance is typically defined as the acceptance of incorrect AI system output, or delegation of decisions to an AI system when inappropriate to do so [29], for example if a doctor were to wrongly accept an AI diagnosis without critically evaluating whether it aligns with the patient's symptoms. Excessive deference to AI advice can lead to errors and severe harms, particularly in high-stakes scenarios such as clinical decision-support [32, 39]. The growing use of LLMs for decision-making and information-seeking in everyday life has also raised concerns about risks of professional deskilling [9], diminished educational outcomes [77] and declines in general problem-solving and critical thinking abilities [18]. Approaches to mitigating overreliance on AI systems primarily fall into two categories: those that encourage users of AI systems to consider the *evidence* for themselves (e.g., cognitive forcing [8] and evaluative AI paradigms [50]), and those that encourage users to develop a better understanding of the AI system and its reliability, through providing *explanations* (e.g., [73]) and/or indications of model certainty (which may be numerical [58] or verbal [34]). In the following subsections, we examine prior work on both of these approaches. Reliance is also related to trust, although they are distinct constructs [15, 28]; trust is commonly defined as a willingness to grant benefit of the doubt, while reliance refers more specifically to a willingness or intent to rely on a third party or system [31]. These attitudinal dispositions may manifest behaviourally in forms such as overreliance, whereby excessive trust is placed in a system (although such behaviour may also stem from user error or attempts to avoid responsibility for decisions or outcomes) [29]. In this study, we measured how often participants chose to use the AI system's predictions, to evaluate the impacts of presenting evidence and explanations for AI system outputs on participants' reliance. We also measured subjective reliance in AI to examine the relationship between subjective and behavioural willingness to rely on AI advice, explanations and evidence. We hypothesised that we would observe a relationship between higher levels of reliance on AI advice and higher subjective reliance in the system (**RQ1:H1**). We also hypothesised that higher usage of AI explanations would be associated with higher subjective reliance (**RQ1:H2**).

### 2.2 The role of evidence in AI reliance

A promising approach to reducing overreliance is including the evidence underpinning the AI system's prediction alongside its output. Prior work has explored how decision-support paradigms can be used to promote cognitive engagement with available evidence and encourage system users to come to their own conclusions. Cognitive forcing is one such technique, in which the user of a decision-support system is presented with the AI system input and must make their own prediction as to the outcome before being shown the AI system's output [8]. The Evaluative AI paradigm is an alternative, which proposes that decision-support systems should provide evidence for and against each hypothetical outcome, rather than recommending a single prediction [50]. Recent work on reliance on LLMs for information-seeking and decision-support has focused on providing evidence in the form

of clickable links to external information (e.g., news articles or government advice). In a question-answering task, LLM responses that listed sources were associated with lower overreliance, and higher user confidence in their own answers [35]. However, the explanations provided by the AI system did not explain *how or why* the sources cited support the AI system's prediction. Moreover, participants only clicked on one or more of these sources 22-28% of the time, suggesting that the positive impacts of sources may be more due to the credibility of the source linked (which were vetted by the authors to contain high-quality, correct information) than the information contained in the links themselves. While source credibility is a critical component of fact-checker decision-making [45, 75], it is notoriously difficult to integrate into automated decision-support systems because it tends to be highly variable and subjective [2]. In this study, we include evidence alongside all explanations, but omit source credibility information to limit impacts of familiarity and subjective bias. We hypothesised that evidence would be utilised more when explanations are not provided (**RQ2:H3**). We also hypothesised that participants would utilise explanations more than other information when they were made available (**RQ2:H4**).

## 2.3 The role of explanations in AI reliance

Explainable AI approaches aim to increase the transparency of AI systems by helping people understand why particular decisions are made by the system [20, 66]. As such, explainable AI methods to support high-stakes information-seeking tasks, such as fact-checking, is an area of considerable interest [23]. Previous work has shown that general explanations of how an AI-based fake-news labelling system works reduced people's intentions to share false news articles, although they had no impact on people's trust in the labels [16]. Natural language explanations that highlight reliability cues in news articles have been found to assist people in identifying unreliable or biased articles [27]. Given natural language explanations of AI decisions, laypeople can detect untrustworthy news articles with equivalent accuracy to professional journalists [64]. People are also more accurate in assessing the credibility of news/social media posts when provided with text-based explanations compared to graphical explanations [61]. However, LLM-generated explanations can also lead to overreliance on AI systems [64, 67]. Prior work has examined how overreliance on AI system advice can be mitigated by communicating that the AI is uncertain in its prediction, for example, through visualisation [58] or linguistic expressions of uncertainty by prepending statements like "I'm not sure, but..." to LLM outputs [34]. Recent work suggests that natural language explanations can potentially enhance user understanding of model uncertainty [69]. For example, providing contrastive information to highlight inconsistency between explanations and model output can help users reason more critically about AI predictions [67]. Given that natural language explanations may help people to assess the credibility of news articles and claims, in this study we examine the utility of natural language explanations to support decision-making in automated fact-checking. However, as they can also lead to overreliance, we evaluate two variants of natural language explanations – a verdict-based explanation that follows the typical approach of justifying system predictions [3, 64], and an uncertainty explanation designed to reduce overreliance by explaining the system's uncertainty in its prediction [70]. We hypothesised that natural language explanations of uncertainty would be more effective in reducing overreliance compared to traditional verdict-focused explanations and control explanations (**RQ3:H5**). We also hypothesised that explanations of uncertainty would be perceived as more useful in a fact-checking scenario (**RQ3:H6**).

## 2.4 AI decision-support for fact-checking

We designed our experiment as a fact-checking task due to its relevance to both information-seeking and decision-support contexts, with potentially high stakes consequences, given that AI systems such as LLMs are increasingly used as sources of information and news [68]. AI tools to support fact-checking are vital to improve fact-checking efficiency. However, recent work with fact-checking practitioners suggests that explanations for how the AI system arrives at its prediction may be key to effective use [75]. The two most common approaches

to explainability in this space are attribution-based explanations, which highlight the key tokens or words that contributed to the AI system's output [56], and natural language explanations in the form of free-text justifications for the AI system's decision [3, 38]. While there is evidence that explanations can enhance understanding of AI systems by increasing model transparency [17], research shows that they also can lead to problematic overreliance on AI advice [5, 52, 57, 78]. For instance, example-based and attribution-based explanations seem to have little effect on people's ability to correctly assess the veracity of claims [42], but can lead to reliance on AI system advice even where it can be demonstrably wrong [41]. In this study, we designed a task in which participants are presented with the input (claim and evidence) and outputs of an AI system (predicted verdict, model uncertainty, AI explanation) and asked to decide whether to use the AI system's prediction or not, and to identify which information they used to make their decision.

## 3 Method

We designed a controlled experiment to understand the role of evidence and explanations in how people make decisions in an AI-supported fact-checking task. Participants were presented with the outputs of an AI system designed to support fact-checking by predicting the veracity of claims (see example interface in Fig. 1). The system provided a numerical estimate of model certainty about the prediction, an explanation, and two evidence documents on which the prediction and explanation were based. The study was designed to assess the impact of explanation type on the fact-checking decision process, where participants could accept or decline the AI prediction. Each participant evaluated eight carefully selected combinations of claims and relevant evidence documents retrieved from online sources using one of three types of explanations (Uncertainty explanations, Verdict explanations, and No explanation) to support the process. We varied model certainty and model correctness counterbalanced within subjects. The study was approved by our institution's Research Ethics Committee.

### 3.1 Experimental setup and design

*3.1.1 Claims and Evidence.* The claims and evidence documents that formed the basis for the materials were drawn from the DRUID [24] dataset. We selected this dataset for its diversity, comprising claims from seven professional fact-checking websites, such as Politifact (https://www.politifact.com/), and FactCheckNI (https://factcheckni.org/) and quality-checked evidence documents retrieved from online sources such as news articles. We selected claims covering science, health and politics (see Table 3) that had been fact-checked by organisations based in the United States, United Kingdom, Sri Lanka and France. The evidence documents (see App. I) consisted of online articles (see [24] for full dataset).

*3.1.2 AI Output: Verdicts, Certainty and Explanations.* For each claim and pair of evidence documents, we used Qwen2.5-14B-Instruct[1], an open-weights, instruction-tuned language model with 14B parameters, to generate a verdict and explanation for the given claim. We chose this model as a representative mid-sized open model that can be run locally (supporting reproducibility without relying on proprietary APIs), and that provides access to token-level logits required for our entropy-based uncertainty estimate. The model's uncertainty was quantified by the entropy of its logits for generating the verdict, i.e., True or False (see App. G.1 for estimation method).

We compared three types of explanations (see Table 1 for examples). Given the role of explanations and expressions of uncertainty in reliance on AI decisions, our aim was to compare the effects of *Uncertainty explanations*, *Verdict-focused explanations*, and a *No explanation* control. The *Uncertainty explanations* provided reasons for the model certainty score. For example, conflicting evidence provided to the model may decrease the model's certainty that the predicted verdict is correct, while corroborating evidence may increase the model's certainty. To generate the uncertainty explanations, we followed the CLUE method [70]. Given a claim and some

[1] https://huggingface.co/Qwen/Qwen2.5-14B-Instruct

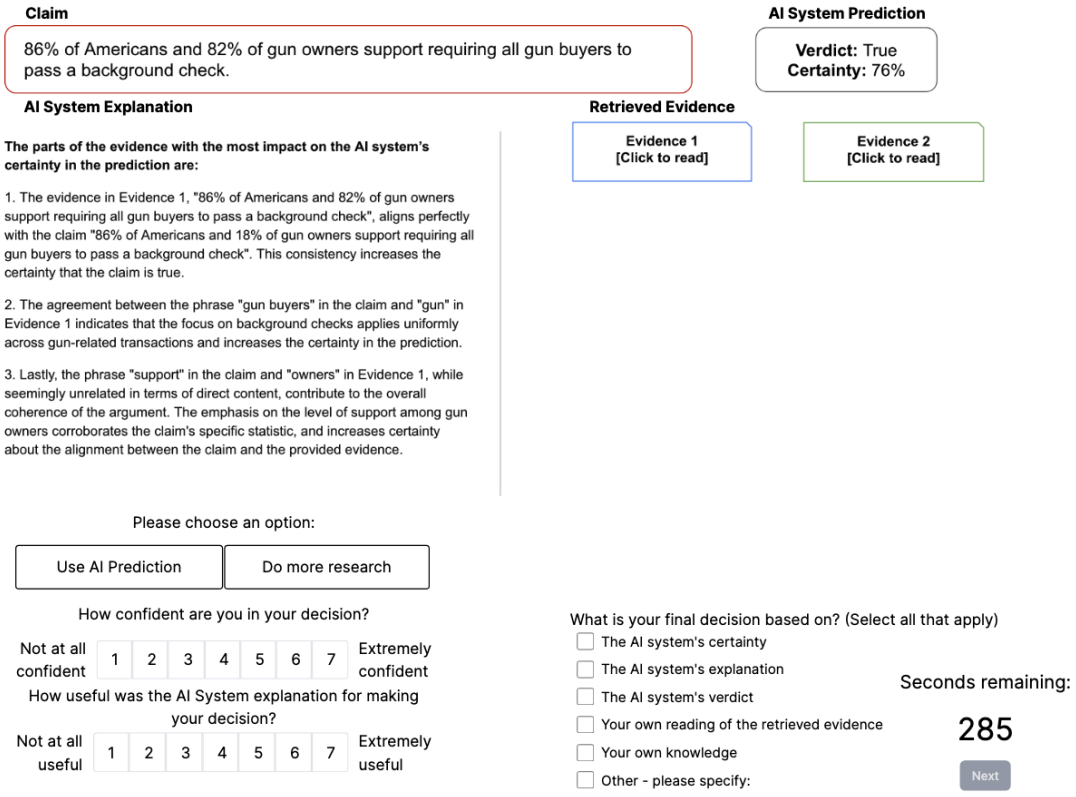



Fig. 1. Example of task interface in Uncertainty Explanation condition

evidence documents as input, this method first identifies key conflicting and concordant spans of text between the claim and evidence pieces that influence model certainty, and then instructs the model to describe how these conflicting or concordant span interactions, such as disagreements between two pieces of evidence, affect the certainty of the verdict prediction (see App. G.3 for the instruction prompt). The *Verdict-focused explanations* were generated using a few-shot prompting method in which the model was instructed to explain why the verdict is predicted by referencing or summarising the provided evidence, and provided with three complete examples within the prompt (see App. G.2). The 'explanations' in the *No explanation* condition were designed as a control condition and merely restated the AI system's verdict prediction and certainty score in text.

We assessed all generated explanations for quality, making minor edits to ensure consistency in structure (e.g., line-breaks and 2-3 numbered paragraphs for readability) and equivalent length in the explanations for the same claim. Where explanations did not explicitly explain the extracted text spans and the resulting AI system certainty percentage as instructed, we made minor edits to include this information to ensure that all explanations met our intended aims. We ensured that all Uncertainty explanations began with "The parts of the evidence with the most impact on the AI system's certainty in the prediction are:" and that all Verdict explanations began with "The parts of the evidence with the most impact on the AI system's predicted verdict are:" to ensure the target of each explanation type was clear.

*3.1.3 Interface design.* Following previous studies examining AI-supported fact-checking (e.g., [42, 51, 64]), we designed the interface to mimic a news dashboard. Participants were presented with a fact-checking claim, the AI system's predicted verdict, certainty, and an explanation for the AI system's output. Participants could access the evidence by clicking on the evidence boxes in the interface to reveal them. We chose to present the evidence within the experiment, rather than as links to external sources (as in e.g., [33, 35]) for several reasons: (i) to mitigate the risk that evidence could be edited or removed, given that the AI outputs were pre-generated, (ii) to maintain participants' focus on the experimental task, rather than becoming distracted by side-research, and hence (iii) to control the content, relevance, and length of information participants were provided, (iv) to allow us to measure the extent to which participants chose to view or hide the evidence, and (v) to study the approach of providing evidence documents 'in-window' used by a number of LLM-based chatbots designed specifically for knowledge intensive information-seeking tasks such as Consensus AI (https://consensus.app/) and Google Gemini's 'Double-check response' feature (https://gemini.google.com/app).

## 3.2 Pre-testing experimental setup through think-aloud study

We pre-tested our initial experimental setup with a think-aloud study to refine our design and materials and to observe in-situ how people reacted to the different types of explanations. We recruited five participants with diverse expertise from our institution (T1-T3) and wider community (T4-T5) (see App. B for demographics). Sessions lasted approximately 1 hour, and were audio recorded and transcribed with participants' permission (see App. C). Each participant saw six claims, two with each type of explanation, in random order. We conducted semi-structured interviews upon task completion. We made several key design changes based on the observations and insights gathered during these sessions.

*3.2.1 Task Design.* Our initial study design asked participants to make a decision about whether the claim was true or false and gave no time limit. After three think-aloud sessions, we observed that people often focused on reading the evidence, to the exclusion of the AI system decisions and explanations. Through participant debriefing, we realised this happened because the decision about whether the claim was true or false did not necessarily require people to engage with the AI system, as the evidence was readily available and people could spend as much time as desired reading the evidence documents to form their decision. We made two changes to the task to encourage participants to engage with the AI system prediction, certainty and explanations. First, we altered the main decision participants made, from "True" vs "False" to "Use AI Prediction" vs "Do more research". In the task instructions (App. H), participants were told to imagine that their goal was to release assessments about whether a claim is true or false, and that their task was to decide whether to rely on the AI prediction, or conduct more research before releasing the assessment. The purpose of this change was to encourage participants to evaluate the AI output, rather than merely evaluating the claim and the evidence. Second, we introduced a time limit of 5 minutes per instance, to create a more realistic situation for decision-making, engender a greater sense of urgency, and discourage any side-research. When the time limit elapsed, evidence and explanations disappeared from the screen, and participants were prompted to make a final decision and move on to the next instance. After introducing these changes, we conducted think-aloud sessions with two additional participants, and observed that these changes shifted participants' focus more towards the AI system.

*3.2.2 Claims.* The think-aloud pre-testing also helped to identify potential issues with the claims presented to participants. After each think-aloud session, we reviewed the comments participants made about the claims, and replaced claims that were too vague, too familiar, or too difficult to understand (e.g., "I don't like the claim [...] it's very vague and I feel like it should just throw it out" – T1).

*3.2.3 Evidence sources.* Participants frequently commented on the importance of knowing the identity of evidence sources: "including the source of the evidence... It will allow me to know how trustworthy the evidence is, who

| |
|---|
| **Claim:** A windmill could spin until it falls apart and never generate as much energy as was invested in building it. |
| **Evidence 1:** No, a professor didn't say a windmill will 'never generate' the energy invested in building it [...] Now, we have a meme roaming social media that claims Canadian author and professor Thomas Homer-Dixon said that windmills will "never" generate as much energy as was expended to build the structures. [...] The meme does get a part of the quote right, but it strategically leaves out large swaths of information in order to give people the impression that all windmills never expend the amount of energy that was invested in it, which is not accurate. The meme's omitted section explains that, while poorly placed windmills may never generate enough energy payback, a good wind site could generate it in three years or less. |
| **Evidence 2:** Factually Inaccurate: Analyzing the energy used in the production of wind turbines does not show that it's greater than the electricity produced over the turbine's working lifetime. |
| **Predicted Verdict:** False **Certainty:** 31% |
| **Uncertainty Explanation:** The parts of the evidence with the most impact on the AI system's certainty in the prediction are:<br>1. The evidence in Evidence 1, "No, a professor didn't say a windmill will 'never generate' the energy invested in building it", directly contradicts the claim's assertion that windmills will never generate sufficient energy. This contradiction increases the AI system's certainty that the claim is False.<br>2. The evidence in Evidence 1, "will 'never generate' the energy invested", aligns with the part of the claim "spin until it falls apart and never generate as much energy as was invested", reducing the AI system's certainty that the claim is False.<br>3. The evidence in Evidence 1, "expended to", aligns with the phrase "invested in building" in the claim, indicating agreement on the energy expenditure aspect, reducing the certainty that the claim is False.. |
| **Verdict Explanation:** The parts of the evidence with the most impact on the AI system's predicted verdict are:<br>1. Evidence 1 explains that the claim comes from a misrepresented quote. The meme states that windmills "will never generate as much energy as was expended to build the structures," but the original quote clarifies that "while poorly placed windmills may never generate enough energy payback, a good wind site could generate it in three years or less." This shows the claim exaggerates and omits important context.<br>2. The same evidence highlights that well-sited windmills can generate enough energy to offset their construction cost quickly, making the claim that all windmills "never generate" enough energy inaccurate. The phrase "a good wind site could generate it in three years or less" directly contradicts the claim.<br>3. Evidence 2 reinforces this, stating that "analyzing the energy used in the production of wind turbines does not show that it's greater than the electricity produced over the turbine's working lifetime." This confirms that, over time, turbines produce more energy than was invested in building them. |
| **No Explanation:** The AI system has judged this claim to be False, with 31% certainty. |

Table 1. Example of claim, evidence, corresponding AI prediction, certainty, and explanations

made the claim" (T2). While we wanted to avoid including specific names of sources to control for potential biases of participants [2, 40], we modified the task instructions to advise that all evidence provided should be considered reasonably reliable.

## 3.3 Controlled Experiment

*3.3.1 Design and Measurements.* The final experiment had a 3 (Explanation type: Uncertainty explanation vs Verdict explanation vs No explanation) x 2 (AI advice: correct vs incorrect) x 2 (AI certainty: high vs low) design

to investigate the effects of explanations and evidence across different levels of AI correctness and AI certainty (see Fig. 2 for an overview). Explanation type was a between-participants factor, while AI correctness and AI uncertainty were within-participant factors. In the main task (see Figure 1), participants were presented with a fact-checking claim, the AI system's predicted verdict, certainty, and an explanation for the AI system's output. Participants could access the evidence by clicking on the evidence boxes in the interface to reveal them. For each claim, participants were asked to decide whether to (a) use the AI system's prediction, or (b) do more research. If a participant chose to 'do more research', a follow-up question asked them to explain their decision given the options 'The AI is incorrect', 'More information is needed to make a decision', or 'Other'. For each decision, participants were asked to indicate what sources of information (i.e., AI verdict, AI certainty, AI explanation, evidence, their own knowledge, and other) their final decision was based on. They were also asked to indicate (i) how confident they were in their own decision, and (ii) how useful the explanation was to their decision on a scale of 1-7. There was a time-limit of 5 minutes for each claim. If this elapsed, the evidence and AI explanations disappeared and participants were forced to make a decision and move on. A post-task questionnaire collected subjective evaluations and free-text responses, see Table 2 and App. D for details.

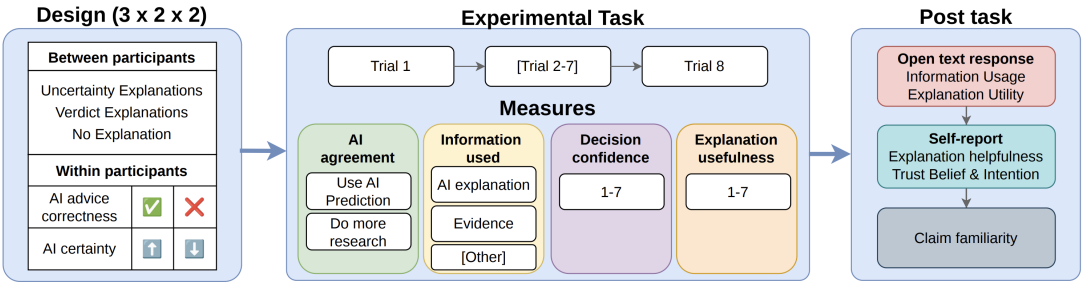



Fig. 2. Overview of study design and flow

*3.3.2 Participants.* A power analysis [19] indicated that 207 participants were required for 90% power for a medium effect (alpha<.05) for two-tailed tests. Participants (N=208) were recruited using Prolific (https://www.prolific.com/), and randomly assigned to one of three conditions: Uncertainty explanation (n=70), Verdict-based explanation (n=69) and No explanation (n=69). Participant demographics are reported in Table 7 (App. E). Participants were pre-screened to be native English speakers from Ireland, the United Kingdom, the United States, Australia, Canada, and New Zealand, who had not participated in previous related studies. Twenty-two participants who failed more than one attention or memory check were excluded prior to data analysis.

*3.3.3 Materials.* Eight unique claims (and corresponding evidence documents) were selected from the DRUID [24] dataset to balance (i) true and false verdicts, (ii) correct and incorrect AI predictions, (iii) high (>50%) and low (<50%) AI certainty (see Table 3). Explanations were consistent with the AI prediction, regardless of whether it was correct or incorrect. For all claims, evidence and explanations see App. I.

*3.3.4 Procedure.* Participants received a link to the study hosted on Gorilla (https://gorilla.sc/). Participants read detailed instructions and completed a practice example (see App. H). They then progressed through the main task, providing responses for the eight claims in random order. They then responded to a post-task questionnaire and optionally provided demographic information. Participants were paid £4.50 for taking part. The average completion time was 29m 54s, corresponding to a payment rate of £9 per hour.

| Task | Measurement | Response |
|---|---|---|
| **Main task** | AI advice use | Use AI prediction vs. Do more research |
| | Information usage | Multiple choice: The AI system's verdict / AI certainty / AI explanation / Evidence / Own knowledge / Other |
| | Decision confidence | 7-point scale: 1="Not at all" to 7="Extremely" |
| | Explanation usefulness | 7-point scale: 1="Not at all" to 7="Extremely" |
| **Post-task** | Information usage (qualitative) | How did you use the information provided by the AI system to make your decisions? |
| | Explanation utility (qualitative) | Please explain why the AI system explanations were/weren't useful for determining how reliable the system was. |
| | Explanation helpfulness (App. D.1) | 5-point scale: 1="Not at all" to 5="Extremely" |
| | Subjective reliance (Trust Belief & Intention [47]) (App. D.2) | 5-point scale: 1="Strongly disagree" to 5="Strongly agree" |
| | Claim familiarity (App. D.3) | 4 options: No knowledge / Limited knowledge / Some knowledge / Full knowledge |

Table 2. Overview of tasks, measurements, and response formats.

| Claim | AI Verdict | AI Advice | AI Certainty |
|---|---|---|---|
| 1. 86% of Americans and 82% of gun owners support requiring all gun buyers to pass a background check. | True | Correct | High |
| 2. A satanic-themed hotel is opening in Texas. | False | Correct | High |
| 3. A scone can equal a third of one's recommended daily calories. | True | Correct | Low |
| 4. A windmill could spin until it falls apart and never generate as much energy as was invested in building it. | False | Correct | Low |
| 5. All tea bags contain harmful microplastics. | True | Incorrect | High |
| 6. Lagan Valley is one of the least wooded areas in Northern Ireland, particularly with regards to ancient woodland | False | Incorrect | High |
| 7. Cristiano Ronaldo was the only European soccer captain who did not wear a rainbow armband at Euro 2020. | True | Incorrect | Low |
| 8. The top issue for college students is the economy. | False | Incorrect | Low |

Table 3. Claims presented to participants during the experiment.

*3.3.5 Data analysis.* Quantitative analysis was performed using R [60]. Alongside traditional quantitative analysis, we analysed themes in our qualitative data from think-aloud interviews and free-text response questions. Two authors open-coded all of the data independently, periodically comparing codes and reconciling differences until achieving full agreement. Open codes were then combined into thematic categories to aid analysis [7].

## 4 Results

The importance of evidence in our study was striking across quantitative and qualitative data. In contrast to prior research [35], 64% of participants opened both evidence documents for every claim they assessed and just three (of 208) did not access either of the evidence documents. Participants reported finding evidence useful

significantly more often than any other available information, regardless of condition. Evidence was also the primary topic of commentary in free-text responses, where participants explained which information they used to make decisions. Explanations were judged to be helpful tools, particularly in identifying when the AI output was less reliable. In this section, we first discuss the original focus of the study, to assess the role of different types of explanations in claim assessment in a fact-checking scenario. We then discuss the role evidence played in this process and the implications of this finding.

### 4.1 The role of explanations in claim assessment

Our initial hypothesis was that uncertainty explanations would be more effective in supporting participants to evaluate whether to agree or disagree with the AI system (**RQ3:H5**). We analysed the extent to which participants used the AI system's prediction, depending on the explanation condition, correctness of the advice, and level of certainty of the system (see Table 4). Predictably, participants were more likely to follow the AI system's advice when it was correct vs. incorrect, F(1, 207)=208.779, p <.001, $\eta_p^2$=.50, and when AI certainty was high vs. low, F(1, 207)=55.55, p <.001, $\eta_p^2$=.21. Yet there was no difference in overreliance, F(2, 205)=1.186, p=.307, or overall reliance based on the explanation condition, F(2, 205)=1.825, p=.164, nor did this factor interact with advice correctness, F(2, 205)=0.1914, p=.9087, or certainty, F(2, 205)=0.5771, p=.749. We did not observe any effect of explanation condition on confidence or trust judgments, nor did prior familiarity with claims affect results (see App. F).

| Measure | All conditions | Uncertainty Explanation | Verdict Explanation | No Explanation |
|---|---|---|---|---|
| **Use of AI prediction (%)** | | | | |
| Overall | 54.56 | 51.79 | 57.79 | 54.17 |
| Correct AI advice | 73.44 | 71.43 | 76.09 | 72.83 |
| Incorrect AI advice | 35.70 | 32.14 | 39.49 | 35.51 |
| High AI certainty | 62.62 | 57.86 | 65.58 | 64.49 |
| Low AI certainty | 46.51 | 45.71 | 50.00 | 43.84 |
| **Explanation usefulness (1–7)** | | | | |
| Overall | 4.93 | 4.98 | 5.12 | 4.70 |
| Correct AI advice | 5.30 | 5.28 | 5.47 | 5.16 |
| Incorrect AI advice | 4.56 | 4.68 | 4.76 | 4.24 |
| High AI certainty | 5.22 | 5.15 | 5.43 | 5.07 |
| Low AI certainty | 4.65 | 4.81 | 4.80 | 4.33 |
| **Explanation helpfulness (1–5)** | | | | |
| Overall | 3.67 | 3.80 | 3.88 | 3.32 |

Table 4. Mean use of AI predictions, and explanation usefulness and helpfulness judgments by participants in each condition. The 'Incorrect AI advice' row corresponds to overreliance rates.

We did observe a main effect of explanation condition on how often people used the explanations (**RQ3:H6**) (see Table 5), F(2, 205)=24.668, p<.001, $\eta_p^2$=.19. Tukey HSD tests showed that participants in the two natural language explanation groups reported using the explanations in their decisions significantly more than the No Explanation group (Uncertainty: p<.001, d=.99, Verdict: p<.001, d=1.07). People judged that both Uncertainty (p=.003, d=.5) and Verdict (p=.013, d=.46) Explanations were more *helpful* than No Explanation and that the Verdict Explanations were more *useful* than No Explanation, p=.01, d=.34. Further, participants reported using the AI explanations more frequently when the AI system was correct (M=55.5%) than when it was incorrect (M=48.1%), F(1, 207)=12.125, p<.001, $\eta_p^2$=.055. No differences emerged between the two explanation groups (p>.05 for all comparisons). There was a main effect of explanation condition on how useful people judged the explanations, F(2, 205)=4.412, p=.013,

$\eta_p^2$ =.02. Bonferroni-corrected post hoc tests indicated that the Verdict Explanation group rated the explanations as more useful than No Explanation, p=.01, d=.34.

**Cross-referencing explanations and evidence to identify AI errors.** In the text responses, participants in the No Explanation condition complained about lack of information, although several mentioned using numerical uncertainty as a cue to be more sceptical. Participants in the explanation conditions often found explanations useful as summaries of evidence when they aligned with evidence. When explanations and evidence were misaligned, however, participants commented that this helped them understand the system's reasoning, aligning with prior research [67]: "Sometimes by reading the explanations I could spot errors in the AI's logic." Qualitative data also suggested reasons for why the natural language explanations did not differ along expected dimensions. Almost all responses in the No Explanation condition mentioned using evidence to assess the claim, suggesting that having access to evidence documents compensated for differences in explanation types. Participants in the Uncertainty condition had mixed opinions on the utility of these explanations. Some found them difficult to understand: "on some of the choices the AI was very helpful but in others I found it very contradictory and confused,"suggesting that more research on explaining uncertainty in natural language is necessary.

| Measure | All conditions | Uncertainty Explanation | Verdict Explanation | No Explanation |
|---|---|---|---|---|
| AI Verdict | 41.10 | 40.00 | 45.83 | 37.50 |
| AI Explanation | 51.80 | 59.46 | 61.41 | 34.42 |
| AI Certainty | 31.00 | 27.14 | 30.98 | 34.78 |
| Evidence | 67.80 | 64.82 | 71.01 | 67.75 |
| Own knowledge | 20.30 | 19.64 | 15.76 | 25.54 |
| Other | 2.88 | 3.21 | 3.80 | 1.63 |

Table 5. Mean percentages of each source of information used.

## 4.2 The role of evidence in claim assessment

Participants reported using the evidence most frequently (67.8% overall), regardless of AI explanation, correctness, or certainty. A one-way repeated measures ANOVA indicated differences in how often participants used each information source, F(5, 1035)=200.6, p<.001, $\eta_p^2$=.49. People in the No Explanation group were more likely to open all evidence documents (83%), compared to the explanation groups (55% Uncertainty, p=.001, d=.62; 57% Verdict, p=.003, d=.58) (**RQ2:H3**). This suggests that natural language explanations may have offered sufficient information to make a judgment based on the explanation alone or a single evidence document. Mirroring the overall trend, participants in the Verdict condition relied more on the evidence than all other information sources, and relied on the explanations second most often, more so than the remaining information sources, p<.001 for all comparisons. Participants in the Uncertainty condition, however, reported using the evidence and explanations equally as often, p=.241 (**RQ2:H4**). They relied on these two information sources significantly more than the others, p<.001 for all comparisons. Participants in the No Explanation condition used evidence significantly more than any other information source, p <.001 for all comparisons. The 'explanations' (restatements of AI verdict and certainty) were no more useful here than the AI verdict (p=.408), AI certainty (p =.911), or their own knowledge (p=.028; not significant on the Bonferroni-corrected alpha). We also conducted exploratory analysis of participants' task responses when they reported using evidence versus when they did not (see App. F.1).

**People inferred evidence quality even in the absence of source identities.** While evidence generally dominated as the most important information for evaluating claims in the qualitative data, a common criticism was that without source information people "found it difficult to fully trust" the evidence. In some cases, people

tried to reverse engineer sources by paying attention to format, language, and, most often, the use of statistics. In fact, the use of statistics was at times perceived as more credible: "if the evidence had actual numbers and data from respectable sources then usually I would side with the AI's verdict." Given that there was no source information in the evidence documents on whether the sources were "respectable" the credibility was clearly gained through the presence of numbers alone. This has implications for what kinds of evidence AI systems might present and suggests caution in how statistical evidence might be interpreted.

### 4.3 The role of subjective reliance in behavioural reliance on AI advice, explanations, and evidence

We investigated the relationships between participants' responses to the trust and belief intention scale and their propensity to rely on (i) AI advice and use (ii) AI explanations and (iii) the underlying evidence in their decisions. We interpret the trust belief and intention scales [47] as essentially measuring subjective reliance on AI output. We found a moderate correlation between people's subjective reliance on the AI system and their tendency to rely on the AI's recommendation (r=.425, p<.001), and subjective reliance predicted higher agreement with AI advice, $\beta$=0.91, SE=.135, p<.001, $R^2$=.181 (**RQ1:H1**). Participants with higher subjective reliance scores were more likely to follow AI advice. We found that all explanation conditions exhibited a negative correlation between subjective reliance and AI agreement (Uncertainty: r=-0.35, p=0.003; Verdict: r=-0.17, p=0.176; No explanation: r=-0.35, p=0.003). Participants' subjective reliance on the AI system was also correlated with tendency to use the AI explanations and underlying evidence in their decisions, although the effects were relatively small (**RQ1:H2)**. Our analysis showed a small positive correlation between participant's subjective reliance judgments and the frequency with which they used the AI explanation to make their final decision (r=.26, p<.001), and reliance predicted higher usage of AI explanations, $\beta$=0.838, SE=0.214, p<.001, $R^2$=.069. On the other hand, subjective reliance in the AI system was negatively correlated with usage of evidence (r=-.29, p<.001), with reliance predicting lower evidence use, $\beta$=-1.015, SE=0.234, p<.001, $R^2$=.084. In other words, use of evidence was associated with lower subjective reliance in the AI system, while use of explanations was associated with higher reliance.

**Using evidence and explanations to assess the reliability of AI advice.** Qualitative findings corroborated the association between low subjective reliance in the AI system and use of evidence: "I didn't trust it so had to check what it claimed". Participants also explained how trust in the AI was shaped by how accurately the outcome reflected the evidence: "I considered whether the AI's decision was based off the evidence, if their decision contradicted the evidence, this makes me feel that the AI was less trustworthy as it had not acknowledged the evidence properly". People in the Explanation conditions also discussed using logical fallacies in explanations as cues to how reliable the system's decision was. Some participants in the Uncertainty condition reported using the sources of uncertainty to "judge when to trust the system less". This suggests that explanations can play an important role in helping to determine whether the system has interpreted available evidence correctly.

## 5 Discussion

Our results show that evidence is a key component of how people make decisions when assessing AI system output, prompting important considerations for developing AI decision-support tools. We observed that having access to evidence enabled participants to evaluate AI content successfully regardless of the explanation provided, although explanations were appreciated. We also show that when people have access to evidence, explanations do not necessarily lead to overreliance (using incorrect AI advice) as previous studies have found [26, 35]. Qualitative data suggest that access to evidence can enable more critical assessment of AI output.

### 5.1 Availability is key for evidence engagement

Previous work has found that when AI systems provide people with links to external sources, relatively few (10-28%) check them [34, 35], and it is unclear to what extent people meaningfully engage with the content of

sources. In contrast, approximately two-thirds of participants in this study opened all of the available evidence documents and almost all opened at least one for each claim. Several factors may account for this difference. First, the evidence was embedded directly within the task interface. Participants could open the evidence alongside the claim, AI prediction, and explanation, instead of following a link to an external webpage (e.g., [35]), which likely reduced friction and effort associated with accessing evidence. Second, the explanations explicitly referenced the evidence, which may have encouraged participants to use it to verify the AI system's interpretation, or made it easier to identify key parts of the evidence more efficiently. However, we observed that people reported using the evidence in their decisions regardless of whether explanations were provided. It may also be the case that the fact-checking context may have been perceived as more high-stakes than the question-answering setups in previous studies [34, 35], motivating participants to consider the information more carefully, although we note that in prior studies click rates were extremely low (~7%) even for potentially high-stakes health stimuli [34]. These differences indicate that more research is needed to understand what kind of evidence is needed, in what context, and what are the best approaches to make it available to people using AI systems.

Participants were sensitive to inconsistencies between the verdict and the explanations, but even where the explanation and the verdict aligned, inconsistencies between explanation and evidence alerted them to critically consider AI system output. We also found that people relied less on the AI system's advice when they used evidence than when they did not, and were more confident in their decisions when using evidence than when they did not. As suggested by Kim et al. [35], the provision of sources themselves may lead to more deliberative reasoning, but attention must be paid to how sources are presented.

### 5.2 Evidence-focused explanations may mitigate overreliance

On the one hand, our findings suggest a more optimistic state of affairs about people's reliance on and critical engagement with AI explanations. Prior studies in AI-supported fact-checking have highlighted that natural language explanations in the form of high-level abstractive summaries, designed to simulate how humans reason about claim veracity, can persuade people to incorrect conclusions [64, 67]. However, providing natural language explanations, which extracted relevant parts of the evidence and explained how they contributed to the AI system's predicted verdict or certainty, did not lead participants to over-rely on incorrect advice in comparison to the No Explanation group. This may indicate that explanations that focus on evidence facilitate people to identify inconsistencies, flawed argumentation or other cues that the explanation may be misleading. However, these explanations also have potential for misuse: cherry-picking evidence or otherwise using evidence out of context to falsely push a certain narrative could be highly persuasive. Although our qualitative data suggested that our participants were relatively alert to when this occurred, this may be less feasible when evidence documents are longer or more complex. Furthermore, we only tested explanations generated using one type of model. Although the Verdict Explanations were prototypical justifications like those produced by many approaches (e.g. [13, 76]), future, more powerful models may have capability to produce more persuasive, misleading explanations. Our findings suggest that providing evidence appears to mitigate this risk to some degree, as users can verify if information is taken out of context, however future work in HCI is needed to improve systems and explanations that allow people to detect when plausible-sounding explanations provide misleading reasoning. Further research should examine evidence and explanation evaluation in more naturalistic, less structured task contexts and with domain experts who may interact differently with explanations than crowdworkers [64, 71].

### 5.3 Design implications for information-seeking and decision-support systems

The role of evidence in AI-assisted decision-making has received relatively little attention in the literature to date. There is some theoretical research on the use of evidence in explanations, such as the evaluative AI framework [50], which builds on an earlier interpretability approach that utilised the information theory concept of weight

of evidence [48]. However, few studies have considered the role of evidence empirically or developed methods that use it along with AI predictions and advice. This indicates a worrying research gap, especially as many commercial LLM chatbots already provide people with sources and evidence documents alongside AI output; some automatically, while others requiring explicit inquiry. As a result, there is little consensus on best practices. When current LLM implementations do give sources, LLMs struggle with reliably producing factual information [4], and often hallucinate references to external sources [1], highlighting key gaps for further development of NLP and information retrieval tools for AI-assisted fact-checking and other complex decision-making tasks.

Research in natural language explanations has largely focused on the production of fluent explanations (e.g., [3, 38]) and how users evaluate them [63]. Our findings call for NLP research to consider the role of evidence in explaining AI output and how to make it available effectively, alongside efforts to produce better explanations. Concurrently, HCI research is needed to gain fine-grained understanding of how evidence should be presented, and how evidence is evaluated and used by people in different AI-supported decision-making contexts.

### 5.4 Limitations and future work

The scope and findings of this study are subject to several additional limitations. Firstly, with the aim of carefully controlling the information available to participants to minimise potential compounds, we designed a fact-checking task using stimuli from a single (though diverse) dataset, consisting of static output and requiring a binary decision, which did not allow participants to engage in open-ended or multi-turn interactions. We selected this approach following similar studies that examine information-seeking and fact-checking (e.g., [35, 64]). However, studying how people utilise explanations and evidence in a more naturalistic task setting is an important direction for future work in this area. Secondly, we defined 'high' and 'low' certainty according to a 50% threshold. In practice, what people consider as high or low certainty is subjective and dependent on individual, cultural, and context-specific factors such as the task or stakes associated with a decision [55, 74]. Nevertheless, in our study participants relied on evidence more than any other information source, irrespective of model certainty, suggesting that the findings apply to a broad range of uncertainty scores. Finally, our experiment only examined the effects of providing natural language explanations and evidence in the context of a fact-checking task, an area which has been underexplored in prior work. Although fact-checking is a prototypical information-seeking task, similar to other question-answering tasks in which the user must make a binary judgment (e.g., [34]), future research should also examine how people use evidence in a broad range of tasks and diverse formats, especially where evidence and explanations may look quite different, such as tabular data features or image pixels.

## 6 Conclusion

We conducted a mixed methods experiment examining how evidence and explanations shape people's use of an LLM-based AI system for fact-checking. We found that the evidence underpinning the AI output was the most important consideration for participants, regardless of whether explanatory information was provided. Quantitative and qualitative data indicated that people found explanations to be useful indicators of where the AI system may be less reliable. Our results highlight the central importance of evidence in AI-assisted information-seeking and decision support. They also show how presenting evidence alongside AI outputs can substantially increase engagement with sources compared to current conventions of linking to external information. Together, these findings suggest promising directions for developing best practices for mitigating overreliance through accessible evidence and evidence-focused explanations.

## Acknowledgments


This research was co-funded by the European Union (ERC, ExplainYourself, 101077481), and supported by the Pioneer Centre for AI, DNRF grant number P1. Views and opinions expressed are those of the authors

only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

The authors wish to thank the participants who generously volunteered to participate in pilot studies, without whom this research would not have been possible.

## References

[1] Hussam Alkaissi and Samy I. McFarlane. 2023. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* 15, 2 (2023). doi:10.7759/cureus.35179

[2] Emily A. Andrews, Nathan Walter, and Erga Atad. 2025. Should We Retire the Concept of Source Credibility? An Experimental Exploration of When Credibility Is (and Is Not) Useful. *Science Communication* (2025), 10755470251334094. doi:10.1177/10755470251334094

[3] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating Fact Checking Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7352–7364. doi:10.18653/v1/2020.acl-main.656

[4] Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence* (2024), 1–12. doi:10.1038/s42256-024-00881-z

[5] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. doi:10.1145/3411764.3445717

[6] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. doi:10.1145/3442188.3445922

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[8] 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 (April 2021), 21 pages. doi:10.1145/3449287

[9] Zana Buçinca, Siddharth Swaroop, Amanda E. Paluch, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2025. Contrastive Explanations That Anticipate Human Misconceptions Can Improve Human Decision-Making Skills. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1024, 25 pages. doi:10.1145/3706598.3713229

[10] Krzysztof Budzyń, Marcin Romańczyk, Diana Kitala, Paweł Kołodziej, Marek Bugajski, Hans O. Adami, Johannes Blom, Marek Buszkiewicz, Natalie Halvorsen, Cesare Hassan, et al. 2025. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *The Lancet Gastroenterology & Hepatology* (2025). doi:10.1016/S2468-1253(25)00133-5

[11] Teodor Chiaburu, Frank Haußer, and Felix Bießmann. 2024. Uncertainty in xai: Human perception and modeling approaches. *Machine Learning and Knowledge Extraction* 6, 2 (2024), 1170–1192. doi:10.3390/make6020055

[12] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis* 16, 1 (2024), 64–93. doi:10.1093/jla/laae003

[13] Matthew R. DeVerna, Harry Yaojun Yan, Kai-Cheng Yang, and Filippo Menczer. 2024. Fact-checking information from large language models can decrease headline discernment. doi:10.48550/arXiv.2308.10800 arXiv:2308.10800 [cs].

[14] Laurence Dierickx, Arjen Van Dalen, Andreas L. Opdahl, and Carl-Gustav Lindén. 2024. Striking the balance in using LLMs for fact-checking: A narrative literature review. In *Multidisciplinary International Symposium on Disinformation in Open Online Media*. Springer, 1–15. doi:10.1007/978-3-031-71210-4_1

[15] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *International journal of human-computer studies* 58, 6 (2003), 697–718. doi:10.1016/S1071-5819(03)00038-7

[16] Ziv Epstein, Nicolo Foppiani, Sophie Hilgard, Sanjana Sharma, Elena Glassman, and David Rand. 2022. Do explanations increase the effectiveness of AI-crowd generated fake news warnings?. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 183–193. doi:10.1609/icwsm.v16i1.19283

[17] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3290605.3300724

[18] Harry Barton Essel, Dimitrios Vlachopoulos, Albert Benjamin Essuman, and John Opuni Amankwa. 2024. ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs). *Computers and Education: Artificial Intelligence* 6 (2024), 100198. doi:10.1016/j.caeai.2023.100198

[19] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160. doi:10.3758/BRM.41.4.1149

[20] Andrea Ferrario and Michele Loi. 2022. How Explainability Contributes to Trust in AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1457–1466. doi:10.1145/3531146.3533202

[21] Shannon K. Gallagher, Jasmine Ratchford, Tyler Brooks, Bryan P. Brown, Eric Heim, William R. Nichols, Scott McMillan, Swati Rallapalli, Carol J. Smith, Nathan VanHoudnos, et al. 2024. Assessing llms for high stakes applications. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice*. 103–105. doi:10.1145/3639477.3639720

[22] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50, 3 (Sept. 2024), 1097–1179. doi:10.1162/coli_a_00524

[23] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206. doi:10.1162/tacl_a_00454

[24] Lovisa Hagström, Sara Vera Marjanovic, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, and Isabelle Augenstein. 2025. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 19691–19730. doi:10.18653/v1/2025.acl-long.968

[25] Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Quantifying Uncertainty in Natural Language Explanations of Large Language Models. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 238)*, Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (Eds.). PMLR, 1072–1080. https://proceedings.mlr.press/v238/harsha-tanneru24a.html

[26] Gaole He, Nilay Aishwarya, and Ujwal Gadiraju. 2025. Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 907–924. doi:10.1145/3708359.3712133

[27] Benjamin D. Horne, Dorit Nevo, John O'Donovan, Jin-Hee Cho, and Sibel Adalı. 2019. Rating reliability and bias in news articles: Does AI assistance help everyone?. In *Proceedings of the international AAAI conference on web and social media*, Vol. 13. 247–256. doi:10.1609/icwsm.v13i01.3226

[28] Yoyo Tsung-Yu Hou and Malte F. Jung. 2021. Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 477 (Oct. 2021), 25 pages. doi:10.1145/3479864

[29] Lujain Ibrahim, Katherine M. Collins, Sunnie S. Y. Kim, Anka Reuel, Max Lamparth, Kevin Feng, Lama Ahmad, Prajna Soni, Alia El Kattan, Merlin Stein, Siddharth Swaroop, Ilia Sucholutsky, Andrew Strait, Q. Vera Liao, and Umang Bhatt. 2025. Measuring and mitigating overreliance is necessary for building human-compatible AI. arXiv:2509.08010 [cs.CY] https://arxiv.org/abs/2509.08010

[30] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (March 2023), 38 pages. doi:10.1145/3571730

[31] A.D. Kaplan, T.T. Kessler, and P.A. Hancock. 2020. How Trust is Defined and its use in Human-Human and Human-Machine Interaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 64, 1 (2020), 1150–1154. doi:10.1177/1071181320641275

[32] Charalampia Xaroula Kerasidou, Angeliki Kerasidou, Monika Buscher, and Stephen Wilkinson. 2022. Before and beyond trust: reliance in medical AI. *Journal of medical ethics* 48, 11 (2022), 852–856. doi:10.1136/medethics-2020-107095

[33] Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024. Can LLMs Produce Faithful Explanations For Fact-checking? Towards Faithful Explainable Fact-Checking via Multi-Agent Debate. arXiv:2402.07401 [cs.CL] https://arxiv.org/abs/2402.07401

[34] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 822–835. doi:10.1145/3630106.3658941

[35] Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 420, 19 pages. doi:10.1145/3706598.3714020

[36] Yoonsu Kim, Jueon Lee, Seoyoung Kim, Jaehyuk Park, and Juho Kim. 2024. Understanding Users' Dissatisfaction with ChatGPT Responses: Types, Resolving Tactics, and the Effect of Knowledge Level. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*

(Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 385–404. doi:10.1145/3640543.3645148

[37] Emir Konuk, Robert Welch, Filip Christiansen, Elisabeth Epstein, and Kevin Smith. 2024. A framework for assessing joint human-AI systems based on uncertainty estimation . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, Vol. LNCS 15010. Springer Nature Switzerland. doi:10.1007/978-3-031-72117-5_1

[38] Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking for Public Health Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 7740–7754. doi:10.18653/v1/2020.emnlp-main.623

[39] Min Hun Lee and Chong Jun Chew. 2023. Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 369 (Oct. 2023), 22 pages. doi:10.1145/3610218

[40] Q. Vera Liao and Wai-Tat Fu. 2014. Expert voices in echo chambers: effects of source expertise indicators on exposure to diverse opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 2745–2754. doi:10.1145/2556288.2557240

[41] Gionnieve Lim and Simon T. Perrault. 2024. XAI in Automated Fact-Checking? The Benefits Are Modest and There's No One-Explanation-Fits-All. In *Proceedings of the 35th Australian Computer-Human Interaction Conference* (Wellington, New Zealand) *(OzCHI '23)*. Association for Computing Machinery, New York, NY, USA, 624–638. doi:10.1145/3638380.3638388

[42] Rhema Linder, Sina Mohseni, Fan Yang, Shiva K Pentyala, Eric D Ragan, and Xia Ben Hu. 2021. How level of explanation detail affects human performance in interpretable intelligent systems: A study on explainable fact checking. *Applied AI Letters* 2, 4 (2021), e49. doi:10.1002/ail2.49

[43] Bingjie Liu. 2021. In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human–AI interaction. *Journal of computer-mediated communication* 26, 6 (2021), 384–402. doi:10.1093/jcmc/zmab013

[44] Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 7001–7025. doi:10.18653/v1/2023.findings-emnlp.467

[45] Xingyu Liu, Li Qi, Laurent Wang, and Miriam J Metzger. 2023. Checking the Fact-Checkers: The Role of Source Type, Perceived Credibility, and Individual Differences in Fact-Checking Effectiveness. *Communication Research* (2023), 00936502231206419. doi:10.1177/00936502231206419

[46] Marta Marchiori Manerba, Karolina Stanczak, Riccardo Guidotti, and Isabelle Augenstein. 2024. Social Bias Probing: Fairness Benchmarking for Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 14653–14671. doi:10.18653/v1/2024.emnlp-main.812

[47] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. The impact of initial consumer trust on intentions to transact with a web site: a trust building model. *The journal of strategic information systems* 11, 3-4 (2002), 297–323. doi:10.1016/S0963-8687(02)00020-3

[48] David Alvarez Melis, Harmanpreet Kaur, Hal Daumé III, Hanna Wallach, and Jennifer Wortman Vaughan. 2021. From human explanation to model interpretability: A framework based on weight of evidence. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9. 35–47. doi:10.1609/hcomp.v9i1.18938

[49] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38. doi:10.1016/j.artint.2018.07.007

[50] Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 333–342. doi:10.1145/3593013.3594001

[51] Sebastião Miranda, David Nogueira, Afonso Mendes, Andreas Vlachos, Andrew Secker, Rebecca Garrett, Jeff Mitchel, and Zita Marinho. 2019. Automated Fact Checking in the News Room. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 3579–3583. doi:10.1145/3308558.3314135

[52] Marvin Pafla, Kate Larson, and Mark Hancock. 2024. Unraveling the Dilemma of AI Errors: Exploring the Effectiveness of Human and Machine Explanations for Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 839, 20 pages. doi:10.1145/3613904.3642934

[53] Ioannis Papantonis and Vaishak Belle. 2025. Why not both? Complementing explanations with uncertainty, and self-confidence in human-AI collaboration. *Frontiers in Computer Science* 7 (2025), 1560448. doi:10.3389/fcomp.2025.1560448

[54] Amalie Brogaard Pauli, Isabelle Augenstein, and Ira Assent. 2025. Measuring and Benchmarking Large Language Models' Capabilities to Generate Persuasive Language. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 10056–10075. doi:10.18653/v1/2025.naacl-long.506

[55] Lawrence D. Phillips and C.N. Wright. 1977. Cultural differences in viewing uncertainty and assessing probabilities. In *Decision Making and Change in Human Affairs: Proceedings of the Fifth Research Conference on Subjective Probability, Utility, and Decision Making, Darmstadt, 1–4 September, 1975*. Springer, 507–519. doi:10.1007/978-94-010-1276-8_34

[56] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 22–32. doi:10.18653/v1/D18-1003

[57] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–67. doi:10.1145/3411764.3445315

[58] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q. Vera Liao, and Nikola Banovic. 2023. Understanding Uncertainty: How Lay Decision-makers Perceive and Interpret Uncertainty in Human-AI Decision Making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 379–396. doi:10.1145/3581641.3584033

[59] Dorian Quelle and Alexandre Bovet. 2024. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence* 7 (2024), 1341697. doi:10.3389/frai.2024.1341697

[60] R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[61] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. 2022. Understanding the Role of Explanation Modality in AI-assisted Decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization* (Barcelona, Spain) *(UMAP '22)*. Association for Computing Machinery, New York, NY, USA, 223–233. doi:10.1145/3503252.3531311

[62] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 410–422. doi:10.1145/3581641.3584066

[63] Michael Schlichtkrull, Nedjma Ousidhoum, and Andreas Vlachos. 2023. The Intended Uses of Automated Fact-Checking Artefacts: Why, How and Who. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 8618–8642. doi:10.18653/v1/2023.findings-emnlp.577

[64] Vera Schmitt, Luis-Felipe Villa-Arenas, Nils Feldhus, Joachim Meyer, Robert P. Spang, and Sebastian Möller. 2024. The Role of Explainability in Collaborative Human-AI Disinformation Detection. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2157–2174. doi:10.1145/3630106.3659031

[65] Claude E. Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27, 3–4 (1948), 379–423, 623–656. doi:10.1002/j.1538-7305.1948.tb01338.x

[66] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International journal of human-computer studies* 146 (2021), 102551. doi:10.1016/j.ijhcs.2020.102551

[67] Chenglei Si, Navita Goyal, Sherry Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III, and Jordan L. Boyd-Graber. 2024. Large Language Models Help Humans Verify Truthfulness – Except When They Are Convincingly Wrong. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Long Papers*. Association for Computational Linguistics, Mexico City, Mexico, 1459–1474. doi:10.18653/v1/2024.naacl-long.81

[68] Felix M. Simon, Rasmus Kleis Nielsen, and Richard Fletcher. 2025. *Generative AI and News Report 2025: How People Think About AI's Role in Journalism and Society*. Technical Report. Reuters Institute for the Study of Journalism, University of Oxford. doi:10.60625/risj-5bjv-yt69

[69] Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W. Mayer, and Padhraic Smyth. 2025. What large language models know and what people think they know. *Nature Machine Intelligence* (2025), 1–11. doi:10.1038/s42256-024-00976-7

[70] Jingyi Sun, Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2025. Explaining Sources of Uncertainty in Automated Fact-Checking (CLUE). *arXiv preprint* (2025). arXiv:2505.17855 https://arxiv.org/abs/2505.17855

[71] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 109–119. doi:10.1145/3397481.3450662

[72] Dennis Ulmer, Jes Frellsen, and Christian Hardmeier. 2022. Exploring Predictive Uncertainty and Calibration in NLP: A Study on the Impact of Method & Data Scarcity. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2707–2735. doi:10.18653/v1/2022.findings-emnlp.198

[73] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proceedings of the ACM on Human-Computer*

*Interaction* 7, CSCW1 (April 2023), 1–38. doi:10.1145/3579605
[74] Thomas S. Wallsten, David V. Budescu, and Ido Erev. 1988. Understanding and using linguistic uncertainties. *Acta Psychologica* 68, 1-3 (1988), 39–52. doi:10.1016/0001-6918(88)90044-3
[75] Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2025. Show Me the Work: Fact-Checkers' Requirements for Explainable Automated Fact-Checking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 197, 21 pages. doi:10.48550/arXiv.2502.09083
[76] Fengzhu Zeng and Wei Gao. 2024. Justilm: Few-shot justification generation for explainable fact-checking of real-world claims. *Transactions of the Association for Computational Linguistics* 12 (2024), 334–354. doi:10.1162/tacl_a_00649
[77] Chunpeng Zhai, Santoso Wibowo, and Lily D Li. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments* 11, 1 (2024), 28. doi:10.1186/s40561-024-00316-7
[78] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. doi:10.1145/3351095.3372852

## A Generative AI usage statement

Large Language Models (ChatGPT-5, Claude Sonnet 4) were used to assist with programming the quantitative statistical analyses, as well as with formatting the tables that appear in the paper.

## B Think-Aloud Study Participants

| ID | Age | Gender | Background |
|---|---|---|---|
| T1 | 24 | Female | Physics |
| T2 | 58 | Female | International Relations |
| T3 | 33 | Male | Computer Science |
| T4 | 66 | Female | Translation |
| T5 | 28 | Male | Sustainability |

Table 6. Demographics of participants in think-aloud study

## C Think-Aloud Study Protocol

### Background knowledge

- How familiar are you with artificial intelligence or AI? Have you used any AI systems or LLMs like ChatGPT?
- How familiar are you with fact-checking?

### Introduction to think-aloud study

- In this experiment, you'll be testing an AI system designed to help with fact-checking. You'll see some examples of claims that the AI system has been given, and its prediction about whether the claim is true or false.
- You'll see some more detailed instructions once you begin the study, and we'd like you to describe what you're thinking as you read through the information you see, for example if something is unclear or confusing.
- Then, during the main study, as you're evaluating the AI system and making decisions about the claims, we'd like you to also describe your impressions and thought processes, and talk through your decisions as you make them. At some points, I may ask you to explain a little more about why you did something.
- Then, at the end of the study I'll ask you some questions about how you found the study overall.

### Think-aloud prompts (to be asked if participant doesn't verbalise their thoughts)

- What was your first impression when you saw this screen?
- Why did you check [the evidence/explanation/AI certainty]?
- How did you decide whether the claim was true or false? What did you base your decision on?
- Why were you confident (or not) in your decision?
- Why did you find the explanation (not) useful? What was (not) useful about it?
- Did the AI's certainty score impact your decision in any way?

### Post-study interview

- How did you find the task overall? Was anything unclear or confusing about what you were asked to do?
- Did you find anything difficult about the task?
- You saw different styles of explanation in the study. Were they helpful? Why/why not?

## D Subjective Evaluation Scales

### D.1 Explanation Helpfulness

To what extent were the AI system explanations helpful in determining how reliable the system was?
(1 = Not at all helpful, 2 = Somewhat helpful, 3 = Neither, 4 = Somewhat helpful, 5 = Extremely helpful)

### D.2 Subjective reliance: Trust Belief and Intention Scales [47]

On this page you are asked to provide your opinion on the AI system you used in the experiment. Please select one answer for each statement (1 = Strongly disagree, 2 = Somewhat disagree, 3 = Neither, 4 = Somewhat agree, 5 = Strongly agree)
*Reverse-scored items

(1) The AI system is competent and effective in fact-checking claims
(2) Overall, the AI system is a capable and proficient information provider
(3) I would characterise the AI system as honest
(4) The AI system is NOT truthful in providing information to me*
(5) I believe that the AI system was developed to act in my (or the users') best interest
(6) The AI system was developed with good intentions to do its best to help me (or the users)
(7) When an important claim arises, I would NOT feel comfortable depending on the information provided by the AI system*
(8) I can always rely on the AI system to find information about claims
(9) I would feel comfortable acting on the information given to me by the AI system
(10) I would not hesitate to use the information the AI system supplied to me

### D.3 Claim familiarity

In the study, you saw 8 different real life claims, some of which you may have come across some of these claims before in your day-to-day life.
Please indicate whether you had read or come across any of these claims **before this experiment**, and if so, how much you knew about them.
**No knowledge** = I had never come across this claim before
**Limited knowledge** = I had not come across this specific claim before but I had previously come across something similar
**Some knowledge** = I had come across this claim before but I could not be sure that I knew whether it was

true/false
**Full knowledge** = I had come across this claim and was confident that I knew it was true/false

## E Main Study Participants

| Age | Gender | % | Education level | % |
|---|---|---|---|---|
| M = 41.44 | Woman | 45.67 | Doctorate degree | 5.28 |
| SD = 13.34 | Man | 51.44 | Master's degree | 13.94 |
| | Non-binary | 1.92 | Bachelor's degree | 35.58 |
| | Prefer not to answer | .96 | Some college/university | 24.52 |
| | | | High school graduate | 17.79 |
| | | | Less than high school | 1.92 |
| | | | Prefer not to answer | 0.96 |

Table 7. Participant demographic details.

## F Additional analyses

### F.1 Participants who did not use evidence

We analysed participants' task responses when they reported using evidence versus when they did not. For trials in which people used evidence, agreement with AI advice was significantly lower (50%) than when they did not (64%), $\beta$ = -0.12, SE = 0.02, t(1130.84) = -4.93, p<.001. People reported than the AI explanation was less useful when they chose to use evidence (5.36) than when they did not (4.73), $\beta$ = -0.36, SE = 0.09, t(1648.93) = -4.04, p<.001. When people used the evidence, they also self-reported higher confidence in their decision (5.82) than when they did not (5.62), ($\beta$ = 0.14, SE = 0.07, t(1623.57) = 2.12, p=.034. Finally, for trials in which participants did not report using evidence, they tended to rely most frequently on the AI explanation on average (59%), followed by AI verdict (48%), AI certainty (40%), their own knowledge (27%) and other (3%). This suggests that in these cases where people did not use evidence, many people viewed the AI output on its own as sufficient to help to make a decision.

### F.2 Claim familiarity

We collected data on how familiar participants were with the claims prior to the experiment, and on participants' confidence and subjective reliance judgements during the experiment. These results did not appear to impact the main findings of the study, but we report them here for completeness and transparency.

On average, people were unfamiliar with the claims prior to the study 58% of the time, had some knowledge 14% of the time, had limited knowledge 24.94% of the time, and had full knowledge of the claims 2.5% of the time. People relied more on their own knowledge when they had some knowledge of the claim than when they did not, t(384)=6.87, p<.001, otherwise, no differences emerged in how people used the information available to them (p>.05 for all comparisons). Verdict explanations were rated as more useful than no explanations when people had no familiarity with the claims, $\chi^2(2)$= 8.22, p=.016. Otherwise, claim familiarity did not impact in how people judged the explanations, or how much they agreed with the AI system (p>.05 for all comparisons, see Table 8).

| Condition | Familiarity | n (#responses) | Agreement (%) | Confidence (1-7) | Usefulness (1-7) |
|---|---|---|---|---|---|
| No Explanation | Any knowledge | 230 | 53.5 | 5.78 | 4.83 |
| | No knowledge | 322 | 55.7 | 5.79 | 4.60 |
| Verdict | Any knowledge | 206 | 55.3 | 5.62 | 5.04 |
| | No knowledge | 338 | 59.2 | 5.73 | 5.24 |
| Uncertainty | Any knowledge | 258 | 50.8 | 5.86 | 4.96 |
| | No knowledge | 302 | 52.6 | 5.71 | 4.99 |

Table 8. Participant AI Agreement, confidence, and usefulness ratings for claims that people had either (a) any prior knowledge of or (b) no prior knowledge of, across explantion conditions.

## F.3 Confidence & Subjective reliance

*F.3.1 Confidence.* There was no impact of explanation condition on how confident people were in their decisions, F(2, 205) = 0.428, p = .652 (see Table 9). Participants were more confident in their decisions when the AI's advice was correct than when it was incorrect, F(1, 207) = 6.122, p = .014, $\eta_p^2$ = .03, and when the AI system's certainty was high rather than low, F(1, 207) = 27.886, p < .001, $\eta_p^2$ = .12.

| Measure | Uncertainty Explanation | Verdict Explanation | No Explanation |
|---|---|---|---|
| **Confidence (1–7)** | | | |
| Overall | 5.78 | 5.70 | 5.78 |
| Correct AI advice | 5.80 | 5.81 | 5.86 |
| Incorrect AI advice | 5.76 | 5.59 | 5.70 |
| High AI certainty | 5.87 | 5.88 | 5.88 |
| Low AI certainty | 5.68 | 5.52 | 5.68 |
| **Subjective reliance (10—50)** | | | |
| Overall | 33.4 | 33.9 | 32.0 |

Table 9. Mean confidence and subjective reliance judgments by participants in each condition.

*F.3.2 Subjective reliance.* The explanation groups did not differ in their judgments for the subjective reliance scales, F(2, 205) = .091, p = 0.40.

# G Uncertainty Estimation and Explanation Generation Method

## G.1 Uncertainty Estimation

By curating the instruction, we guided the model to generate the verdict along with the explanations, see App. G.2 and G.3 for the prompts. Based on the model's verdict prediction, we derived its uncertainty as follows. To quantify model uncertainty for a verdict label on input $X$ (one claim plus two evidence passages), we compute the predictive entropy [65, 72] of the softmax distribution over logits, which requires only a single forward pass and is widely used. For candidate labels $\mathcal{Y} = \{\text{True}, \text{False}, \text{NEI}\}$ with logits $\ell(y_i)$, the probability of label $y_i$ is

$$P(y_i \mid X) = \frac{\exp\big(\ell(y_i)\big)}{\sum_{j=1}^{|\mathcal{Y}|} \exp\big(\ell(y_j)\big)}. \tag{1}$$

The uncertainty score for $X$ is the entropy of this distribution,

$$u(X) \;=\; -\sum_{y_i \in \mathcal{Y}} P(y_i \mid X)\, \log P(y_i \mid X), \tag{2}$$

### G.2 Few-shot Prompt for Verdict Explanation Generation

For the verdict explanation generation, we leveraged a three-shot prompt for instructing the model to generate the verdict along with the reasons towards this verdict, see Table 10.

### G.3 Few-shot Prompt for Uncertainty Explanation Generation

For the uncertainty explanation generation, we used a three-shot prompt to instruct the model to generate the verdict, along with the reasons to forward its verdict prediction certainty by referring to the key span interactions extracted in the previous step, following [70], see the prompts in Table 11.

## H Task Instructions

### Screen 1

**Welcome to the experiment!**
**What is this study about?**
This study is about AI systems and fact-checking. Fact-checking is the process of verifying the truthfulness of claims, statements, or stories by comparing them against credible evidence and sources.
In this study, you will see examples of an AI system designed to help people with fact-checking. Your task will be to carefully read and evaluate the information provided by the AI system.
You will see and evaluate 8 different claims in total. After you have evaluated these claims, you will answer some questions about your experiences of and opinions on the AI system and demographics.
You will have limited time to evaluate each claim. The main study will take 25-30 minutes to complete, while the questionnaire will take 5-10 minutes.
Before starting, make sure your browser window is set to full screen. Please make sure you are sitting somewhere comfortable and quiet so that you can complete the experiment in one sitting!
Click 'Next' to read the instructions for the task.

### Screen 2

**What do I have to do?**
Imagine your job is to release assessments about whether a claim is true or false. You will see claims, an AI system's prediction about whether this claim is true or false, how certain the system is about its label, and an explanation of the AI system's prediction. These explanations are intended to help you decide how to interpret the AI prediction. You have the option to see the evidence that the AI system used to make its prediction. Your task is to decide whether you can rely on the AI system prediction or need to do more research before releasing the assessment of the claim.

**Important:** Please only consider the provided information — claim, verdict, uncertainty, explanations & evidence documents (if applicable) — when evaluating explanations. Sometimes you will be familiar with the claim, but we ask you to approach each claim as new, whether or not you have seen it before. It doesn't matter whether you personally agree or disagree with the claim or evidence. We are asking you to evaluate what the AI produces: if you were to see this claim for the first time, would you find the explanation provided by the AI useful? Would you be willing to rely on the AI verdict to make a statement about the claim?

Screen 3

**About the AI System**

The AI System you will evaluate in this study is based on a Large Language Model (LLM). This AI System has been designed to help people with fact-checking. It has been trained to predict whether the claim is true or false based on the evidence it has been provided. The AI System also provides explanations for its outputs, which you will evaluate in this study.

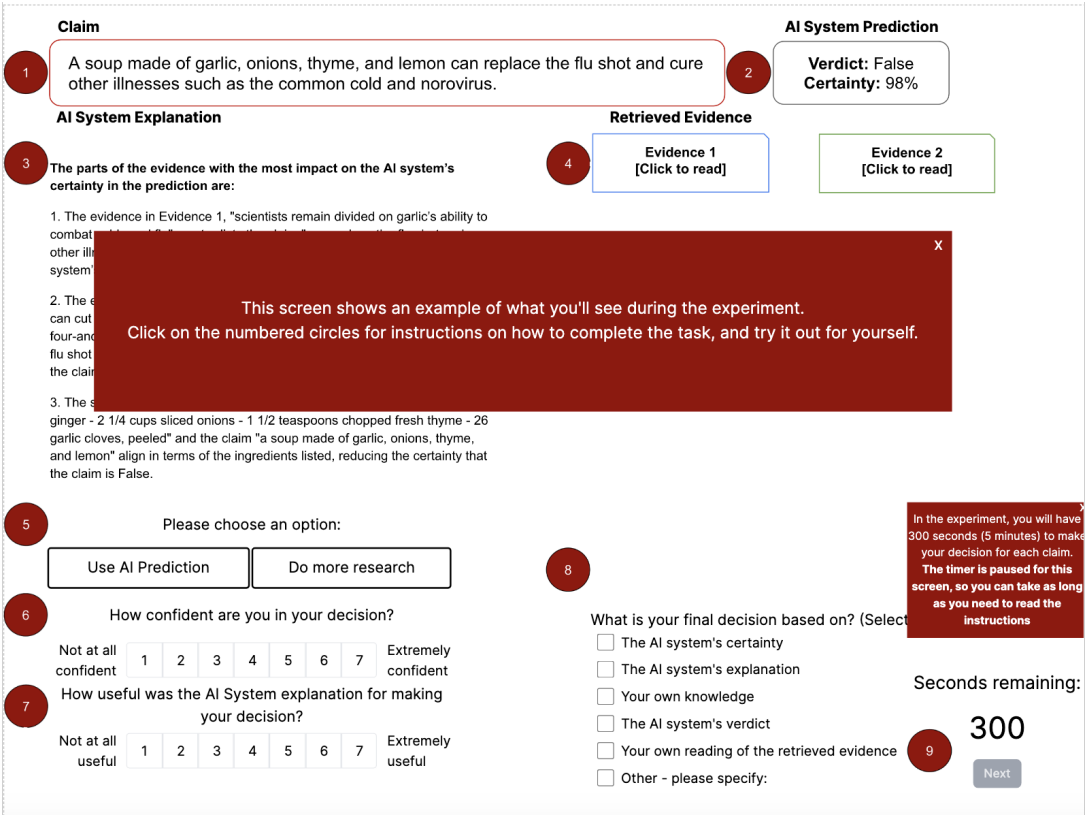
**What information will I see? (Definitions)**

- A **claim** is some statement about the world. It may be true, false, or somewhere in between.
- **Evidence** is additional information is typically necessary to verify the truthfulness of a claim. An evidence document consists of one or several sentences extracted from an external source for the particular claim. In this study, you will have the option of reading two evidence document excerpts that have been retrieved for a claim. These evidence document excerpts may or may not agree with each other. Please assume that all evidence documents are reasonably reliable.
- A **verdict** is reached based on the available evidence, regarding whether a claim is true or false.
- A **certainty score** indicates how confident an AI system about the correctness of its prediction. For example, a Certainty score of 99% indicates very high confidence that the predicted verdict is correct, but a score of 1% indicates very low confidence that the predicted verdict is correct.
- An **explanation** for the AI system's prediction explains the sources of uncertainty regarding the verdict. In other words, the explanation highlights parts of the evidence that increase or reduce its certainty that its verdict is correct.
- Your **TASK** is to decide whether you can rely on the verdict about the claim produced by the AI system, based on the information available, and how confident you are in your judgment, and decide how helpful the explanation is in making your decision.

On the next page, you will see an example of the task.

Screen 4

(1) First, read the claim that the AI system has been given.
(2) Next, read the verdict output by the AI system, along with its certainty that the output is correct.
(3) Next, read the explanation for the AI system's output. Here, the explanation highlights parts of the evidence that increase or reduce its certainty that its verdict is correct.
(4) If you want to, you can check the evidence that the AI system used to make its prediction, by clicking on 'Evidence 1' and/or 'Evidence 2'.
(5) Based on the information provided, decide whether you can rely on the verdict produced by the AI system. Try to disregard any prior knowledge or opinions you may have and focus only on the information presented here. Based on the information provided, decide whether you can rely on the verdict produced by the AI system. Try to disregard any prior knowledge or opinions you may have and focus only on the information presented here.
(6) Indicate how confident you are about your decision whether to rely on the AI system verdict or not.
(7) Rate how useful the AI system explanation above was for making your decision.
(8) Select the sources of information that you based you final decision on. There is no limit to how many you can select. If you used other sources not listed here, you can type them in 'Other'.
(9) When you have completed all these steps, press 'Next' to move on to the next claim. You have 300 seconds (5 minutes) to complete your decision for each claim. If the timer runs out before you finish, your response will not be counted.




Screen 5

Thank you for completing the practice example!
Now, you're ready to start the study. You will see 8 different claims in total.
When you're ready, click 'Next' to begin the study.

# I Study Materials

## I.1 Claim 1 (True, Correct, High Certainty)

**Claim:** 86% of Americans and 82% of gun owners support requiring all gun buyers to pass a background check.
**AI Prediction:** True
**AI Certainty:** 76%

*Evidence 1.* Statistic: 86% of Americans and 82% of gun owners support requiring all gun buyers to pass a background check, no matter where they buy the gun and no matter who they buy it from". [...] One of the survey's questions was: "Do you support or oppose requiring a background check on all gun buyers?" 86 percent of respondents said support, 7 percent said oppose, and 8 percent said unsure, according to the Giffords report on the poll. [...] A March 2019 poll conducted by Quinnipiac University found that 87 percent of national gun owners support "requiring background checks for all gun owners". The poll surveyed 1,120 national voters and had a margin of error of plus or minus 3.4 percentage points.

*Evidence 2.* Iowa voters will decide in 2022 whether to add a state Constitution amendment affirming the right to bear arms. Support for the amendment has fallen along party lines. Opposing Democrats refer to a poll that says an overwhelming number of even gun owners support background checks for gun buyers.

*Uncertainty Explanation.* The parts of the evidence with the most impact on the AI system's certainty in the prediction are:
1. The evidence in Evidence 1, "86% of Americans and 82% of gun owners support requiring all gun buyers to pass a background check", aligns perfectly with the claim "86% of Americans and 18% of gun owners support requiring all gun buyers to pass a background check". This consistency increases the certainty that the claim is true.
2. The agreement between the phrase "gun buyers" in the claim and "gun" in Evidence 1 indicates that the focus on background checks applies uniformly across gun-related transactions and increases the certainty in the prediction.
3. Lastly, the phrase "support" in the claim and "owners" in Evidence 1, while seemingly unrelated in terms of direct content, contribute to the overall coherence of the argument. The emphasis on the level of support among gun owners corroborates the claim's specific statistic, and increases certainty about the alignment between the claim and the provided evidence.

*Verdict Explanation.* The parts of the evidence with the most impact on the AI system's predicted verdict are:
1. Evidence 1 directly supports the claim, stating that "86% of Americans and 82% of gun owners support requiring all gun buyers to pass a background check, no matter where they buy the gun and no matter who they buy it from". The poll details show that when asked, "Do you support or oppose requiring a background check on all gun buyers?", 86% responded "support". This aligns exactly with the percentages in the claim.
2. Evidence 1 cites a Quinnipiac University poll reporting that "87 percent of national gun owners support requiring background checks for all gun owners", which reinforces the claim's assertion of broad support among both the general public and gun owners.
3. Evidence 2 does not give exact percentages but echoes the same conclusion. It states that "an overwhelming number of even gun owners support background checks for gun buyers", which is consistent with the data presented in Evidence 1. While it is less specific, it corroborates the general finding of widespread support.

*No Explanation.* The AI system has judged this claim to be True, with 76% certainty.

### I.2 Claim 2 (False, Correct, High Certainty)

**Claim:** A satanic-themed hotel is opening in Texas.
**AI Prediction:** False
**AI Certainty:** 67%

*Evidence 1.* Baphomet bedside buddies, upside down crosses, spewing Satan sinks and a devilish welcome, one Facebook user claimed a satanic hotel would soon be opening in Plano, Texas. But don't pull out the rosaries and holy water yet, the hotel is not actually real. On Feb. 23, a Facebook user claimed that a satanic hotel was supposedly opening in an abandoned office building in the heart of the Downtown Plano Art District. [...] But the Baphomet-themed hotel won't be coming anywhere near Plano. The images were created using an AI art platform. According to The Buzz, an Instagram account posted photos showing how AI art can be made using a specific algorithm. In this case, a satanically-spooky hotel.

*Evidence 2.* A video shared on Facebook allegedly shows images of a satanic-themed hotel that is set to open in Plano, Texas. [...] A Facebook video allegedly shows photos of a Satanic-themed hotel that is set to open in Plano on June 6. The post shares a photo of a red bedroom with a goat-like figure. "This is a satanic hotel located in Plano, Texas" the caption reads. "This brand new hotel will open June 6 at 6pm (666)." The claim is fabricated.

Check Your Fact found no credible news reports about a satanic-themed hotel opening in Plano, Texas. The photo originates from Ink Poisoning, an apparel brand. The company posted the images on Facebook with a disclaimer stating the images are "AI concept art created by us".

*Uncertainty Explanation.* The parts of the evidence with the most impact on the AI system's certainty in the prediction are:
1. The evidence in Evidence 1, "Baphomet bedside buddies, upside down crosses, spewing Satan sinks and a devilish welcome, one Facebook user claimed a satanic hotel would soon be opening in Plano, Texas", is somewhat related to the claim "A satanic-themed hotel is opening in Texas". This reduces the AI system's certainty that the claim is false.
2. The evidence in Evidence 1, "the hotel is not actually real", is in direct contradiction to the claim "A satanic-themed hotel is opening in Texas". This statement increases the certainty that the claim is false by clarifying that the images and descriptions of the hotel are merely AI-generated concepts and not an actual establishment.
3. The agreement between the evidence in Evidence 1, that "the images were created using an AI art platform", and the phrase "supposedly opening" increase AI system uncertainty that the claim is false.

*Verdict Explanation.* The parts of the evidence with the most impact on the AI system's predicted verdict are:
1. Evidence 1 makes clear that the idea of a satanic-themed hotel opening in Plano, Texas, comes from a social media post and is not real. It explains that a Facebook user claimed "a satanic hotel would soon be opening in Plano, Texas", but the article quickly clarifies that "the hotel is not actually real". The supposed images of the "Baphomet bedside buddies, upside down crosses, spewing Satan sinks and a devilish welcome" were not photos of a real hotel but instead "created using an AI art platform".
2. The Buzz further explained that the Instagram account that posted them did so to show "how AI art can be made using a specific algorithm". This confirms the images were fictional concept art, not evidence of an actual hotel.
3. Evidence 2 supports this conclusion. It describes how a Facebook video claimed "a satanic-themed hotel… is set to open in Plano on June 6" and even added details like "open June 6 at 6pm (666)". However, the evidence directly states that "the claim is fabricated" and that "no credible news reports" support the idea of such a hotel opening. It also identifies the true origin of the images: they came from Ink Poisoning, an apparel brand, which shared them with a disclaimer noting they were "AI concept art created by us".

*No Explanation.* The AI system has judged this claim to be False, with 67% certainty.

### I.3 Claim 3 (True, Correct, Low Certainty)

**Claim:** A scone can equal a third of one's recommended daily calories.
**AI Prediction:** True
**AI Certainty:** 35%

*Evidence 1.* An average scone equals a fifth of recommended daily calories for females, and a sixth for males. The number of calories tends to be proportionate to the size of a scone rather than the luxuriousness of its filling.

*Evidence 2.* On 11 February 2019, a headline of the Irish News and the Belfast Telegraph stated that "One large scone can equal a third of recommended daily calories". Both articles used a Press Association report that referred to "a survey". [...] The report states that "the highest calorie scone, without the addition of spread or jam, provides over a third (38%) of the recommended daily calorie intake". [...] For people aged 19 to 64, the Government Dietary Recommendations suggest a daily intake of 2,000kcal for females and 2,500kcal for males. The largest scone of the survey, then, indeed equals a third of recommended calories; taking up 37.8% of a daily female diet, and 30.2% of a daily male diet.

*Uncertainty Explanation.* The parts of the evidence with the most impact on the AI system's certainty in the prediction are: 1. The evidence in Evidence 1, "An average scone equals a fifth of recommended daily calories for females, and a sixth for males", aligns with the claim "A scone can equal a third of one's recommended daily calories". Although the specific fractions differ, the core message that a scone can be a substantial contributor to daily calorie consumption is consistent, increasing the certainty that the claim is true.
2. The evidence in Evidence 2, "without the addition of spread or jam, provides over a third (38%) of the recommended daily calorie intake", corroborates the claim by providing a concrete example where a larger scone meets or exceeds the claim's assertion. This specific instance strengthens the certainty that the claim is true.
3. The evidence in Evidence 2, "the largest scone of the survey, then, indeed equals a third of recommended calories; taking up 37.8% of a daily female diet, and 30.2% of a daily male diet", further reinforces the certainty that the claim is true by offering precise figures that closely match the claim's statement.

*Verdict Explanation.* The parts of the evidence with the most impact on the AI system's predicted verdict are: 1. Evidence 1 states that an average scone provides "a fifth of recommended daily calories for females, and a sixth for males". This shows that even a regular-sized scone contains a significant share of daily energy needs.
2. Evidence 1 also emphasizes that calorie content is more closely tied to the size of the scone than to its filling, suggesting that larger scones will contribute an even higher proportion of daily calories.
3. Evidence 2 confirms that in some cases a single scone can equal about a third of recommended calories. It explains that "the highest calorie scone, without the addition of spread or jam, provides over a third (38%) of the recommended daily calorie intake". This is further broken down to "37.8% of a daily female diet, and 30.2% of a daily male diet", based on standard intake levels of 2,000 kcal for women and 2,500 kcal for men.

*No Explanation.* The AI system has judged this claim to be True, with 35% certainty.

### I.4 Claim 4 (False, Correct, Low Certainty)

**Claim:** A windmill could spin until it falls apart and never generate as much energy as was invested in building it.
**AI Prediction:** False
**AI Certainty:** 31%

*Evidence 1.* No, a professor didn't say a windmill will 'never generate' the energy invested in building it [...] Now, we have a meme roaming social media that claims Canadian author and professor Thomas Homer-Dixon said that windmills will "never" generate as much energy as was expended to build the structures. [...] The meme does get a part of the quote right, but it strategically leaves out large swaths of information in order to give people the impression that all windmills never expend the amount of energy that was invested in it, which is not accurate. The meme's omitted section explains that, while poorly placed windmills may never generate enough energy payback, a good wind site could generate it in three years or less.

*Evidence 2.* Factually Inaccurate: Analyzing the energy used in the production of wind turbines does not show that it's greater than the electricity produced over the turbine's working lifetime.

*Uncertainty Explanation.* The parts of the evidence with the most impact on the AI system's certainty in the prediction are:
1. The evidence in Evidence 1, "No, a professor didn't say a windmill will 'never generate' the energy invested in building it", directly contradicts the claim's assertion that windmills will never generate sufficient energy. This contradiction increases the AI system's certainty that the claim is False.
2. The evidence in Evidence 1, "will 'never generate' the energy invested", aligns with the part of the claim "spin until it falls apart and never generate as much energy as was invested", reducing the AI system's certainty that

the claim is False.
3. The evidence in Evidence 1, "expended to", aligns with the phrase "invested in building" in the claim, indicating agreement on the energy expenditure aspect, reducing the certainty that the claim is False.

*Verdict Explanation.* The parts of the evidence with the most impact on the AI system's predicted verdict are:
1. Evidence 1 explains that the claim comes from a misrepresented quote. The meme states that windmills "will never generate as much energy as was expended to build the structures", but the original quote clarifies that "while poorly placed windmills may never generate enough energy payback, a good wind site could generate it in three years or less". This shows the claim exaggerates and omits important context.
2. The same evidence highlights that well-sited windmills can generate enough energy to offset their construction cost quickly, making the claim that all windmills 'never generate' enough energy inaccurate. The phrase "a good wind site could generate it in three years or less" directly contradicts the claim.
3. Evidence 2 reinforces this, stating that "analyzing the energy used in the production of wind turbines does not show that it's greater than the electricity produced over the turbine's working lifetime". This confirms that, over time, turbines produce more energy than was invested in building them.

*No Explanation.* The AI system has judged this claim to be False, with 31% certainty.

### I.5 Claim 5 (True, Incorrect, High Certainty)

**Claim:** All tea bags contain harmful microplastics.
**AI Prediction:** True
**AI Certainty:** 73%

*Evidence 1.* At ORGANIC INDIA, we believe in plastic-free tea bags to protect your body and the earth; and preserve the pure, organic tea experience that you deserve. Unfortunately, the industry norm is to include plastic in the mesh tea bag, which releases billions of microplastics and nanoplastics into your infusion. Consuming microplastics, as you can imagine, can have negative impacts on your health and on the earth. [...] Yes, but not all tea bags. The vast majority of brands on the shelves have mesh tea bags that are composed of 20-30% plastic. According to studies, a single standard tea bag releases 11.6 billion microplastics and 3.1 billion nanoplastics into every cup of tea. Consumers are increasingly becoming aware of this fact and opting either for loose teas, or selecting tea brands that have no plastic in the teabag. ORGANIC INDIA tea bags do not contain plastic, and are made instead of unbleached filter paper that is verified home compostable, biodegradable, Non-GMO, allergen-free and chemical-free. [...] In short, yes. Microplastics in tea bags can be harmful not only for the environment, but for human health as well. Microplastics may enter the human system through ingestion, inhalation or dermal absorption. Studies have shown that microplastics have potential to cause damage to the cells in the quantities currently found in food and beverages like tea.

*Evidence 2.* Therefore, the use of plastic in tea bags has come under scrutiny due to the potential release of microplastics when exposed to heat. As we mentioned earlier, many tea bag manufacturers use nylon or polyethylene terephthalate (PET) in their tea bags, which are not biodegradable and can release harmful microplastics when exposed to heat. These microplastics can end up in your drink, and studies have shown that they may pose a risk to human health. [...] Another study published in the journal Environmental Science & Technology in 2020 found that plastic tea bags could release up to 16 micrograms of microplastics per liter of tea when brewed at a temperature of 95°C. These studies highlight the importance of choosing plastic-free tea bags to minimize exposure to potentially harmful plastics. [...] Yes, some tea bags can contain microplastics, particularly those made with plastic materials like nylon or polypropylene. However, The Tea Spot's sachets are made from sugar

cane fibers, which are biodegradable, compostable, and completely free from petroleum-based plastic. Therefore, you can rest assured that when you brew our tea, you're not consuming any harmful plastics or microplastics.

*Uncertainty Explanation.* The parts of the evidence with the most impact on the AI system's certainty in the prediction are:
1. The statement in Evidence 2, "Therefore, the use of plastic in tea bags has come under scrutiny due to the potential release of microplastics when exposed to heat", does not support the blanket statement that "All tea Bags contain harmful microplastics". This disagreement reduces the AI system certainty that the claim is True.
2. The statement in Evidence 1, "Yes, but not all tea bags. The vast majority of brands on the shelves have mesh tea bags that are composed of 10-30% plastic", contradicts the claim by indicating that not all tea bags contain harmful microplastics, reducing certainty that the claim is True.
3. The statement in Evidence 1, "ORGANIC INDIA tea bags do not contain plastic, and are made instead of unbleached filter paper that is verified home compostable, biodegradable, Non-GMO, allergen-free and chemical-free", undermines the claim by providing a specific example of a brand that produces tea bags without harmful microplastics, further reducing certainty that the claim is True.

*Verdict Explanation.* The parts of the evidence with the most impact on the AI system's predicted verdict are:
1. Evidence 1 states that "the vast majority of brands on the shelves have mesh tea bags that are composed of 20–30% plastic". This shows that most commercially available tea bags contain plastic, making the exposure to microplastics nearly universal for consumers using standard tea bags.
2. Evidence 1 notes that "a single standard tea bag releases 11.6 billion microplastics and 3.1 billion nanoplastics into every cup of tea", and Evidence 2 confirms that materials like nylon or PET "can release harmful microplastics when exposed to heat". Both sources directly link the presence of plastic in tea bags to microplastic contamination in brewed tea.
3. Evidence 1 explains that microplastics "can be harmful not only for the environment, but for human health as well" and "may cause damage to the cells in the quantities currently found in food and beverages like tea". Evidence 2 echoes this, stating that microplastics "may pose a risk to human health".

*No Explanation.* The AI system has judged this claim to be True, with 73% certainty.

### I.6 Claim 6 (False, Incorrect, High Certainty)

**Claim:** Lagan Valley is one of the least wooded areas in Northern Ireland, particularly with regards to ancient woodland.
**AI Prediction:** False
**AI Certainty:** 78%

*Evidence 1.* The Lagan Valley (Irish: Cluain an Lagáin, Ulster Scots: Glen Lagan) is an area of Northern Ireland between Belfast and Lisburn. The Lagan is a famous river that flows into Belfast Lough. For a section, the river forms part of the border between the counties of Antrim and Down. It has a number of interesting features including a towpath which runs alongside the River Lagan. The towpath is popular with walkers, runners, cyclists, dog owners etc. It is a very scenic and peaceful area and is ideal for walking, cycling etc. The towpath begins in the Stranmillis area of south Belfast and runs all the way to Lisburn. The cycle route forms part of National Cycle Route 9. There are a number of "off route" mountain bike trails along the towpath. Lagan Valley is an Area of Outstanding Natural Beauty (AONB). The AONB was established in 1965 and the greater part of it lies within the Greater Belfast area. [...] Belvoir Park Forest Education Centre

*Evidence 2.* Therefore, it is fair to say that Lagan Valley is one of the least wooded areas in NI. [...] Lisburn & Castlereagh is the least densely wooded council (3.7%) in Northern Ireland, followed by Armagh, Banbridge &

Craigavon (4.0%) and Antrim & Newtownabbey (4.4%). [...] Based on all this, Ms Eastwood's claim that "Lagan Valley is one of the least wooded areas in NI" is backed by evidence – although the precise figures are available broken down by council areas, rather than Westminster/Assembly constituencies.

*Uncertainty Explanation.* The parts of the evidence with the most impact on the AI system's certainty in the prediction are:
1. The evidence in Evidence 1, "Lagan Valley is an Area of Outstanding Natural Beauty (AONB)", does not directly address the claim about the level of woodlands in the area, reducing the certainty of the AI system's prediction.
2. The evidence in Evidence 2, "Therefore, it is fair to say that Lagan Valley is one of the least wooded areas in NI", directly supports the claim that Lagan Valley is one of the least wooded areas in Northern Ireland, particularly with regards to ancient woodland. This agreement reduces the certainty that the claim is False.
3. The evidence in Evidence 2 that "Lisburn & Castlereagh is the least densely wooded council (3.7%) in Northern Ireland" aligns with the evidence in Evidence 1 that the Lagan Valley "is an area of Northern Ireland between Belfast and Lisburn". This reduces the certainty that the claim is False.

*Verdict Explanation.* The parts of the evidence with the most impact on the AI system's predicted verdict are:
1. Evidence 1 describes Lagan Valley as "an Area of Outstanding Natural Beauty (AONB)" and mentions that it "has a number of interesting features including a towpath which runs alongside the River Lagan". It also emphasizes that the area is "very scenic and peaceful" and "ideal for walking, cycling etc.", suggesting the presence of natural green spaces rather than a sparsely wooded landscape.
2. Evidence 2 notes that "Lisburn & Castlereagh is the least densely wooded council (3.7%) in Northern Ireland, followed by Armagh, Banbridge & Craigavon (4.0%) and Antrim & Newtownabbey (4.4%)". It adds that Ms Eastwood's claim that "Lagan Valley is one of the least wooded areas in NI" is "backed by evidence – although the precise figures are available broken down by council areas, rather than Westminster/Assembly constituencies", showing that the claim is not directly supported for Lagan Valley itself. 3. Neither piece of evidence provides specific data about ancient woodland in Lagan Valley. The lack of precise figures for the area means the claim about ancient woodland cannot be verified.

*No Explanation.* The AI system has judged this claim to be False, with 78% certainty.

### I.7 Claim 7 (True, Incorrect, Low Certainty)

**Claim:** Cristiano Ronaldo was the only European soccer captain who did not wear a rainbow armband at Euro 2020.
**AI Prediction:** True
**AI Certainty:** 29%

*Evidence 1.* Cristiano Ronaldo is refusing to wear the "One Love" armband, which advocates for the LGBTQ+ community, during the 2022 World Cup. The world-renowned soccer player, Cristiano Ronaldo, revealed in a press conference that he will not be participating in wearing the "One Love" armband. The armband consists of a heart with a rainbow inside of it and the number one. This "One Love" band is being used to advocate inclusivity, being that the 2022 World Cup is being held in Qatar, a country where same-sex relationships are illegal. [...] Players who wear the armband are threatened with receiving a yellow card, according to FIFA's regulations. This yellow card can result in a suspension from the games if the player were to receive another yellow card during the game. Obviously, any mention of social issues is discouraged on the field. Due to the consequences, none of the captains from any of the other national teams have worn the armband, but none have faced backlash like Ronaldo has.

*Evidence 2.* There is evidence that three team captains — from Germany, England, and Denmark — wore rainbow or pro-LGBTQ+ armbands during this tournament, which included 24 teams. Germany's captain, Manuel Neuer, was briefly investigated by the Union of European Football Associations for his rainbow armband. Cristiano Ronaldo was among many other captains who wore normal armbands.

*Uncertainty Explanation.* The parts of the evidence with the most impact on the AI system's certainty in the prediction are:
1. The evidence in Evidence 1, "team captains — from Germany, England, and Denmark — wore rainbow or pro-LGBTQ+ armbands during this tournament", is not consistent with the part of the claim "Cristiano Ronaldo was the only European soccer captain who did not wear a rainbow armband at Euro 2020". This indicates that there were multiple captains who did not wear the armband, not just Ronaldo and reduces AI system certainty that the claim is True.
2. The evidence in Evidence 1, "Cristiano Ronaldo was among many other captains who wore normal armbands", is not consistent with the part of the claim "Cristiano Ronaldo was the only European soccer captain.", implying he was not the only captain without a rainbow armband, reducing certainty that the claim is True.
3. The evidence in Evidence 1, "during this tournament", is not consistent with the part of the claim "Euro 2020", as it refers to a different event. This discrepancy highlights that the evidence discusses a different tournament, possibly the 2022 World Cup, rather than Euro 2020 and reduces certainty that the claim is True.

*Verdict Explanation.* The parts of the evidence with the most impact on the AI system's predicted verdict are:
1. Evidence 1 notes that Ronaldo "is refusing to wear the 'One Love' armband, which advocates for the LGBTQ+ community," and describes how he revealed in a press conference that he "will not be participating in wearing the 'One Love' armband." The evidence emphasizes that while "players who wear the armband are threatened with receiving a yellow card", Ronaldo's decision not to wear it drew notable attention, suggesting he stood out among team captains. The passage also clarifies that "none of the captains from any of the other national teams have worn the armband" at other tournaments, highlighting Ronaldo's high-profile refusal.
2. Evidence 2 provides additional support, stating that "three team captains — from Germany, England, and Denmark — wore rainbow or pro-LGBTQ+ armbands during this tournament". It also specifically mentions that "Cristiano Ronaldo was among many other captains who wore normal armbands", directly confirming that he did not wear the rainbow armband while other European captains did.

*No Explanation.* The AI system has judged this claim to be True, with 29% certainty.

### I.8 Claim 8 (False, Incorrect, Low Certainty)

**Claim:** The top issue for college students is the economy.
**AI Prediction:** False
**AI Certainty:** 49%

*Evidence 1.* "The everyday concerns of Americans are the economy — even young people that we work with at Young America's Foundation, our polling shows nationwide, their top issue for college students is the economy", Walker said. [...] Is Walker right that "The top issue for college students is the economy"? [...] The economy was the top issue for college students who participated in the group's poll. Wisconsin pollster Charles Franklin said the polling company is reputable and the questions were overall evenhanded.

*Evidence 2.* The midterm election is right around the corner and the economy and abortion rights are the top issues of concern for college students according to a new survey by BestColleges.com. These issues are also ranked as the most important for Americans overall. BestColleges.com surveyed over 1000 students about their political beliefs and plans to vote. Thirty-seven percent (37%) of those surveyed noted the economy, inflation, and

employment as the most important issues in this election. Students are worried about finding jobs, their future finances, and the growing cost of everything from housing to peanut butter. White college men, in particular, are overwhelmingly concerned about the economy, employment, and inflation, with almost half (49%) selecting these three issues as their top political issues — a significantly higher percentage than other demographic groups surveyed. Abortion rights, which have been the subject of heated national and local political campaigns, were ranked as a top concern by college students with 33% of those surveyed ranking these rights as a top issue. College women were particularly concerned about legislation having an impact on their bodies. While college men chose the economy as their top issue (44%), abortion rights were the most important issue for college women (43%), with women of color expressing the deepest concern. Only 21% of college men considered abortion rights an important issue. For college men, the economy, gun policy, healthcare, racial and ethnic inequality, and climate change all came out ahead of abortion rights as leading issues.

*Uncertainty Explanation.* The parts of the evidence with the most impact on the AI system's certainty in the prediction are:
1. The evidence in Evidence 1, "The economy was the top issue for college students who participated in the group's poll", is consistent with the claim "The top issue for college students is the economy", reducing AI system certainty that the claim is False. 2. The evidence in Evidence 2, "abortion rights were the most important issue for college women (43%)" contrasts with the claim, increasing certainty that the claim is False.
3. The statement in Evidence 1, "The everyday concerns of Americans are the economy", is consistent with the part of Evidence 2, "Students are worried about finding jobs, their future finances, and the growing cost of everything", reducing AI system certainty that the claim is False.

*Verdict Explanation.* The parts of the evidence with the most impact on the AI system's predicted verdict are:
1. Evidence 1 reports that, according to polling by Young America's Foundation, "the economy was the top issue for college students who participated in the group's poll". While this appears to support the claim, the evidence also raises the question: "Is Walker right that 'The top issue for college students is the economy'?" suggesting uncertainty about whether this finding is representative of all college students.
2. Evidence 2 provides more detailed data, showing that college students are concerned about multiple issues. It states that "the economy and abortion rights are the top issues of concern for college students", with 37% ranking the economy and 33% ranking abortion rights as their primary concern. The evidence further breaks this down by gender: "college men chose the economy as their top issue (44%)", whereas "abortion rights were the most important issue for college women (43%), with women of color expressing the deepest concern". This demonstrates that the economy is not universally the top issue across all college students.

*No Explanation.* The AI system has judged this claim to be False, with 49% certainty.

| | |
|---|---|
| **Task Description:** You are a helpful assistant. Here are your tasks: 1. Supply the verdict for the claim based on the relationship between the claim and the two evidence passages. 2. Explain your reasoning for the verdict prediction by identifying the three most influential span interactions from Claim-Evidence 1, Claim-Evidence 2, and Evidence 1-Evidence 2, and describing how each interaction (agree, disagree, or unrelated) affects your overall prediction. **Return format:** [ Verdict Prediction ] [ Explanation ] | |
| **SHOT 1 Input** | **Claim:** Winston Churchill once stated, "A pessimist sees the difficulty in every opportunity; an optimist sees the opportunity in every difficulty".<br>**Evidence 1:** … "A pessimist", says the Rev. F. W. Cole, president of the Cardiff Free Church Council, …<br>**Evidence 2:** Winston Churchill's quote,"A pessimist sees the difficulty in every opportunity; an optimist sees the opportunity in every difficulty", … |
| **SHOT 1 Output** | **Verdict Prediction:** False **Explanation:** The parts of the evidence with the most impact on my predicted verdict are: The evidence in Evidence 1, "says the Rev. F. W. Cole", conflicts with the part of the claim "Winston Churchill once stated"; the evidence in Evidence 2 that "Winston Churchill's quote, 'A pessimist sees the difficulty in every opportunity …' " aligns with the claim's attribution; taken together, the attribution in Evidence 1 ("Rev. F. W. Cole") versus Evidence 2 ("Winston Churchill") is inconsistent, which weighs against the claim. |
| **SHOT 2 Input** | **Claim:** During the last hundred years the temperature is increased about 0.1°C because of carbon dioxide. The human contribution was about 0.01°C.<br>**Evidence 1:** "The likely range of total human-caused global surface temperature increase … is 0.8°C to 1.3°C (best estimate 1.07°C [2.01 °F]). Over this period, well-mixed greenhouse gases (GHGs) contributed a warming of 1.0°C to 2.0°C, and other human drivers (principally aerosols) contributed a cooling of 0.0°C to 0.8°C, natural (solar and volcanic) drivers changed global surface temperature by −0.1°C to +0.1°C, and internal variability changed it by −0.2°C to +0.2°C."<br>**Evidence 2:** "After using a bunch of nonsense calculations to say that 0.1 °C warming is from CO2, they say that 90% of the change in CO2 is caused by the oceans. … the oceans are actually absorbing CO2 … their conclusions cannot be supported." |
| **SHOT 2 Output** | **Verdict Prediction:** False **Explanation:** The parts of the evidence with the most impact on my predicted verdict are: Evidence 1 reports human-caused warming near 1°C overall and 1.0–2.0°C from well-mixed greenhouse gases, contradicting the claim's "about 0.1°C because of carbon dioxide" and "about 0.01°C" human contribution; Evidence 2 challenges the assertion that "0.1 °C warming is from CO2", which also goes against the claim; the segment in Evidence 1 about aerosol cooling and natural drivers (−0.1°C to +0.1°C) is unrelated to Evidence 2's point about oceans absorbing CO2. |
| **SHOT 3 Input** | **Claim:** Vitamin D appears increase COVID-19 mortality rates<br>**Evidence 1:** If vitamin D does in fact reduce the severity of COVID-19 in regard to pneumonia/ARDS, inflammation, inflammatory cytokines and thrombosis, … supplements would offer a relatively easy option to decrease the impact of the pandemic.<br>**Evidence 2:** A Vitamin D deficiency that is not sufficiently treated is associated with COVID-19 risk. |
| **SHOT 3 Output** | **Verdict Prediction:** False **Explanation:** The parts of the evidence with the most impact on my predicted verdict are: Evidence 1 indicates that vitamin D reduces COVID-19 severity, which opposes the claim that vitamin D increases mortality; Evidence 2 links untreated vitamin D deficiency with higher COVID-19 risk, which also runs against the claim; Evidence 1's indication of benefit from adequate vitamin D is consistent with Evidence 2's association between deficiency and risk, jointly weighing against the claim. |
| **NEW INSTANCE** | **Claim:** { CLAIM } **Evidence 1:** { E1 } **Evidence 2:** { E2 } **Your answer:** |

Table 10. Verdict explanation generation prompt

| | |
|---|---|
| **Task Description:** You are a helpful assistant. Here are your tasks : 1. Supply the verdict for the claim based on the relationship between the claim and the two evidence passages. 2. Explain your certainty for your verdict prediction by referring to the three span interactions provided below ( Claim-Evidence 1 , Claim-Evidence 2 , Evidence 1-Evidence 2) and describing how each interaction's relation (agree, disagree, or unrelated) affects your overall certainty . **Return format:** [ Verdict Prediction ] [ Explanation ] | |
| **SHOT 1 Input** | **Claim:** Winston Churchill once stated, "A pessimist sees the difficulty in every opportunity; an optimist sees the opportunity in every difficulty".<br>**Evidence 1:** … " A pessimist", says the Rev. F. W. Cole, president of the Cardiff Free Church Council, …<br>**Evidence 2:** Winston Churchill's quote, "A pessimist sees the difficulty in every opportunity; an optimist sees the opportunity in every difficulty", …<br>**Span interactions (Pre-filled):** (1) "Winston Churchill once stated" - "says the Rev. F. W. Cole" (C - E1) relation: [disagree]; (2) "Winston Churchill once stated" - "Winston Churchill's quote, 'A pessimist sees the difficulty in every opportunity' " (C - E2) relation: [agree]; (3) " 'A pessimist', says the Rev. F. W. Cole, president of the Cardiff Free Church Council" - "Winston Churchill's quote, 'A pessimist sees the difficulty in every opportunity' " (E1 - E2) relation: [disagree] |
| **SHOT 1 Output** | **Verdict Prediction:** False **Uncertainty Explanation:** The parts of the evidence with the most impact on my certainty in the prediction are: the claim's phrase "Winston Churchill once stated" conflicts with Evidence 1's attribution to "the Rev. F. W. Cole", which decreases my certainty in the claim; the same claim wording matches Evidence 2, which introduces the line as "Winston Churchill's quote", increasing my certainty; since one passage attributes the quote to Rev. Cole and the other to Churchill, the evidences disagree with each other, and that inconsistency lowers my overall certainty and supports predicting False. |
| **SHOT 2 Input** | **Claim:** During the last hundred years the temperature is increased about 0.1°C because of carbon dioxide. The human contribution was about 0.01°C.<br>**Evidence 1:** … well-mixed greenhouse gases (GHGs) contributed a warming of 1.0°C to 2.0°C …<br>**Evidence 2:** … "0.1 °C warming is from CO2 … cannot be supported". …<br>**Span interactions (Pre-filled):** (1) "temperature is increased about 0.1°C because of carbon dioxide" - "well-mixed greenhouse gases (GHGs) contributed a warming of 1.0°C to 2.0°C" (C - E1) relation: [disagree]; (2) "0.1°C warming is from CO2" - "0.1 °C warming is from CO2 … cannot be supported" (C - E2) relation: [disagree]; (3) "well-mixed greenhouse gases … 1.0°C to 2.0°C" - "0.1 °C warming is from CO2 … cannot be supported" (E1 - E2) relation: [agree] |
| **SHOT 2 Output** | **Verdict Prediction:** False **Uncertainty Explanation:** The parts of the evidence with the most impact on my certainty in the prediction are: the claim's "about 0.1°C because of carbon dioxide" contrasts with Evidence 1's "1.0°C to 2.0°C" contribution from well-mixed greenhouse gases, which pushes my certainty toward False; Evidence 2 explicitly states that "0.1 °C warming is from CO2 ... cannot be supported", further moving my certainty toward False; together, the larger range in Evidence 1 and the rejection of the 0.1 °C figure in Evidence 2 reinforce each other, increasing my certainty in the False prediction. |
| *SHOT 3 omitted for space (structure identical to SHOT 1 and SHOT 2).* | |
| **NEW INSTANCE** | **Claim:** { CLAIM }<br>**Evidence 1:** { E1 }<br>**Evidence 2:** { E2 }<br>**Span interactions (Pre-filled):** (1) '{ SPAN1 - A }' - '{ SPAN1 - B }' (C - E1) relation: { REL1 }; (2) '{ SPAN2 - A }' - '{ SPAN2 - B }' (C - E2) relation: { REL2 }; (3) '{ SPAN3 - A }' - '{ SPAN3 - B }' (E1 - E2) relation: { REL3 }<br>**Your answer:** |

Table 11. Uncertainty explanation generation prompt