Universiteit Leiden

Master Creative Intelligence & Technology
Leiden Institute of Advanced Computer Science
Leiden University

How XAI Explanation Formats Influence Human Judgment in AI-Assisted Misinformation Evaluation

Master's Thesis

Name	Yuning Yao
Student ID	4244435
Date	July 13, 2026
1st supervisor	Derya Soydaner
2nd supervisor	Peter van der Putten

Leiden Institute of Advanced Computer Science
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

Abstract

AI credibility tools are increasingly used to support judgments about misinformation, but explanation interfaces do not always help users decide when to follow AI advice and when to question it. We examine token-level SHAP attribution cues in an online misinformation-judgment task with five interface conditions: no AI, prediction-only advice, static SHAP, sequential SHAP, and sequential SHAP with a boundary cue. We ask whether these formats change judgment updating, final alignment with AI advice, confidence change, and whether confidence better reflects actual performance.

AI assistance does not produce an overall accuracy benefit. Prediction-only advice and static SHAP show similar final alignment with incorrect AI advice, at about 61%, so static SHAP offers no clear protective advantage over a bare recommendation. Sequential reveal shows fewer descriptive shifts toward incorrect AI advice, while the boundary cue shows lower final alignment with incorrect AI advice but also a possible trade-off when AI advice is correct. Model-based evidence for these condition differences is limited. These findings suggest that evaluating XAI interfaces requires distinguishing between advice-induced updating, final alignment, accuracy, and confidence-related performance, as explanation quality cannot be captured by a single metric.

Contents

1 Introduction **4**

1.1 Research objective and questions 4

1.2 Hypotheses 5

1.3 Contributions 5

2 Related Work **6**

2.1 Misinformation judgment and truth discernment 6

2.2 Appropriate reliance in human–AI decision making 7

2.3 Behavioral limits of AI explanations 7

2.4 Attribution explanations and interaction design 8

2.5 Research gap and positioning 9

3 Method **9**

3.1 Procedure 9

3.2 Study design and problem formulation 10

3.3 Participants 11

3.4 Background questionnaire 11

3.5 Stimuli 11

3.6 AI advisor and explanations 12

3.7 Interface conditions 12

3.8 Counterbalancing 16

3.9 Outcome measures 17

3.10 Statistical analysis plan 18

4 Results **19**

4.1 Participant sample and data quality 20

4.2 Overall performance and baseline judgment stability 20

4.3 RQ1: Judgment change across interface conditions 21

4.4 RQ2: Correctness-conditional advice use 22

4.5 RQ3: Confidence and chosen-answer Brier score 24

4.6 Hypothesis-oriented model summary and sensitivity analyses 25

4.7 Item heterogeneity and sensitivity analyses 26

5 Discussion **27**

5.1 Principal findings . . . 27
5.2 Prediction-only advice and static SHAP . . . 27
5.3 Sequential reveal, boundary cue, and selectivity . . . 28
5.4 Methodological contribution . . . 28
5.5 Design implications . . . 29
5.6 Limitations . . . 29
5.7 Future work . . . 29
5.8 Implementation, reproducibility, and ethics . . . 30

6 Conclusion **30**

Appendix overview **32**

A Participant background characteristics **32**

B Full background questionnaire items **32**

C Task-flow interface screenshots **33**

D Stimulus items and AI predictions **35**

E Exploratory mixed-effects model details **37**
E.1 M1: Final correctness on all trials . . . 38
E.2 M2: Final correctness on AI-present trials . . . 38
E.3 M3: Final alignment on AI-present trials . . . 38
E.4 M4: Final chosen-answer Brier score on AI-present trials . . . 39
E.5 M5: Advice adoption on initial non-alignment trials . . . 39
E.6 Planned contrasts from M3, M4, and M5 . . . 39
E.7 Participant-aggregated chosen-answer Brier robustness check . . . 40
E.8 Sensitivity model with displayed model score . . . 40
E.9 Sensitivity check (exclude sessions under 8 minutes) . . . 41
E.10 Exploratory participant-level moderator checks . . . 41

1 Introduction

In online misinformation tasks, AI support often appears as a predicted label with a confidence score, sometimes accompanied by an explanation linked to the prediction. Such support can help users correct an initial mistake, but it can also move users away from a correct judgment when the AI is wrong.

A relevant design objective is therefore not greater agreement with AI. It is appropriate reliance: users should benefit from correct recommendations while resisting incorrect ones (Lee & See, 2004). This distinction is especially important in misinformation evaluation. An intervention may help users reject false content, but it may also make them less willing to accept accurate content (Hoes et al., 2024). Aggregate agreement or accuracy can therefore conceal different effects on correct and incorrect judgments.

Prior work presents algorithmic support through prediction labels and confidence scores (Zhang et al., 2020), model-level accuracy information (de Jong et al., 2025; Yin et al., 2019), feature-attribution explanations such as LIME and SHAP (Lundberg & Lee, 2017; Ribeiro et al., 2016), and interaction techniques such as cognitive forcing or progressive disclosure (Buçinca et al., 2021; Muralidhar, 2024). Although these interfaces differ, each is intended to help users accept useful recommendations and resist misleading ones.

However, comparable human–AI decision-making work shows that explanations do not necessarily make advice use more sensitive to AI correctness. Explanations can increase acceptance of AI recommendations without making acceptance more selective, and greater transparency can improve users' ability to simulate model predictions without necessarily helping them detect model errors or make better final decisions (Bansal et al., 2021; Poursabzi-Sangdeh et al., 2021). This means that explanation success should not be evaluated by final agreement alone, but by whether users benefit from correct advice without also following incorrect advice (Schemmer et al., 2023; Schoeffer et al., 2025).

Prior work commonly compares advice with and without explanations or compares different explanation types. It has paid less attention to whether the same attribution content produces different judgment updates when shown all at once or revealed sequentially. We vary the timing and framing of SHAP attribution cues while keeping the item-level predictions and SHAP values fixed.

To examine this presentation contrast, participants judge whether short news-style items are true or false before and after exposure to one of five advice-interface conditions: no AI-driven advice, prediction-only advice, prediction plus static SHAP explanation, prediction plus sequentially revealed SHAP explanation, or prediction plus sequential SHAP with a boundary cue. Across the four AI-present conditions, the item-level AI recommendation and SHAP values remain fixed. This lets us compare explanation presentation formats while holding the underlying AI advice and explanation content constant. We treat attribution cues as model-derived interface information, not direct representations of the model's internal reasoning.

We adopt a decision-dynamics perspective. Final agreement alone does not establish reliance because a participant may already have agreed with the AI before seeing it. Our analysis therefore focuses on changes between initial and final judgments. It distinguishes correction, preservation, and deterioration, and separates responses to correct and incorrect AI advice.

1.1 Research objective and questions

The research objective is to examine how AI explanation presentation shapes selective reliance and confidence–correctness correspondence in AI-assisted misinformation evaluation.

Our main research question is:

> *How do AI explanation presentation formats affect judgment updating, confidence change, and confidence–correctness correspondence in misinformation evaluation, particularly when AI advice is incorrect?*

Our study addresses three sub-questions:

RQ1. How do different AI interface formats affect changes from initial to final credibility judgments?

RQ2. How do these formats affect movement toward and final alignment with AI advice, especially when the AI recommendation is incorrect?

RQ3. How do explanation formats affect confidence change and confidence–correctness correspondence across AI-correct and AI-incorrect trials?

1.2 Hypotheses

Prior work on AI advice, explanations, and interaction design informs the following hypotheses. We use them to organize the analysis and interpretation of the observed judgment-updating patterns.

H1a: AI advice and judgment change. We expect AI-present conditions to increase the probability of changing the initial judgment compared with the no-AI baseline.

H1b: Direction of judgment change. We expect judgment changes within AI-present conditions to occur more often toward than away from the AI recommendation.

H2: Static explanations. We expect static token-level explanations to increase final alignment with the AI recommendation compared with prediction-only advice.

H3: Incorrect-AI alignment. We expect static explanations to increase final alignment with incorrect AI recommendations compared with prediction-only advice.

H4: Sequential reveal. We expect sequential reveal to reduce movement toward incorrect AI advice and final alignment with incorrect recommendations compared with static explanations.

H5: Boundary cue. We expect the boundary cue to reduce final alignment with incorrect AI recommendations and limit confidence increases on incorrect final judgments compared with sequential reveal alone. We do not expect it to substantially reduce alignment when the AI is correct.

We report descriptive trial-level patterns and use mixed-effects models as the principal inferential analyses. We interpret the model results cautiously because the study has a modest number of AI-incorrect trials per condition and does not directly measure processing fluency, cognitive effort, perceived understanding, attention allocation, or awareness of model limitations. We therefore treat mechanism claims as interpretations rather than tested mediators.

1.3 Contributions

The thesis makes four contributions.

1. We compare static and sequential presentations of the same token-level SHAP attribution content in an AI-assisted misinformation judgment task. This allows us to examine whether changing presentation format, while holding prediction and attribution values constant, relates to different judgment-updating patterns.
2. We combine initial-to-final judgment transitions with correctness-conditional reliance measures. This avoids treating final agreement with AI as reliance when participants may already have held the AI-aligned judgment before advice exposure.
3. We provide descriptive evidence that interface conditions can show different outcome profiles across harmful shifts, final incorrect-AI alignment, beneficial reliance, and confidence–correctness correspondence. Researchers therefore need to evaluate explanation formats with more than one global agreement or accuracy metric.
4. We make the study code and analysis data available through a public GitHub repository.[1]

2 Related Work

This section reviews the work needed to motivate the study design. It first discusses misinformation judgment as a truth-discernment problem, then reviews appropriate reliance in human–AI decision making, behavioral limits of AI explanations, and interaction-design choices for presenting attribution cues.

2.1 Misinformation judgment and truth discernment

Misinformation evaluation requires users to distinguish true from false content rather than becoming uniformly more skeptical. This distinction is commonly described as truth discernment. Changes in overall belief do not necessarily indicate improved discernment because an intervention may reduce belief in both true and false information (Pennycook & Rand, 2021).

Misinformation research identifies several routes through which discernment can fail. Cognitive, social, and affective factors shape belief in misinformation, including familiarity, prior beliefs, source cues, and emotional responses (Ecker et al., 2022). Lower analytic reasoning is associated with greater susceptibility to partisan false news, and poor discernment is also linked to limited relevant knowledge and familiarity-based heuristics (Pennycook & Rand, 2019, 2021).

The same logic applies to intervention design. Accuracy incentives improve truth discernment, whereas incentives tied to political in-group approval reduce it (Rathje et al., 2023). Inoculation against emotional manipulation improves manipulation detection, but improves truth discernment only when paired with an accuracy prompt (Pennycook et al., 2024). Several misinformation interventions can also reduce belief in false information while lowering the perceived credibility of factual information (Hoes et al., 2024). These findings matter because an intervention can appear successful if only false-item belief is measured, while still weakening acceptance of true information.

Correction and warning cues raise a related measurement problem. Political corrections can fail among the subgroup most likely to hold a misperception and can sometimes backfire, although later work cautions against treating strong backfire effects as the usual outcome (Nyhan & Reifler, 2010; Swire-Thompson et al., 2020). Recent misinformation-interface work similarly treats warning labels as behavioral interventions whose effects depend on presentation and context (Gruzd et al., 2024). For our

[1] https://github.com/itsyuning/xai-ui-study

analysis, the requirement is clear: we evaluate AI support separately on true and false items, so that correction, preservation, and deterioration remain visible rather than being absorbed into an overall belief score.

2.2 Appropriate reliance in human–AI decision making

The goal of AI-assisted decision making is not maximal trust or agreement, but appropriate reliance. Lee and See (2004) argue that users should rely on automation in ways that reflect its capabilities and limitations. Trust may influence reliance, but the two constructs are not equivalent. Hoff and Bashir (2015) further show that trust depends on dispositional, situational, and learned factors.

Inappropriate automation use can take opposite forms: misuse and disuse (Parasuraman & Riley, 1997). Evidence from automated-aid and algorithm-advice settings shows why this balance is unstable. Observed aid errors can reduce trust and lead users to ignore even superior automated aids (Dzindolet et al., 2002). Observed algorithm errors can also reduce willingness to use algorithms, even when the algorithm remains superior to a human forecaster (Dietvorst et al., 2015). In other numerical-estimation tasks, participants weight algorithmic advice more strongly than comparable human advice (Logg et al., 2019). These findings caution against treating reliance as a stable user tendency; it shifts with task context and prior performance information.

Case-specific studies provide a stronger basis for evaluation. Zhang et al. (2020) find that model confidence helps users differentiate cases with different levels of model certainty, but improved trust calibration does not automatically improve joint performance. Related work on accuracy disclosure shows that communicating model-level accuracy can also shape reliance, but this differs from showing an item-level prediction score because it sets a general expectation about the system rather than explaining a specific recommendation (de Jong et al., 2025).

Schoeffer et al. (2024) define reliance through adherence to or override of AI recommendations, conditioned on recommendation correctness. Their results show that explanations can change override behavior without improving users' ability to distinguish correct from incorrect advice.

Chaleshtori et al. (2024) emphasize the need to record an initial unaided judgment. Final agreement with AI does not necessarily indicate reliance because users may already have held the same judgment. We use the initial response as part of the outcome structure, which lets us distinguish correction, preservation, and deterioration.

2.3 Behavioral limits of AI explanations

Human-centered XAI research argues that evaluations should judge explanation quality relative to a specific task and user rather than infer it from technical properties alone (Abdul et al., 2018; Doshi-Velez & Kim, 2017).

Behavioral evaluations complicate the assumption that explanation presence improves decision quality. Explanations can increase acceptance of AI recommendations regardless of correctness, without improving performance beyond showing the recommendation and confidence (Bansal et al., 2021). Greater transparency can also improve model simulatability without consistently improving error detection or model use (Poursabzi-Sangdeh et al., 2021).

Across tasks, explanation effects appear conditional rather than uniformly beneficial. Benefits emerge only in limited method–task combinations, and subjective explanation ratings do not reliably predict simulatability (Hase & Bansal, 2020). Explanation effects also vary with task knowledge, model complexity, explanation type, representation, and conceptualization, which can affect comprehension,

confidence, trust, and reliance in different ways (Delaunay et al., 2025; Pareek et al., 2024; Wang & Yin, 2022).

User knowledge shapes how people interpret explanations. Chen et al. (2023) identify three forms of intuition used to evaluate AI support: intuition about the task outcome, relevant features, and AI limitations. In their experiments, example-based explanations produce less overreliance than the feature-based explanations tested, although the result does not generalize beyond those tasks by itself.

Technical experience also does not guarantee correct interpretation. Kaur et al. (2020) find that data scientists using InterpretML and SHAP sometimes over-trust and misuse the tools. Morrison et al. (2025) further show that explanation correctness can vary independently of prediction correctness and can affect reliance and team performance.

The common thread is that explanations often change acceptance or perceived understanding without reliably improving judgment quality. We target this behavioral gap by evaluating explanation formats through judgment transitions, error detection, and final performance rather than through explanation presence or subjective clarity alone.

2.4 Attribution explanations and interaction design

We use token-level SHAP attributions because the experiment requires stable, rankable textual cues. We can present these cues in different formats while holding explanation content constant. SHAP provides additive feature attributions grounded in Shapley values (Lundberg & Lee, 2017). This makes it suitable for constructing both a static heatmap and a sequential reveal from the same attribution values.

Researchers should not treat SHAP outputs as direct representations of model reasoning. Jacovi and Goldberg (2020) distinguish plausibility to users from faithfulness to model behavior. Adebayo et al. (2018) show that some saliency methods fail sanity checks, while Slack et al. (2020) demonstrate that post-hoc explanations can be manipulated to conceal undesirable model behavior.

Automatic explanation metrics also do not guarantee human usefulness. Nguyen (2018) reports only moderate correspondence between deletion-based metrics and human judgments of local text explanations. Chaleshtori et al. (2024) similarly argue that realistic decision tasks with strong baselines should test explanation utility. In their legal claim-verification study, highlights and influential examples do not consistently improve decisions beyond AI predictions and confidence.

Interaction design may change how participants use explanation cues. Buçinca et al. (2021) find that cognitive-forcing interfaces reduce overreliance, but also receive less favorable subjective evaluations. Their results show that reducing overreliance may require additional effort.

One relevant design choice is whether users form an initial judgment before seeing AI advice. This sequencing preserves an observable unaided response. It also lets us distinguish later correction, preservation, and deterioration. Sequentially revealing explanation cues is not equivalent to prior cognitive-forcing interfaces. Delayed advice, user-controlled disclosure, waiting requirements, and incremental feature reveal impose different forms of effort.

Prior work explores progressive disclosure as a transparency-oriented interface principle (Muralidhar, 2024). The available evidence relies mainly on prototype evaluation and subjective feedback. It provides a design rationale for staged presentation, but not established evidence that sequential explanations improve reliance, confidence–correctness correspondence, or accuracy.

Sequential reveal may produce competing effects. Presenting attribution cues incrementally may reduce simultaneous information load and encourage smaller evidence evaluations. Conversely, it may increase anchoring on the earliest or strongest cues. This unresolved tension motivates the C2–C3 comparison:

the two conditions use identical attribution information but organize it differently over time.

2.5 Research gap and positioning

Explanations can change reliance without making users more sensitive to AI correctness. Their effects also depend on user knowledge, explanation type, model properties, and interface design.

Most studies compare different explanation methods or explanation presence versus absence. The temporal presentation of identical attribution information has received less direct attention. The contrast is important here because a static heatmap and a sequential reveal can contain the same attribution values while creating different interaction patterns.

We isolate this presentation contrast by holding item-level AI predictions and SHAP attribution values constant across C2 and C3. We also examine whether adding a boundary cue to sequential reveal relates incrementally to different outcomes in C4. The analysis treats C4 as sequential reveal plus an added cue, not as an independent explanation format.

The design records participants' initial judgments, the direction of each judgment change, and whether the AI recommendation is correct. We evaluate transitions, correctness-conditional alignment, deterioration, final accuracy, confidence change, and confidence–correctness correspondence. Increased agreement, perceived transparency, or confidence does not by itself indicate a successful explanation. The relevant question is whether participants update differently when the AI is correct and when it is wrong.

3 Method

We use a within-subject experimental design to examine how different AI explanation formats affect human credibility judgments. Stimulus groups are counterbalanced across the five interface conditions to reduce item–condition confounding. We combine a staged judgment task, fixed item-level AI advice, and trial-level measures of judgment updating and confidence.

3.1 Procedure

Participants complete the study online in two stages. They first open a Qualtrics survey, which begins with an information and consent page. Participants who consent complete the background questionnaire. At the end of the questionnaire, Qualtrics redirects them to the custom web experiment.[2] The redirect applies the counterbalancing-list assignment described later in this section.

In the web experiment, each participant completes 25 trials. On each trial, participants read a short news-style item and select whether they believe it is TRUE or FAKE. They then rate confidence in that initial judgment. They next enter the interface stage. In C0, the interface shows no AI information, but participants still continue to the final judgment stage. In AI-present conditions, participants see the AI prediction and, depending on the condition, explanation cues or a boundary cue. After the interface stage, all participants make a final TRUE/FAKE judgment and again rate their confidence. The task presents the 25 trials in randomized order within each participant's assigned counterbalancing list, so condition trials are intermingled rather than blocked.

[2]Public demonstration deployment: https://xaiui.vercel.app. This link documents the interface flow. The analyses use the archived matched participant-task dataset described in the Method and Results sections.

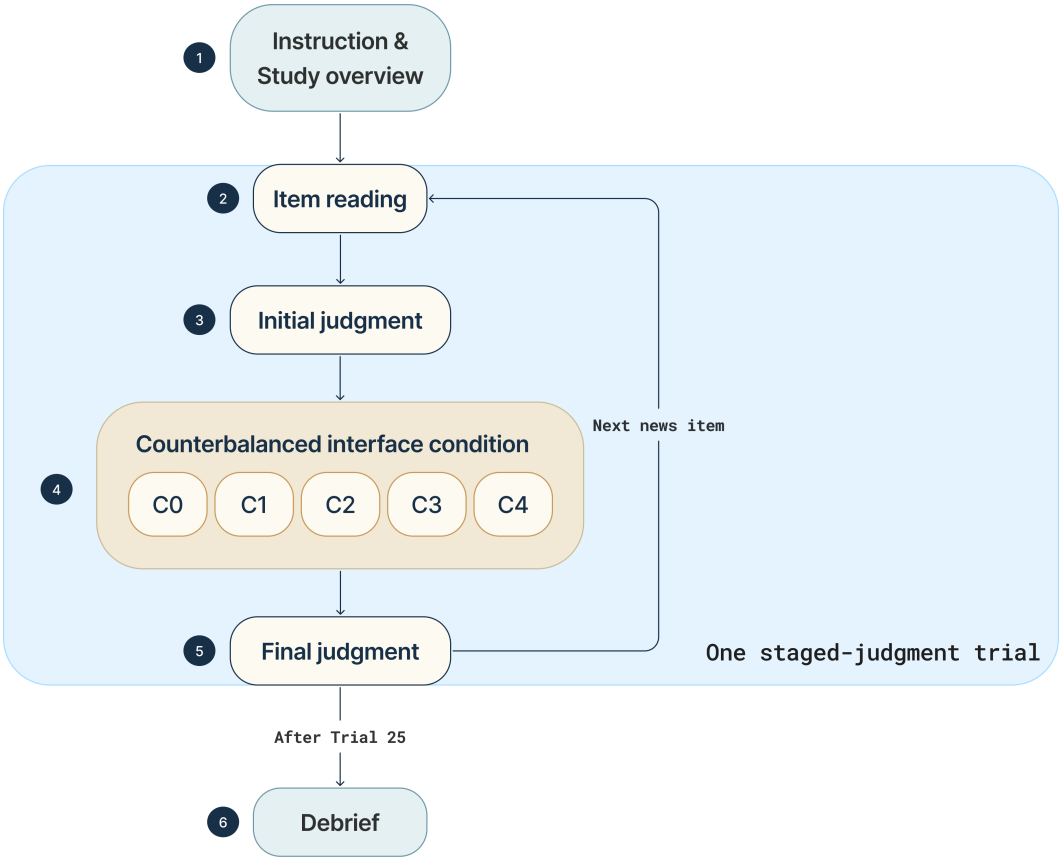


Figure 1: Overview of the web-experiment flow after the Qualtrics survey redirect. Numbered markers indicate (1) instruction and study overview, (2) item reading, (3) initial label and confidence judgment, (4) counterbalanced interface condition, (5) final label and confidence judgment, and (6) debrief. The loop indicates repetition across 25 news items before the debrief page.

After the final trial, participants see a debrief page. The debrief explains why the study includes misleading content and imperfect AI advice, provides contact information, and offers optional item clarification.

Figure 1 summarizes the web-experiment flow after the Qualtrics survey redirect. Appendix C provides screenshots of the instruction, item reading, initial judgment, and final judgment stages.

3.2 Study design and problem formulation

Formally, each trial represents a transition from an initial human judgment $J_{ij}^{(0)}$ to a final human judgment $J_{ij}^{(1)}$ after exposure to an interface condition C_{ij}, where i indexes participants and j indexes trials. In AI-present conditions, the participant also receives an AI recommendation A_{ij}. This structure lets us measure whether the participant stays with their initial judgment, shifts toward the AI, shifts away from the AI, or aligns with the AI at the final judgment stage.

The staged design treats AI output as advice that may change a participant's judgment rather than as a simple performance intervention. It captures whether $J_{ij}^{(1)}$ is correct, how it differs from $J_{ij}^{(0)}$, whether

any change follows the AI recommendation, and how confidence changes after the interface stage. The same initial-to-final structure in C0 helps separate AI-related updating from ordinary reconsideration.

3.3 Participants

Participants are recruited online through convenience sampling via personal networks, student communities, and relevant online forums. Eligible participants are adults who can read English and complete the study remotely. Participant background variables describe the sample and support interpretation of the results; they are not primary predictors in the main hypothesis tests.

3.4 Background questionnaire

The background questionnaire collects demographic and task-relevant variables. These include age group, gender, country or cultural background, first or primary language, education level, English reading comfort, and online news-reading frequency. They also include familiarity with misinformation and fact-checking, perceived usefulness of AI-generated predictions or explanations, likelihood of verifying AI advice, and general caution toward AI-generated outputs. We export questionnaire responses from Qualtrics and link them to task records using anonymized participant IDs. The full questionnaire items are provided in Appendix B.

We use the questionnaire to describe the final participant sample and check whether participants have sufficient English reading comfort for the task. These variables are not part of the experimental manipulation, so we treat them as descriptive background variables.

3.5 Stimuli

The stimulus set consists of 25 short news-style items from a curated item pool that we prepare before data collection. We collect candidate items from public fact-checking and news-analysis sources, including AFP Fact Check, Snopes, PolitiFact, and other topic-specific institutional outlets. We then select items to cover politics, health, science and technology, business, and society. Each item includes a headline-like title and a short content summary. Each item has a fixed ground-truth label, either TRUE or FAKE, based on the originating source material or dataset label. Appendix D provides the full list of item titles, topic categories, ground-truth labels, AI predictions, and item-level source provenance.

We keep the stimuli short enough for repeated trial-based evaluation while still supporting a credibility judgment. This choice reflects online misinformation contexts, where users often encounter condensed claims, headlines, summaries, or short posts rather than full articles. The format also imposes a limitation. Some claims may require background knowledge, source checking, or visual evidence that is not fully available in the short text. The Discussion returns to this limitation.

The final stimulus pool includes both AI-correct and AI-incorrect cases because appropriate reliance means following AI advice when it is correct and resisting it when it is wrong. We purposively construct the pool to contain a relatively high proportion of AI errors (10 of 25 items). This enables analysis of behavior under incorrect advice. We therefore do not interpret aggregate accuracy and reliance rates as estimates of performance under the model's natural error distribution.

3.6 AI advisor and explanations

The AI advisor is a misinformation-classification system that produces one binary recommendation for each item. It uses the publicly available `hamzab/roberta-fake-news-classification` checkpoint, based on the RoBERTa architecture (Liu et al., 2019).[3] The checkpoint is fine-tuned on the approximately balanced Kaggle Fake and Real News dataset.[4] We apply the same binary decision rule to every stimulus before participant data collection: the model recommends FAKE when $P(\text{FAKE}) \geq P(\text{TRUE})$ and TRUE otherwise.

After participants submit their initial judgment, AI-present conditions display the fixed item-level recommendation as TRUE or FAKE together with a model score. The score is the model probability for the displayed label: $P(\text{FAKE})$ for a FAKE recommendation and $1 - P(\text{FAKE})$ for a TRUE recommendation. It is presented as an indication of model output rather than as an independently calibrated probability of real-world truth. For a given item, the recommendation and score remain identical across all AI-present conditions. The experimental manipulation changes whether attribution cues are shown, how they are presented, and whether the boundary cue appears. Appendix D reports the fixed item-level recommendations and scores.

Implementation. The study interface is implemented as a static web application using HTML, CSS, and JavaScript and deployed through Vercel. Before data collection, Python scripts and a notebook using PyTorch, Hugging Face Transformers, and SHAP generate the fixed model predictions and token-level attribution values used by the interface. Precomputing these outputs ensures that participants assigned to different interface conditions receive the same item-level model output.

For explanation-based conditions, we compute token-level attributions with SHAP, an additive feature-attribution method based on Shapley values (Lundberg & Lee, 2017).[5] We use the partition explainer to explain the logit of $P(\text{FAKE})$, with `max_evals=300` and `fixed_context=1`. Delimiter tokens required by the checkpoint are removed before display. Highlight color indicates whether a token pushes the model output toward FAKE or TRUE, and intensity scales with the absolute attribution value. C2 displays all available token highlights simultaneously, whereas C3 and C4 reveal up to the 24 tokens with the largest absolute SHAP values. The goal is not to evaluate the technical faithfulness of SHAP itself, but to examine how these explanation interfaces affect human judgment updating.

3.7 Interface conditions

The independent variable is the AI interface condition. We use five conditions. Figure 2 shows the full-screen layout using a C3 sequential-reveal trial as an example. Figures 3, 4, 5, 6, and 7 then show the condition-specific interface stages.

[3] Model card: https://huggingface.co/hamzab/roberta-fake-news-classification
[4] https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset
[5] SHAP software repository: https://github.com/shap/shap

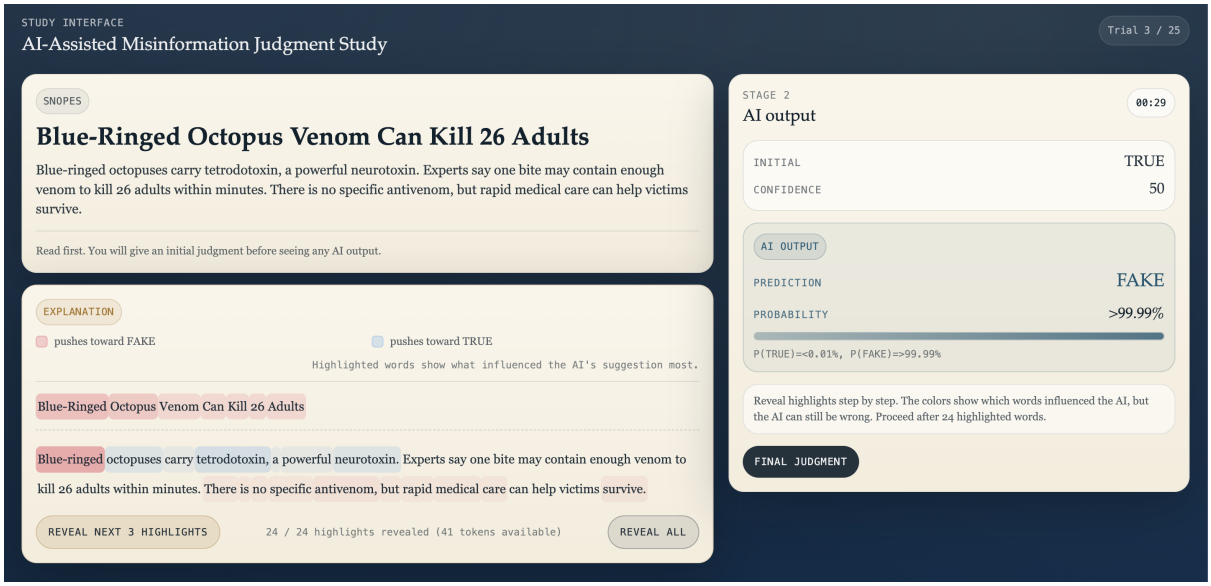


Figure 2: Full-screen example of the study interface in the C3 sequential-reveal condition. The example shows the stimulus text, the participant's initial judgment, the AI output, and the token-level explanation panel after all highlights have been revealed.

C0: No AI baseline. Participants make an initial judgment and report confidence. They then briefly review the item without AI advice before making a final judgment and reporting confidence again. This condition measures natural reconsideration and judgment stability without AI assistance.

Figure 3: C0 no-AI baseline interface. Participants complete the same staged judgment task without seeing an AI recommendation.

C1: Prediction-only. Participants see the AI's binary prediction, either TRUE or FAKE, together with the displayed model score. This condition isolates the influence of the AI recommendation itself and provides the comparison point for testing whether explanation cues add anything beyond prediction-only advice.

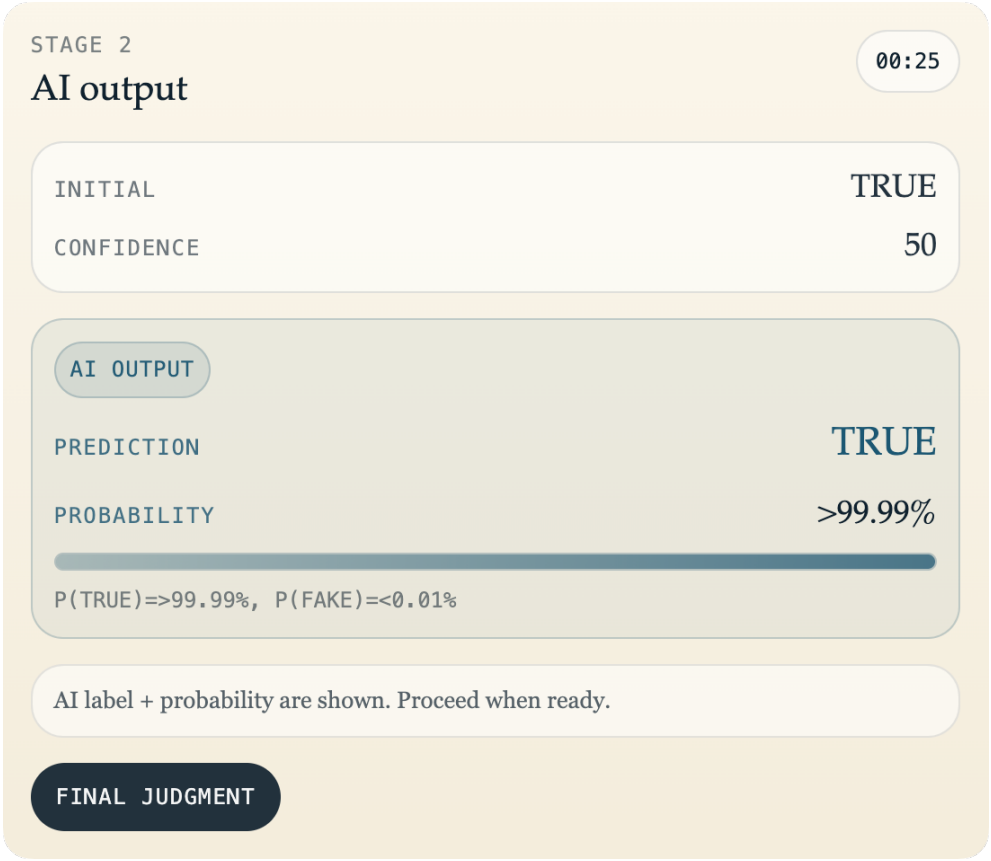


Figure 4: C1 prediction-only interface. The interface shows the AI recommendation and displayed model score without token-level explanation cues.

C2: Static explanation. Participants see the AI prediction and displayed model score together with a static token-level explanation. The explanation appears as highlighted textual evidence associated with the AI output.

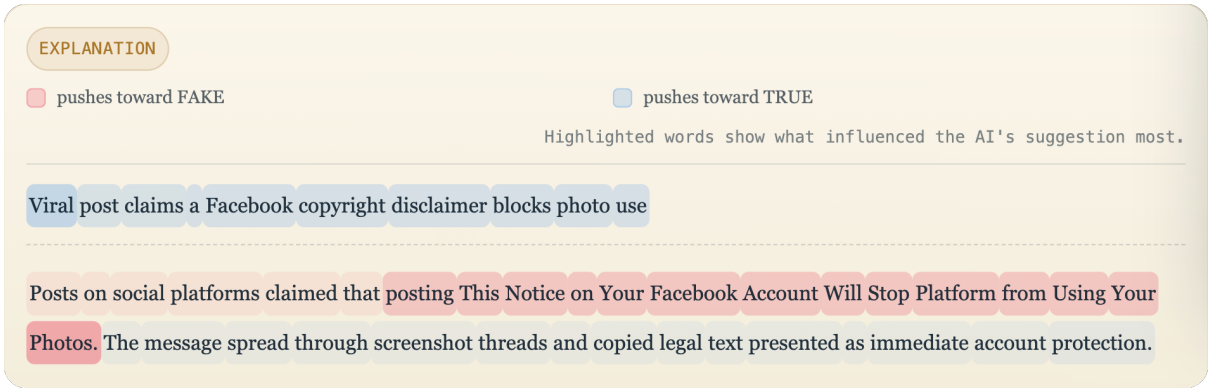


Figure 5: C2 static explanation interface. All token-level SHAP highlights appear simultaneously with the AI recommendation.

C3: Sequential reveal. Participants first see the AI prediction without the full explanation. They then click the *Reveal Next 3 Highlights* button to reveal three additional token-level SHAP highlights at each step until all 24 explanation cues become visible. Highlights are revealed in descending order of absolute SHAP value, so participants first see the tokens with the strongest influence on the model

output. This condition tests whether incremental presentation affects judgment updating relative to displaying all highlights simultaneously.

(a) Early reveal: 3 of 24 highlights shown.

(b) Complete reveal: 24 of 24 highlights shown.

Figure 6: C3 sequential-reveal interface at an early step and after all highlights have been revealed. The two views illustrate how attribution cues accumulate across the interaction.

C4: Sequential reveal with boundary cue. Participants see the sequential explanation interface together with the boundary cue *This model may fail on satire, ambiguous framing, or unfamiliar events. Treat it as one input among others.* This condition tests whether making model fallibility salient relates to lower incorrect-AI alignment or better confidence–correctness correspondence under AI error.

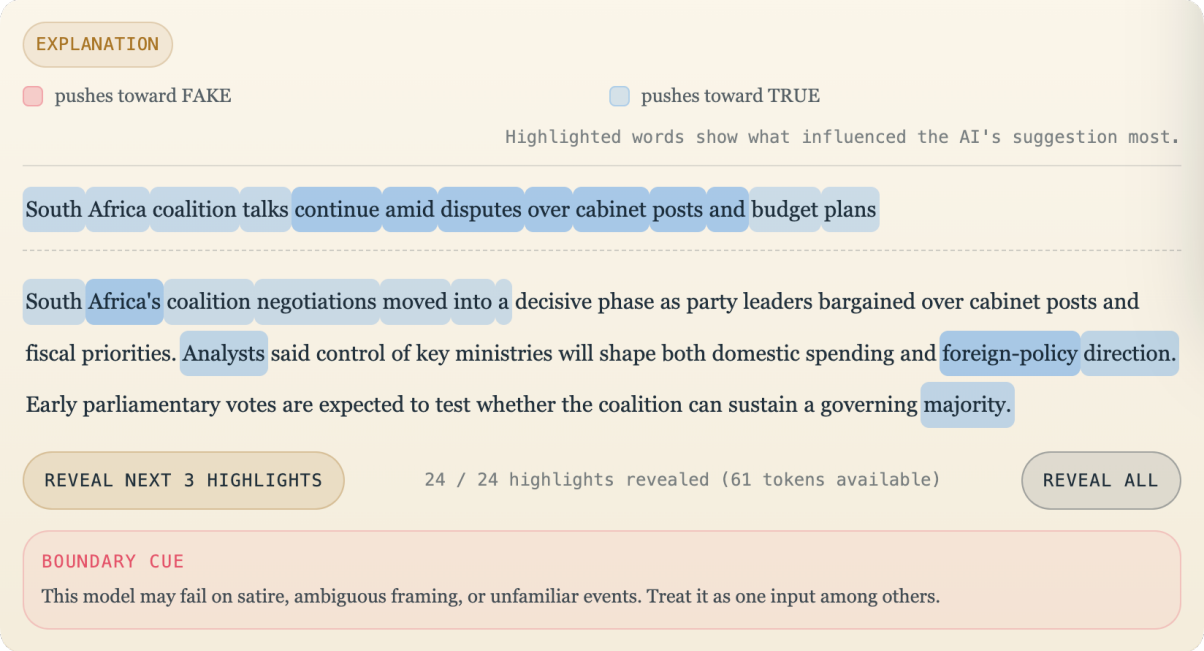


Figure 7: C4 sequential-reveal interface with boundary cue. The cue reminds participants that the AI may fail on satire, ambiguous framing, or unfamiliar events.

The adjacent comparisons examine the addition of a static explanation (C2 versus C1), the move from a full static display to sequential reveal of the 24 strongest highlights (C3 versus C2), and the addition of a boundary cue (C4 versus C3). In C0, AI correctness is an item-level counterfactual property used only for matched accuracy comparisons; it does not represent advice exposure or reliance.

3.8 Counterbalancing

We use a counterbalanced within-subjects design so that each participant experiences all five conditions, each news item appears only once per participant, and item properties are not tied to a single condition. The 25 items are divided into five fixed groups (A–E) of five items. Five counterbalancing lists rotate these groups across the five interface conditions, giving each participant five trials per condition and 25 trials in total. Auto mode assigns participants to lists through a server-side round-robin procedure and retains the assignment for the participant identifier.

The group-to-condition assignment rotates across five lists. Thus, each participant sees each news item only once, while across the five lists each item group appears once in every interface condition. This design allows us to compare interface conditions within participants while reducing the risk that item-specific properties, such as topic, difficulty, or plausibility, are confounded with a particular condition. Table 1 shows the item-group-to-condition mapping.

AI correctness is also distributed across item groups, but not perfectly evenly. The groups contain different numbers of AI-wrong items: A = 2, B = 2, C = 3, D = 2, and E = 1. Because groups rotate across conditions, the number of AI-wrong trials in a given condition varies by list. Each participant still completes five trials per condition, but a condition can contain between one and three AI-wrong trials depending on the assigned list. Aggregated across the evenly filled lists, each AI-present condition contains 140 AI-wrong trials and 210 AI-correct trials.

The implemented web task randomizes trial order within each participant's assigned list using a seed-based order. This intermingles trials from different conditions rather than presenting conditions in blocks.

Table 1: Item-group-to-condition mapping in the five-list counterbalancing scheme. Each participant completes one list, with five trials per condition and 25 trials in total.

List	Group A	Group B	Group C	Group D	Group E
List 1	C0	C1	C2	C3	C4
List 2	C1	C2	C3	C4	C0
List 3	C2	C3	C4	C0	C1
List 4	C3	C4	C0	C1	C2
List 5	C4	C0	C1	C2	C3

Note. C0 = no-AI baseline; C1 = prediction only; C2 = static explanation; C3 = sequential reveal; C4 = sequential reveal with boundary cue.

The design therefore counterbalances item-to-condition assignment but does not fully counterbalance condition order; condition order is not treated as a primary target of the study.

3.9 Outcome measures

We use several trial-level outcome measures to capture decision dynamics.

Initial and final accuracy. Initial accuracy indicates whether the participant's first judgment matched the ground-truth label. Final accuracy indicates whether the participant's final judgment matched the ground truth after the interface stage.

Judgment shift. A judgment shift occurs when the participant's final judgment differs from their initial judgment. This measure captures whether the participant changes their mind during the trial.

Final AI alignment. In AI-present conditions, final AI alignment indicates whether the participant's final judgment matched the AI recommendation. We do not treat alignment as inherently good or bad. It helps when the AI is correct and becomes problematic when the AI is incorrect. We interpret final alignment on AI-incorrect trials as final alignment with incorrect AI advice, not as evidence that AI exposure caused that alignment. Some participants already agreed with the AI before exposure. Trials in which an initially correct response changed to an incorrect response matching the AI recommendation provide stronger evidence of harmful advice adoption.

Shift toward AI. A shift toward AI occurs when the participant's initial judgment differs from the AI recommendation but their final judgment matches the AI recommendation. This measure captures adoption of AI advice within the relevant risk set. We operationalize harmful advice adoption as a change from an initial judgment that disagrees with an incorrect AI recommendation to a final judgment that matches that recommendation.

Shift away from AI. A shift away from AI occurs when the participant's initial judgment matches the AI recommendation but their final judgment no longer matches it. This measure captures resistance or movement away from AI advice.

Correction and deterioration. A correction occurs when a participant moves from an initially incorrect judgment to a finally correct judgment. A deterioration occurs when a participant moves from an initially correct judgment to a finally incorrect judgment. These measures indicate whether judgment updating improves or harms accuracy.

Confidence change. We calculate confidence change as final confidence minus initial confidence. Positive values indicate increased confidence after the interface stage, while negative values indicate reduced confidence.

Chosen-answer Brier score. We assess confidence–correctness correspondence with a chosen-answer Brier score, adapted from the quadratic scoring rule for probabilistic predictions (Brier, 1950). Because we elicit confidence in the chosen answer rather than a direct probability that the item is TRUE, we do not interpret the score as a standard truth-probability Brier score. We scale final confidence ratings from 0–100 to 0–1 and interpret them as the probability assigned to the participant's own final answer. For example, if a participant chooses TRUE with 80% confidence, this corresponds to $p(\text{TRUE}) = 0.80$. If the participant chooses FAKE with 80% confidence, this corresponds to $p(\text{TRUE}) = 0.20$. With $y = 1$ for TRUE items and $y = 0$ for FAKE items, we write the score as:

$$\text{Brier}_{ij} = (p_{ij}(\text{TRUE}) - y_{ij})^2,$$

where $p_{ij}(\text{TRUE})$ is participant i's confidence-derived probability that item j is true, and y_{ij} is the ground-truth label. This formulation is mathematically equivalent to scoring confidence in the chosen answer against final-response correctness. Lower chosen-answer Brier scores indicate better confidence–correctness correspondence. However, this score jointly reflects final accuracy and confidence extremity, so we do not interpret it as a pure calibration measure. We use it alongside accuracy and descriptive reliability analyses to assess whether confidence corresponds to response correctness.

3.10 Statistical analysis plan

We first report descriptive trial-level results across all 70 participants: initial and final accuracy, judgment-shift rates, confidence change, final AI alignment, override rates, and chosen-answer Brier scores. For AI-present trials, we separate AI-correct from AI-incorrect trials because alignment has opposite implications across these two strata. C0 appears in performance and stability analyses, but not in reliance or advice-adoption analyses because participants in C0 do not receive AI advice. Process logs record interface-stage duration as the elapsed time between entering the interface/advice stage and entering the final-judgment stage. We use this timing variable only as an interpretive sensitivity check, not as a primary outcome.

We then use mixed-effects models to ask whether outcomes differ across interface conditions while accounting for repeated judgments from the same participants and items. Table 2 summarizes the trials included, dependent variable, main predictors, and purpose of each model. The main independent variable is interface condition. In AI-present models, we also include AI correctness and the condition-by-AI-correctness terms because agreement with AI advice has different implications when the AI is correct versus incorrect. All models include crossed random intercepts for participant and item to account for repeated observations from the same participants and repeated use of the same items.

We use Bayesian logistic mixed-effects models with a logit link for binary outcomes: final correctness (M1 and M2), final AI alignment (M3), and advice adoption (M5). We use a linear mixed-effects model

Table 2: Summary of trial-level statistical models.

Model	Trials	Dependent variable	Main predictors	Purpose
M1	C0–C4	Final correctness	Condition, initial correctness, initial confidence	Compare final correctness in each condition with the no-AI baseline.
M2	C1–C4	Final correctness	Condition, AI correctness, condition × AI correctness, initial correctness, initial confidence	Test whether explanation conditions differ from prediction-only advice in final correctness.
M3	C1–C4	Final AI alignment	Condition, AI correctness, condition × AI correctness, initial alignment, initial confidence	Test whether participants end with the AI recommendation, separately considering AI-correct and AI-incorrect trials.
M4	C1–C4	Chosen-answer Brier score	Condition, AI correctness, condition × AI correctness, initial chosen-answer Brier score	Test whether confidence corresponds better to correctness across interface conditions.
M5	Initial non-alignment trials in C1–C4	Advice adoption	Condition, AI correctness, condition × AI correctness, initial confidence	Test whether participants move toward the AI recommendation when they initially disagreed with it.

for the chosen-answer Brier score (M4) because it is a continuous trial-level score bounded between 0 and 1. Analyses are conducted in Python 3.9.6 using pandas 2.3.3, NumPy 1.26.4, and statsmodels 0.14.6. The Bayesian logistic mixed-effects models are fitted with statsmodels' `BinomialBayesMixedGLM` using variational Bayes estimation (`fit_vb`). Fixed-effect coefficients are reported as posterior means, and approximate 95% posterior intervals are computed as posterior mean $\pm$ 1.96 posterior standard deviations from the variational posterior. Odds ratios are obtained by exponentiating the log-odds coefficients and interval limits. The chosen-answer Brier score model is fitted with statsmodels `MixedLM` using maximum likelihood.

We then conduct planned pairwise comparisons to determine whether each added interface feature changes the outcome relative to the immediately simpler condition: C2 versus C1 for adding static SHAP highlights to prediction-only advice, C3 versus C2 for moving from the full static display to sequential reveal of the 24 strongest highlights, and C4 versus C3 for adding the boundary cue to sequential reveal. When a model includes AI correctness, we report the same comparison separately for trials where the AI recommendation is correct and where it is incorrect. We report interval estimates as the main inferential evidence and Holm-adjusted approximate two-sided p-values as supplemental summaries. For Bayesian logistic models, intervals refer to approximate 95% posterior intervals for odds ratios; for the linear chosen-answer Brier model, intervals refer to 95% Wald confidence intervals for coefficient differences. We also add the displayed model score to the final-alignment model as a sensitivity check because that score varies at the item level.

4 Results

We organize the results around the research questions. We first describe the final matched sample and data-quality checks, then report overall performance and baseline judgment stability. We then examine judgment change across interface conditions (RQ1), correctness-conditional advice use (RQ2), and confidence change with the chosen-answer Brier score (RQ3). Descriptive trial-level results show the behavioral patterns. Exploratory mixed-effects models provide the principal inferential analyses while accounting for repeated observations from participants and items.

Table 3: Overall initial-to-final judgment transitions.

Transition type	Count	Percentage of all trials
Initial correct → final correct	1135	64.9%
Initial incorrect → final incorrect	327	18.7%
Initial incorrect → final correct	121	6.9%
Initial correct → final incorrect	167	9.5%
Total	1,750	100.0%

4.1 Participant sample and data quality

The final matched dataset includes 70 participants and 1,750 trial-level observations. Each participant contributes 25 complete trials, with five trials in each interface condition, and each counterbalancing list contributes 14 participants. All 70 matched sessions are retained in the primary analyses, leaving a final sample of $N = 70$.

Six sessions take less than eight minutes and are flagged as potentially fast responses. Excluding these sessions does not materially change the main descriptive patterns, so they are retained in the primary analyses.

The final sample consists mostly of young adults with university-level education and high self-reported English reading comfort. Appendix A reports participant background characteristics in Table 11.

4.2 Overall performance and baseline judgment stability

Overall, final judgments become less accurate while confidence increases. Across all conditions, initial accuracy is 74.4%, while final accuracy is 71.8%, a decrease of 2.6 percentage points. The transition counts in Table 3 show why this occurs: 167 trials move from an initially correct to a finally incorrect judgment, whereas 121 trials move from an initially incorrect to a finally correct judgment; 1135 stay correct and 327 stay incorrect. An unadjusted McNemar check gives $\chi^2(1) = 7.03$ and $p = .008$. Because this aggregate check does not account for repeated observations from participants and items, it is treated as descriptive; the mixed-effects models provide the main exploratory inference.

Mean confidence nevertheless increases by 6.06 points from initial to final judgment. Participants become more confident overall even though final accuracy does not improve. The chosen-answer Brier score changes only slightly, from 0.222 at the initial stage to 0.221 at the final stage. The confidence increase therefore does not produce a clear aggregate improvement in confidence–correctness correspondence.

The second judgment stage is associated with lower final accuracy, higher confidence, and different patterns of judgment revision. The following analyses separate these outcomes by interface condition and AI correctness.

Table 4 summarizes performance by interface condition. In this table, shift is calculated as the percentage of trials within a condition on which the final judgment differs from the initial judgment, regardless of direction, correctness, or AI alignment. The no-AI baseline condition (C0) has the highest final accuracy at 73.4%; participants' unaided judgments are therefore already relatively strong in this item set. Among the AI-present conditions, C1 and C2 show the clearest descriptive decline from initial to final accuracy. In C1, accuracy decreases from 74.9% to 70.6%; in C2, it decreases from 75.4% to 70.0%. These two conditions also produce relatively high judgment-shift rates, at 21.1% and 20.9% respectively.

C3 and C4 show different patterns. C3 has a smaller decline than C1 and C2, moving from 74.0% to

Table 4: Judgment performance by interface condition.

Condition	N	Initial acc.	Final acc.	Shift	ΔConf.	Chosen-answer Brier
C0 No AI	350	74.3%	73.4%	3.1%	+2.23	0.207
C1 Prediction-only	350	74.9%	70.6%	21.1%	+5.09	0.245
C2 Static SHAP	350	75.4%	70.0%	20.9%	+8.12	0.232
C3 Sequential reveal	350	74.0%	73.1%	17.4%	+10.27	0.215
C4 Sequential + boundary cue	350	73.4%	71.7%	20.0%	+4.62	0.208

Table 5: AI alignment and direction of judgment change in AI-present conditions.

Condition	AI correct	Initial align	Final align	Toward AI	Away from AI
C1 Prediction-only	60.0%	58.6%	79.1%	20.9%	0.3%
C2 Static SHAP	60.0%	58.3%	79.1%	20.9%	0.0%
C3 Sequential reveal	60.0%	64.3%	80.0%	16.6%	0.9%
C4 Sequential + boundary cue	60.0%	56.3%	73.4%	18.6%	1.4%

73.1% and remaining closer to baseline. C4 has lower final accuracy, at 71.7%, but has the lowest final chosen-answer Brier score among AI-present conditions. Accuracy and the chosen-answer Brier score therefore do not point to the same condition-level ranking.

Confidence increases in every condition. The largest increase occurs in C3, where confidence rises by 10.27 points on average. C2 shows a larger descriptive confidence increase than C1, despite similar or lower final accuracy.

4.3 RQ1: Judgment change across interface conditions

RQ1 examines whether interface format affects how often participants revise their initial judgments and whether these revisions move toward or away from the AI recommendation.

To examine reliance on AI advice more directly, the next analysis focuses only on AI-present trials (C1–C4; $N = 1{,}400$). Table 5 reports initial alignment, final alignment, and the direction of judgment shifts relative to the AI recommendation.

The most stable pattern concerns the direction of judgment revision. C1 and C2 show almost no away-from-AI shifts (0.3% and 0.0%), while C3 and C4 show small away-from-AI rates (0.9% and 1.4%). Most observed judgment changes move toward rather than away from the AI recommendation.

Final alignment is highest in C3, at 80.0%, followed by C1 and C2 at 79.1%, and C4 at 73.4%. Higher alignment does not imply better performance by itself, because alignment is beneficial only when the AI is correct. High alignment is therefore ambiguous: it may reflect useful reliance when the AI is correct, but it may also reflect incorrect-AI alignment when the AI is wrong.

Taken together, AI-present conditions show substantially more judgment revision than C0, and the observed revisions are strongly asymmetric: participants mostly move toward rather than away from AI advice. These comparisons are descriptive because the planned models do not directly test overall shift frequency.

Table 6: AI alignment and confidence outcomes by AI correctness.

AI correctness	N	Final align	Final acc.	Shift	Toward AI	ΔConf.
AI wrong	560	58.2%	41.8%	30.0%	28.6%	+1.33
AI correct	840	91.1%	91.1%	13.1%	13.0%	+10.81

4.4 RQ2: Correctness-conditional advice use

RQ2 examines whether participants use AI advice differently when the recommendation is correct versus incorrect and whether this pattern varies across interface formats.

Table 6 separates AI-present trials according to whether the AI recommendation is correct. Final alignment is descriptively higher on AI-correct than AI-wrong trials. When the AI is correct, participants align with it on 91.1% of trials. When the AI is wrong, final alignment falls to 58.2%. Because AI correctness is an item-level property rather than a randomized manipulation, this difference may also reflect item difficulty.

The two rates differ descriptively. However, the AI-wrong alignment rate remains high. On trials where the AI recommendation is incorrect, participants still end with the AI-consistent answer in more than half of cases. This is high final alignment with incorrect AI advice. It does not by itself show that AI exposure causes the final AI-aligned answer.

The confidence pattern adds a second layer to this interpretation. On AI-correct trials, confidence increases by 10.81 points on average. On AI-wrong trials, confidence increases by 1.33 points. The average confidence increase is therefore much smaller on AI-wrong trials than on AI-correct trials. Nevertheless, final choices still often move into alignment with incorrect AI recommendations. In other words, the problematic pattern emerges more strongly at the level of judgment direction than at the level of confidence inflation.

Interface format under AI error. The main research question concerns whether explanation format changes advice use differently when the AI is correct versus incorrect. Table 7 reports condition-level outcomes separated by AI correctness.

The highest rates of incorrect-AI alignment and harmful movement appear in C1 and C2. When the AI is wrong, prediction-only advice produces a final alignment rate of 60.7% (85/140) and a toward-AI shift rate of 32.1% (45/140). Static SHAP explanations produce a very similar pattern, with final alignment at 61.4% (86/140) and toward-AI shift at 32.9% (46/140) under AI error. These one-trial differences are negligible. Thus, static explanations do not show a clear protective advantage over prediction-only advice, but they also do not provide clear evidence of greater harmful advice adoption.

C3 and C4 show comparatively favorable descriptive values on different indicators. C3 has the lowest toward-AI shift rate under AI error, at 22.1% (31/140), compared with 32.9% (46/140) in C2. This gives a 15-trial difference within the AI-wrong condition cells. C4 has the lowest final alignment with incorrect AI advice, at 52.1% (73/140), compared with 58.6% (82/140) in C3. It also has the lowest descriptive chosen-answer Brier score under AI error, at 0.329. These descriptive differences do not establish why the conditions differ.

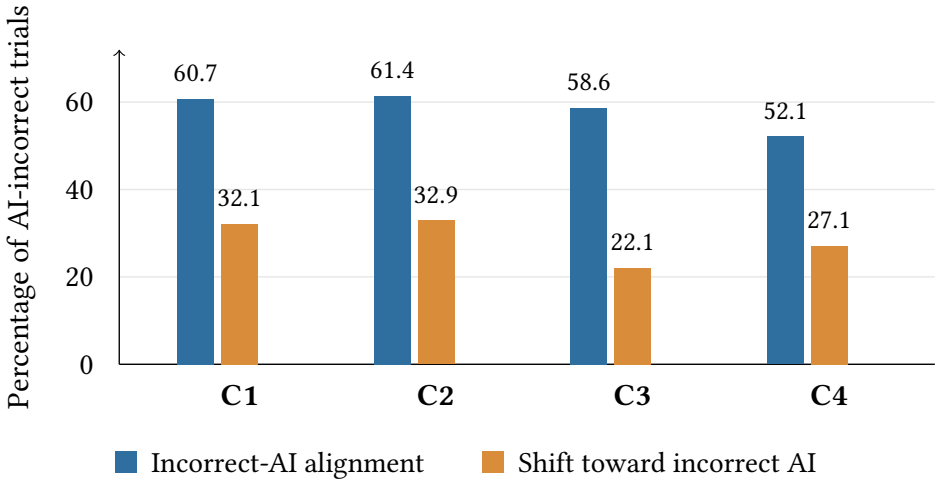


Figure 8: Descriptive reliance outcomes on AI-incorrect trials. C3 has the lowest shift-toward-incorrect-AI rate, while C4 has the lowest final incorrect-AI alignment.

Table 7: Interface-condition outcomes by AI correctness.

Condition	AI	*N*	Final align	Toward AI	Chosen-answer Brier
C1 Prediction-only	Wrong	140	60.7% (85/140)	32.1% (45/140)	0.407
C1 Prediction-only	Correct	210	91.4% (192/210)	13.3% (28/210)	0.136
C2 Static SHAP	Wrong	140	61.4% (86/140)	32.9% (46/140)	0.413
C2 Static SHAP	Correct	210	91.0% (191/210)	12.9% (27/210)	0.111
C3 Sequential reveal	Wrong	140	58.6% (82/140)	22.1% (31/140)	0.399
C3 Sequential reveal	Correct	210	94.3% (198/210)	12.9% (27/210)	0.092
C4 Sequential + boundary cue	Wrong	140	52.1% (73/140)	27.1% (38/140)	0.329
C4 Sequential + boundary cue	Correct	210	87.6% (184/210)	12.9% (27/210)	0.128

Note. Percentages for final alignment and toward-AI shifts include event counts in parentheses.

Table 7 and Figure 8 show that the condition ranking depends on the reliance measure. Descriptively, C3 has fewer shifts toward incorrect AI advice than C2, while C4 has lower final incorrect-AI alignment than C3. The corresponding model contrasts remain inconclusive: C3 versus C2 harmful advice adoption gives OR = 0.71 with interval [0.40, 1.27], and C4 versus C3 final incorrect-AI alignment gives OR = 0.93 with interval [0.54, 1.61].

Correction and deterioration by condition. To further characterize the cost-benefit profile of each interface, Figure 9 reports corrections and deteriorations by condition. A correction refers to a trial where the participant moves from an initially incorrect judgment to a finally correct judgment. A deterioration refers to the opposite pattern: initially correct to finally incorrect.

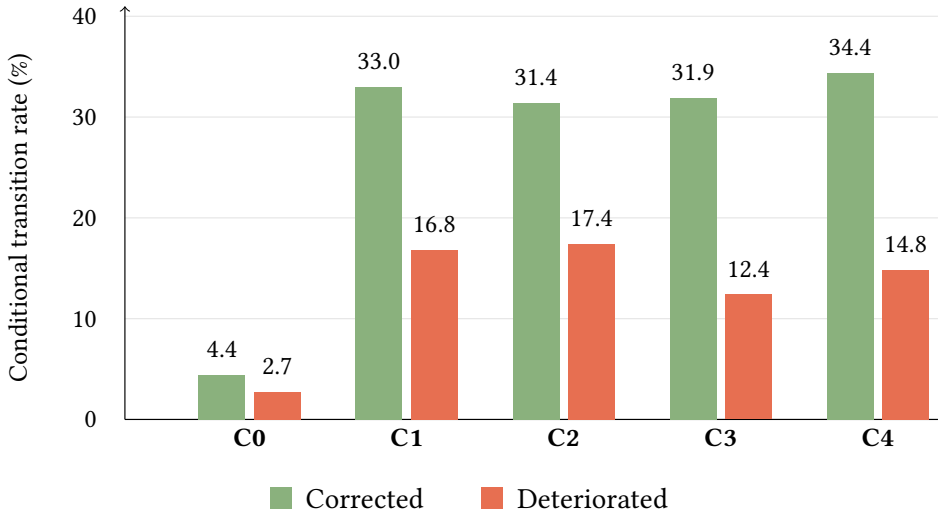


Figure 9: Conditional correction and deterioration rates by interface condition. Correction rate is $P(\text{final correct} \mid \text{initial incorrect})$, and deterioration rate is $P(\text{final incorrect} \mid \text{initial correct})$. Counts for corrections/deteriorations are C0: 4/7, C1: 29/44, C2: 27/46, C3: 29/32, and C4: 32/38.

In absolute counts, deteriorations exceed corrections in every condition. C1 has 29 corrections and 44 deteriorations; C2 has 27 corrections and 46 deteriorations. C3 and C4 show smaller absolute gaps, with 29 corrections versus 32 deteriorations in C3 and 32 corrections versus 38 deteriorations in C4. By contrast, C0 has few changes overall, as expected, because participants do not receive AI advice.

Because corrections and deteriorations have different risk sets, Figure 9 reports conditional rates. The correction rate is the percentage of initially incorrect trials that became correct. The deterioration rate is the percentage of initially correct trials that became incorrect.

The transition analysis shows that the cost-benefit profile depends on the starting judgment. Raw counts describe the overall trial burden, whereas conditional rates show how often initially incorrect judgments become correct and initially correct judgments deteriorate. C1 and C2 have the highest conditional deterioration rates. C3 and C4 have lower deterioration rates while maintaining similar or higher correction rates. These patterns describe asymmetric updating, but they require interpretation alongside initial-state denominators, alignment, and chosen-answer Brier scores rather than as standalone evidence of a harmful mechanism.

4.5 RQ3: Confidence and chosen-answer Brier score

RQ3 examines whether explanation format affects confidence and whether confidence corresponds to the correctness of participants' final answers.

Confidence increases in every condition, but the increase does not track final accuracy. C3 produces the largest mean confidence increase, at +10.27 points, while C4 produces a smaller confidence increase, at +4.62 points. C2 also increases confidence more than C1 (+8.12 versus +5.09) even though C2 does not produce higher final accuracy. Confidence change alone is therefore not a sufficient indicator of improved AI-assisted judgment.

Table 8 summarizes the confidence and chosen-answer Brier outcomes used for RQ3. Overall chosen-answer Brier scores are lowest in C0 and C4 (0.207 and 0.208), followed by C3 (0.215), C2 (0.232), and C1 (0.245). Among AI-present conditions, C4 has the lowest descriptive Brier score when the AI is wrong (0.329), whereas C3 has the lowest descriptive Brier score when the AI is correct (0.092). These patterns

do not give the same ranking as final accuracy or advice adoption.

Table 8: Confidence change and chosen-answer Brier score by interface condition.

Condition	ΔConf.	Overall Brier	Brier, AI wrong	Brier, AI correct
C0 No AI	+2.23	0.207	–	–
C1 Prediction-only	+5.09	0.245	0.407	0.136
C2 Static SHAP	+8.12	0.232	0.413	0.111
C3 Sequential reveal	+10.27	0.215	0.399	0.092
C4 Sequential + boundary cue	+4.62	0.208	0.329	0.128

Note. Lower chosen-answer Brier scores indicate better confidence–correctness correspondence for the selected answer. C0 has no AI-wrong or AI-correct strata because participants do not receive AI advice.

Because the chosen-answer Brier score jointly reflects final accuracy and confidence extremity, we interpret it as confidence–correctness correspondence rather than pure calibration. The descriptive C4 advantage over C3 under AI error is not clearly supported by the mixed model ($b = -0.036$, 95% interval [-0.087, 0.014]). Thus, confidence change and chosen-answer Brier score do not provide clear evidence that one explanation format performs best across AI-correct and AI-incorrect trials.

4.6 Hypothesis-oriented model summary and sensitivity analyses

Table 9 summarizes the planned contrasts. It separates descriptive patterns from model-estimated contrasts and marks unsupported or inconclusive results explicitly. The adoption model (M5) is especially relevant for harmful movement toward AI because it excludes trials where participants already agreed with the AI before exposure.

Appendix Figure 11 visualizes selected coefficients from the exploratory models. The large AI-correctness coefficient should be read cautiously because it is partly structural: when participants align with correct advice, their final response is correct, whereas alignment with incorrect advice produces an incorrect response. Under AI error, static SHAP does not differ clearly from prediction-only advice in final alignment or advice adoption. Sequential reveal shows lower odds of adopting incorrect AI advice than static SHAP in direction (M5: OR = 0.71), but the posterior interval crosses 1. The boundary-cue condition does not show lower odds of incorrect-advice adoption than sequential reveal (M5: OR = 1.14). The direct C4 versus C3 contrast in the trial-level Brier model also does not clearly support its lower descriptive chosen-answer Brier score under AI error (M4: $b = -0.036$, interval crossing 0).

Thus, the results do not support the expectation that static SHAP would increase AI alignment beyond prediction-only advice; instead, the two conditions produce nearly identical descriptive patterns.

Table 9: Hypothesis-oriented decision table.

Hyp.	Primary estimand	Estimate	Interval	Interpretation
H1a	Judgment shift in AI-present conditions versus C0	19.9% (278/1400) vs. 3.1% (11/350)	Descriptive summary	Descriptive only; planned models focus on reliance/adoption.
H1b	Direction of AI-present judgment shifts	Toward AI: 19.2% (269/1400); away: 0.6% (9/1400)	Descriptive summary	Descriptive only; planned models focus on reliance/adoption.
H2	C2 versus C1 final AI alignment	79.1% vs. 79.1%	Descriptive summary	Not clearly supported.
H3	C2 versus C1 final alignment under AI error (M3)	OR = 1.04	[0.72, 1.50]	Not clearly supported.
H4	C3 versus C2 harmful advice adoption under AI error (M5)	OR = 0.71	[0.40, 1.27]	Consistent with expectation, but inconclusive.
H4	C3 versus C2 final alignment under AI error (M3)	OR = 0.64	[0.37, 1.12]	Consistent with expectation, but inconclusive.
H5	C4 versus C3 final incorrect-AI alignment (M3)	OR = 0.93	[0.54, 1.61]	Not clearly supported.
H5	C4 versus C3 beneficial advice adoption when AI is correct (M5)	OR = 0.35	[0.11, 1.12]	Possible trade-off; inconclusive.
H5*	C4 versus C3 chosen-answer Brier score under AI error (M4)	$b = -0.036$	[-0.087, 0.014]	Related confidence outcome; inconclusive.

Note. The table summarizes H1a, H1b, and H2 descriptively because the planned inferential contrasts focus on correctness-conditional reliance and advice adoption. M3 and M5 are Bayesian logistic mixed-effects models; entries report odds ratios with approximate 95% posterior intervals. For odds ratios, values below 1 indicate lower odds for the first condition relative to the comparison condition, whereas values above 1 indicate higher odds. Intervals that include 1 do not provide clear evidence of a directional difference. M5 uses only trials where the initial judgment does not match the AI recommendation. M4 is a linear mixed-effects model; entries report coefficient differences with 95% Wald confidence intervals. For M4, intervals that include 0 do not provide clear evidence of a coefficient difference. *The M4 contrast is related to H5's confidence component, but it does not directly test confidence increases on incorrect final judgments; that component of H5 is therefore not directly tested. None of the planned Holm-adjusted contrasts in Appendix E.6 meet the conventional $p < .05$ threshold, so we do not describe these model-based contrasts as statistically significant. Appendix E reports full coefficients for M1, M2, M3, M4, and M5 (Appendix E.1, E.2, E.3, E.4, and E.5), along with the planned contrasts in Appendix E.6.

The sensitivity checks do not change the main interpretation. A participant-condition aggregated Brier check shows the same C4 direction under AI error, but not a uniformly favorable pattern across AI correctness; Appendix E.7 reports this check. Adding the displayed model score to the final-alignment model leaves the planned condition contrasts almost unchanged, as shown in Appendix E.8. Process logs also show longer interface-stage duration in sequential conditions (median 7.8 seconds in C3 and 8.8 seconds in C4, compared with 2.6 seconds in C1 and 4.0 seconds in C2). Appendix E.9 reports the duration-exclusion sensitivity check, and Appendix E.10 reports exploratory participant-level moderator checks. These timing differences matter for interpretation: C2 versus C3 and C3 versus C4 are not pure format contrasts.

4.7 Item heterogeneity and sensitivity analyses

Finally, we inspect item-level results to identify whether reliance patterns concentrate in particular stimulus types. Table 10 lists the items with the lowest final accuracy. These items do not automatically indicate invalid stimuli. Rather, they indicate which types of claims are most susceptible to AI-driven updating.

The lowest-accuracy items vary in topic and claim type. Because item characteristics are identified post hoc and are not independently coded or experimentally manipulated, this table is treated as a descriptive sensitivity check rather than evidence of item-level moderators.

Table 10: Lowest final-accuracy items by internal task ID.

Internal ID	Short description	Truth	AI	Initial acc.	Final acc.
13	Canada's personal-data rules	FAKE	TRUE	51.4%	27.1%
16	Harris' father as Marxist economist	TRUE	FAKE	50.0%	28.6%
22	Eclipse image behind CN Tower	FAKE	TRUE	64.3%	38.6%
20	Canadian tourist-pass conspiracy	FAKE	TRUE	71.4%	44.3%
21	Blue-ringed octopus venom	TRUE	FAKE	67.1%	44.3%
0	The Orbanisation of America	TRUE	FAKE	70.0%	47.1%
12	Hurricane Helene geoengineering claim	FAKE	TRUE	72.9%	52.9%
6	Covid vaccines and cancer claim	FAKE	TRUE	74.3%	57.1%

Note. Internal IDs follow the archived task records and match the ordering in Appendix D.

5 Discussion

Our results do not fit a simple account in which more transparency improves performance. In this task, the presence of an explanation says less than the pattern of judgment changes, final alignment, and confidence-related performance when advice is correct or incorrect. No single ranking of the five interfaces captures all of these outcomes.

The discussion focuses on three implications: final agreement is not a sufficient measure of reliance, static SHAP does not clearly protect against incorrect advice, and sequential reveal and the boundary cue involve different trade-offs.

5.1 Principal findings

AI advice carries considerable weight in the second judgment stage, and most judgment changes in AI-present trials move toward the AI recommendation. Yet final agreement alone does not identify reliance because some participants already agree with the AI before seeing the recommendation. Final alignment therefore needs to be read together with toward-AI shifts, deterioration, and advice adoption among initially non-aligned trials. AI correctness is equally important: alignment with correct advice can support performance, whereas alignment with incorrect advice can harm it. This distinction keeps the interpretation focused on selective reliance rather than global agreement with AI (Lee & See, 2004; Schemmer et al., 2023).

Judgment quality and subjective certainty move in different directions: final judgments become slightly less accurate, while confidence increases. The second judgment stage gives participants another opportunity to reconsider, but here it leaves them more certain without making them more correct. Greater confidence in an incorrect final judgment could make later revision more difficult, although the present study does not test this possibility.

5.2 Prediction-only advice and static SHAP

Prediction-only advice and static SHAP lead to the same interpretation: static token-level attribution cues do not function as a clear safeguard against incorrect AI recommendations in this interface and task context. Under AI error, C1 and C2 show very similar final-alignment and harmful-shift patterns. This does not mean that static SHAP adds harm beyond prediction-only advice. It means that adding a static explanation does not, by itself, make reliance more selective.

The static-attribution finding narrows the claim that can be made about these displays. In this task, the

static display gives participants more model-derived information, but it does not clearly help them resist incorrect advice. This pattern is consistent with prior behavioral evaluations showing that explanations can increase acceptance or model simulatability without reliably improving error detection or final decisions (Bansal et al., 2021; Poursabzi-Sangdeh et al., 2021). Explanation content therefore needs behavioral evaluation; researchers should not assume that making model output more visible makes judgment more selective.

5.3 Sequential reveal, boundary cue, and selectivity

Sequential reveal and the boundary cue raise different interpretive possibilities. C3 looks most favorable on harmful movement toward incorrect AI advice. C4 looks more favorable on final incorrect-AI alignment and the chosen-answer Brier score under AI error. We should not treat these patterns as evidence of distinct cognitive mechanisms, because we do not directly measure processing depth, attention, perceived AI authority, understanding of attribution cues, or uptake of the boundary cue.

The C4 pattern matters because it suggests a selectivity trade-off. Appropriate reliance requires users to resist incorrect recommendations without broadly rejecting useful ones (Lee & See, 2004). Lower alignment under AI error is desirable only if the interface preserves alignment with correct AI advice. C4 shows lower final alignment when the AI is wrong, but it also shows lower alignment than C3 when the AI is correct. The model results do not clearly establish this difference as an effect, but the pattern still cautions against treating the boundary cue as straightforwardly beneficial. It may encourage generalized resistance rather than calibrated resistance.

The timing evidence introduces a separate interpretation problem. Cognitive-forcing interfaces are designed to prompt more deliberate engagement and have been shown to reduce overreliance, while progressive disclosure provides a rationale for presenting information in stages (Buçinca et al., 2021; Muralidhar, 2024). In our study, sequential conditions require additional reveal interactions and involve longer interface-stage duration. Any C2 versus C3 or C3 versus C4 difference may therefore reflect pacing, interaction effort, or time-on-task rather than presentation sequence alone. The lower harmful-shift rate in C3 could come from staged visual presentation, added friction, longer deliberation time, or a combination of these factors. Our design cannot separate these explanations.

C3 and C4 should therefore not be read as two increasingly stronger safeguards. They appear to change different parts of the decision process and may introduce different costs. For design, the useful question is whether an interface helps users resist incorrect advice without discouraging appropriate use of correct advice.

5.4 Methodological contribution

Our clearest contribution is methodological. By separating final AI agreement, advice-induced movement, final accuracy, deterioration, confidence change, and the chosen-answer Brier score, we show why these outcomes should not be treated as interchangeable measures of explanation quality or appropriate reliance. Using final agreement as the sole reliance measure may classify pre-existing agreement as advice-induced reliance.

In practice, the same interface can look favorable on one outcome but not another. A single global metric would obscure this outcome dependence and could make a design look beneficial even when it shifts the trade-off rather than improving selectivity.

5.5 Design implications

Our findings suggest three cautious design implications. First, designers should not assume that attribution explanations protect users merely because they provide more information. Static token highlights show no clear protective advantage over prediction-only advice.

Second, explanation timing warrants further investigation together with interaction time and pacing. Sequential reveal may change how users engage with advice, but this experiment cannot separate presentation sequence from time-on-task and interaction effort. A follow-up study could compare sequential reveal with a time-matched static interface or a static interface with an equivalent enforced delay. Such a comparison would help isolate sequence from friction.

Third, researchers should evaluate boundary cues for selectivity, not only for reduced agreement with AI. A cue that lowers alignment with incorrect advice may still be problematic if it also lowers beneficial reliance when the AI is correct. Generic warnings may reduce harmful reliance while encouraging unnecessary skepticism toward useful AI support. Context-sensitive uncertainty cues may be more useful than broad disclaimers, but this possibility requires direct testing.

5.6 Limitations

Several limitations guide interpretation. We use short headline-and-summary stimuli rather than full news articles, so readers should understand the task as a controlled veracity-judgment task based on limited text. The sample is moderate: 70 participants complete 1,750 trials, but each participant sees only a small number of AI-incorrect trials per condition. We therefore interpret small condition differences cautiously, especially when model intervals are wide.

The stimulus labels and fixed AI predictions also constrain interpretation. The stimulus pool intentionally includes AI errors, and AI-correct and AI-wrong items may differ in difficulty, truth structure, or topic. The item-level analysis suggests that low-accuracy items often appear counterintuitive, politically charged, visually dependent, or institutionally complex, but these characteristics come from post hoc interpretation. Future work should treat them as hypotheses rather than established moderators.

We do not directly measure the proposed cognitive mechanisms. We also do not include manipulation checks for whether participants notice, understand, or use the SHAP highlights. Some participants may have treated the highlighted tokens as decorative emphasis, ignored them, or misunderstood what the colors and reveal sequence indicated. This limits interpretation of the explanation-format comparisons: observed differences cannot show that participants processed the attribution cues as intended. The same issue applies to the boundary cue, which participants may not have read closely or interpreted as a reminder of model fallibility. The analysis therefore treats SHAP highlights and the boundary cue as interface cues presented to participants, not as verified inputs to participants' reasoning.

Finally, the outcome measures have limits. The binary TRUE/FAKE response captures categorical judgment revision rather than continuous belief updating. The task measures confidence for the chosen answer, not as a direct probability that the claim is true. The chosen-answer Brier score therefore reflects confidence–correctness correspondence for the selected response, so we do not treat it as a pure calibration measure.

5.7 Future work

Future studies should test these explanation formats with larger and more diverse samples, which would provide more precise estimates of responses to AI errors and allow comparisons across user groups.

Richer news materials, including longer articles, source cues, images, and platform context, could test whether the observed judgment patterns extend beyond a controlled headline-and-summary task.

A second direction is to study explanation processing more directly. Follow-up studies could add manipulation checks, attention measures, or post-task comprehension questions to assess whether users notice and understand SHAP highlights, sequential reveal steps, and boundary cues. This would help distinguish interface exposure from actual explanation use.

Finally, future work could compare token-level SHAP explanations with other forms of AI support, such as example-based explanations, uncertainty displays, model-level accuracy information, explicit warnings about possible AI error, or generative-AI explanations that interpret evidence in an audience-targeted narrative. Such systems may make explanations easier to read, but they could also introduce persuasive framing or unsupported rationales. These comparisons would clarify whether the present findings are specific to attribution-based explanations or reflect broader patterns in AI-assisted misinformation evaluation.

5.8 Implementation, reproducibility, and ethics

The experiment uses a custom online interface and fixed item-level AI advice. The appendix reports the questionnaire items, interface screenshots, stimulus provenance, and model summaries, and the Procedure section links the public interface demonstration. The code and data used for the analyses are available in a public GitHub repository.[6]

XAI also has an ethical role because explanations can shape trust, resistance, and confidence in automated advice. In misinformation judgment, an explanation may be technically faithful while still encouraging agreement with an incorrect recommendation. We therefore treat XAI as a user-facing intervention and use the study to test whether explanation formats support calibrated reliance, complementing model-centered XAI evaluation with evidence about decisions, confidence, and exposure to AI error.

The procedure includes informed consent before data collection and a debrief after the task, as described in Section 3.1. We collect no names, email addresses, visual data, audio data, biometric data, or other directly identifying information. Because the task involves potentially misleading news-style items and fallible AI advice, the debrief explains why the study includes misleading content and imperfect AI advice, offers optional item clarification, and provides a downloadable or savable debrief record.

6 Conclusion

We examine how AI advice and token-level explanation presentation relate to judgment updating, alignment, confidence change, and confidence–correctness correspondence in a controlled misinformation-evaluation task. The results do not show a simple overall accuracy benefit of AI assistance, nor do they show that static SHAP improves resistance to incorrect advice relative to prediction-only output.

Explanation formats produce different patterns across outcomes. Sequential reveal and the boundary cue look comparatively favorable on different measures, but the exploratory analyses do not establish that either format improves selective reliance. Final AI agreement, advice-induced movement, accuracy, deterioration, and confidence-related performance should therefore remain separate outcomes.

The conclusions are limited to the model, SHAP presentation, stimulus set, participant sample, and online task used here. Broader recommendations about explanation timing or boundary cues require

[6] https://github.com/itsyuning/xai-ui-study

larger samples, direct measures of the proposed cognitive mechanisms, and richer item-level controls. XAI evaluation should distinguish increased agreement from appropriate updating: users need to benefit from correct advice without becoming more likely to follow incorrect advice.

Appendix overview

The appendices provide supporting material for the Method and Results sections. Appendix A describes the final participant sample. Appendix B lists the background questionnaire items. Appendix C shows the shared task-flow screens. Appendix D reports stimulus provenance and fixed AI predictions. Appendix E provides the exploratory mixed-effects model details, planned contrasts, and sensitivity checks.

A Participant background characteristics

Table 11 summarizes the background characteristics of the final matched participant sample.

Table 11: Participant background characteristics.

Characteristic	Summary
Final matched sample	$N = 70$ participants
Age group	18–24: 25.7%; 25–34: 38.6%; 35–44: 22.9%; 45–54: 7.1%; 55–64: 5.7%.
Gender	Female: 50.0%; Male: 47.1%; Non-binary / third gender: 2.9%.
Education level	Bachelor's: 52.9%; Master's: 24.3%; other or lower levels: 21.4%; prefer not to say: 1.4%.
English reading comfort	$M = 6.57$, $SD = 0.86$ on a 7-point scale.
Online news-reading frequency	At least 3 days per week: 74.3%; less often: 25.7%.
Familiarity with misinformation/fact-checking	$M = 5.13$, $SD = 1.32$ on a 7-point scale.
Caution toward AI-generated outputs	$M = 5.46$, $SD = 1.20$ on a 7-point scale.

B Full background questionnaire items

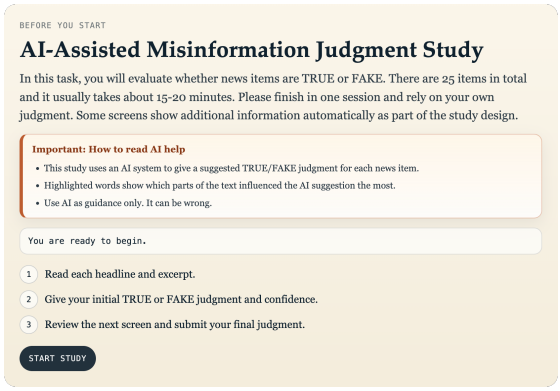
This appendix reproduces the participant background questionnaire items from the study, excluding the consent item.

Variable	Question wording	Response format
Age	How old are you?	Single choice: 18–24 years old; 25–34 years old; 35–44 years old; 45–54 years old; 55–64 years old.
Gender	How do you describe yourself?	Single choice: Male; Female; Non-binary / third gender; optional self-describe free-text field.
Culture	List of Countries	Single choice from country list.
First language	What is your first or primary language?	Open text.

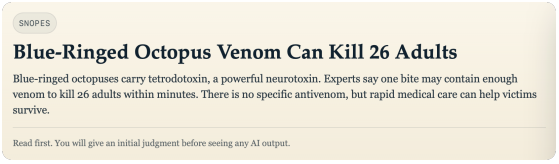
Variable	Question wording	Response format
Education	What is the highest level of education you have completed?	Single choice: Secondary; Vocational or Similar; Some University but no degree; University - Bachelors Degree; Master's degree (MA, MSc, MPhil, etc.); Prefer not to say.
English comfort	How comfortable are you reading short English news texts?	7-point scale; higher values indicate greater comfort (labeled points in export include 4 = Moderately comfortable, 7 = Very comfortable).
English reading ability	How would you rate your English reading ability?	7-point scale; higher values indicate stronger ability (labeled points in export include 4 = Moderate, 7 = Very strong).
News frequency	How often do you read news online?	Single choice: Never; Less than once a week; 1 to 2 days a week; 3 to 6 days a week; Every day.
Misinformation familiarity	How familiar are you with misinformation, fake news, and fact-checking practices?	7-point scale: 1 = Not at all familiar; 4 = Moderately familiar; 7 = Very familiar.
AI usefulness	How helpful do you generally find AI-generated predictions or explanations when making judgments?	7-point scale: 1 = Not at all helpful; 4 = Moderately helpful; 7 = Very helpful.
AI verification tendency	When AI-generated predictions or explanations are provided, how likely are you to verify them before relying on them?	7-point scale: 1 = Very unlikely; 4 = Moderately likely; 7 = Very likely.
AI caution	How cautious are you when using AI-generated outputs to make judgments?	7-point scale; higher values indicate greater caution (labeled points in export include 4 = Moderately cautious, 7 = Very cautious).

C Task-flow interface screenshots

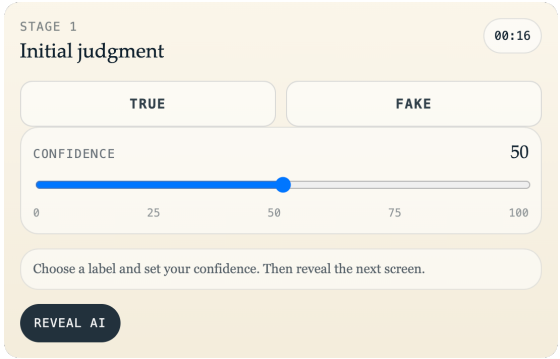
Figure 10 documents the shared task-flow screens that participants saw before and after the condition-specific interface stage.



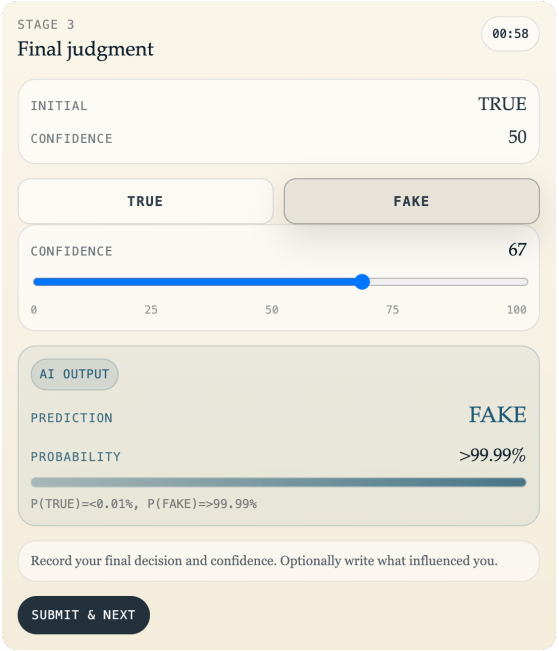
(a) Instruction screen



(b) Stimulus reading view



(c) Initial judgment stage



(d) Final judgment stage

Figure 10: Task-flow interface screenshots for the instruction, item reading, initial judgment, and final judgment stages.

D Stimulus items and AI predictions

Table 13 lists the 25 stimulus items in the experiment, ordered by internal task ID. The table includes topic category, ground-truth label, AI prediction, AI correctness, and source provenance.

Table 13: Stimulus provenance and AI correctness by internal task ID.

Internal ID	Topic	Truth	AI pred.	AI correct	Source/provenance	Title
0	POLITICS	TRUE	FAKE	No	European Council on Foreign Relations	The Orbanisation of America: Hungary's lessons for Donald Trump
1	POLITICS	TRUE	TRUE	Yes	Center on Budget and Policy Priorities	Congress begins new budget talks as shutdown deadlines approach
2	POLITICS	FAKE	FAKE	Yes	Snopes	Mormon Founder Joseph Smith Was 1st Assassinated US Presidential Candidate
3	POLITICS	FAKE	FAKE	Yes	Snopes	Viral post claims a Facebook copyright disclaimer blocks photo use
4	POLITICS	TRUE	TRUE	Yes	APRI - Africa Policy Research Institute	South Africa coalition talks continue amid disputes over cabinet posts and budget plans
5	HEALTH	TRUE	TRUE	Yes	The World Economic Forum	Researchers report advances across several cancer treatment approaches
6	HEALTH	FAKE	TRUE	No	AFP Fact Check	Covid vaccines linked to study on rising cancer cases
7	HEALTH	TRUE	FAKE	No	University of Bristol	August: COVID-19 and mental illnesses
8	HEALTH	TRUE	TRUE	Yes	Cancer News	NHS expands enrollment pathways for personalized cancer vaccine trials
9	HEALTH	TRUE	TRUE	Yes	Frontiers	German hospitals expand sustainability initiatives in clinical operations
10	SCI_TECH	TRUE	TRUE	Yes	Rice University	Artemis II delay leads NASA to review mission-readiness plans
11	SCI_TECH	TRUE	TRUE	Yes	cio.com	NASA appoints its first chief AI officer
12	SCI_TECH	FAKE	TRUE	No	AFP Fact Check	Hurricane Helene 'geoengineered' to expand lithium mining
13	SCI_TECH	FAKE	TRUE	No	AFP Fact Check	Canada's personal-data rules were quietly expanded
14	SCI_TECH	TRUE	TRUE	Yes	Cybernews	NASA says AI tools could help plan future life-detection missions
15	BUSINESS	FAKE	FAKE	Yes	Snopes	Trump Said He Would Not Defend NATO Allies If Russia Attacks Them

Internal ID	Topic	Truth	AI pred.	AI correct	Source/provenance	Title
16	BUSINESS	TRUE	FAKE	No	Snopes	Kamala Harris' Father Has Been Described as a Marxist Economist
17	BUSINESS	TRUE	TRUE	Yes	Nature	Saudi labour-market study analyses links between inflation, oil prices, trade, and unemployment
18	BUSINESS	TRUE	TRUE	Yes	PolitiFact	Crime is down in Venezuela, not as much as Donald Trump says
19	BUSINESS	TRUE	FAKE	No	Investopedia	What Is Stagflation, What Causes It, and Why Is It Bad?
20	SOCIETY	FAKE	TRUE	No	AFP Fact Check	Posts claim canadian tourist pass reignites surveillance conspiracy theories
21	SOCIETY	TRUE	FAKE	No	Snopes	Blue-Ringed Octopus Venom Can Kill 26 Adults
22	SOCIETY	FAKE	TRUE	No	AFP Fact Check	Viral posts claim eclipse image behind CN Tower was real
23	SOCIETY	FAKE	FAKE	Yes	AFP Fact Check	Posts Claim Australia Listed George Soros as a 'Global Terrorist'
24	SOCIETY	FAKE	FAKE	Yes	AFP Fact Check	Posts Claim Australia Signed WEF Treaty to Close Banks

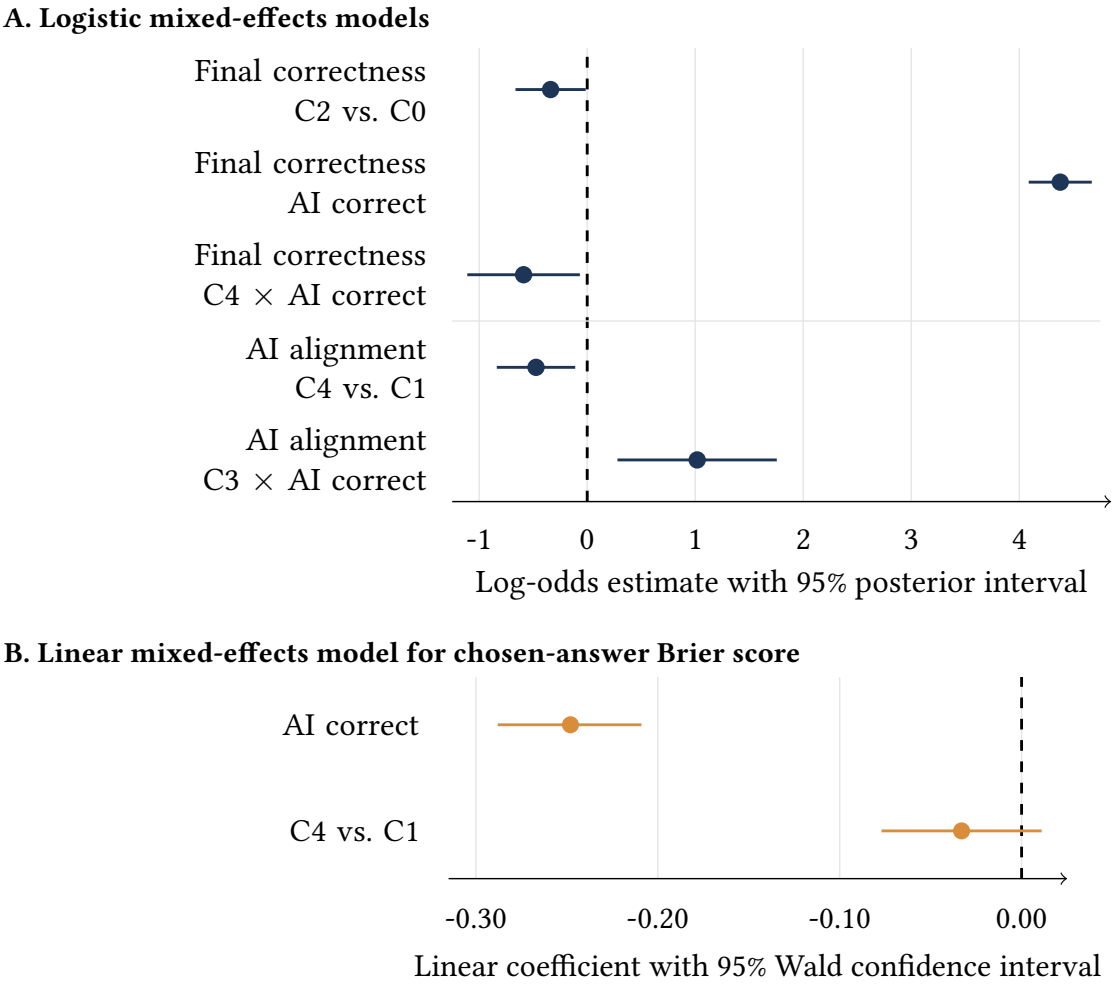


Figure 11: Selected fixed-effect estimates from the exploratory mixed-effects analyses. Points indicate coefficient estimates and horizontal lines indicate 95% intervals. Panel A shows log-odds coefficients with approximate posterior intervals for final correctness and AI alignment. Panel B shows linear coefficients with Wald confidence intervals for chosen-answer Brier score. Negative values in Panel B indicate lower chosen-answer Brier scores and better confidence–correctness correspondence. Dashed vertical lines mark the null value of zero.

E Exploratory mixed-effects model details

This appendix reports full fixed-effect coefficients for the exploratory mixed-effects analyses in the Results section. Figure 11 first gives a compact visual summary of focal coefficients. Appendices E.1–E.5 then report the five model coefficient tables. Appendix E.6 reports the pre-specified pairwise contrasts, and the remaining subsections report robustness, sensitivity, and exploratory moderator checks.

For Bayesian logistic mixed models (M1–M3 and M5), coefficients are posterior means from statsmodels `BinomialBayesMixedGLM` variational Bayes estimation, with approximate 95% posterior intervals (mean $\pm$ 1.96 posterior SD) and odds ratios (OR). For the linear mixed model (M4), fitted with statsmodels `MixedLM`, we report coefficients with standard errors, 95% Wald confidence intervals, and p-values.

E.1 M1: Final correctness on all trials

Term	β	95% posterior interval	OR	OR 95% posterior interval
Intercept	-0.975	[-1.125, -0.824]	0.38	[0.32, 0.44]
C1 vs C0	-0.279	[-0.606, 0.048]	0.76	[0.55, 1.05]
C2 vs C0	-0.339	[-0.664, -0.014]	0.71	[0.52, 0.99]
C3 vs C0	0.055	[-0.309, 0.419]	1.06	[0.73, 1.52]
C4 vs C0	-0.139	[-0.468, 0.189]	0.87	[0.63, 1.21]
Initial correctness	3.988	[3.802, 4.175]	53.97	[44.79, 65.02]
Initial confidence (z)	0.223	[0.075, 0.372]	1.25	[1.08, 1.45]

E.2 M2: Final correctness on AI-present trials

Term	β	95% posterior interval	OR	OR 95% posterior interval
Intercept	-3.653	[-3.823, -3.484]	0.03	[0.02, 0.03]
C2 vs C1	-0.161	[-0.492, 0.170]	0.85	[0.61, 1.18]
C3 vs C1	0.255	[-0.115, 0.625]	1.29	[0.89, 1.87]
C4 vs C1	0.292	[-0.034, 0.617]	1.34	[0.97, 1.85]
AI correct	4.379	[4.087, 4.671]	79.74	[59.56, 106.77]
C2 × AI correct	0.245	[-0.333, 0.824]	1.28	[0.72, 2.28]
C3 × AI correct	0.244	[-0.432, 0.921]	1.28	[0.65, 2.51]
C4 × AI correct	-0.589	[-1.110, -0.068]	0.55	[0.33, 0.93]
Initial correctness	4.010	[3.804, 4.216]	55.15	[44.89, 67.75]
Initial confidence (z)	0.219	[0.051, 0.387]	1.24	[1.05, 1.47]

E.3 M3: Final alignment on AI-present trials

Term	β	95% posterior interval	OR	OR 95% posterior interval
Intercept	-0.219	[-0.408, -0.030]	0.80	[0.66, 0.97]
C2 vs C1	0.038	[-0.330, 0.406]	1.04	[0.72, 1.50]
C3 vs C1	-0.402	[-0.816, 0.012]	0.67	[0.44, 1.01]
C4 vs C1	-0.474	[-0.837, -0.111]	0.62	[0.43, 0.89]
AI correct	0.935	[0.610, 1.259]	2.55	[1.84, 3.52]
C2 × AI correct	0.071	[-0.566, 0.709]	1.07	[0.57, 2.03]
C3 × AI correct	1.018	[0.280, 1.755]	2.77	[1.32, 5.78]
C4 × AI correct	0.271	[-0.304, 0.846]	1.31	[0.74, 2.33]
Initial alignment	4.834	[4.187, 5.480]	125.65	[65.83, 239.85]
Initial confidence (z)	-0.598	[-0.789, -0.408]	0.55	[0.45, 0.67]

E.4 M4: Final chosen-answer Brier score on AI-present trials

Term	b	95% Wald confidence interval	SE	p
Intercept	0.253	[0.221, 0.285]	0.016	$< .001$
C2 vs C1	0.015	[-0.029, 0.058]	0.022	.502
C3 vs C1	0.003	[-0.047, 0.054]	0.026	.899
C4 vs C1	-0.033	[-0.077, 0.011]	0.022	.140
AI correct	-0.248	[-0.288, -0.209]	0.020	$< .001$
C2 × AI correct	-0.037	[-0.096, 0.022]	0.030	.217
C3 × AI correct	-0.054	[-0.119, 0.011]	0.033	.106
C4 × AI correct	0.021	[-0.039, 0.080]	0.031	.502
Initial chosen-answer Brier	0.601	[0.557, 0.644]	0.022	$< .001$

E.5 M5: Advice adoption on initial non-alignment trials

Term	β	95% posterior interval	OR	OR 95% posterior interval
Intercept	-0.220	[-0.419, -0.021]	0.80	[0.66, 0.98]
C2 vs C1	-0.008	[-0.395, 0.379]	0.99	[0.67, 1.46]
C3 vs C1	-0.348	[-0.785, 0.088]	0.71	[0.46, 1.09]
C4 vs C1	-0.220	[-0.601, 0.161]	0.80	[0.55, 1.18]
AI correct	0.674	[0.319, 1.028]	1.96	[1.38, 2.80]
C2 × AI correct	0.169	[-0.522, 0.860]	1.18	[0.59, 2.36]
C3 × AI correct	1.221	[0.423, 2.018]	3.39	[1.53, 7.52]
C4 × AI correct	0.034	[-0.601, 0.669]	1.03	[0.55, 1.95]
Initial confidence (z)	-0.649	[-0.850, -0.448]	0.52	[0.43, 0.64]

E.6 Planned contrasts from M3, M4, and M5

Model / outcome	Trials	Contrast	Estimate	95% interval	Holm p
M3 final alignment	AI wrong	C2 vs C1	OR = 1.04	[0.72, 1.50]	1.000
M3 final alignment	AI wrong	C3 vs C2	OR = 0.64	[0.37, 1.12]	.359
M3 final alignment	AI wrong	C4 vs C3	OR = 0.93	[0.54, 1.61]	1.000
M3 final alignment	AI correct	C2 vs C1	OR = 1.12	[0.53, 2.33]	.772
M3 final alignment	AI correct	C3 vs C2	OR = 1.66	[0.54, 5.09]	.751
M3 final alignment	AI correct	C4 vs C3	OR = 0.44	[0.15, 1.31]	.418
M5 advice adoption	AI wrong	C2 vs C1	OR = 0.99	[0.67, 1.46]	1.000
M5 advice adoption	AI wrong	C3 vs C2	OR = 0.71	[0.40, 1.27]	.760
M5 advice adoption	AI wrong	C4 vs C3	OR = 1.14	[0.64, 2.03]	1.000
M5 advice adoption	AI correct	C2 vs C1	OR = 1.17	[0.53, 2.59]	.690
M5 advice adoption	AI correct	C3 vs C2	OR = 2.04	[0.61, 6.80]	.495
M5 advice adoption	AI correct	C4 vs C3	OR = 0.35	[0.11, 1.12]	.231
M4 chosen-answer Brier	AI wrong	C2 vs C1	$b = 0.014$	[-0.036, 0.065]	1.000
M4 chosen-answer Brier	AI wrong	C3 vs C2	$b = -0.012$	[-0.062, 0.038]	1.000
M4 chosen-answer Brier	AI wrong	C4 vs C3	$b = -0.036$	[-0.087, 0.014]	.487
M4 chosen-answer Brier	AI correct	C2 vs C1	$b = -0.022$	[-0.063, 0.019]	.361
M4 chosen-answer Brier	AI correct	C3 vs C2	$b = -0.028$	[-0.069, 0.013]	.361
M4 chosen-answer Brier	AI correct	C4 vs C3	$b = 0.037$	[-0.004, 0.078]	.231

For M3 and M5, intervals are approximate posterior intervals from the variational posterior. For M4, intervals are Wald confidence intervals. Holm-adjusted p-values are approximate two-sided values computed within each model and AI-correctness stratum. Odds ratios below 1 indicate lower odds for the first condition named in the contrast relative to the second condition.

E.7 Participant-aggregated chosen-answer Brier robustness check

Trials	Condition or contrast	Mean or difference	95% interval	p
AI wrong	C1	0.417	[0.350, 0.485]	–
AI wrong	C2	0.437	[0.371, 0.503]	–
AI wrong	C3	0.416	[0.347, 0.486]	–
AI wrong	C4	0.336	[0.269, 0.403]	–
AI wrong	C2 – C1	0.020	[-0.062, 0.101]	.632
AI wrong	C3 – C2	-0.021	[-0.102, 0.060]	.612
AI wrong	C4 – C3	-0.080	[-0.157, -0.004]	.040
AI correct	C1	0.139	[0.104, 0.174]	–
AI correct	C2	0.116	[0.080, 0.153]	–
AI correct	C3	0.096	[0.061, 0.130]	–
AI correct	C4	0.131	[0.095, 0.168]	–
AI correct	C2 – C1	-0.023	[-0.054, 0.009]	.156
AI correct	C3 – C2	-0.021	[-0.055, 0.014]	.232
AI correct	C4 – C3	0.036	[-0.002, 0.074]	.061

Means are participant-condition averages. Intervals and p-values for contrasts come from paired participant-level differences ($N = 70$). This robustness check is descriptive and does not replace the trial-level mixed-effects model.

E.8 Sensitivity model with displayed model score

Term	β	95% posterior interval	OR	OR 95% posterior interval
Intercept	-0.136	[-0.325, 0.053]	0.87	[0.72, 1.05]
C2 vs C1	0.020	[-0.348, 0.387]	1.02	[0.71, 1.47]
C3 vs C1	-0.419	[-0.835, -0.003]	0.66	[0.43, 1.00]
C4 vs C1	-0.468	[-0.831, -0.106]	0.63	[0.44, 0.90]
AI correct	0.764	[0.442, 1.087]	2.15	[1.56, 2.97]
C2 × AI correct	0.068	[-0.567, 0.702]	1.07	[0.57, 2.02]
C3 × AI correct	1.019	[0.287, 1.751]	2.77	[1.33, 5.76]
C4 × AI correct	0.257	[-0.315, 0.828]	1.29	[0.73, 2.29]
Displayed model score (z)	0.285	[0.118, 0.452]	1.33	[1.13, 1.57]
Initial alignment	4.844	[4.198, 5.490]	126.94	[66.52, 242.22]
Initial confidence (z)	-0.607	[-0.799, -0.415]	0.54	[0.45, 0.66]

Trials	Contrast	Base M3 OR	With displayed model score OR
AI wrong	C2 vs C1	1.04 [0.72, 1.50]	1.02 [0.71, 1.47]
AI wrong	C3 vs C2	0.64 [0.37, 1.12]	0.65 [0.37, 1.12]
AI wrong	C4 vs C3	0.93 [0.54, 1.61]	0.95 [0.55, 1.65]
AI correct	C2 vs C1	1.12 [0.53, 2.33]	1.09 [0.52, 2.27]
AI correct	C3 vs C2	1.66 [0.54, 5.09]	1.67 [0.55, 5.10]
AI correct	C4 vs C3	0.44 [0.15, 1.31]	0.44 [0.15, 1.31]

We z-standardise the displayed model score across AI-present trials. The covariate is item-level, so we do not interpret it as a causal score effect.

E.9 Sensitivity check (exclude sessions under 8 minutes)

We re-estimate the final-alignment model (M3) after excluding six sessions with duration under 8 minutes. Key terms remain directionally consistent:

- C3 vs C1 (AI-wrong baseline): $\beta = -0.421$ (SD $= 0.218$)
- C4 vs C1 (AI-wrong baseline): $\beta = -0.434$ (SD $= 0.192$)
- AI correct: $\beta = 0.933$ (SD $= 0.173$)
- C3 $\times$ AI correct: $\beta = 0.795$ (SD $= 0.399$)
- C4 $\times$ AI correct: $\beta = 0.060$ (SD $= 0.305$)

E.10 Exploratory participant-level moderator checks

As an appendix-only exploratory analysis, we examine whether participant-level survey variables relate to lower alignment with incorrect AI advice. These checks are not part of the primary research question, and we interpret them cautiously given the participant sample size ($N = 70$). For survey items exported with mixed numeric/label strings (e.g., "7 - Very likely"), we use the leading numeric value to recover the 1–7 scale.

First, we estimate a Bayesian logistic mixed model on AI-present trials with participant- and item-level random intercepts:

$$\texttt{final\_align} \sim \texttt{condition} * \texttt{ai\_correct} + \texttt{initial\_align} + \texttt{z\_initial\_conf} + \texttt{ai\_caution\_z} + \texttt{verify\_ai\_z} + \texttt{ai\_usefulness\_z}.$$

Term	β	95% posterior interval	OR	OR 95% posterior interval
AI caution (z)	0.051	[-0.144, 0.247]	1.05	[0.87, 1.28]
Verify AI before relying (z)	-0.184	[-0.379, 0.010]	0.83	[0.68, 1.01]
Perceived AI usefulness (z)	0.109	[-0.088, 0.306]	1.11	[0.92, 1.36]

None of these exploratory participant-level coefficients has a 95% interval excluding 0.

Second, to examine incorrect-AI alignment at the participant level, we compute each participant's alignment rate on AI-wrong trials (7–9 AI-present trials per participant, depending on list; mean rate $= 0.584$, $SD = 0.235$), then correlate this rate with selected survey variables using Spearman correlations:

Variable	n	Spearman ρ	p
AI caution	70	-0.072	.556
Likelihood of verifying AI	70	-0.106	.385
AI usefulness	70	0.059	.626
Misinformation familiarity	70	0.030	.807
English reading ability	70	0.144	.235

The exploratory pattern is directionally consistent with the idea that higher caution and verification tendency may relate to lower AI-wrong alignment, but the effects are small and uncertain in this sample.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3173574.3174156

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps [NeurIPS 2018]. *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/1810.03292

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? the effect of AI explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3411764.3445717

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–21. https://doi.org/10.1145/3449287

Chen, V., Liao, Q. V., Vaughan, J. W., & Bansal, G. (2023). Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), 1–32. https://doi.org/10.1145/3610219

de Jong, S., Bos, M. W., van Berkel, N., & Lamers, M. H. (2025). Algorithm appreciation or aversion: The effects of accuracy disclosure on users' reliance on algorithmic suggestions. *Behaviour & Information Technology*, 1–20. https://doi.org/10.1080/0144929X.2025.2535732

Delaunay, J., Galarraga, L., Largouet, C., & van Berkel, N. (2025). Impact of explanation technique and representation on users' comprehension and confidence in explainable ai. *Proceedings of the ACM on Human-Computer Interaction*, *9*(2), 1–28. https://doi.org/10.1145/3711011

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning [arXiv:1702.08608]. *arXiv*. https://arxiv.org/abs/1702.08608

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79–94. https://doi.org/10.1518/0018720024494856

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29. https://doi.org/10.1038/s44159-021-00006-y

Gruzd, A., Mai, P., & Soares, F. B. (2024). To share or not to share: Randomized controlled study of misinformation warning labels on social media. *Disinformation in Open Online Media: MISDOOM 2024*, *15175*, 46–69. https://doi.org/10.1007/978-3-031-71210-4_4

Hase, P., & Bansal, M. (2020). Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 5540–5552. https://doi.org/10.18653/v1/2020.acl-main.491

Hashemi Chaleshtori, F., Ghosal, A., Gill, A., Bambroo, P., & Marasovic, A. (2024). On evaluating explanation utility for human-AI decision making in NLP. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 7456–7504. https://doi.org/10.18653/v1/2024.findings-emnlp.439

Hoes, E., Aitken, B., Zhang, J., Gackowski, T., & Wojcieszak, M. (2024). Prominent misinformation interventions reduce misperceptions but increase scepticism. *Nature Human Behaviour*, *8*, 1545–1553. https://doi.org/10.1038/s41562-024-01884-x

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 4198–4208. https://doi.org/10.18653/v1/2020.acl-main.386

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach [arXiv:1907.11692]. *arXiv*. https://arxiv.org/abs/1907.11692

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*. https://arxiv.org/abs/1705.07874

Muralidhar, D. (2024). The effect of progressive disclosure in the transparency of explainable artificial intelligence systems. *Proceedings of the 2024 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, 382–383. https://doi.org/10.1109/VL/HCC60511.2024.00057

Nguyen, D. (2018). Comparing automatic and human evaluation of local explanations for text classification. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1069–1078. https://doi.org/10.18653/v1/N18-1097

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*, 303–330. https://doi.org/10.1007/s11109-010-9112-2

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Pareek, S., van Berkel, N., Velloso, E., & Goncalves, J. (2024). Effect of explanation conceptualisations on trust in ai-assisted credibility assessment. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW2), 1–31. https://doi.org/10.1145/3686922

Pennycook, G., Berinsky, A. J., Bhargava, P., Lin, H., Cole, R., Goldberg, B., Lewandowsky, S., & Rand, D. G. (2024). Inoculation and accuracy prompting increase accuracy discernment in combination but not alone. *Nature Human Behaviour*, *8*(12), 2330–2341. https://doi.org/10.1038/s41562-024-02023-2

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–52. https://doi.org/10.1145/3411764.3445315

Rathje, S., Roozenbeek, J., Van Bavel, J. J., & van der Linden, S. (2023). Accuracy and social motivations shape judgements of (mis)information. *Nature Human Behaviour*, *7*, 892–903. https://doi.org/10.1038/s41562-023-01540-w

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Schemmer, M., Kühl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 410–422. https://doi.org/10.1145/3581641.3584066

Schoeffer, J., De-Arteaga, M., & Kühl, N. (2024). Explanations, fairness, and appropriate reliance in human-AI decision-making. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3613904.3642621

Schoeffer, J., Jakubik, J., Vössing, M., Kühl, N., & Satzger, G. (2025). AI reliance and decision quality: Fundamentals, interdependence, and the effects of interventions. *Journal of Artificial Intelligence Research*, *82*, 471–501. https://doi.org/10.1613/jair.1.15873

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 180–186. https://doi.org/10.1145/3375627.3375830

Spitzer, P., Morrison, K., Turri, V., Feng, M., Perer, A., & Kühl, N. (2025). Imperfections of XAI: Phenomena influencing AI-assisted decision-making. *ACM Transactions on Interactive Intelligent Systems*. https://doi.org/10.1145/3750052

Swire-Thompson, B., DeGutis, J., & Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition*, *9*(3), 286–299. https://doi.org/10.1016/j.jarmac.2020.06.006

Wang, X., & Yin, M. (2022). Effects of explanations in AI-assisted decision making: Principles and comparisons. *ACM Transactions on Interactive Intelligent Systems*, *12*(4), 1–36. https://doi.org/10.1145/3519266

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300509

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3351095.3372852