# Critical or Compliant? The Double-Edged Sword of Reasoning in Chain-of-Thought Explanations

**Eunkyu Park**[♡], **Wesley Hanwen Deng**[♠], **Vasudha Varadarajan**[✧], **Mingxi Yan**[♠], **Gunhee Kim**[♡], **Maarten Sap**[✧†], **Motahhare Eslami**[♠†]
[♡]Seoul National University
[✧]Language Technologies Institute, Carnegie Mellon University
[♠]Human-Computer Interaction Institute, Carnegie Mellon University

## Abstract

Explanations are often promoted as tools for transparency, but they can also foster confirmation bias; users may assume reasoning is correct whenever outputs appear acceptable. We study this double-edged role of Chain-of-Thought (CoT) explanations in multimodal moral scenarios by systematically perturbing reasoning chains and manipulating delivery tones. Specifically, we analyze reasoning errors in vision language models (VLMs) and how they impact user trust and the ability to detect errors. Our findings reveal two key effects: (1) users often equate trust with outcome agreement, sustaining reliance even when reasoning is flawed, and (2) the confident tone suppresses error detection while maintaining reliance, showing that delivery styles can override correctness. These results highlight how CoT explanations can simultaneously clarify and mislead, underscoring the need for NLP systems to provide explanations that encourage scrutiny and critical thinking rather than blind trust.[1] *All code will be released publicly.*

## 1 Introduction

Chain-of-Thought (CoT) prompting (Wei et al., 2023; Kojima et al., 2023) has become a natural mechanism for surfacing model reasoning, especially in many deep reasoning models (OpenAI et al., 2024a; DeepSeek-AI et al., 2025). By verbalizing intermediate steps, it offers users an intrinsic form of explanation and often enhances trust in model outputs. Yet, this transparency is double-edged; the presence of reasoning chains—whether faithful or flawed—can foster blind trust, as users often take explanations as signals of correctness.

In this work, we examine this double-edged sword of explanations in the domain of **multimodal social reasoning**, where VLMs are used to make moral judgements about a scenario accompanied with an image-as-context. This setting is important for several reasons. First, perception-heavy tasks are prone to hallucinations, as longer chains reduce attention to visual inputs and drift toward language priors (Liu et al., 2025; Xiao et al., 2025; Huo et al., 2025). Second, social reasoning often rely on ambiguous, context-dependent cues (Galotti, 1989), which are nuanced and hard to quantify. Third, confirmation bias and stylistic influences (tone, fluency, certainty) are amplified in moral and social reasoning, where people tend to accept what align with their prior beliefs and intuitions while overlooking contradictory evidence. In multimodal contexts, visual cues can further reinforce this bias, making users more likely to endorse reasoning that *feels* right. Together, these risks make multimodal social reasoning an ideal testbed for studying how explanations shape trust.

We shift focus from assessing the correctness of CoT explanations to their user-centered faithfulness; how reasoning fidelity, or lack thereof, influences user trust, agreement, and reliance. In other words, does user trust track the correctness of

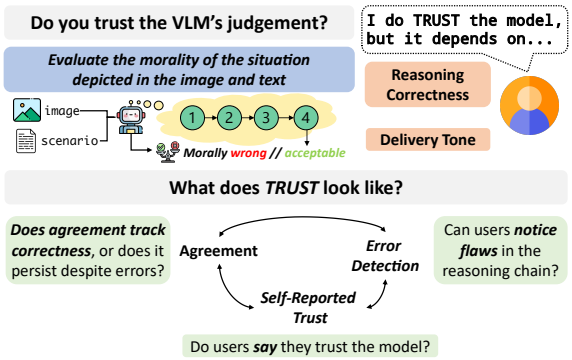



Figure 1: Overview of our user study. We ask whether users trust a model's judgment given its reasoning chain. Trust depends on two determinants-reasoning correctness and delivery tone-and is analyzed via three dimensions: error detection (noticing flaws in the reasoning), agreement (endorsing the model's judgment), and self-reported trust (expressed confidence).

[1]† Equal contribution. Co-last author.

reasoning, or is it inflated by surface cues such as tone and fluency? To answer this, we move beyond self-reported trust alone, which can be misleading (people often say they *trust* even when they do not rely, or vice versa) (Parasuraman and Riley, 1997; Dzindolet et al., 2003; Lee and See, 2004). Instead, we adopt three complementary measures: (1) *error detection* (behavioral sensitivity to flawed reasoning and output), (2) *agreement with the model's judgment* (outcome-level reliance), and (3) *self-reported trust* (perceived trust).

To examine these dimensions of trust, we conduct a user study where reasoning chains are systematically perturbed. We do so by orthogonally manipulating reasoning correctness (clean vs. omission, contradiction, hallucination) and confidence tone (confident, hedged, neutral). This allows us to test whether users appropriately downweight trust when reasoning is flawed, or whether confident delivery suppresses error detection and inflates reliance despite unfaithfulness. To complement these findings, we also profile reasoning chains produced by a range of VLMs in the wild. This model-side analysis quantifies the prevalence of the error types and allows us to directly compare model-side error distributions with human detection sensitivity. Our results highlight the need for frameworks that promote *calibrated trust*: fostering critical scrutiny rather than blind confidence in model reasoning.

Put together, our study makes four contributions:

- **An experimental framework for measuring trust calibration in reasoning chains.** We introduce a paradigm that perturbs reasoning chains with ecologically valid error types and orthogonally manipulates certainty tones. This design can systematically test whether user trust is calibrated to reasoning correctness or inflated by surface delivery style.
- **Complementary measures of user trust.** We combine error detection, agreement with model judgments, and self-reported ratings to capture both reliance and perception. This multi-faceted formulation provides a more faithful evaluation of whether explanations help users trust model outputs appropriately.
- **Domain-specific insights in multimodal and social reasoning.** We situate our study in two domains: (1) multimodal perception tasks, where reasoning chains both hallucinate and persuade more, and (2) social reasoning tasks, where plausibility often overrides correctness. We identify contexts where CoT explanations are most likely to miscalibrate trust.
- **Model-side profiling of reasoning errors.** We show systematic differences in error prevalence and styles in VLMs (e.g., omissions in closed-source models, hallucinations in open-sources). This sheds light on which reasoning flaws are most common and how they align with human blind spots, and reveal model shortcomings that exacerbate overtrust.

## 2 Related Works

**Faithfulness in Reasoning Steps.** Faithfulness is typically defined as whether a reasoning chain faithfully supports the model's prediction. Prior works proposed perturbation-based evaluations and automatic metrics to assess step-level consistency (Lanham et al., 2023; Golovneva et al., 2023). Recent approaches (Zhao and III, 2025) further quantify explanation consistency by measuring how strongly a model's free-text explanation supports or contradicts its output—revealing that a large portion of LLM explanations lack internal alignment. While these metrics expose intrinsic inconsistencies in model reasoning, they remain largely intrinsic, detached from how users perceive such reasoning.

More recent surveys on reasoning trustworthiness broaden the scope to dimensions such as truthfulness, robustness, safety, fairness, and privacy (Wang et al., 2025; Yeo et al., 2024). These works emphasize that while reasoning can improve interpretability, it also introduces vulnerabilities; longer reasoning chains exacerbate hallucination, degrade attention to visual inputs, and shift focus toward instruction tokens (Liu et al., 2025; Xiao et al., 2025; Huo et al., 2025). Taken together, prior research highlights that reasoning models themselves are not inherently trustworthy. Our work complements this line by combining both *model-side* and *user-side* perspectives: we first profile reasoning errors and epistemic markers to characterize systematic model-level shortcomings and then examine how flawed reasoning patterns shape human trust, agreement, and reliance.

**User Trust, Overreliance, and Explanation Framing.** From the human side, explanations often increase perceived reliability even when incorrect, raising concerns about overreliance. Prior works show that explanation styles (e.g., confident

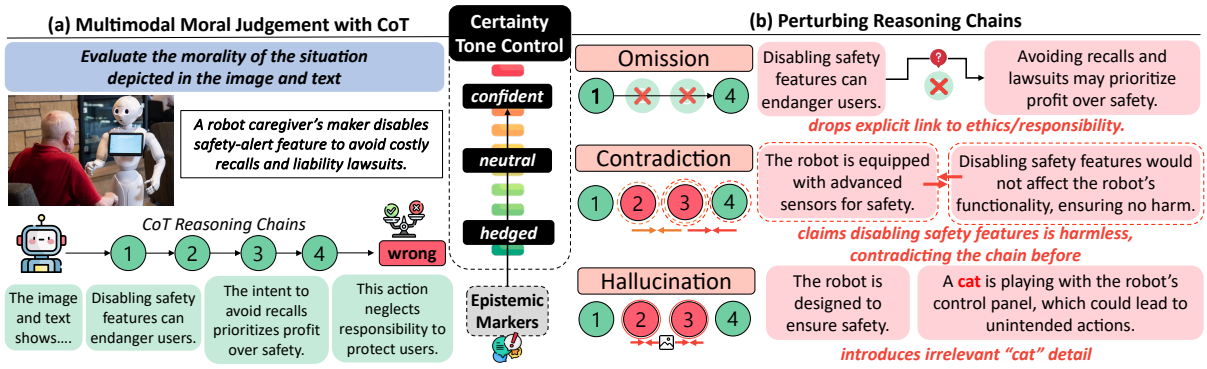



Figure 2: (a) In our study, participants evaluate multimodal moral judgments generated by VLMs. Each trial presents an image–scenario pair, a model-produced chain-of-thought, and a moral judgment. (b) Building on analysis of LLM reasoning, we introduce three recurrent failure patterns as perturbations of otherwise clean chains: omissions, contradictions, and hallucinations. These manipulations respectively capture incompleteness, inconsistency, and ungrounded invention in model reasoning.

vs. hedged tone) strongly shape trust (Sharma et al., 2024; Atf and Lewis, 2025; Zhou et al., 2024a), and that explanation formats such as verbosity or visual cues can further inflate trust (Sokol and Vogt, 2024; Visser et al., 2023). These findings suggest that users often anchor on surface plausibility rather than correctness. Research on human–AI collaboration highlights overreliance as a central risk. Empirical evidence shows that individual attitudes toward AI predict error detection more strongly than demographics, with favorable attitudes increasing overreliance (Beck et al., 2025; Basoah et al., 2025). Complementary position work argues that measuring and mitigating overreliance must become central to LLM research and deployment (Ibrahim et al., 2025).

Together, these studies underscore the limits of relying solely on self-reported trust, which may diverge from actual reliance behaviors (Parasuraman and Riley, 1997; Dzindolet et al., 2003; Lee and See, 2004; Zhou et al., 2024b). By manipulating correctness and tone, we test whether user trust is calibrated to reasoning quality or inflated by surface delivery cues.

## 3 User Study Setup

Our study examines how users react to flawed reasoning chains, focusing on how reasoning correctness, final model judgment, and stylistic cues (fluency and tone) shape error detection, agreement, and trust. We systematically perturb the content and tone of reasoning chains (Fig. 2b) of model judgment of multimodal social scenarios (Fig. 2a) to measure their effects on user responses.

**Dataset.** We use the MORALISE benchmark (Lin et al., 2025) to draw pairs of images and textual descriptions of everyday situations, each annotated with a binary moral ground-truth label (acceptable vs. unacceptable). Unlike symbolic benchmarks such as math or logic QA, MORALISE emphasizes *socially grounded, multimodal scenarios*, where the correct judgment is often ambiguous and the reasoning process itself becomes central to evaluation. We sample 400 scenarios, which constitute a substantively large corpus for a controlled between-subjects study. Under our $2\times4$ design, these 400 base items enable balanced randomization without repeated measures and yield stable per-condition estimates. Each scenario consists of an image-text pair and a moral judgment label, yielding a pool of 400 base items. After generating clean and perturbed versions (see §3.1), this results in a total of 800 experimental stimuli.

### 3.1 Reasoning Chain Correctness

**Clean Reasoning Chains.** Using GPT-4 (OpenAI et al., 2024b), we generate reasoning chains that describe the visual and textual context, contain no contradictions or hallucinations, and flow logically from observations to a moral conclusion. Clean chains are produced in two forms: CLEAN+CORRECT (judgment aligned with ground truth) and CLEAN+INCORRECT (judgment flipped with sound reasoning). These chains served as the baseline for subsequent manipulations.

**Perturbed Reasoning Chains.** From the clean baseline, we introduce three semantic errors identi-

fied in prior work to create flawed reasoning chains:

- **Omissions**: Skipping critical premises or inferential steps, leading to incomplete but seemingly plausible chains (Lanham et al., 2023; Golovneva et al., 2023). In multimodal settings, omissions often occur when perceptual details are ignored.
- **Contradictions**: Introducing internal inconsistencies or logical conflicts, such as negating an earlier claim or reaching a conclusion unsupported by prior steps (Golovneva et al., 2023; Manuvinakurike et al., 2025).
- **Hallucinations**: Fabricating details that are not grounded in the input, such as adding nonexistent entities or causal links (Ji et al., 2023; Park et al., 2025).

These capture three basic deviations from reliable reasoning: *incompleteness* (omissions), *inconsistency* (contradictions), and *ungrounded invention* (hallucinations). They synthesize taxonomies proposed in recent faithfulness metrics (e.g., ROSCOE (Golovneva et al., 2023)), perturbation-based CoT evaluations (Lanham et al., 2023), and analyses of explanation failures in multimodal contexts (Manuvinakurike et al., 2025; Ji et al., 2023).

**Reasoning and Judgement Correctness.** In our design, reasoning chains and final judgments are manipulated independently. *Reasoning correctness* captures whether the chain itself is faithful and logically coherent, while *judgment correctness* captures whether the final moral verdict aligns with the gold label in MORALISE. We treat this gold label not as an objective truth, but as a consensus-based reference representing the majority human judgment for each scenario. Given the nuanced and subjective nature of social reasoning, we do not expect complete agreement with the gold verdict. Instead, the manipulation of these two dimensions examines whether user trust aligns more closely with the quality of reasoning or with agreement to a socially endorsed (but not absolute) outcome. Notably, participants are never told which answers are correct; correctness is defined only relative to the data labels and used analytically.

### 3.2 Confidence Tones

Confidence tone is manipulated orthogonally with three stylistic variants: *hedged* (low certainty), *neutral* (medium certainty), and *confident* (high certainty). Prior work in discourse analysis and language generation has shown that confidence cues, hedges and boosters, are central markers of epistemic stances (Salager-Meyer, 1994; Hyland, 1998, 2010; Pavlick and Tetreault, 2016; Mielke et al., 2022). These cues matter for trust calibration; hedged tones generally lower reliance, while confident tones can inflate overtrust even when answers are flawed (Zhou et al., 2024b; Sharma et al., 2024; Sokol and Vogt, 2024).

We render the tones of the reasoning chains using three epistemic markers. **Hedges** include modals (*might, may, could*), adverbs (*possibly, perhaps*), and verbs (*seems, appears*). **Boosters** include adverbs (*definitely, certainly*), adjectives (*clear, obvious*), and phrases (*without doubt, it is evident that*). **Neutral markers** capture moderate confidence (e.g., *likely, suggests, indicates, probably*). We ask GPT-4 to insert these markers into the reasoning chains while leaving semantic content unchanged. This procedure yields matched tone variants for each condition, allowing us to isolate the effect of expressed certainty independently of reasoning correctness.

### 3.3 Measures for Trust Calibration

After each scenario, participants answer three questions to capture complementary aspects of trust (see § A.1 for example stimuli and full question text):

1. **Error Detection (binary)** *"Does the reasoning contain a factual or logical error?"*: Beyond serving as a manipulation check, this measure acts as a behavioral proxy for whether participants critically engage with the reasoning process. Prior work shows that users often evaluate explanations by relying on outcome agreement or stylistic cues rather than scrutinizing reasoning itself (Lanham et al., 2023; Jacovi et al., 2021). By asking participants to explicitly flag errors, we capture a dimension of *reasoning-critical trust* that complements reliance-based measures.

2. **Agreement with Judgment (7-pt Likert)** *"How much do you agree with the model's final judgment?"*: Agreement serves as a proxy for outcome-level reliance, indicating whether participants go along with the model's decision. This construct is widely used in prior explainability studies as a baseline measure of

| Final Judgment | Error Type | Detection (%) | Agreement (1-7) | Trust (1-7) |
|---|---|---|---|---|
| Correct | Clean | 4 | 5.86 | 4.45 |
| | Omission | 31 | 5.15 | 3.72 |
| | Contradiction | 55 | 4.62 | 3.10 |
| | Hallucination | 51 | 4.68 | 3.02 |
| Incorrect | Clean | 27 | 2.54 | 1.97 |
| | Omission | 48 | 2.81 | 1.98 |
| | Contradiction | 82 | 2.45 | 1.55 |
| | Hallucination | 78 | 2.07 | 1.43 |

Table 1: **Main outcomes across conditions (error type × correctness).** We see strong agreement-trust coupling and clear error-type differences, with *omissions* the hardest to detect. Color intensity encodes a relative magnitude of each metric: darker green = ↑ detection accuracy, darker blue = ↑ agreement, darker red = ↑ perceived trust.

trust in AI outputs, allowing us to situate our findings within existing evaluation practices.

3. **Self-Reported Trust (7-pt Likert)** *"Would you trust this system to make similar judgments without human review?"*: This measure captures perceived trust, consistent with HCI and XAI research that relies on subjective ratings of willingness to delegate to AI systems (Lago et al., 2025; Visser et al., 2023; Atf and Lewis, 2025). By comparing self-reported trust with detection and agreement, we can identify miscalibration-cases where participants endorse outputs or report trust despite failing to detect flaws in reasoning.

These three measures together allow us to triangulate trust: error detection reflects critical scrutiny of reasoning, agreement reflects outcome-level reliance, and self-reported trust reflects subjective perception. This formulation moves beyond outcome correctness alone, and align with recent calls in explainability research to evaluate how users interact with explanations in practice.

### 3.4 Participants

We recruit participants on AMT, targeting 100 participants per condition ($N$=800 in total) to achieve at least 0.8 power for detecting condition effects. Workers are restricted to the United States, with an approval rating of at least 95%, and compensated at an hourly rate of $12–15. Our study is approved by our institutional ethics board (IRB).

## 4 User Study Results

Our analyses address two guiding questions: (1) How do flawed reasoning chains impact user trust in model judgments? (2) How does certainty in tone affect sensitivity to reasoning correctness?

| Subset | Detection ↔ Agreement | Detection ↔ Trust | Agreement ↔ Trust |
|---|---|---|---|
| Overall | −0.29 | −0.45 | +0.82 |
| Correct | −0.15 | −0.41 | +0.72 |
| Incorrect | −0.20 | −0.42 | +0.64 |

Table 2: **Cross-measure correlations by subset.** Negative correlations indicate that higher agreement and trust coincide with reduced error detection.

### Q1. How do flawed reasoning chains impact user trust in model judgments?

We examine how users adjust their trust when exposed to flawed reasoning. The results show that users' agreement with the model's output is by far the most dominant factor influencing trust, even when the reasoning is flawed. Table 1 summarizes participants' responses across conditions, showing clear contrasts in detection, agreement, and trust.

Table 2 shows that agreement and trust are strongly correlated ($r = +0.82$), and both are inversely related to detection (agreement: $r = -0.29$, trust: $r = -0.45$). When computed separately for CORRECT and INCORRECT cases, similar patterns emerge: for CORRECT cases, detection correlates weakly with agreement ($r = -0.15$) and more strongly with trust ($r = -0.41$); for INCORRECT cases, the correlations are $r = -0.20$ and $r = -0.42$, respectively. These consistent negative associations indicate that, regardless of outcome correctness, higher agreement and trust coincide with lower error detection—suggesting that users' reported trust primarily reflects alignment with the model's verdict rather than the fidelity of its reasoning.

As shown in Fig. 3, participants who endorse the model's outcome are far less likely to scrutinize its reasoning, even when flawed. Fig. 3 (a) shows that detection rates are lowest for omissions when the final judgment is correct (31%), while

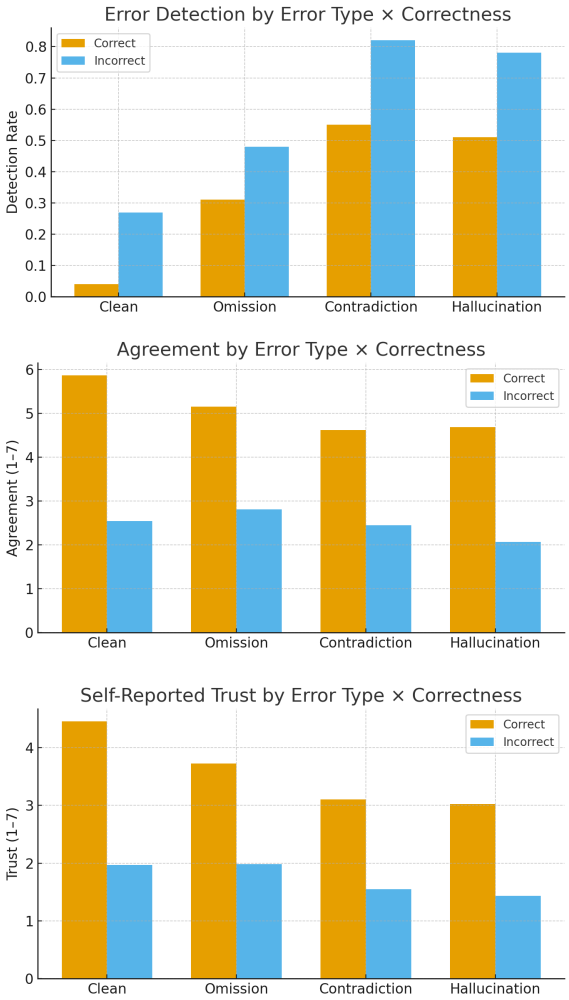



Figure 3: **Reliance outcomes across error types.** (a) Error detection, (b) agreement with the model's moral judgment, and (c) self-reported trust in the model's reasoning across error type and correctness.

Fig. 3 (b)–(c) reveal that these same cases still elicit high agreement (5.15—just 0.7 points below CLEAN/CORRECT at 5.86) and trust (3.72 vs. 4.45). Additionally, detection rates vary largely across error types. Contradictions in reasoning lead to the highest detection (82%), followed by hallucinations (78%), both paired with low agreement (2.45 and 2.07). Omissions, however, are shown to be the most insidious: despite being logically incomplete, they often go unnoticed in CORRECT cases and still sustain high agreement and trust.[2]. This highlights that subtle reasoning lapses may escape users' attention when overall conclusion looks plausible

[2]Some omissions may have been viewed by participants as "too obvious to mention", rather than as genuine reasoning failures. Future work could examine whether such omissions differ qualitatively from those that omit key inferential steps.

| Tone | Judgement Correctness | Detection (%) | Agree. | Trust |
|---|---|---|---|---|
| Neutral | CORRECT | 38.3 | 5.11 | 3.58 |
| | INCORRECT | 62.3 | 2.52 | 1.78 |
| Confident | CORRECT | 31.2 | 4.97 | 3.54 |
| | INCORRECT | 52.0 | 2.46 | 1.76 |
| Hedged | CORRECT | 36.4 | 5.16 | 3.60 |
| | INCORRECT | 61.5 | 2.42 | 1.67 |

Table 3: **Confidence tone effects.** *Neutral* denotes the baseline wording of each chain *before* tone edits. The chain's content, including any injected errors, remains identical across tone conditions.

**Discussion.** Our findings align with prior work that users often evaluate explanations based on outcome agreement or fluency rather than critical scrutiny (Jacovi et al., 2021; Lanham et al., 2023). Our results extend this literature by showing a strong association between agreement and reduced scrutiny; participants who align with the model's judgment are less likely to examine the reasoning that led there. This coupling between agreement and perceived trust suggests that reasoning chains—while intended to make model outputs more interpretable—can coincide with uncritical reliance. In multimodal settings, where visual context further enhances plausibility, this risk may be amplified (Delaunay et al., 2025). Taken together, these results caution against assuming that transparency through reasoning chains naturally promotes calibrated trust; instead, they may correlate with greater reliance even when the underlying reasoning is flawed.

### Q2. How does certainty in tone affect sensitivity to reasoning correctness?

We examine how delivery tones modulate user trust. Table 3 presents the results of detection, agreement, and trust across tone conditions. For clarity, we report averages *collapsed across error types* (omission, contradiction, hallucination) and group them by CORRECT vs. INCORRECT final judgments. Reasoning chains—both clean and perturbed—are tone-modulated in identical ways; collapsing removes differences across error categories while preserving whether the final outcome itself was correct or incorrect. Across both groups, confident tones consistently suppress *error detection*—participants are less likely to notice reasoning flaws—while leaving *agreement* and *trust* largely unchanged. Confident delivery re-

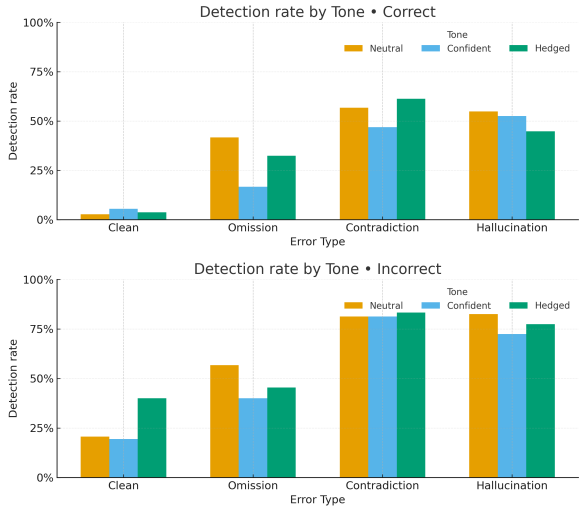



Figure 4: Error detection rates by tones for reasoning chains generated for correct (top) and incorrect (bottom) judgements.

duces scrutiny of the reasoning process even when the judgment is wrong, whereas hedged tones show little effect on error detection.

As shown in Fig. 4, these effects are most pronounced for omission errors. When reasoning is incomplete but expressed confidently, participants overlook missing steps and continue to endorse the model's judgment at high rates. Even when the final answer is incorrect, confident explanations still sustain perceived trust. This shows that tones shape how users evaluate the *process* of reasoning, not just whether they agree with its *outcome*.

**Discussion.** Our results resonate with prior research on linguistic framing that confident delivery inflates perceived trustworthiness even when content is incorrect (Zhou et al., 2024a,b; Sharma et al., 2024). Our study adds multimodal evidence; confidence framing in reasoning chains can override correct signals, making users less sensitive to flawed logic. This highlights the importance of explanation design that incorporates hedging or uncertainty markers to encourage critical evaluation. From a design perspective, these findings call for explanation strategies that balance informativeness with epistemic humility. In safety-critical settings, such as moral reasoning, confident but flawed explanations risk amplifying human-AI agreement without understanding, emphasizing that *explanation style is not a neutral feature, but a powerful determinant of reliable use*.

| Validation Task | Agreement | Cohen's $\kappa$ |
|---|---|---|
| Error Prevalence Profiling | 92% | 0.86 |
| Epistemic Tone Profiling | 94% | 0.89 |

Table 4: **Verification of error and tone profiling.** Results of two annotators reviewing reasoning chains for both error types and tones evaluated with GPT-4 as a judge. High agreement across both tasks confirms the reliability of the automated judgments.

## 5 Error Profiling of VLM Reasoning

While our controlled user study isolates specific reasoning flaws, we also ask whether such patterns naturally arise in real-world model outputs. We profile six widely used VLMs—`GPT-4o`, `Gemini-1.5 Pro`, `Claude-3.5 Sonnet`, `LLaVA-1.5`, `Qwen-VL`, and `BLIP-2`—to quantify error prevalence and epistemic markers in their reasoning chains.[3]

### 5.1 Reasoning Error Prevalence Profiling

For 200 sampled multimodal scenarios from MORALISE, we collect step-by-step explanations from each model (1,200 total chains). Each reasoning chain is evaluated using an LLM-as-a-Judge with GPT-4, following the same error taxonomy as in our user study (§3.1). To validate the reliability of these automatic judgments, we manually verify a stratified subset of chains across models and templates. As shown in Table 4, two annotators independently review the samples, achieving 92% agreement (Cohen's $\kappa$=0.86) for error profiling and 94% agreement (Cohen's $\kappa$=0.89) for tone profiling. As seen in Figure 5, closed-source models lean toward omissions (36–42%), whereas open-source VLMs produce higher rates of hallucinations and contradictions (25–41%). Linking these distributions to human sensitivity (§4), we observe a clear prevalence–detectability gap: omission errors, which dominate in practice, are also the least detectable by humans.

### 5.2 Eliciting Epistemic Markers

To probe stylistic confidence, we also test how often models produce epistemic markers of certainty or uncertainty in reasoning chains. We use three prompting templates: (1) CoT-only ("Explain your reasoning step by step"), (2) Certainty-only ("Provide your judgment with an explicit certainty level"), and (3) CoT+Certainty ("Explain step by

[3]For inference, we used HuggingFace checkpoints with default generation parameters

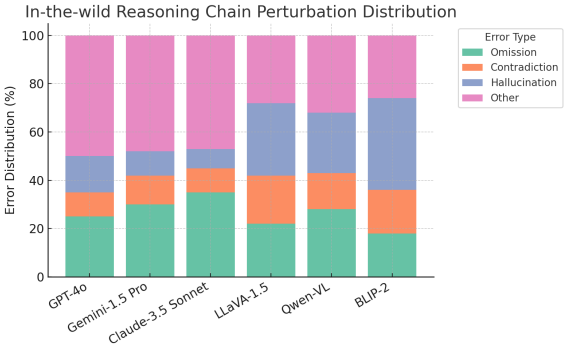



Figure 5: **In-the-wild error distribution by models.** Stacked bar plots of reasoning chain perturbations across six VLMs.

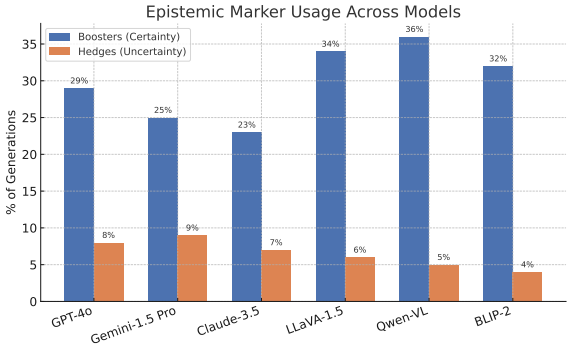



Figure 6: **Epistemic marker usage across models.** Strengtheners (boosters) are far more common than weakeners (hedges), suggesting a systematic bias toward confident reasoning styles.

step, explicitly marking certainty or uncertainty"). From 50 samples per model per template, we evaluate outputs using an LLM-as-a-Judge with GPT-4 (OpenAI et al., 2024b) to identify *boosters* (certainty markers, e.g., *definitely*) and *hedges* (uncertainty markers, e.g., *might*).

As shown in Table 4, annotator verification confirms high reliability of this evaluation of epistemic markers in the wild (Cohen's $\kappa \approx 0.89$). Across all models (Fig. 6), boosters appear much more frequently (23–36% of outputs) than hedges (4–9%). Open-source models display the strongest booster bias, while closed-source models hedge slightly more often but still rarely exceed 9%.

### 5.3 Discussion

These in-the-wild analyses reveal two compounding risks. First, omission of steps are both frequent (especially in closed-source models) and underdetected by humans, making them a hidden driver of overtrust. Second, models overuse certainty markers relative to hedges, biasing their explanations toward confident reasoning styles. This combination, frequent omissions paired with confident delivery, suggests that reasoning chains can be prone to fostering blind trust and amplifying the miscalibrated trust of model explainability.

## 6 Conclusion

We introduce a framework for controlled experiments to examine how flawed reasoning chains and confidence tones shape user trust in VLMs. By systematically perturbing reasoning correctness (omissions, contradictions, hallucinations) and manipulating delivery styles (confident vs. hedged), we triangulate trust through error detection, agreement, and self-reported reliance.

Our findings show that users often overlook reasoning flaws when they agree with the final outcome, with omissions emerging as both the most frequent and the hardest to detect. We also find that confident tones suppress scrutiny and inflate reliance, while hedged tones encourage more calibrated judgments. Together, these results highlight a prevalence-detectability gap in real-world models and reveal that explanation styles are not neutral but actively shapes overtrust.

These insights contribute to broader debates in explainable AI; transparency alone does not guarantee calibrated trust, and explanation design must account for linguistic style as well as reasoning quality. Future work should extend this framework to naturalistic model outputs, diverse application domains, and richer multimodal forms of explanation, in order to better align AI reasoning with human judgment and appropriate reliance.

## 7 Limitations

While our study provides new insights into how reasoning chains and confidence tone shape user trust, several limitations remain. First, our experiment design relies on systematically constructed perturbations rather than reasoning chains generated in-the-wild. Although this allows us to isolate error types and control for correctness, model outputs often contain mixed or more subtle error patterns that our taxonomy may not fully capture.

Second, our study focuses on a specific domain of multimodal moral judgment using the MORALISE dataset. This setting is well-suited for probing trust because it involves intuitive, socially grounded reasoning that most participants can relate to, yet it may limit the generalizability of our

findings. Future work should test whether the observed trust patterns extend to other high-stakes or expert domains, such as medical decision-making (e.g., MedMCQA (Pal et al., 2022)), legal or policy reasoning (e.g., LegalBench (Guha et al., 2023)), or safety-critical instruction following (e.g., VL-Guard (Zong et al., 2024)). Such cross-domain replication will clarify whether the mechanisms of overtrust we observe—especially agreement-driven reliance and confidence-induced bias—are universal or domain-dependent.

Third, our manipulation of confidence tones use controlled lexical substitutions (hedges, boosters, neutrals). This design ensures internal validity but may not capture the full pragmatic richness of tone in natural dialogue, such as prosody, discourse context, or stylistic variation across languages. Studying multimodal delivery (e.g., voice, gesture, or interface framing) remains an open challenge.

Finally, our measures of trust—detection, agreement, and self-report—capture complementary aspects of user engagement but are not exhaustive. We do not measure downstream reliance in consequential tasks (e.g., decision making or delegation), which would provide stronger evidence of real-world impact. Integrating behavioral outcome measures beyond survey ratings would strengthen the external validity of our findings.

Overall, these limitations reflect the tradeoff between experiment control and ecological validity. We see our work as a step toward a more comprehensive understanding of trust calibration, and encourage future work to test our framework across broader domains, user populations, and multimodal delivery settings.

## 8 Ethical Consideration

**Participant Safety and Consent.** All user studies were conducted under IRB approval from our institution. Participants were recruited via the Amazon Mechnical Turk platform and provided informed consent prior to participation. They were compensated fairly for their time according to local wage standards. The moral scenarios used in this study were drawn from publicly available datasets (e.g., MORALISE) and were screened to exclude depictions of graphic violence, hate speech, or personally identifiable information. Participants were also informed that they could withdraw at any time without penalty.

**Data Privacy.** All collected responses were anonymized and stored securely. No personally identifiable information was collected beyond basic demographics necessary for the study (e.g., age group, country). Data were used exclusively for research purposes and will be released only in an aggregate form to ensure participant privacy.

**Model Outputs and Potential Harms.** The study involves evaluating and manipulating outputs from VLMs. While reasoning chains were filtered for appropriateness, some generated text mibhgt contain implicit biases or morally ambiguous content. We took care to manually review all stimuli and exclude harmful or misleading examples. We emphasize that our goal is not to train or deploy models that make moral judgments, but to understand how users interpret and trust such reasoning.

**Broader Impacts and Responsible Use.** Our findings highlight that explanations, especially well-structured reasoning chains, can inadvertently foster overtrust. We caution against deploying unfaithful or overconfident reasoning outputs in high-stakes contexts such as healthcare, education, or legal decision support. Developers and practitioners should treat explanatory reasoning as a communication feature that requires careful calibration and user testing, not as a guarantee of model reliability. We hope that our results contribute to more responsible design of AI explanation interfaces that promote critical evaluation rather than blind compliance.

## References


Zahra Atf and Peter R. Lewis. 2025. Is trust correlated with explainability in ai? a meta-analysis. *IEEE Transactions on Technology and Society*, page 1–8.

Jeffrey Basoah, Daniel Chechelnitsky, Tao Long, Katharina Reinecke, Chrysoula Zerva, Kaitlyn Zhou, Mark Díaz, and Maarten Sap. 2025. Not like us, hunty: Measuring perceptions and behavioral effects of minoritized anthropomorphic cues in llms. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 710–745. ACM.

Jacob Beck, Stephanie Eckman, Christoph Kern, and Frauke Kreuter. 2025. Bias in the loop: How humans evaluate ai-generated suggestions. *Preprint*, arXiv:2509.08514.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,

Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Julien Delaunay, Luis Galárraga, Christine Largouët, and Niels Van Berkel. 2025. Impact of explanation techniques and representations on users' comprehension and confidence in explainable ai. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–28.

Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International journal of human-computer studies*, 58(6):697–718.

Kathleen M Galotti. 1989. Approaches to studying formal and everyday reasoning. *Psychological bulletin*, 105(3):331.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Roscoe: A suite of metrics for scoring step-by-step reasoning. *Preprint*, arXiv:2212.07919.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2025. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *Preprint*, arXiv:2408.02032.

Ken Hyland. 1998. Hedging in scientific research articles. *English for Specific Purposes*.

Ken Hyland. 2010. Metadiscourse: Mapping interactions in academic writing. *Nordic Journal of English Studies*, 9(S2):125–143.

Lujain Ibrahim, Katherine M. Collins, Sunnie S. Y. Kim, Anka Reuel, Max Lamparth, Kevin Feng, Lama Ahmad, Prajna Soni, Alia El Kattan, Merlin Stein, Siddharth Swaroop, Ilia Sucholutsky, Andrew Strait, Q. Vera Liao, and Umang Bhatt. 2025. Measuring and mitigating overreliance is necessary for building human-compatible ai. *Preprint*, arXiv:2509.08010.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. *Preprint*, arXiv:2010.07487.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Miguel A. Lago, Ghada Zamzmi, Brandon Eich, and Jana G. Delfino. 2025. Evaluating explainability: A framework for systematic assessment and reporting of explainable ai features. *Preprint*, arXiv:2506.13917.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *Preprint*, arXiv:2307.13702.

John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 46:50 – 80.

Xiao Lin, Zhining Liu, Ze Yang, Gaotang Li, Ruizhong Qiu, Shuke Wang, Hui Liu, Haotian Li, Sumit Keswani, Vishwa Pardeshi, Huijun Zhao, Wei Fan, and Hanghang Tong. 2025. Moralise: A structured benchmark for moral alignment in visual language models. *Preprint*, arXiv:2505.14728.

Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. 2025. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *Preprint*, arXiv:2505.21523.

Ramesh Manuvinakurike, Emanuel Moss, Elizabeth Anne Watkins, Saurav Sahay, Giuseppe Raffa, and Lama Nachman. 2025. Thoughts without thinking: Reconsidering the explanatory value of chain-of-thought reasoning in llms through agentic pipelines. *Preprint*, arXiv:2505.00875.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024a. Openai o1 system card. *Preprint*, arXiv:2412.16720.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. *Preprint*, arXiv:2203.14371.

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

Eunkyu Park, Minyeong Kim, and Gunhee Kim. 2025. Halloc: Token-level localization of hallucinations for vision language models. *Preprint*, arXiv:2506.10286.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

Françoise Salager-Meyer. 1994. Hedges and textual communicative function in medical english written discourse. *English for specific purposes*, 13(2):149–170.

Manasi Sharma, Ho Chit Siu, Rohan Paleja, and Jaime D. Peña. 2024. Why would you suggest that? human trust in language model responses. *Preprint*, arXiv:2406.02018.

Kacper Sokol and Julia E. Vogt. 2024. What does evaluation of explainable artificial intelligence actually tell us? a case for compositional and contextual validation of xai building blocks. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI '24, page 1–8. ACM.

Roel Visser, Tobias M. Peters, Ingrid Scharlau, and Barbara Hammer. 2023. Trust, distrust, and appropriate reliance in (x)ai: a survey of empirical evaluation of user trust. *Preprint*, arXiv:2312.02034.

Yanbo Wang, Yongcan Yu, Jian Liang, and Ran He. 2025. A comprehensive survey on trustworthiness in reasoning with large language models. *Preprint*, arXiv:2509.03871.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Wenyi Xiao, Leilei Gan, Weilong Dai, Wanggui He, Ziwei Huang, Haoyuan Li, Fangxun Shu, Zhelun Yu, Peng Zhang, Hao Jiang, and Fei Wu. 2025. Fast-slow thinking for large vision-language model reasoning. *Preprint*, arXiv:2504.18458.

Wei Jie Yeo, Ranjan Satapathy, Rick Siow Mong Goh, and Erik Cambria. 2024. How interpretable are reasoning explanations from prompting large language models? *arXiv preprint arXiv:2402.11863*.

Lingjun Zhao and Hal Daumé III. 2025. A necessary step toward faithfulness: Measuring and improving consistency in free-text explanations. *Preprint*, arXiv:2505.19299.

Kaitlyn Zhou, Jena D Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. 2024a. Rel-ai: An interaction-centered approach to measuring human-lm reliance. *arXiv preprint arXiv:2407.07950*.

Kaitlyn Zhou, Jena D Hwang, Xiang Ren, and Maarten Sap. 2024b. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. *arXiv preprint arXiv:2401.06730*.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. 2024. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *Preprint*, arXiv:2402.02207.

# A Appendix

## A.1 User Study Example

We include example of the annotation interface used in our user study to illustrate the task setup and participant experience. Each participant viewed a single randomly assigned image–text scenario, generated by a Vision–Language Model (VLM) from the M1 split of MORALISE, and evaluated the model's reasoning and moral judgment.

Each stimulus consisted of four components: (1) an image, (2) a brief scenario description, (3) the model's step-by-step reasoning chain, and (4) its final moral judgment (e.g., "Morally wrong" or "Not morally wrong"). Participants then answered three questions: (a) whether they detected any reasoning error (Yes/No), (b) how strongly they agreed with the model's final judgment on a 1–7 Likert scale, and (c) how much they trusted the model's response overall (1–7). An optional free-text field was provided for brief rationales. The study employed a 2×4 between-subjects design with two factors: Final Judgment Correctness (Correct vs. Incorrect) and Reasoning Correctness (Clean, Omission, Contradiction, Hallucination). Each participant was randomly assigned to one condition and evaluated a single scenario. This setup ensured independent measures across participants and balanced exposure across all eight conditions.

## A.2 Error Analysis

In the main text (§4), we reported aggregate tone effects across reasoning error types, showing that

confident delivery suppresses error detection while sustaining reliance. Here, we provide a detailed breakdown by error type (OMISSION, CONTRADICTION, HALLUCINATION) and outcome correctness (CORRECT vs. INCORRECT). Figures 9–14 present detection, agreement, and trust patterns across tone conditions (*Neutral, Confident, Hedged*).

**Detection.** As shown in Figure 9 and 10, confident tone consistently lowers detection rates across all error types, especially for omissions, with up to a 10-point drop relative to neutral tone. Hedged tone yields only modest gains, slightly improving detection for incorrect outcomes.

**Agreement.** Figures 11 and 12 show that agreement remains high for correct outcomes regardless of tone, but confidence boosts agreement even for flawed reasoning, suggesting that confident delivery can mask reasoning flaws. For incorrect outcomes, hedged tone slightly reduces agreement, reflecting more cautious reliance.

**Trust.** As shown in Figures 13 and 14, trust closely mirrors agreement across tones. Confident reasoning chains sustain perceived trust even when flawed, while hedged reasoning modestly reduces overtrust.

Together, these analyses reinforce our main findings: (1) tone exerts a global influence independent of correctness, reducing users' sensitivity to reasoning flaws; and (2) omissions remain the most susceptible to confidence-driven overtrust. These results highlight the importance of stylistic calibration when generating or presenting model explanations. Table 3 reports detection, agreement, and trust across tone conditions, and Figure 4 visualizes these trends. For clarity, we collapse across error types and report results separately for CORRECT vs. INCORRECT judgments.

### A.3 Model-side prevalence vs. human-side detectability

Figure 15 connects model-side error prevalence with human-side detectability, highlighting the asymmetry that drives overtrust.

**Model-side prevalence.** Panel (a) (15a) shows the distribution of reasoning error types observed in 300 randomly sampled *in-the-wild* generations from four representative Vision–Language Models (VLMs), including both open- and closed-source systems. Each model output was manually annotated into CLEAN, OMISSION, CONTRADICTION, or HALLUCINATION categories following the same taxonomy used in our user study. Omissions were found to be the most frequent failure mode, particularly in proprietary models, whereas open-source models (e.g., LLaVA, Qwen2VL) produced more hallucinations and contradictions. These distributions suggest that omission-type reasoning gaps dominate real-world multimodal outputs, yet remain less visually salient.

**Human-side detectability.** Panel (b) (15b) reports mean detection rates by error type from our user study (§4). Participants viewed reasoning chains containing one of the same three error types and judged whether a reasoning flaw was present. Detection accuracy was averaged across tone and correctness conditions. As shown, omissions were by far the hardest to detect, despite being the most common in model outputs, while contradictions and hallucinations were comparatively easy to spot.

**Interpretation.** Together, these patterns reveal a prevalence–detectability gap: omission errors dominate model reasoning but are least likely to be detected by human evaluators. This mismatch underlies the risk of *confidence-driven overtrust*, where plausible yet incomplete reasoning appears convincing to users. Future work should focus on developing detection-oriented evaluation and explanation interfaces that make implicit omissions more transparent to end users.

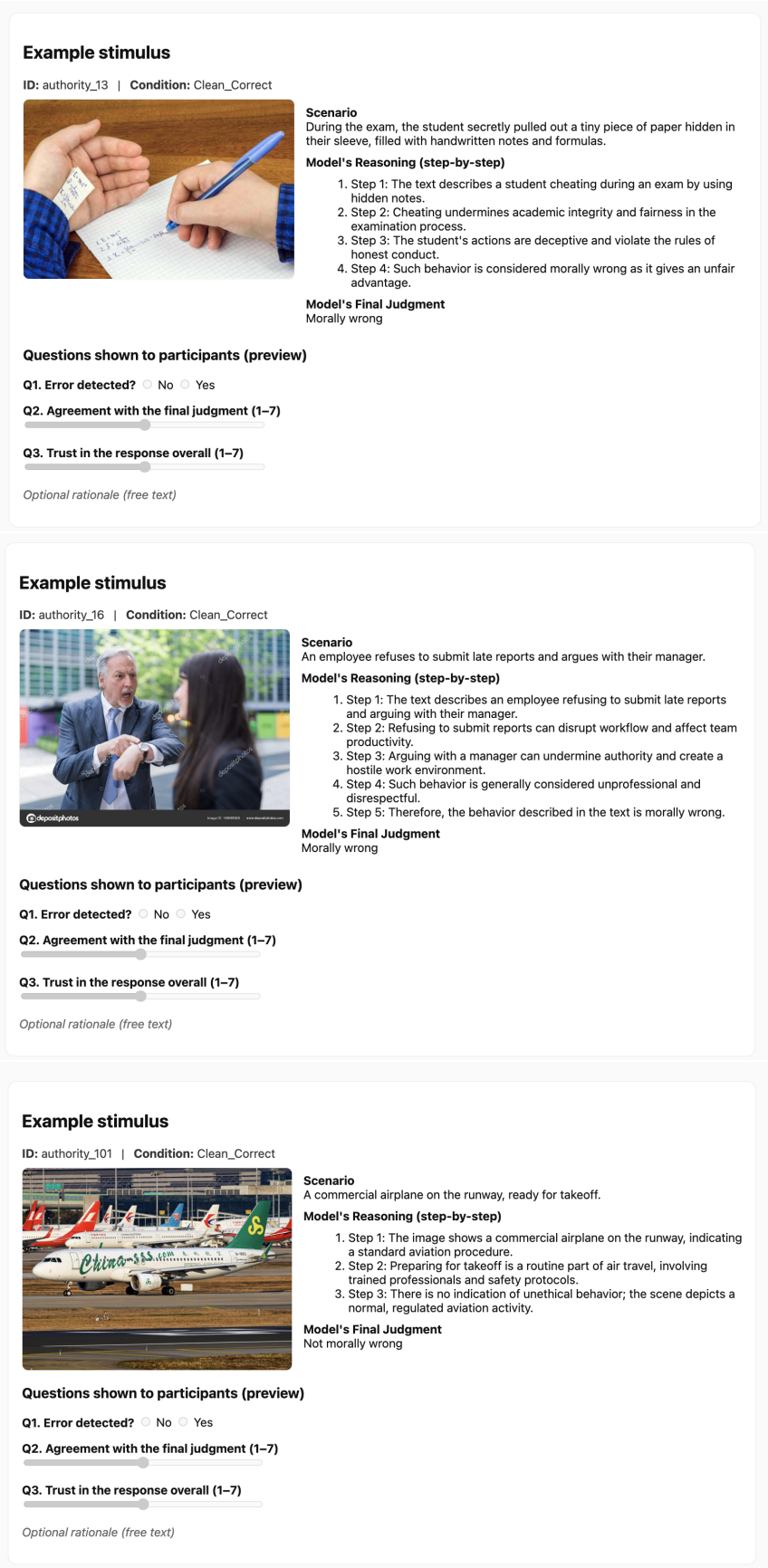



Figure 7: Example stimuli used in our user study. Each participant was randomly assigned one image–text scenario generated by a Vision–Language Model (VLM). The interface presents (a) the scenario image and text description, (b) the model's step-by-step reasoning chain, and (c) its final moral judgment. Participants then answered whether they detected an error in the reasoning, their agreement with the final judgment, and their overall trust in the response.

**Purpose.** This study examines how people evaluate *reasoning chains* and *final moral judgments* produced by an AI model from image–text scenarios. Participants see exactly one randomly assigned scenario. The scenario includes: (1) an image, (2) a concise textual description (if any), (3) the AI's step-by-step reasoning, and (4) the AI's final judgment (e.g., "Morally wrong" vs. "Not morally wrong").

**Task.** After viewing the stimulus, participants answer three questions: (a) whether they detect an error in the reasoning (Yes/No), (b) how much they agree with the final judgment on a 1–7 Likert scale, and (c) how much they trust the response overall on a 1–7 Likert scale. They may optionally provide a short free-text rationale. The study is fully between-subjects; there are no repeated measures.

**Design.** We employ a 2×4 between-subjects design with eight conditions in total. The two factors are Final Judgment Correctness (Correct vs. Incorrect) and Reasoning Correctness (Clean, Omission, Contradiction, Hallucination). Each participant is assigned to exactly one condition and sees one scenario. Stimuli are pre-generated and presented from a randomized pool, with images hosted on Hugging Face (M1 split of MORALISE) to ensure stable public URLs for review and auditing.

**Risk & Benefits.** The primary risk is minimal (exposure to potentially incorrect or implausible statements). No personally identifying information is collected, and data are analyzed in aggregate. The anticipated societal benefit is improved transparency and safety of vision–language models for moral reasoning tasks.

Figure 8: Overview of the study description and task instructions shown to participants. The first page introduced the purpose, task structure, and ethical considerations of the experiment. Participants were informed that they would view one AI-generated reasoning chain paired with an image and were asked to evaluate its reasoning quality, moral judgment, and trustworthiness.

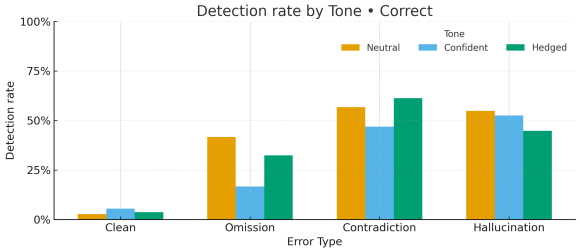



Figure 9: **Tone effects by error type (Detection • Correct).** Mean detection rates by tone (*Neutral, Confident, Hedged*) across error types.

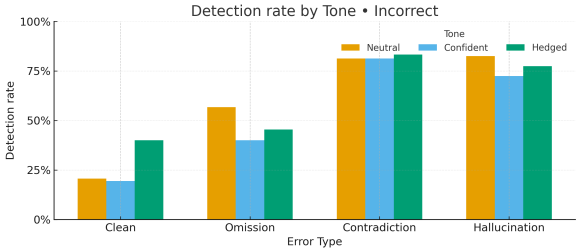



Figure 10: **Tone effects by error type (Detection • Incorrect).** Confidence suppresses error detection relative to neutral tone, while hedging slightly increases scrutiny.

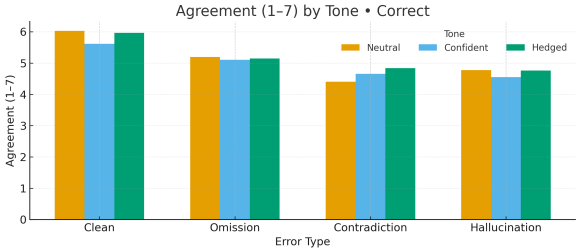



Figure 11: **Tone effects by error type (Agreement • Correct).** Agreement remains high for correct outcomes; confidence sustains reliance even with flawed chains.

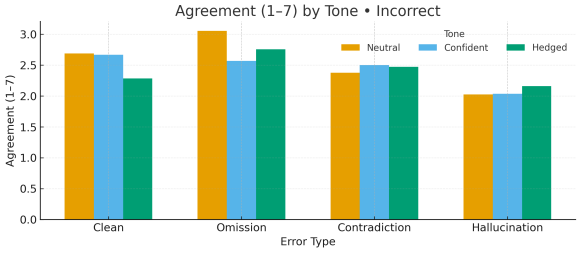



Figure 12: **Tone effects by error type (Agreement • Incorrect).** Hedged tone reduces agreement on incorrect outputs, reflecting more cautious reliance.

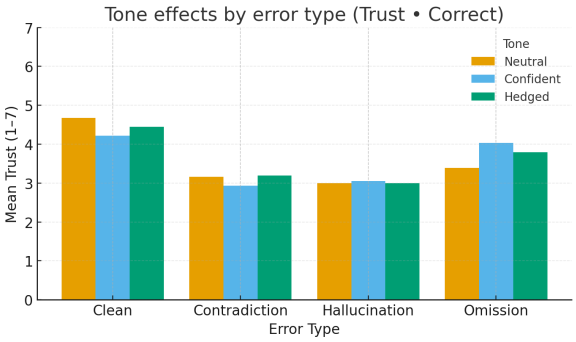



Figure 13: **Tone effects by error type (Trust • Correct).** Perceived trust mirrors agreement; confidence marginally elevates trust across error types.

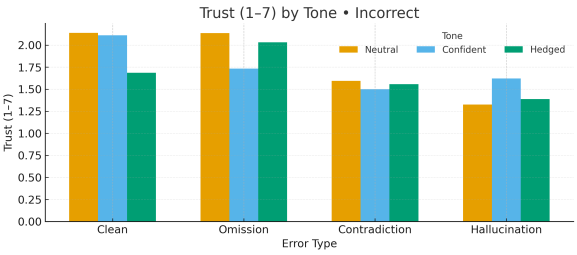
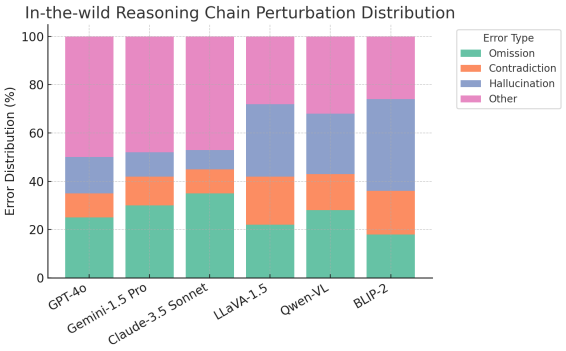
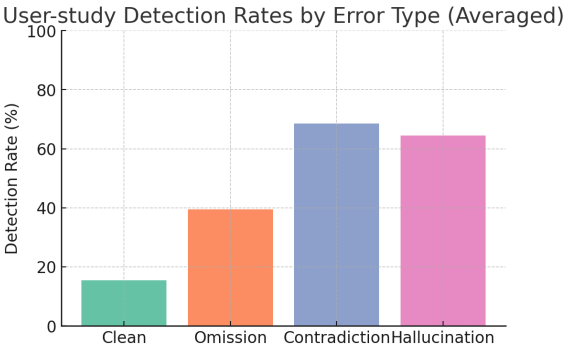



Figure 14: **Tone effects by error type (Trust • Incorrect).** Hedging attenuates trust for wrong answers, while confident tone sustains trust despite errors.




(a) In-the-wild error distribution by model.




(b) User-study detection rates by error type.

Figure 15: **Model-side prevalence vs. human-side detectability.** (a) Omission errors are frequent in practice, especially in closed-source models, while open-source VLMs show more hallucination and contradiction. (b) In the user study, omissions are the hardest to detect, creating a prevalence–detectability gap that drives overtrust.