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

Anonymous ACL submission

Abstract

Explanations are often promoted as tools for transparency, but they can also foster blind trust; users may assume reasoning is correct whenever model 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 perturb reasoning chains generated by vision-language models (VLMs) with three error types (omissions, contradictions, hallucinations) and measure their effects on user trust and error detection through a study with 800 participants, alongside a parallel detection task with models. Our findings reveal two key effects: (1) users often equate trust with outcome agreement regardless of flaws in the model's reasoning, and (2) confident tones suppress scrutiny while leaving agreement and self-reported trust largely unchanged. These results highlight how CoT explanations can simultaneously clarify and mislead in tandem, and point to user-side scrutiny, not just model-side faithfulness, as a critical dimension for calibrated trust in AI reasoning. *All code will be released publicly upon acceptance.*

1 Introduction

Chain-of-Thought prompting (Wei et al., 2023; Kojima et al., 2023) has become a standard tool for surfacing model reasoning, especially in deep reasoning models (OpenAI et al., 2024a; DeepSeek-AI et al., 2025). By verbalizing intermediate steps, CoT offers users an intrinsic form of explanation and often enhances trust in model outputs. Yet, this transparency is double-edged: reasoning chains, whether faithful or flawed, can foster blind trust when readers fail to detect errors that are present.

In this work, we examine error detection in CoT explanations: *when presented with a model's reasoning trace, do human users critically scrutinize its internal logic?* While prior work has focused on enhancing the *faithfulness* of CoT explanations with respect to outputs (Lightman et al., 2023; Lanham et al., 2023; Sharma et al., 2024), we examine an orthogonal question: how CoT explanations, regardless of faithfulness, are *interpreted* as a *user-facing artifact*, and whether users critically engage with their content or defer to surface cues such as confident delivery. We instantiate this question in **multimodal moral reasoning**, a domain where the costs of uncritical trust are especially high: unlike math or factual QA, moral scenarios have ambiguous ground truth that makes correctness hard to externally verify, readers' own moral intuitions invite confirmation bias, and visual context amplifies the narrative coherence of a chain even when its reasoning is flawed (Delaunay et al., 2025). These properties make multimodal moral reasoning a *worst-case testbed* for whether users, and the automated judges increasingly used to oversee them, can flag broken reasoning when nothing else in the task signals that something is wrong.

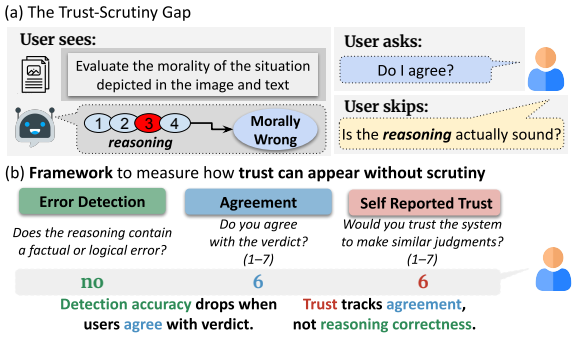

Figure 1: **The trust-scrutiny gap.** (a) *What we study*: when reading a model's judgment on subjective topics, users often bypass flawed reasoning when they agree with the verdict. (b) *Framework to decompose trust*: we measure the three signals (error detection, agreement, and self-reported trust) in response to controlled perturbations of the reasoning chain.

To study this, we conduct parallel error-detection studies on the same set of perturbed reasoning

chains. On the human side, we run a controlled between-subjects user study on 400 multimodal moral scenarios, systematically perturbing reasoning chains with three error types (OMISSION, CONTRADICTION, HALLUCINATION) and altering confidence tone. We complement subjective self-reports of trust—which are known to dissociate from behavior (Parasuraman and Riley, 1997; Dzindolet et al., 2003; Lee and See, 2004)—with two more behaviorally grounded measures: whether users flag an error when one is present, and whether they agree with the model's verdict. On the model side, we probe models to perform the same error detection task on the same stimuli to compare human and model sensitivity to each error type.

Together, trust and agreement appear decoupled from scrutiny. Across our triangulated measures, agreement and self-reported trust move together ($r \approx 0.8$), but both correlate weakly with error detection ($r = -0.29$ to -0.45): users endorse and trust flawed reasoning chains. Confident tone in explanations further suppresses scrutiny, inflating agreement and trust without improving detection. Error detection is also selective; users catch contradictions and hallucinations at moderate-to-high rates, but spot OMISSIONs only 40% of the time. When we conduct the same error-detection task with models in place of humans, the same selectivity emerges: VLMs catch contradictions and hallucinations with high accuracy but detect omissions far less reliably. Linking these patterns to in-the-wild profiling, we also identify a *prevalence–detectability gap*; omissions are simultaneously the most frequent error type in in-the-wild VLM reasoning chains. Because users sustain high agreement and trust even when omissions go undetected, these patterns have direct implications for downstream reliance on flawed reasoning, and underscore the need for evaluation frameworks that promote *calibrated trust*; encouraging critical scrutiny rather than blind confidence in model reasoning. Our key contributions are:

- **A controlled framework for measuring error detection in reasoning chains.** We introduce a paradigm that perturbs CoT explanations with three ecologically valid error types (omission, contradiction, hallucination) and orthogonally manipulates certainty tones.
- **Triangulated user-side trust measurement.** Combining error detection, agreement with model judgments, and self-reported ratings, we find that agreement and self-reported trust are strongly coupled but both are decoupled from error detection, revealing cases where users sustain perceived trust despite failing to identify flawed reasoning.
- **The prevalence–detectability gap.** Linking detection rates to in-the-wild profiling of six VLMs, omissions emerge as the most frequent error type in real VLM reasoning chains and the error type that neither human users nor models reliably catch.

2 Related Work

Faithfulness in Reasoning Steps. Faithfulness is typically defined as whether a reasoning chain reliably supports the model's output. Prior works proposed perturbation-based evaluations and automatic metrics to assess step-level consistency (Lanham et al., 2023; Golovneva et al., 2023). Recent approaches (Agarwal et al., 2024) 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. However, these metrics 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; Wei Jie et al., 2024). They 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 models themselves are not inherently trustworthy. Our work complements this by combining both model-side and user-side perspectives; we first profile reasoning errors and epistemic markers to characterize model-level shortcomings and then examine how flawed reasoning patterns shape human trust, agreement, and reliance.

User Trust, Overreliance, and 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 vs. hedged tone) strongly shape trust (Sharma et al., 2024; Atf and Lewis, 2025; Zhou et al., 2025), and formats

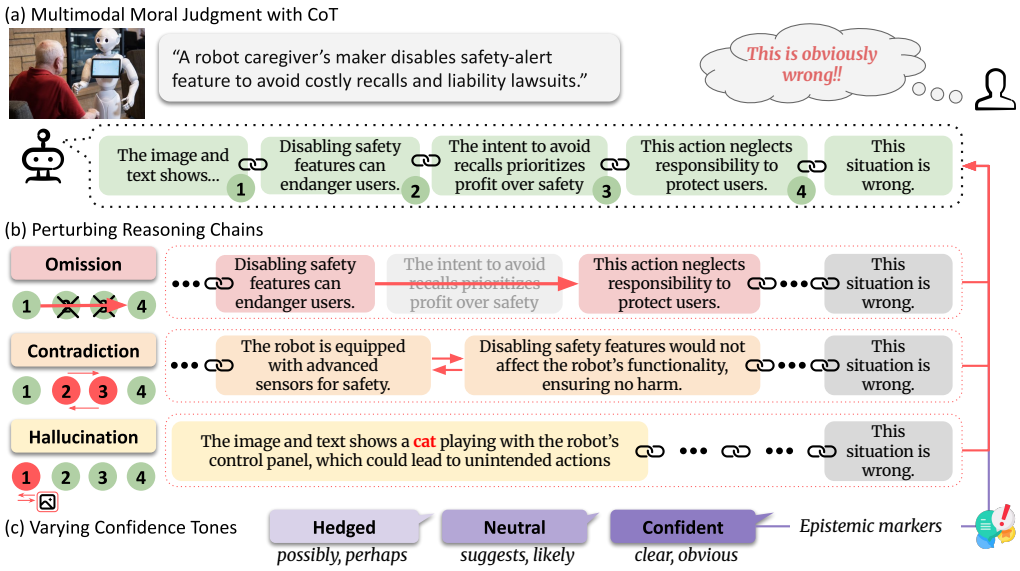

Figure 2: **Stimulus Construction Pipeline.** (a) Participants evaluate multimodal moral judgments generated by VLMs. Each trial presents an image–scenario pair, a model-produced CoT, and a binary moral judgment. (b) We introduce three recurrent failure patterns as perturbations of otherwise clean chains: omissions, contradictions, and hallucinations. These respectively capture incompleteness, inconsistency, and ungrounded invention in model reasoning. (c) Confidence tones in the reasoning chains are also manipulated with specific epistemic markers.

such as verbosity or visual cues can further inflate trust (Sokol and Vogt, 2024; Visser et al., 2023). They find users often anchor on surface plausibility rather than correctness. Research on human–AI collaboration presents overreliance as a big risk. 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). Others argue that measuring and mitigating overreliance must be central to LLM research and deployment (Ibrahim et al., 2025). 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., 2024). By manipulating correctness and tone, we test whether user trust is calibrated to reasoning correctness or inflated by surface delivery cues.

3 User Study Setup

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

Scenario Selection. We use the MORALISE benchmark (Lin et al., 2025) to draw multimodal social scenarios, each consisting of an image and a text annotated with a binary label (ACCEPTABLE vs. UNACCEPTABLE). Unlike domains such as math or logic, MORALISE emphasizes *socially grounded, multimodal situations* where judgments can be ambiguous and explanations can be persuasive even when imperfect. We exclude scenarios containing graphic violence or hate. From the remaining set, we sample 200 ACCEPTABLE and 200 UNACCEPTABLE scenarios, yielding 400 scenarios. For each scenario, we construct two MORALISE-relative verdict conditions: a *correct* verdict, where the model's final judgment matches the MORALISE label, and an *incorrect* verdict, with respect to the MORALISE label. Since our main question concerns how reasoning errors and confidence tone affect user scrutiny, we average over verdict conditions in the main analysis and report verdict-condition-stratified results in § A.4. Each scenario is assigned one reasoning-error condition at the scenario level. Crossing 400 scenarios with two verdict conditions yields 800 stimuli.

Study Setup. We employ a fully between-subjects design. Each scenario is assigned one reasoning-error condition (CLEAN, OMISSION, CONTRADICTION, or HALLUCINATION) at the scenario level, and each stimulus is independently assigned

one confidence tone (§3.2). Each participant is assigned to a single stimulus—one image–scenario pair, one explanation, one verdict, and one tone—and responds to three measures in a fixed order: reasoning-error detection, verdict agreement, and trust/willingness to rely on the model. This design avoids carry-over effects that could arise from exposing the same participant to multiple error types or tones (Buçinca et al., 2021). We recruit 800 AMT workers, compensated at $12–15/hour.

3.1 Correctness of Reasoning Chains

Clean Reasoning Chains. We first construct *faithful and coherent reasoning chains* that describe the visual and textual context without contradiction or hallucinations and flow logically from observations to a moral conclusion, which serve as a control for studying user trust. We use GPT-4V (OpenAI et al., 2024b) with a prompt that enforces grounded description and coherent step-by-step reasoning (see §A.7.1). This can be applied to reasoning chains produced by any CoT-capable model. To confirm that these chains are clean—i.e., free of the error types in our taxonomy—we apply a two-stage verification. First, GPT-5.2 (OpenAI, 2025) screens each chain against our error taxonomy, flagging any omission, contradiction, or hallucination. Then, the chains are verified by two independent annotators over all clean chains for each scenario (Cohen's κ=0.87). We also validate robustness with Gemini-2.5 Pro (Comanici et al., 2025), obtaining high cross-judge agreement (91.5%). Since perturbations are introduced by design, the judge's role is simply to verify that the intended error is present and no unintended errors were introduced, a far more constrained task than open-ended evaluation.

Perturbed Reasoning Chains. From the clean chains, we induce errors of three types to create flawed reasoning chains. These capture *incompleteness* (OMISSION), *inconsistency* (CONTRADICTION), and *ungrounded invention* (HALLUCINATION), observing taxonomies proposed in faithfulness metrics (e.g., ROSCOE; Golovneva et al., 2023), perturbation-based model reasoning evaluations (Lanham et al., 2023), and analyses of explanation failures in multimodal contexts (Manuvinakurike et al., 2025).

- **OMISSION**: 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.
- **CONTRADICTION**: 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).
- **HALLUCINATION**: Fabricating details that are not grounded in the input (image+scenario), such as adding nonexistent entities or causal links (Ji et al., 2023; Park et al., 2025).

With the verified CLEAN chains, we construct localized errors via controlled perturbations, which lets us isolate the causal effect of each error type on trust—model outputs in the wild contain mixed, confounded patterns (§5). We generate OMISSION, CONTRADICTION, and HALLUCINATION variants using standardized edit rules (Alg. 1 in §A.7), and re-run the judge to confirm that (i) the target error is present and (ii) no additional error types are introduced (§A.7.3). Finally, two annotators confirm the presence of the intended error type in **95.8%** of cases. Unintended secondary errors occur in 3.3% of chains, most commonly involving borderline omissions following contradiction edits. Overall coherence is rated as preserved in **97.5%** of cases. Inter-annotator agreement is high (Cohen's κ=0.84). Details to the validations are in § A.7.3.

3.2 Confidence Tones

Confidence tone is assigned to each stimulus at random, independently of its error condition, in one of three stylistic variants rendered by corresponding markers: **hedged** (*low certainty* induced by modals such as *might, may, could*, adverbs such as *possibly, perhaps*, and verbs such as *seems, appears*), **neutral** (*medium certainty* induced with markers like *likely, suggests*), and **confident** (high certainty induced by adverbs like *definitely*, adjectives like *clear, obvious*, and phrases like *without doubt, it is evident that*). These markers are well-established cues for trust calibration, with hedged tones lowering perceived trust and confident tones inflating trust even for wrong answers (Hyland, 1998; Pavlick and Tetreault, 2016; Mielke et al., 2022; Zhou et al., 2024; Sharma et al., 2024; Sokol and Vogt, 2024). Our lexical substitution approach isolates this single dimension to enable clean causal attribution; we verify that tone edits do not introduce surface-level confounds (e.g., length) via checks in § A.7.4.

3.3 Measures for Trust Calibration

Our approach operationalizes trust as a multi-layered construct rather than a single attitude. Lee and See (2004) describe how trust in automation operates as a causal chain from beliefs to attitudes to intentions to behaviors, and these layers can diverge: self-reported trust attitudes do not reliably predict behavioral engagement (Visser et al., 2023; Körber, 2018), and behavioral measures capture dimensions that attitudinal measures miss (Buçinca et al., 2021). Kohn et al. (2021) also note that behavioral measures better reflect actual engagement. We therefore triangulate trust across three measures, separating *behavioral* signals (what users do in response to a reasoning chain) from *perceived* trust (what users report feeling), with each measure capturing a different layer of this causal chain (Example stimuli are in § A.6):

1. **Error Detection (binary)** *"Does the reasoning contain a factual or logical error?"*: A behavioral proxy for critical engagement with the reasoning. Behavioral measures of critical engagement diverge meaningfully from self-reported trust (Visser et al., 2023; Buçinca et al., 2021); by asking participants to explicitly flag errors, we capture *reasoning-critical scrutiny* that purely attitudinal measures can miss.
2. **Agreement with Judgment (7-pt Likert)** *"How much do you agree with the model's final judgment?"*: A behavioral proxy at the outcome level, indicating whether participants endorse the model's decision. Agreement is widely used in explainability studies (Lee and See, 2004; Bansal et al., 2021; Vereschak et al., 2024) as an indicator of how users respond to AI outputs.
3. **Self-Reported Trust (7-pt Likert)** *"Would you trust the system to make similar judgments?"*: A perceived-trust measure, consistent with HCI and XAI research using subjective ratings of willingness to delegate (Lago et al., 2025; Visser et al., 2023; Atf and Lewis, 2025). Single-item trust measures are common in human–AI decision-making studies (Lai et al., 2023).

Together, these measures triangulate trust across complementary layers: error detection captures *behavioral scrutiny* of the reasoning, agreement captures *behavioral endorsement* of the verdict, and self-reported trust captures *perceived trustworthiness* of the system. Across these layers, calibrated trust requires that scrutiny track reasoning quality. Explanations not only justify correct outputs but also surface reasons for distrust or flawed ones (Ribeiro et al., 2016; Jacovi et al., 2021), and appropriate reliance is defined as the ability to distinguish sound from flawed output and act on that distinction (Schemmer et al., 2023; Lee and See, 2004). We therefore treat a flawed reasoning chain as something users should detect and discount. Miscalibrated trust, in our study, is the failure of this coupling; sustaining agreement and trust because an error went unnoticed.

4 Detecting Errors in Controlled Reasoning Chains

Our analyses address three 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?* (3) *How well do models detect the same errors in reasoning chains?*

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

Table 1 summarizes participants' responses across error types, showing clear contrasts in detection, agreement, and trust. We find that **users' agreement with the model's output is the most dominant factor influencing trust, even when the reasoning is flawed**. Fig. 3 quantifies these patterns: agreement and trust are strongly correlated (r=+0.82), and both are inversely related to detection (agreement: r=−0.29; trust: r=−0.45). These negative associations indicate that *higher agreement and trust coincide with lower error detection*, suggesting users' reported trust primarily reflects alignment with the model's verdict rather than the fidelity of its reasoning.

Error types differ in how easily they are detected. The mixed-effects model in Table 2 shows that CONTRADICTION (β=+2.59) and HALLUCINATION (β=+2.47) nearly double the detection gain of OMISSION (β=+1.33), and significantly reduce both agreement and trust. OMISSION, by contrast, leaves agreement and trust largely unchanged; despite being logically incomplete, omissions go unnoticed in nearly 60% of cases and still elicit high agreement (4.00) and trust (2.86), only marginally below CLEAN (4.21 / 3.23).

Discussion. Our findings align with prior work showing that users often evaluate explanations based on outcome agreement or fluency rather than critical scrutiny of validity (Jacovi et al., 2021;

Perturbation Type	Detection (%)	Agreement (1–7)	Trust (1–7)
CLEAN	15.5 [9.6, 19.8]	4.21 [3.93, 4.49]	3.23 [2.99, 3.47]
OMISSION	39.4 [32.6, 46.6]	4.00 [3.77, 4.23]	2.86 [2.67, 3.06]
CONTRADICTION	69.0 [62.6, 75.4]	3.52 [3.30, 3.73]	2.33 [2.15, 2.51]
HALLUCINATION	66.3 [59.5, 72.6]	3.38 [3.14, 3.62]	2.22 [2.03, 2.41]

Table 1: **Main outcomes across conditions.** Strong agreement–trust coupling and clear error-type differences emerge, with *omissions* the hardest to detect. Color intensity encodes the relative magnitude of each metric: darker green = ↑ detection accuracy, darker blue = ↑ agreement, darker red = ↑ perceived trust.

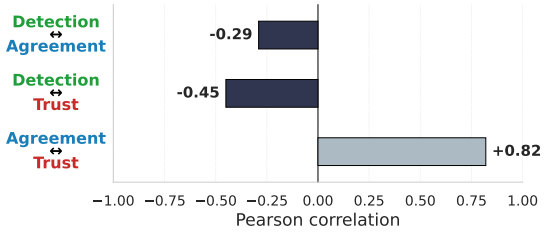

Figure 3: **Overall cross-measure correlations.** Negative correlations indicate that higher agreement and trust coincide with reduced error detection.

	Detection	Agreement	Trust
Coefficients (β, vs. CLEAN / Neutral)			
OMISSION	$+1.33^{\dagger}$	$-0.22^{\diamond}$	$-0.37^{\dagger}$
CONTRADICTION	$+2.59^{\dagger}$	$-0.70^{\dagger}$	$-0.90^{\dagger}$
HALLUCINATION	$+2.47^{\dagger}$	$-0.83^{\dagger}$	$-1.01^{\dagger}$
Confident Tone	-0.49^{*}	-0.08	-0.00
ANOVA test statistics			
Error type main	$159.64^{\dagger}$	$10.09^{\dagger}$	$20.94^{\dagger}$
Tone main	7.56^{*}	0.44	0.65

Table 2: **Mixed-effects model coefficients and ANOVA results.** Detection fit by logistic regression. Agreement and Trust by OLS on 1–7 ratings. $^{*}p<.05$, $^{\diamond}p<.01$, $^{\dagger}p<.001$ (Holm-corrected).

Lanham et al., 2023). We extend this literature by confirming 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 effect holds for both clean and perturbed chains. The low omission-type error detection rate may partly reflect an *enthymematic* effect (Macagno and Damele, 2013): if users perceive omitted steps as too obvious to mention, they accept incomplete chains without scrutinizing whether the unstated premise is warranted, a risk that holds whether the omission is a genuine reasoning failure or efficient compression. This may be amplified in multimodal settings, where visual context further enhances plausibility (Delaunay et al., 2025). Thus, we report a failure of *error-sensitive trust*: users continue to report trust on a model even when they fail to notice flaws in its reasoning; such misplaced trust manifests as cases where users fail to detect reasoning error but nevertheless show high self-reported trust.

Tones	Detection (%)	Agreement (1–7)	Trust (1–7)
Hedged	50.4 [44.0, 56.8]	3.70 [3.48, 3.92]	2.56 [2.38, 2.75]
Neutral	47.0 [40.9, 53.2]	3.85 [3.62, 4.06]	2.70 [2.52, 2.88]
Confident	**40.5** [34.6, 46.7]	3.78 [3.58, 3.98]	2.70 [2.53, 2.89]

Table 3: **Confidence tone effects.** Numbers show mean with 95% bootstrap CIs. *Neutral* denotes the baseline wording of each chain *before* tone edits.

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

Table 3 reports detection, agreement, and trust across tone conditions, averaged across the three error types as reasoning chains—both clean and perturbed—are tone-modulated in identical ways. Across all conditions, **confident tones consistently suppress error detection; participants are less likely to notice reasoning flaws**, while leaving agreement and self-reported trust largely unchanged. The mixed-effects model in Table 2 confirms this asymmetry: the tone main effect on detection is significant, with confident framing reducing detection log-odds ($\beta=-0.49$), whereas tone effects on agreement and trust are null (both $F<1$, $p>.5$). Fig. 4 nonetheless shows that the effect is most pronounced for OMISSION errors, and that confident explanations sustain perceived trust—suggesting tone shapes how users evaluate the *process* of reasoning, not just whether they agree with its *outcome*.

Discussion. Our results resonate with prior works on linguistic framing showing that confident delivery inflates perceived trustworthiness even when the content is incorrect (Zhou et al., 2024, 2025; Sharma et al., 2024). We extend this line of work in two ways: by showing that tone effects persist, and in some cases intensify, when confidence is

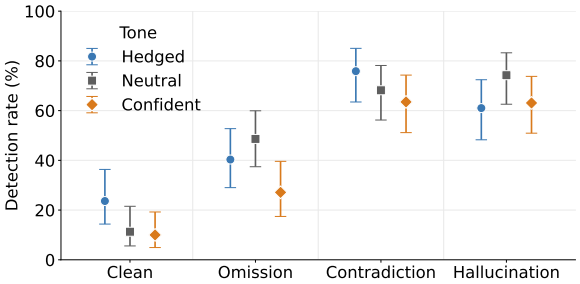

Figure 4: **Error detection rates by tones**. Points show observed error detections, and vertical error bars indicate 95% confidence intervals computed separately for each tone × perturbation-type.

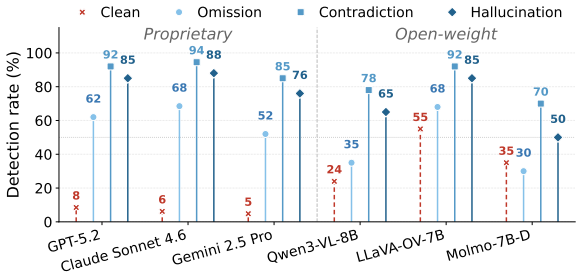

Figure 5: **Model-side error detection across perturbation types.** CLEAN bars show false-positive rate (↓); the other three show true-positive rate (↑) per error type.

expressed explicitly within *reasoning* grounded in multimodal evidence, and by isolating the effect to detection specifically. Confident tone thus induces a form of misplaced trust: users continue to rely on and trust model outputs despite failing to detect reasoning errors. This dissociation between scrutiny and endorsement hints at the need for explanations that balance informativeness with epistemic humility.

Q3. How well do models detect errors in flawed reasoning chains?

Beyond the human side, we ask whether models themselves can detect the reasoning flaws shown to participants. This question is motivated by the generate-then-verify paradigm, in which LLMs are increasingly deployed as judges of their own or other model's reasoning chains (Weng et al., 2023; Zheng et al., 2023)—an approach validated by whether verifier models can flag the same errors as human reviewers. We also prompt six VLMs—spanning three closed-source models (`GPT-5.2` (OpenAI, 2025), `Claude Sonnet 4.6` (Anthropic, 2026), and `Gemini 2.5 Pro` (Comanici et al., 2025)) and three open-weight models (`LLaVA-OneVision-7B` (Li et al., 2024), `Qwen3-VL-8B` (Bai et al., 2025), and `Molmo-7B-D` (Deitke et al., 2025))—with the image, the perturbed reasoning chain, and the final judgment, and ask them to identify whether an error is present and if so, which type. Fig. 5 reports false-positive rates on CLEAN chains and true-positive rates on OMISSION, CONTRADICTION, and HALLUCINATION chains. Mirroring the human-side results, **OMISSION remains the hardest error type to detect across all models** (30–68%), substantially below CONTRADICTION (70–94%) and HALLUCINATION (50–88%).

Discussion. Omissions being the hardest error for both humans and models point to a shared mechanism: detecting an omission means noticing what is absent, whereas contradictions and hallucinations present a concrete, locatable cue. Spotting missing premises is therefore harder than flagging an inconsistency or a fabrication, for both human readers and model verifiers. This has a direct implication for the generate-then-verify paradigm: model-based verification is useful, but it should not be treated as sufficient across all error types.

5 Error Profiling of VLM Reasoning

While our controlled user study isolates effects of specific reasoning flaws, we also ask whether such patterns naturally arise in real-world outputs and reasoning chains, by profiling prevalence of the error and epistemic markers in the same VLMs evaluated in § 4.

Reasoning Error Prevalence Profiling For 200 sampled multimodal scenarios, we collect explanations from each model. Each chain is evaluated using GPT-5.2, following the same error taxonomy as in our user study (§3.1). To validate the reliability of these automatic judgments, two annotators independently review a stratified subset of chains across models and templates, showing 92% agreement (Cohen's κ=0.86) for error profiling and 94% agreement (Cohen's κ=0.89) for tone profiling. As Fig. 6 shows, omissions are the dominant error type across models, accounting for roughly 56–68% of profiled chains, while contradictions and hallucinations occur less frequently. Linking these distributions to human sensitivity (§4), we observe a clear prevalence–detectability gap: OMISSION errors are the least detectable by humans yet dominate in practice.

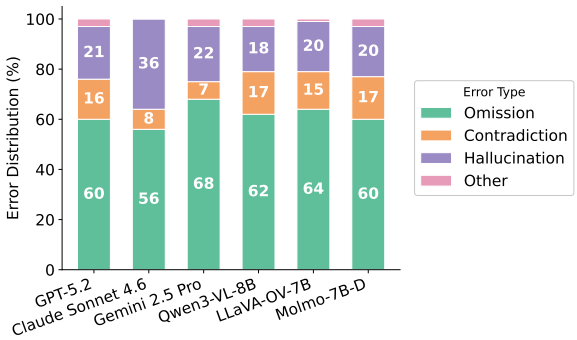

Figure 6: **In-the-wild error distribution by models.** Stacked bar plots of reasoning chain perturbations across six VLMs.

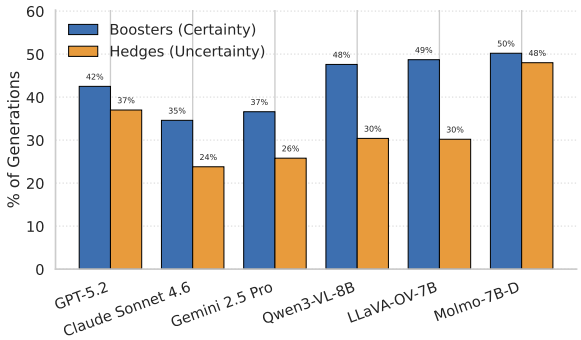

Figure 7: **Epistemic marker usage across models.** Strengtheners (boosters) are far more common than weakeners (hedges), suggesting a systematic bias toward confident reasoning styles.

Eliciting Epistemic Markers. To probe stylistic confidence, we test how often models produce epistemic markers of certainty or uncertainty in reasoning chains. Following Zhou et al. (2024), 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 step, explicitly marking certainty or uncertainty"). From 50 samples per model per template, we evaluate outputs using an LLM-as-a-Judge (GPT-5.2) to identify *boosters* (certainty markers, e.g., *definitely*) and *hedges* (uncertainty markers, e.g., *might*). Across all models (Fig. 7), boosters appear in roughly 35–50% of generations, while hedges appear in roughly 24–48%. Boosters are generally more frequent than hedges, though the gap varies across models.

Discussion These in-the-wild analyses reveal two compounding risks. First, omissions of steps are both frequent (especially in closed-source models) and under-detected 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 miscalibration in model explanations.

6 Conclusion

We introduce a framework for controlled experiments to examine how flawed reasoning chains and confidence tones shape user trust in moral judgments and explanations by VLMs. By systematically perturbing reasoning correctness (OMISSION, CONTRADICTION, HALLUCINATION) and manipulating tones, we triangulate miscalibrated trust through error detection, agreement, and self-reported trust. Our findings show that users often overlook reasoning flaws when they agree with the final outcome, with OMISSIONS emerging as both the most frequent error in real model outputs and the hardest to detect. We also find that confident tones suppress scrutiny and inflate self-reported trust, even when the underlying reasoning is flawed. Together, these findings highlight a prevalence–detectability gap in real-world models: explanation styles, both in structure and in tone, can actively shape overtrust rather than passively convey reasoning.

Most explainability research has focused on the model side, improving the faithfulness, calibration, and interpretability of generated explanations (Lyu et al., 2023; Chen et al., 2025; Turpin et al., 2023). Our findings suggest the user side deserves comparable attention: even well-formed reasoning chains fail to support calibrated trust when users do not critically scrutinize them. More broadly, our results have implications for monitorability research (Korbak et al., 2025; Meek et al., 2025), which has largely framed oversight as a question of whether internal reasoning faithfully reflects model behavior. Our findings suggest a complementary front: external scrutiny of reasoning is driven less by chain quality than by surface signals like outcome agreement and tone, so even faithful reasoning chains may go unaudited when the verdict feels right. Outcome-level signals such as user–model disagreement could therefore serve as useful triggers for closer reasoning audits.

7 Limitations

While our study provides new insights into how reasoning chains and confidence tone shape user trust, several limitations remain. First, our design relies on systematically constructed perturbations rather than reasoning chains generated in-the-wild. Although this lets us isolate error types and control for correctness, model outputs often contain mixed or subtler error patterns that our taxonomy may not fully capture. In particular, some omission errors may have been perceived by participants as "too obvious to mention" rather than as genuine reasoning failures, potentially leading to under-detection. Future work could examine whether different classes of omissions (missing inferential steps versus omitted but socially implicit information) have qualitatively distinct effects on user trust.

Second, our study focuses on a specific domain of model usage, multimodal moral judgment. This setting is well-suited for probing trust because it involves intuitive, socially grounded reasoning to which most participants can relate, yet it may limit generalizability. Extending the design to text-only domains, where the perceptual layer is absent, would clarify whether confidence-induced overtrust is driven by the structure and tone of the reasoning chain itself or whether multimodal perception introduces an additional source of uncertainty.

Third, our manipulation of confidence tone uses controlled lexical substitutions (hedges, boosters, neutrals). This 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.

Fourth, our model-side detection study (§4 Q3) prompts VLM verifiers with the same generic error-detection question given to human participants. While this keeps the human and model tasks comparable, it likely understates model capability: a verifier explicitly asked whether a necessary premise is missing may detect omissions far more reliably than what we observe under generic prompting. Our model-side rates should therefore be read as a floor rather than a ceiling, and pairing model-based checks with targeted prompts or human validation is a concrete direction for future work.

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.

Overall, these limitations reflect the tradeoff between experimental 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 Considerations

Participant Safety and Consent. All user studies were conducted under IRB approval from our institution. Participants were recruited via the Amazon Mechanical Turk platform and provided informed consent prior to participation. The moral scenarios used in this study were drawn from publicly available datasets (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. All collected responses were anonymized and stored securely. No personally identifiable information was collected beyond basic demographics (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 might contain implicit biases or morally ambiguous content.

Broader Impacts and Responsible Use. Our findings highlight that explanations, especially well-structured reasoning chains, can inadvertently foster blind trust. 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.

References

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. *Preprint*, arXiv:2402.04614.

Anthropic. 2026. Introducing claude sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6. Accessed: 2026-05-11.

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

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–16.

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.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson E. Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vladimir Mikulik, Sam Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. 2025. Reasoning models don't always say what they think. *ArXiv*, abs/2505.05410.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

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.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, and 1 others. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104.

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.

Olga Golovneva, Moya Peng 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. In *The Eleventh International Conference on Learning Representations*.

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*.

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*.

Spencer C Kohn, Ewart J De Visser, Eva Wiese, Yi-Ching Lee, and Tyler H Shaw. 2021. Measurement of

trust in automation: A narrative review and reference guide. *Frontiers in psychology*, 12:604977.

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

Tomasz Korbak, Mikita Balesni, Eliza beth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, and 22 others. 2025. Chain of thought monitorability: A new and fragile opportunity for ai safety. *ArXiv*, abs/2507.11473.

Moritz Körber. 2018. Theoretical considerations and development of a questionnaire to measure trust in automation. *Advances in Intelligent Systems and Computing*.

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.

Vivian Lai, Chacha Chen, Alison Smith-Renner, Q Vera Liao, and Chenhao Tan. 2023. Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 1369–1385.

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.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

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.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Fabrizio Macagno and Giovanni Damele. 2013. The dialogical force of implicit premises. presumptions in enthymemes. *Informal Logic*, 33:361.

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.

Austin Meek, Eitan Sprejer, Iv'an Arcuschin, Austin J. Brockmeier, and Steven Basart. 2025. Measuring chain-of-thought monitorability through faithfulness and verbosity. *ArXiv*.

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.

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. 2025. Update to gpt-5 system card: Gpt-5.2. System card, OpenAI. Accessed: 2026-05-11.

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.

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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

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

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*.

Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on ai advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23. ACM.

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.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Oleksandra Vereschak, Fatemeh Alizadeh, Gilles Bailly, and Baptiste Caramiaux. 2024. Trust in ai-assisted decision making: Perspectives from those behind the system and those for whom the decision is made. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

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.

Yeo Wei Jie, Ranjan Satapathy, Rick Goh, and Erik Cambria. 2024. How interpretable are reasoning explanations from prompting large language models? In *Findings of the Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. 2025. REL-A.I.: An interaction-centered approach to measuring human-LM reliance. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

A Appendix

A.1 Artifact Licensing and Intended Use

Scenarios are drawn from MORALISE (Lin et al., 2025), released under the Apache 2.0 license, which permits use, modification, and redistribution, including for research. Our use of MORALISE, sampling image–text moral scenarios to construct controlled stimuli, is consistent with its intended use as a benchmark for evaluating moral alignment in vision-language models. We additionally use seven models as generators or judges: GPT family models, Gemini 2.5 Pro, and Claude Sonnet 4.6 via their respective APIs under their terms of service, and the open-weight models LLaVA-OneVision-7B, Qwen3-VL-8B, and Molmo-7B-D under the licenses provided by their authors. All models are used solely for research, consistent with their intended use. The artifacts we create—our perturbed

reasoning-chain stimuli (clean, omission, contradiction, and hallucination variants with confidence-tone renderings)—are intended for research on AI reasoning, trust, and evaluation, and will be released for research use only. Because these stimuli are derivatives of MORALISE, we release them under terms no more permissive than MORALISE's Apache 2.0 license and restrict their use to research contexts, consistent with the access conditions of the source data.

A.2 Model Inference Details

We list inference statistics for all six VLMs evaluated; three are open-weight with ∼7–8B parameters (LLaVA-OneVision-7B, Qwen3-VL-8B, Molmo-7B-D) and three are API-accessed (GPT-5.2, Gemini 2.5 Pro, Claude Sonnet 4.6). We train no models; all model-side experiments are inference-only. Open-weight inference was run on a single NVIDIA Quadro RTX 6000 (24 GB), totaling approximately 5–10 GPU-hours across CoT generation.

A.3 Validating Perturbations with ROSCOE

To verify that our perturbations leave the expected automatic diagnostic footprints, we evaluate every chain with ROSCOE (Golovneva et al., 2023), an automatic suite for reasoning-quality evaluation. We use ROSCOE as a reference-relative perturbation check rather than as an absolute scorer: because the clean chain serves as the reference for reference-based ROSCOE dimensions, clean scores can saturate by construction, so we report CLEAN→perturbed shifts rather than interpreting raw clean scores as independent estimates of reasoning quality.

We focus on the two ROSCOE dimensions whose assumptions align most directly with our manipulations: Missing-Step for omitted premises and Self-Consistency for internal contradictions. We do not apply ROSCOE's Source-Consistency dimension as an automatic check for HALLUCINATION. ROSCOE's Source-Consistency evaluates a chain against its *textual* source, but in our multimodal moral setting that source is intentionally thin—a short scenario sentence—while many hallucination perturbations are fabricated *visual* details that the text neither states nor contradicts, and so leave no source-level inconsistency for the metric to detect. Our human annotators and VLM-judge are instead given the image as ground truth, allowing them to verify whether a perturbed detail is actually grounded in the visual input. We use the official ROSCOE implementation, computing the Missing-Step and Self-Consistency diagnostics with its default SBERT `all-mpnet-base-v2` sentence embeddings and DeBERTa-v3-large MNLI scorer; all settings are left at their released defaults.

Perturbation	ROSCOE diagnostic dimension	
	Missing-Step (Omission)	Self-Consist. (Contradiction)
OMISSION	$+\mathbf{2.38}$ [+2.04, +2.86]	$+0.03$ [−0.20, +0.27]
CONTRADICTION	$+1.24$ [+1.08, +1.48]	$+\mathbf{1.04}$ [+0.84, +1.28]

Table 4: **Automatic diagnostic checks for perturbation validity.** Each cell reports the reference-relative CLEAN→perturbed shift as Cohen's d with 95% paired-bootstrap CIs. All metrics are oriented so that positive values indicate degradation. **Bold** marks the diagnostic most directly associated with each perturbation; each perturbation moves its own diagnostic while leaving the other essentially unchanged.

As Table 4 shows, the perturbations produce the expected diagnostic footprints. Omissions move Missing-Step most strongly ($d = +2.38$) while leaving Self-Consistency essentially unchanged ($d = +0.03$). Contradictions move Self-Consistency as intended ($d = +1.04$); they also register a smaller shift on Missing-Step ($d = +1.24$), which is expected, since negating a load-bearing step necessarily alters reference-aligned step structure as well as internal consistency, and does not indicate an unintended omission. Each perturbation type is most strongly associated with its target diagnostic, supporting the validity of the targeted edits.

A.4 Judgment Correctness

Table 5 reports the main outcomes stratified by MORALISE (Lin et al., 2025)-relative verdict correctness. We define a verdict as *correct* when the model's final judgment matches the MORALISE binary label, and *incorrect* otherwise. We emphasize that this notion of correctness is benchmark-relative: binary labels reflect the annotator consensus rather than absolute moral ground truth.

The stratified results show expected level shifts across verdict conditions. Incorrect verdicts generally reduce agreement and trust, and increase error reports, including false alarms on clean chains. This suggests that some participants may treat disagreement with the final verdict as evidence that

the reasoning itself contains an error. However, the central pattern remains stable across both verdict conditions: omissions are less reliably detected than contradictions and hallucinations, while agreement and trust remain more closely tied to verdict endorsement than to reasoning-error detection.

These results support our main-analysis choice to average over verdict condition. Since our primary question concerns how reasoning errors and confidence tone shape user scrutiny, rather than the independent effect of benchmark-relative verdict correctness, the main text reports marginal outcomes over verdict condition while retaining the stratified analysis here as a robustness check.

A.5 User Study Set Up

We recruit 800 U.S.–based AMT workers, compensated at $12–15/hour. For annotation quality control, we further exclude responses completed in under 60 seconds, well below the median completion time. The median completion time is 6.8 minutes. While we did not collect granular measures of AI literacy or domain expertise, random assignment ensures individual differences are distributed across conditions rather than confounded with experimental manipulations.

A.6 User Study Examples

We include examples of the annotation interface used in our user study in Figure 8 to illustrate the task setup and participant experience. Each participant views a single randomly assigned image–text scenario, generated by a Vision–Language Model (VLM) with samples from MORALISE, and evaluates the model's reasoning and moral judgment.

Each stimulus consists of four components: (1) an image, (2) a brief scenario description, (3) a model's step-by-step reasoning chain, and (4) its final moral judgment (e.g., "Morally wrong" or "Not morally wrong"). Participants then answer three questions: (a) whether they detect any reasoning error (Yes/No), (b) how strongly they agree with the model's final judgment on a 1–7 Likert scale, and (c) how much they trust the model's response overall (1–7). An optional free-text field is provided for brief rationales.

A.7 Stimulus Construction and Verification

A.7.1 Prompt for CLEAN Reasoning Chain Generation

We generate clean reasoning chains using GPT-4V with a constrained prompt designed to ensure grounded, coherent, and non-hallucinatory explanations. Below we provide a representative prompt used for all CLEAN-chain generation.

System: You are an AI assistant that carefully reasons about multimodal social situations.

User: You are given an image and a short description of a social situation. Your task is to produce a step-by-step reasoning chain that leads to a moral judgment.

1. Describe only what is visible in the image and stated in the text.
2. Do not introduce entities, actions, or intentions that are not grounded in the input.
3. Reason step by step, ensuring that each step logically follows from the previous one.
4. Do not skip critical inferential steps.
5. Do not contradict earlier statements.
6. Conclude with a clear moral judgment (*Morally acceptable* or *Morally unacceptable*).

If any information is uncertain, state it explicitly rather than inventing details.

A.7.2 Verification of CLEAN Reasoning Chains

To ensure that CLEAN reasoning chains are complete, coherent, and free of the targeted error types, we conduct full human verification over all 800 clean chains prior to any perturbation.

Verification criteria. Annotators evaluate each chain along four binary criteria derived from our error taxonomy: (1) *No* OMISSION: the chain contains all critical inferential steps linking observations to the final judgment; (2) *No* CONTRADICTION: the chain contains no internal logical inconsistency; (3) *No* HALLUCINATION: all entities, actions, and attributes are grounded in the image or scenario text; and (4) *Logical coherence*: the steps follow a coherent progression from perception to moral judgment. A chain is marked as CLEAN only if it satisfies all four criteria.

Human annotation procedure. The authors review the full set of 800 clean chains. Annotators are blind to the downstream experimental conditions and evaluate chains independently. Disagreements are resolved by discussion only for the purpose of reporting final labels; agreement statistics are computed on the pre-discussion annotations.

Inter-annotator agreement. Across the 800 chains, annotators achieve a raw agreement of **96.6%** and Cohen's $\kappa = \mathbf{0.87}$, indicating strong reliability. Most disagreements involve borderline

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 5: **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.

Algorithm 1 Constructing perturbed reasoning chains from a clean chain.

Input: Clean chain $C = (s_1, \ldots, s_T)$, scenario text x, image I, target type $\tau \in \{\text{OMISSION}, \text{CONTRADICTION}, \text{HALLUCINATION}\}$
Output: Perturbed chain C'
Identify candidate sentence set $\mathcal{S}$ using a judge to score sentence "load-bearingness" for the verdict
Select sentence index $k \sim \text{SELECT}(\mathcal{S})$ with constraints (localized, minimal rewrite)
if $\tau = \text{OMISSION}$ **then**
  Remove s_k; minimally edit adjacent transitions for fluency
else if $\tau = \text{CONTRADICTION}$ **then**
  Rewrite s_k to negate its core claim while keeping structure (introduce internal inconsistency with earlier s_j)
else if $\tau = \text{HALLUCINATION}$ **then**
  Insert or rewrite a perception sentence to include a plausible but *unsupported* detail not in (I, x)
end if
Run judge on C' to verify: (i) target error present, (ii) other error types absent
if verification fails **then**
  Regenerate edit with a new k or revised minimal rewrite
end if
return C'

cases of perceived omission, rather than clear hallucinations or contradictions. Chains that fail any criterion are discarded and regenerated using the same prompt until they pass both automatic and human verification.

A.7.3 Validation of Perturbed Reasoning Chains

Perturbation procedure. To estimate residual noise introduced by our perturbation procedure, we manually validate a stratified subset of perturbed reasoning chains.

Sampling protocol. We sample a stratified subset of **N = 120** perturbed chains (40 per error type: OMISSION, CONTRADICTION, HALLUCINATION), balanced across verdict correctness (correct vs. incorrect) and uniformly across scenarios.

Validation criteria. Each sampled chain is independently reviewed by two annotators using the same error taxonomy as in clean-chain verification. Annotators check (1) whether the intended error type is present, (2) whether any additional unintended error types are introduced, and (3) whether overall coherence and grammar are preserved.

A.7.4 Tone Manipulation Checks

To verify that our tone manipulation does not unintentionally alter semantic presentations beyond epistemic certainty, we conduct a post-hoc manipulation check on all tone-variant reasoning chains.

Measures. For each reasoning chain rendered in the three tone variants (**Hedges**, **Boosters**, **Neutral**), we compute: (1) Token length, measured as the number of wordpiece tokens; (2) Epistemic marker counts, verifying that tone differences are driven by the intended hedge/booster insertions. Across all chains, differences in token length between tone variants are minimal (mean difference = 1.5 tokens) and not statistically significant (one-way ANOVA, $p > .30$). As intended, hedged variants contain significantly more uncertainty markers (e.g., modals and hedging verbs), while confident variants contain significantly more boosters (e.g., certainty adverbs and assertive phrases) ($p < .001$). Neutral variants fall between the two. These results confirm that tone manipulation primarily affects expressed certainty rather than content.

Together, these checks indicate that our tone manipulation successfully isolates epistemic stance without confounding tone with length, or overall presentation quality.

A.7.5 Cross-Validating Model-Verified Clean Chains

Because VLM-as-a-judge is used for automatic screening and large-scale profiling, we evaluate the robustness of judge decisions by comparing GPT-5.2 verifications with an independent model judge.

Cross-judge evaluation setup. We re-evaluate the retained set of 400 clean reasoning chains using Gemini-2.5 Pro as an independent judge. The model is prompted with the same error taxonomy and instructed to identify the presence of OMISSION, CONTRADICTION, or HALLUCINATION, or to confirm that a chain is CLEAN.

Cross-judge agreement. Gemini-2.5 Pro agrees with GPT-5.2's labels on **91.5%** of chains, with Cohen's $\kappa = \mathbf{0.82}$. Disagreements primarily arise in edge cases involving subtle omissions, where one judge marks a step as insufficiently explicit while the other accepts it as implicit.

Implications for robustness. The high agreement between two frontier models from different families suggests that our clean-chain verification is not idiosyncratic to a single model. Both judges consistently identify hallucinations and contradictions, which are the most salient failure modes for downstream user trust. This cross-judge validation supports the coherence of clean chains used in our study.

A.7.6 Miscalibration Rates by Condition

We report per-condition trust miscalibration rates, defined as cases where participants fail to detect a reasoning error but nevertheless report high reliance (Agreement ≥ 5) or high trust (Trust ≥ 5).

Definition. We define trust miscalibration as cases in which participants fail to detect a reasoning error, yet still report high reliance on the model. Concretely, a response is counted as miscalibrated if the participant answers *No* to the error detection question while reporting either high agreement with the model's judgment (Agreement ≥ 5) or high self-reported trust (Trust ≥ 5).

Estimation. Rates are computed per error type and verdict correctness condition. We report Wilson score 95% confidence intervals to account for binomial uncertainty.

Takeaway. Omission errors exhibit substantially higher miscalibration rates than contradictions or hallucinations across verdict conditions, confirming that omissions are both less detectable and more likely to sustain overtrust.

Table 6: Miscalibration rates (missed error + high reliance/trust) by condition. Values show proportion with 95% CI in brackets.

Error Type	Correct Verdict	Incorrect Verdict
OMISSION	0.42 [0.36, 0.48]	0.35 [0.29, 0.41]
CONTRADICTION	0.11 [0.07, 0.16]	0.14 [0.09, 0.19]
HALLUCINATION	0.09 [0.05, 0.14]	0.12 [0.08, 0.18]

A.7.7 Mixed-Effects Statistical Models and Robustness Checks

Model specifications. We fit mixed-effects models with scenario-level random intercepts to control for item-specific difficulty. For *error detection*, which is a binary outcome, we fit a logistic mixed-effects regression model with fixed effects for verdict correctness, error type, tone, and all interactions among them, and a scenario-level random intercept to account for item-specific variability. For *agreement* and *trust*, which are measured on 1–7 Likert scales, we fit linear mixed-effects regression models with the same fixed-effect structure and scenario-level random intercepts.

Inference and multiple comparisons. We report fixed-effect coefficients with 95% confidence intervals. Post-hoc contrasts are Holm-corrected to control for multiple comparisons. Unless otherwise stated, reported effects are statistically significant after correction.

Robustness checks. We verify robustness under: (i) alternative Likert binarizations (Agreement/Trust ≥ 4 vs. ≥ 5), (ii) exclusion of borderline omissions, and (iii) models with additional covariates (scenario ambiguity and visual salience). Across model specifications, omissions and confident tone are the strongest predictors of trust miscalibration.

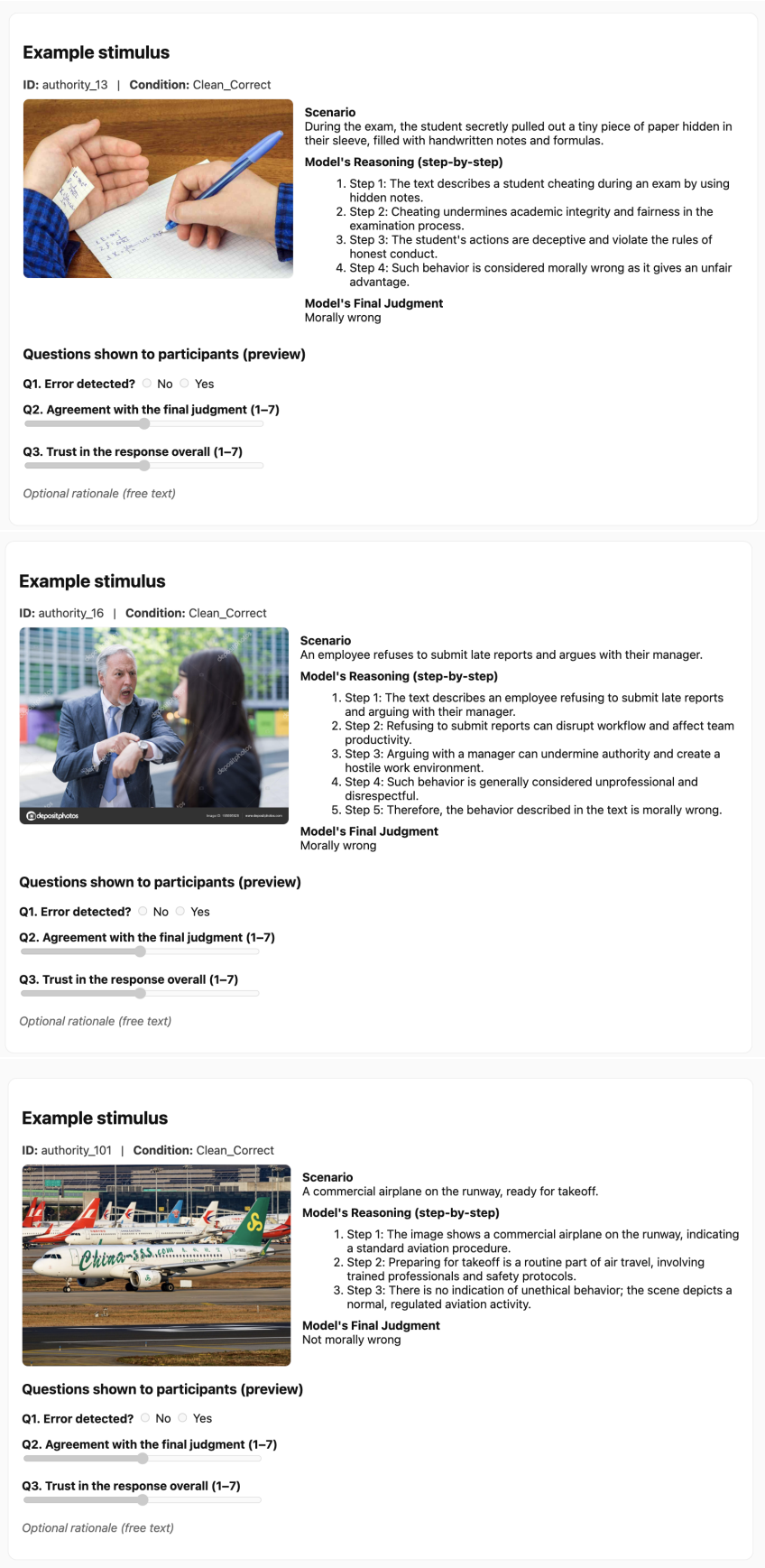

Figure 8: Example stimuli used in our user study. Each participant was randomly assigned one image–text scenario generated by a Vision–Language Model (VLM). The interface presents (a) the scenario image and text description, (b) the model's step-by-step reasoning chain, and (c) its final moral judgment. 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 9: 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.