DO FRONTIER MODELS KNOW WHAT THEY DON'T KNOW? A UNIFIED PROTOCOL FOR TRUSTWORTHINESS EVALUATION ACROSS VISION, RETRIEVAL, AND AGENTIC TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Frontier models are increasingly capable, yet deployment hinges on a different question: do they know when they are wrong? We argue that trustworthiness — abstaining under insufficient or corrupted evidence, flagging conflicts, and diagnosing failures — should be measured *directly* rather than inferred from accuracy. We introduce a unified, defect-conditioned protocol that injects controlled defects into otherwise-solvable instances and rewards models for the *appropriate response* (answer, abstain, or conflict-aware), with programmatic gold labels, multi-judge voting (Fleiss κ=0.76), and stratified failure audits. We instantiate the protocol in three scenarios spanning modality, evidence, and action: visual claim verification (S1, 1,582 shared image–claim probes), conflicting-retrieval calibration (S2, 3,000 instances), and agentic trace diagnosis (S3, 300 multi-step tasks) — all under an identical roster of six frontier models. Across all three we find that (i) self-reported confidence is a weak error detector, *inversely* calibrated on S2 (error-ranking AUROC < 0.5 on every contested slice) and only mildly informative on S1 (AUROC 0.645), while cross-model disagreement is the stronger signal in all three scenarios and strongest under agentic action (S3 AUROC 0.853 vs. self-confidence 0.764, where half the agents self-report near-maximal confidence at chance-level AUROC even when they fail); (ii) a *consensus blind spot* exists where models fail *together* on plausible-but-wrong inputs (stale and adversarial evidence), so cross-model disagreement — otherwise the strongest signal — collapses below chance; and (iii) neither introspection nor self-consistency resampling rescues this blind spot, though a lightweight evidence-critique prompt closes the adversarial half of it (misled-answer rate $0.72 \rightarrow 0.00$) while the stale half stays open. The unified protocol exposes a shared failure axis invisible to any single-scenario benchmark: models are reliable when defects are overt and confidently wrong when defects are plausible. We release the protocol, data, and judge harness.

1 INTRODUCTION

The dominant story of the last two years is capability: frontier models pass demanding evaluations and operate tools autonomously (Yue et al., 2024; Liu et al., 2024). Deployment, however, turns on a complementary property that benchmarks rarely isolate — *trustworthiness*: knowing when the available evidence is insufficient, flagging when sources conflict, and recognising when a plan has gone wrong (Sun et al., 2024). A model that answers a poisoned retrieval as confidently as a clean one, or that narrates a broken agent trajectory as a success, is dangerous precisely because it is fluent. We ask a single question across very different tasks: *do frontier models know what they don't know?*

Existing evaluations make this hard to answer. Accuracy on clean benchmarks conflates capability with calibration (Guo et al., 2017); hallucination leaderboards score surface correctness but not the meta-decision to abstain (Lin et al., 2022; Li et al., 2023b; Guan et al., 2024); and existing LLM calibration studies are modality-specific and use heterogeneous metrics, so it is impossible to tell whether a failure is a quirk of one task or a property of the model (Xiong et al., 2024; Geng et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

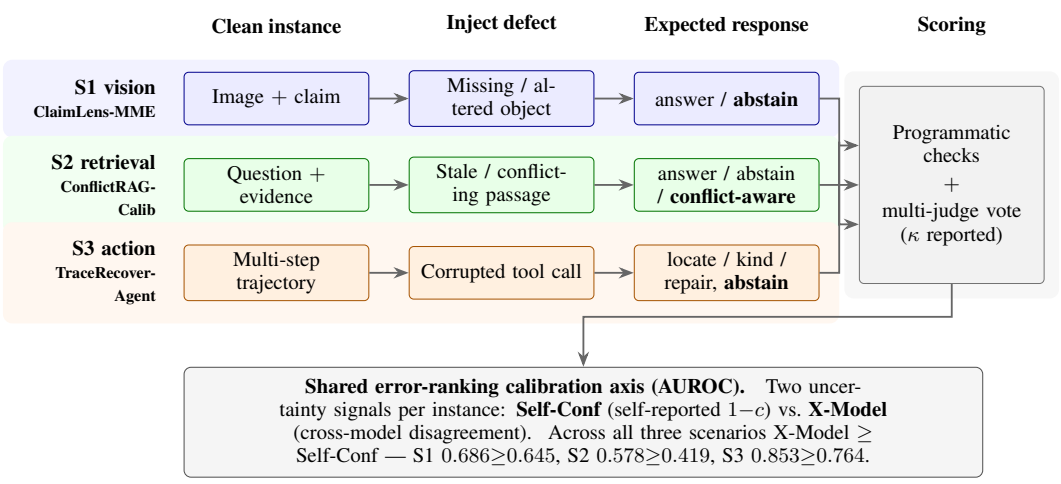


Figure 1: Overview of the unified defect-conditioned protocol. A single template—clean instance → injected defect → expected response class (answer / abstain / conflict-aware) → programmatic and multi-judge scoring—is instantiated across vision (S1), retrieval (S2), and action (S3), and every scenario is read out on the same error-ranking calibration axis contrasting self-reported confidence with cross-model disagreement.

2024). What is missing is a *common protocol* that poses the same trustworthiness question — "respond appropriately to a defective input" — across modalities and task types, with comparable metrics.

We propose such a protocol (Figure 1). Its core is *defect conditioning*: we start from instances a capable model can solve, inject a controlled defect (a missing object, stale or contradictory evidence, a corrupted tool call), and define the trustworthy behaviour as a specific response class (answer, abstain, or conflict-aware) with a *programmatic* gold label. Subjective outputs are scored by a multi-judge vote (Fleiss κ=0.76), and every aggregate is backed by a stratified audit of failures. Crucially, the same protocol yields the same metric family — decision accuracy, abstention F1, calibration error, and the error-ranking AUROC of candidate uncertainty signals — in every scenario, so cross-scenario comparison is meaningful.

We instantiate the protocol in three scenarios chosen to span the deployment surface: **S1 (vision)** verifies natural-language claims about images, including under-determined claims that should be refused; **S2 (retrieval)** probes calibration under six evidence-defect conditions over 3,000 instances; and **S3 (action)** asks an agent to diagnose *why* a recorded tool-use trajectory failed and propose a repair. Studying them together, rather than in three separate papers, is what lets us separate scenario-specific noise from shared model behaviour.

Our contributions are: (1) a *unified, defect-conditioned trustworthiness protocol* with programmatic gold, multi-judge scoring, and a common metric family, instantiated in three scenarios spanning modality, evidence, and action; (2) *three cross-scenario findings* — self-reported confidence is inversely calibrated everywhere, a consensus blind spot makes models fail together on plausible-but-wrong inputs, and agreement-based signals beat introspection but do not close the blind spot; and (3) an *open release* of data, programmatic checkers, and the judge harness, enabling reproducible trustworthiness evaluation and extension to new scenarios.

2 Related Work

Calibration and selective prediction. Classical calibration and selective prediction ask whether confidence tracks accuracy and when a model should abstain (Guo et al., 2017; Geifman & El-Yaniv, 2017). Extensions cover domain shift, structured generation, and selective generation (Kamath et al., 2020; Malinin & Gales, 2021; Ren et al., 2023). LLM studies elicit verbal confidence and probe self-knowledge (Tian et al., 2023; Xiong et al., 2024; Zhou et al., 2024a; Geng et al., 2024). We instead rank errors by confidence under controlled input defects, isolating where calibration matters.

Scenario	Modality / Task	Injected input defect	Expected appropriate response
S1 ClaimLens-MME	Multimodal free-form → claim extraction	Image unsupported / inferable / OCR-prone	Do not over-claim (low over-claim rate)
S2 ConflictRAG-Calib	Retrieval three-way decision	Missing / irrelevant / conflicting / stale / adversarial passages	Correct `answer` / `abstain` / `conflict-aware`
S3 TraceRecover-Agent	Tool-call trace diagnosis	Missing-step / out-of-order / wrong-arg / wrong-tool fault	Locate + classify + repair the failing step

Table 1: The unified protocol instantiated across three scenarios: a controlled input defect paired with the response a trustworthy model should produce. The question/task is held fixed while input quality is varied.

Hallucination and abstention. Truthfulness and hallucination evaluations test factuality, grounding, and self-detection (Lin et al., 2022; Manakul et al., 2023; Li et al., 2023b; Guan et al., 2024; Yin et al., 2024); broad multimodal suites emphasize perception and expert reasoning (Fu et al., 2023; Yue et al., 2024). Our protocol uses one three-way *answer / abstain / conflict-aware* decision across vision, retrieval, and action. Free-form outputs use majority-vote LLM judges (Zheng et al., 2023) with inter-judge agreement reported.

RAG under imperfect evidence. RAG augments generation for knowledge-intensive tasks (Lewis et al., 2020), but retrieved passages may be noisy, conflicting, counterfactual, stale, or adversarial (Chen et al., 2024; Longpre et al., 2021; Zhou et al., 2024c; Amirshahi et al., 2025). We recast this as calibration: the target is the appropriate decision and confidence given six controlled evidence defects, including a stale consensus blind spot.

Agent evaluation and introspection. Agent benchmarks measure success in tool-use, web, and interactive environments (Liu et al., 2024; Yao et al., 2024; Qin et al., 2024; Li et al., 2023a; Deng et al., 2023; Zhou et al., 2024b). ReAct and Toolformer study reasoning-action interleaving and tool invocation (Yao et al., 2023; Schick et al., 2023); Reflexion critiques trajectories (Shinn et al., 2023). Our S3 instead programmatically scores fault localisation, kind, and repair-tool choice on the same trustworthiness axis as the non-agentic scenarios.

Agreement as an uncertainty signal. Sample consistency and model self-evaluation provide uncertainty proxies (Kadavath et al., 2022; Wang et al., 2023); semantic entropy clusters meaning-equivalent samples (Farquhar et al., 2024). We show when cross-model disagreement works and expose its consensus blind spot on plausible-but-wrong inputs. Unlike broad scorecards such as TrustLLM (Sun et al., 2024), our defect-conditioned error-ranking axis is identical across vision, retrieval, and action.

3 Unified Trustworthiness Protocol

Capability benchmarks ask *whether a model produces the correct output*. We instead ask whether a model produces the *appropriately uncertain* output when its input or evidence is defective — whether it "knows what it does not know." This section defines the shared protocol that the three scenarios in §4 instantiate; the protocol, not any single scenario, is the contribution that makes a *cross-scenario* study possible.

Defect-conditioned design. Every instance is built around a controlled *input defect* paired with an *expected appropriate response*. We hold the underlying question (or task) fixed and vary only the quality of the input, so that the difference in model behaviour is attributable to the defect rather than to task difficulty. Table 1 maps each scenario onto this template.

Programmatic gold. Because defects are injected by construction, the expected response is known programmatically: a pack type implies an `answer`/`abstain`/`conflict-aware` decision (S2), an image–claim entailment label implies whether a claim is supported (S1), and an injected fault implies the offending step and its repair (S3). No human gold is required for the primary

metric, which keeps the benchmark cheap to scale and free of annotation drift on the "easy" slices. An independent three-judge panel drawn from outside the evaluation roster confirms this gold is reliable where defects are overt (agreement 0.940 for S1 and 0.947 for S2 on the objective categories); on the blind-spot slices the panel is lured into the very same errors the models make, third-party evidence that those slices are genuinely hard (supplementary material).

Multi-judge voting with agreement. Wherever a model output is free-form (a generated paragraph in S1, a rationale in S2), we extract atomic decisions with an ensemble of three frontier LLM judges and take a majority vote, reporting Fleiss' κ as a reliability check (Zheng et al., 2023). To avoid self-evaluation bias, the judge ensemble is drawn from model families *disjoint* from the systems under test, so no model grades its own outputs. High κ on easy slices and low κ on hard slices is itself a signal: it certifies that a model-performance gap reflects intrinsic task difficulty rather than label noise (see the per-pack κ analysis in §4).

Stratified failure audit. For each scenario we draw a stratified sample of wrong predictions (balanced across defect types and models) and characterise recurring failure patterns qualitatively. This audit is what turns a leaderboard into a diagnosis.

Trustworthiness metrics. Beyond raw accuracy we report, where applicable, calibration error (ECE), area under the risk–coverage curve (AURC), abstention F1, and — central to §5 — the discriminative power (area under the ROC curve, AUROC) of three candidate error signals: (i) self-reported confidence, (ii) cross-model disagreement, and (iii) single-model self-consistency under resampling. A trustworthy model is one whose uncertainty signal actually *ranks* its own errors; §5 shows that self-reported confidence systematically fails this test across all three scenarios.

Models under test. All three scenarios evaluate the *same* six frontier systems: closed (GPT-5.5, Claude-Opus-4.8, Gemini-3.5-Flash, Doubao-Seed-2.0-Pro) and open-weight (Qwen3.7-Plus, GLM-5V-Turbo). Because every scenario probes an identical model roster on shared instances, our cross-scenario claims concern the *same* models rather than incomparable populations, and the per-instance disagreement and self-confidence signals in Section 5 are directly commensurable across the vision, retrieval, and action settings.

4 Three Scenarios

We instantiate the protocol of §3 in three scenarios chosen to span distinct modalities and failure surfaces: free-form multimodal generation (S1), retrieval-grounded decision making (S2), and agentic tool-use diagnosis (S3) (Li et al., 2023b; Chen et al., 2024; Liu et al., 2024). Each subsection states the scenario-specific defect taxonomy, reports per-model results, and feeds its trustworthiness signals into the cross-scenario analysis of §5.

4.1 S1: Visual Claim Consistency (ClaimLens-MME)

Defect and gold. A model is prompted to describe an image in free form; we decompose the response into atomic visual claims and ask three judges whether each claim is *supported*, *contradicted*, or *insufficiently grounded* by the image. The appropriate behaviour under an under-determined image is to refrain from asserting unsupported claims. The defect is therefore implicit in the image content (objects absent, only inferable, or OCR-prone text).

Metrics. Claim-support rate (CSR), its complement the unsupported-claim rate (UCR), contradicted-claim rate (CCR), insufficient-support rate (ISR), and an over-claim index (OCI, mean unsupported claims per response). Lower UCR/CCR/ISR/OCI indicate a more trustworthy describer.

Findings. Table 2 reports six models over 500 images (three prompts/model; three-judge majority vote on an $\sim$ 9k stratified claim sample). CSR spans 0.869–0.933: even the strongest describer leaves roughly one visual claim in fifteen unsupported, and the over-claim index doubles across the roster (0.09–0.18). Critically, the dominant error is *insufficient grounding* (ISR up to 0.104)

Table 2: ClaimLens-MME results (500 images, three prompts/model, six models, three-judge majority vote over an ∼9k stratified claim sample; per (image×model) 3-claim stratum). CSR = claim-support rate; UCR = unsupported-claim rate (1−CSR); CCR = contradicted-claim rate; ISR = insufficient-support rate; OCI = over-claim index (mean unsupported visual claims per response). Judge agreement: pairwise raw agreement 0.74; Fleiss κ=0.21 (compressed by the high prevalence of *supported*, $P_e \approx 0.69$); three-judge unanimity 0.62.

Model	#Claims	CSR	UCR ↓	CCR ↓	ISR ↓	OCI ↓
Claude-Opus-4.8	1500	0.933	0.067	0.030	0.036	0.09
GPT-5.5	1500	0.925	0.075	0.031	0.044	0.10
Gemini-3.5-Flash	1500	0.922	0.078	0.024	0.054	0.11
GLM-5V-Turbo	1497	0.891	0.109	0.033	0.076	0.15
Qwen3.7-Plus	1500	0.881	0.119	0.038	0.081	0.17
Doubao-Seed-2.0-Pro	1500	0.869	0.131	0.027	0.104	0.18

Table 3: ClaimLens-MME *discriminative* probe: 1,582 balanced image–claim pairs (791 supported / 791 unsupported, majority gold over three judges), each posed to six models as a yes/no support decision with a self-reported confidence. We report per-model accuracy and the error-ranking AUROC of two uncertainty signals (positive label = wrong decision): **Conf** = self-reported confidence ($1 - c$); **X-Model** = cross-model disagreement (fraction of the six models that dissent from the majority on that pair). Both signals are *above* chance here, and cross-model disagreement is at least as strong as introspection on every model — the same ordering that holds in S2 and S3.

Model	Acc	Conf	X-Model
GPT-5.5	0.711	0.741	0.664
Claude-Opus-4.8	0.735	0.674	0.596
Gemini-3.5-Flash	0.682	0.692	0.675
Doubao-Seed-2.0-Pro	0.674	0.701	0.675
GLM-5V-Turbo	0.645	0.679	0.746
Qwen3.7-Plus	0.624	0.670	0.749
Overall	0.679	0.645	0.684

rather than outright contradiction (CCR ≤ 0.038): models over-assert under-determined content far more often than they state outright falsehoods, driven by the visually demanding claim types (count, OCR, inferential) — the visual analogue of the over-answer behaviour in S2. Reported CSR is a stratified-sample estimate; judge reliability is discussed in Table 2.

Discriminative probe. To make S1 commensurable with S2 and S3 we additionally run a balanced yes/no probe: 1,582 image–claim pairs (791 supported / 791 unsupported), each answered by all six models with a self-reported confidence (Table 3). Accuracy spans 0.624–0.735 and, crucially, both uncertainty signals carry information here: self-confidence reaches AUROC 0.645 pooled and cross-model disagreement 0.684, with disagreement at least matching introspection on every model (and clearly stronger on the weaker open-weight models, X-Model 0.75 for Qwen3.7-Plus and GLM-5V-Turbo). Unlike the retrieval setting, visual support decisions do *not* exhibit inverse calibration — the failure axis is present but milder, which we return to in the cross-scenario analysis.

4.2 S2: Retrieval Conflict Calibration (ConflictRAG-Calib)

Pack types. Each instance is a (question, passages, gold) triple in one of six pack types: **sufficient** (one gold passage + distractors), **missing** (no answer present), **irrelevant** (off-topic passages), **conflicting** (gold + a contradicting passage), **stale** (gold + an outdated passage), and **adversarial** (near-duplicate distractor, no gold). Passage order is shuffled before pids are assigned, so the gold pid is not fixed. For *sufficient*/*conflicting*/*stale* the gold decision is `answer`; otherwise `abstain`. Scale: 500 base QAs × 6 pack types = 3,000 instances × 6 models = 18,000 rollouts.

Label reliability. Three frontier LLMs re-annotate a stratified $n = 150$ sub-sample; overall Fleiss' $\kappa = 0.76$ (substantial). Easy slices reach near-perfect agreement while *stale*/*adversarial* show κ near zero — the raters themselves disagree, confirming that the model-performance gap on those slices reflects intrinsic difficulty, not annotation noise.

Per-pack decision accuracy (Figure 3) confirms the cliff between overt and plausible defects; the full per-pack breakdown is tabulated in the appendix.

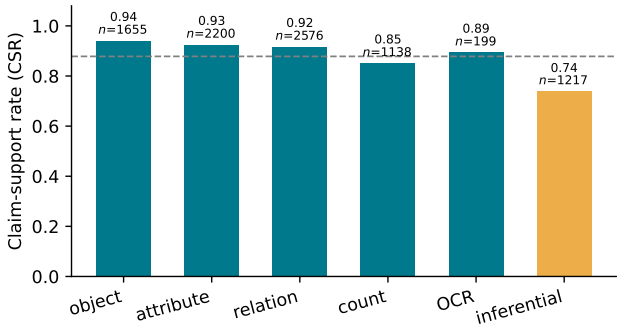


Figure 2: ClaimLens-MME claim-support rate (CSR) decomposed by claim type, pooled over six models. Visually demanding types — *count*, *OCR*, and *inferential* — fall well below the dashed mean, while *object*/*attribute*/*relation* claims are comparatively well supported: over-assertion concentrates exactly where grounding requires reading, counting, or inference. Bar labels give per-type sample sizes n.

Table 4: Main results on ConflictRAG-Calib ($n = 3000$, six pack types, six models). DecAcc = decision (answer/abstain/conflict-aware) accuracy; AbsF1 = abstain F1; CAR = conflict-aware recall; OvA = over-answer rate; ECE = expected calibration error; AURC = area under risk–coverage curve. ↓ = lower is better.

Model	DecAcc	AbsF1	CAR	OvA ↓	ECE ↓	AURC ↓
Claude-Opus-4.8	0.745	0.886	0.532	0.276	0.160	0.203
Qwen3.7-Plus	0.703	0.825	0.555	0.341	0.274	0.281
Gemini-3.5-Flash	0.698	0.841	0.504	0.361	0.558	0.433
Doubao-Seed-2.0-Pro	0.676	0.856	0.405	0.386	0.559	0.408
GPT-5.5	0.661	0.853	0.367	0.405	0.278	0.255
GLM-5V-Turbo	0.654	0.843	0.371	0.415	0.364	0.342

Headline. Table 4 ($n = 3{,}000$, six models) shows **Claude-Opus-4.8** leads on decision accuracy (0.745) *and* is the best calibrated: lowest ECE (0.160), lowest AURC (0.203). Open-weight Qwen3.7-Plus is a close second on accuracy (0.703) and posts the highest conflict-aware recall (0.555). Gemini-3.5-Flash and Doubao-Seed-2.0-Pro pair competitive accuracy with the worst calibration (ECE > 0.55). Calibration quality and raw accuracy are thus only weakly coupled (supplementary material).

Per-pack breakdown. The per-pack numbers (appendix) localise the gap: all six models score ≥ 0.99 on the three trivial slices (sufficient, missing, irrelevant) and stay above 0.62 on conflicting, but collapse on *stale* (0.112–0.294) and *adversarial* (0.112–0.284). No model clears 0.30 on either plausible-defect slice; Qwen3.7-Plus tops stale while Claude-Opus-4.8 tops adversarial. Different families fail on different axes — invisible to pooled metrics.

Failure audit. A stratified 278-sample audit of the stale/conflicting slices surfaces three patterns: *recency blindness* (preferring the longer, confident older passage over the gold recency cue), *conflict commitment* (siding with the contradicting passage even when gold is authoritative), and a universal *single-citation* behaviour (each answer cites one passage; wrong-source citation is confined to the stale/conflicting slices).

4.3 S3: Agentic Trace Recovery (TraceRecover-Agent)

Tasks. A *supervisor* reads a failed multi-step tool-use trace and must (a) locate the offending step, (b) classify the fault kind, and (c) propose the correct repair tool. Four fault types are injected: `missing_step`, `out_of_order`, `wrong_arg`, `wrong_tool` ($n = 106$/model). A paired end-to-end split has the agent solve 300 multi-step tasks in a 5-tool sandbox (ReAct, ≤ 12 turns), each judged by an execution-grounded validator, and each agent additionally self-reports a completion confidence.

Supervisor (read-only) diagnosis results, per-fault breakdowns, and the Gemini format-compliance fix are reported on an earlier five-model roster in the supplementary material; diagnosis is largely solved and clusters tightly across models there, which sets up the diagnosis-to-action gap below.

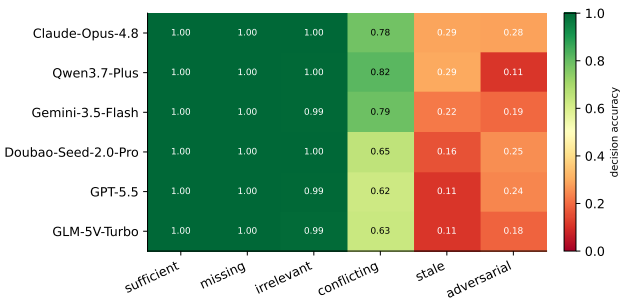


Figure 3: Per-pack decision accuracy on S2 (rows = models sorted by overall accuracy, columns = evidence-defect packs). Every model is near-perfect on overt defects (*missing*, *irrelevant*) yet collapses on the plausible-but-wrong *stale* and *adversarial* packs.

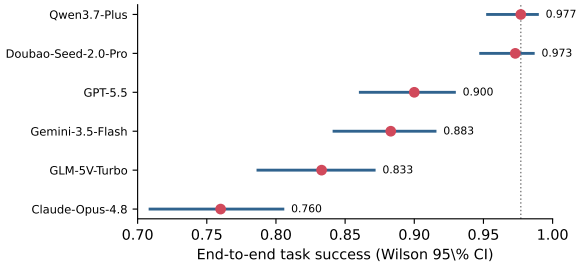


Figure 4: TraceRecover base-agent end-to-end task success (n=300 multi-step tasks, all six models), with Wilson 95% CIs. A residual capability gradient ($0.977 \rightarrow 0.760$) emerges under acting that read-only diagnosis — near-saturated and tightly clustered across models (appendix) — does not predict; the dotted line marks the top score.

Diagnosis-to-action gap. Figure 4 contrasts read-only diagnosis (supplementary material) with acting on an end-to-end probe (300 tasks). End-to-end success ranges from 0.977 (Qwen/Doubao) to 0.760 (Claude-Opus-4.8), a spread the tightly-clustered read-only diagnosis does not anticipate; the ranking is also *reordered* — the worst actor is not the worst diagnoser. Acting reveals a difficulty axis diagnosis does not.

Dual-signal calibration. We apply the shared calibration protocol of S1/S2 to the 1,800 agentic episodes (Table 5): for each task we rank episodes by two uncertainty signals and measure how well each ranks the failures (positive label = failed task). Cross-model disagreement is the strongest error detector we observe in *any* scenario (overall AUROC 0.853), while self-reported confidence is markedly weaker: it collapses to exactly chance (0.500) for three of the six agents ($\bar{c}$=0.99 even on tasks they fail), and only GPT-5.5 and Claude-Opus-4.8 extract any self-signal (overall Conf 0.764). The introspection-below-consensus ordering of S1 and S2 therefore not only persists but *widens* under sequential tool use, completing the shared failure axis across all three scenarios.

5 Cross-Scenario Findings

Studying three scenarios under one protocol lets us test whether trustworthiness failures are scenario-specific quirks or shared properties of frontier models. Three findings hold across scenarios and are the payoff of the merged design.

Finding 1: self-reported confidence is inversely calibrated. Figure 5 (and its per-slice breakdown in the supplementary material) measures how well each uncertainty signal ranks a model's own errors. Self-reported confidence (Conf) sits *below* the random baseline of 0.5 on every contested slice — 0.200 (conflicting), 0.079 (stale), 0.294 (adversarial), and 0.419 overall — meaning models are, if anything, *more* confident when wrong (the near-1.0 value on *sufficient* is a degenerate artefact of $\sim$ 3 errors in $\sim$ 3,000 rollouts). On S1 the discriminative calibration probe (Table 3) tells a milder but consistent story: pooled self-confidence reaches an error-ranking AUROC of only 0.645 (no model exceeds 0.741) — informative but far from reliable, and always weaker than the

Table 5: TraceRecover-Agent *dual-signal* calibration on 300 multi-step file-operation tasks, each attempted by six agents (1,800 episodes). Task success is decided by an execution-grounded validator. We report per-model success accuracy and the error-ranking AUROC of the same two uncertainty signals used in S1/S2 (positive label = failed task): **Conf** = the agent's self-reported completion confidence ($1 - c$); **X-Model** = cross-model disagreement (fraction of the six agents that dissent from the majority outcome on that task). Cross-model disagreement is by far the strongest signal in the agentic setting (overall 0.853), whereas self-confidence collapses to chance for three of six models — these agents report near-maximal confidence ($\bar{c}$=0.99) even on tasks they fail. This reproduces the S1/S2 ordering (Table 3) and completes the shared failure axis across all three scenarios.

Model	Acc	Conf	X-Model
GPT-5.5	0.900	0.775	0.958
Gemini-3.5-Flash	0.883	0.500	0.964
Qwen3.7-Plus	0.977	0.500	0.844
Doubao-Seed-2.0-Pro	0.973	0.500	0.864
GLM-5V-Turbo	0.833	0.498	0.790
Claude-Opus-4.8	0.760	0.842	0.854
Overall	0.888	0.764	0.853

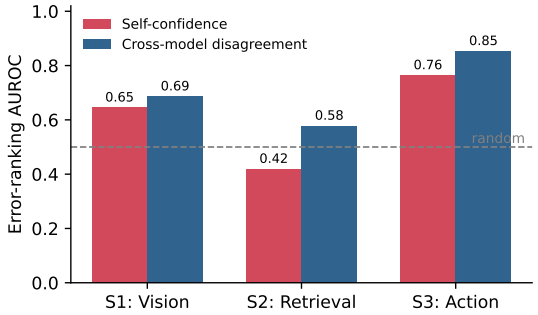


Figure 5: Error-ranking AUROC of two uncertainty signals (self-confidence, red; cross-model disagreement, blue), six models pooled, across S1 vision, S2 retrieval, and S3 agentic action. Cross-model disagreement matches or beats self-confidence in every scenario (strongest in S3, 0.853), while self-confidence stays at or below the random 0.5 line on contested slices and collapses to chance for half the S3 agents.

cross-model signal discussed next. *Asking a model how sure it is does not, on its own, surface its errors.*

Finding 2: a consensus blind spot. Cross-model disagreement (X-Model) is a strong error signal on *conflicting* (0.681) and *irrelevant* (0.783) slices, but it too collapses on *stale* (0.310) and *adversarial* (0.314) — below chance. On these slices the models do not merely fail individually; they fail *together*, agreeing on the same wrong decision, so disagreement carries no signal. This consensus blind spot reappears across scenarios: S2's stale/adversarial packs, S3's `missing_step` fault (the hardest category for every supervisor, per the supplementary material), and S1's insufficient-grounding claims are all cases where the defect is *plausible* rather than overt, and frontier models share the same inductive bias toward the plausible-but-wrong reading.

Finding 3: agreement-based signals beat introspection — until consensus fails. Per-slice (supplementary material), "ask several models whether they agree" dominates "ask one model how sure it is" on every slice where any signal works. We also tested whether single-model self-consistency (Wang et al., 2023) (resampling at $T = 0.7$) recovers signal inside the consensus blind spot; it does not (sample-agreement AUROC indistinguishable from random on stale and adversarial; see supplementary material), confirming that the blind spot is a stable property of the models rather than sampling noise. The practical implication is a hierarchy of error detectors: *cross-model disagreement* where models can disagree, but *neither* introspective confidence *nor* resampling rescues the consensus blind spot — which is therefore the open frontier this benchmark isolates.

Finding 4: a first partial solution — evidence critique closes the adversarial blind spot but only dents the stale one. A pure benchmark that ends at "all models fail" invites the question of whether

the failure is at all remediable. We therefore test one lightweight, model-agnostic intervention on the S2 blind-spot slices (500 queries $\times$ 6 models per slice): before answering, the model is instructed to *critique each passage* for temporal staleness and corroboration, emit a per-passage reliability and a scalar `min_reliability`, and is forbidden from answering off a single uncorroborated passage. This evidence-critique prompt *eliminates* the adversarial blind spot: the misled answer rate drops from 0.721 to 0.000 (paired bootstrap $\Delta = -0.721$, 95% CI $[-0.737, -0.704]$) and decision error from 0.800 to 0.003, while the new `min_reliability` signal becomes a usable failure detector (AUROC 0.553 vs. the baseline confidence's 0.294). On stale evidence it still helps substantially — the misled answer rate falls from 0.791 to 0.255 ($\Delta = -0.536$, CI $[-0.556, -0.517]$) — but the decision-match error is unchanged ($0.804 \rightarrow 0.806$, CI $[-0.011, +0.015]$, n.s.): the model learns *not to be misled* into answering, yet still cannot tell a "stale-but-in-conflict" pack (flag conflict-aware) from one warranting abstention. The blind spot is thus partly, not fully, closable by prompting: adversarial staleness is a surface cue models can distrust, whereas quiet temporal drift remains open — the residual we expect external, freshness-grounded signals (not introspection) to address.

A shared failure axis. Across modality (vision), evidence (retrieval), and action (agentic tool use), the same axis separates trustworthy from untrustworthy behaviour: models are reliable when the defect is *overt* (missing/irrelevant evidence, observable wrong tool, absent object) and collapse — confidently and in agreement — when the defect is *plausible* (stale evidence, an inferable missing step, an under-determined image). The three commensurable probes make the axis quantitative: cross-model disagreement matches or exceeds self-confidence as an error detector in all three scenarios — S1 (AUROC 0.686 vs. 0.645), S2 (0.578 vs. 0.419), and S3 (0.853 vs. 0.764 over $1{,}800$ agentic episodes; all three paired differences exclude zero), with both signals degrading together on the plausible-defect slices. The gap is widest under sequential tool use: acting amplifies disagreement into our strongest detector (0.853) while half the agents' self-confidence collapses to chance, staying near-maximal even on failures. No single-scenario benchmark exposes this regularity; the unified protocol makes it visible.

6 Discussion

Why a unified protocol matters. In isolation each scenario yields a local story — "this VLM over-claims," "this LLM trusts stale evidence," "this agent mis-diagnoses missing steps." Only under a fixed protocol do these resolve into one phenomenon: frontier models share an inductive bias toward the *plausible* reading of a defective input, correlated across models, which is why both introspection and cross-model agreement fail in the same place.

Implications for deployment. The error-detector hierarchy is actionable: disagreement is a cheap abstention trigger where inputs can be routed to several models — except on the plausible-but-wrong slices, where a wrong answer is costliest and self-reported confidence must never gate safety. Closing the blind spot needs external signals (Zhou et al., 2024c): an evidence-critique prompt eliminates the adversarial blind spot and halves the stale one (Finding 4), yet quiet temporal drift persists, pointing to evidence-freshness grounding as the remaining lever.

Limitations. S1 CSR is a stratified-sample estimate; the modest Fleiss κ=0.21 is a high-prevalence artefact, not unstable judging; and injected defects sample, rather than exhaust, the defect space (details in the supplementary material). The protocol nonetheless extends easily, needing only a defect generator, a programmatic checker, and the shared metric family.

Conclusion. Under one defect-conditioned protocol spanning vision, retrieval, and agentic tasks, confidence is inversely calibrated everywhere and a consensus blind spot makes frontier models fail together on plausible-but-wrong inputs — they do not yet reliably know what they don't know. We release the protocol, data, and judge harness as a first-class trustworthiness target.

Reproducibility Statement

All numbers are recomputed from released logs by aggregation scripts. Each scenario stores its per-instance records to disk—S1 (vision) defect-injected image–question pairs, S2 (retrieval) clean/stale/adversarial evidence slices, and S3 (agentic) fault-injected traces—together with model outputs, judge labels, and the calibration read-out contrasting self-reported confidence with cross-model disagreement. The unified protocol, roster of six frontier models, defect taxonomies, and scoring definitions are specified in the main text; the scripts regenerate every reported table, bootstrap confidence interval, and statistical test from the stored records. Code, data, and run scripts sufficient to reproduce all results are provided as supplementary material.

Ethics Statement

This work is a diagnostic benchmark that measures whether frontier models know what they do not know across vision, retrieval, and agentic tasks. It introduces no new human-subjects data and no personally identifiable information; all underlying instances are either synthetically constructed or drawn from publicly available datasets under their respective licenses. The measurements characterize model calibration and failure detection rather than enhancing capability, and we report negative and blind-spot results honestly (e.g. the consensus blind spot on adversarial retrieval slices) to avoid overclaiming.

Use of Large Language Models

Large language models were used only as a general-purpose writing and coding assistant: to polish prose, refactor plotting and aggregation scripts, and check LaTeX formatting. The benchmark design, defect taxonomies, metric definitions, analysis, and scientific claims are the authors' own. Every number, table, and statistical test was produced by the released code on the released data and independently verified by the authors; no experimental result or citation was generated by a language model without such verification.

References

Shakiba Amirshahi, Amin Bigdeli, Charles L. A. Clarke, and Amira Ghenai. Evaluating the robustness of retrieval-augmented generation to adversarial evidence in the health domain. *arXiv preprint arXiv:2509.03787*, 2025.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. *arXiv preprint arXiv:2006.09462*, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. API-Bank: A comprehensive benchmark for tool-augmented LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023b.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations (ICLR)*, 2024.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2021.

Potsawee Manakul, Adian Liusie, and Mark J.F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. ToolLLM: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations (ICLR)*, 2024.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. Out-of-distribution detection and selective generation for conditional language models. *arXiv preprint arXiv:2209.15558*, 2023.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. 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 (ACL)*, 2024a.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations (ICLR)*, 2024b.

Yujia Zhou, Wenbo Zhang, Jingying Shao, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, and Zheng Liu. Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102*, 2024c.

Table 6: TraceRecover supervisor evaluation (n=106 per model, single-fault diagnosis). LocAcc = step localisation; KindAcc = fault-kind classification; ToolAcc = correct repair tool; ParseFail = JSON parse / format errors (counted as wrong on all three metrics).

Model	LocAcc	KindAcc	ToolAcc	ParseFail ↓
Claude-Opus-4.7	0.991	0.972	0.943	0.000
Claude-S4.5	0.972	0.953	0.906	0.000
Gemini-2.5-Pro	0.934	0.840	0.764	0.000
GPT-4o	0.934	0.896	0.868	0.000
GPT-5.1	0.915	0.915	0.868	0.000

Table 7: LocAcc by fault type across all five supervisors (n per cell shown in parentheses; full LocAcc/KindAcc/ToolAcc reported in Table 6). `missing_step` remains the hardest category for every model because the supervisor must *infer* an absent step from the task description rather than diagnose an observable one.

Model	`missing_step` (29)	`out_of_order` (29)	`wrong_arg` (29)	`wrong_tool` (19)
Claude-Opus-4.7	0.966	1.000	1.000	1.000
Claude-S4.5	0.931	1.000	0.966	1.000
Gemini-2.5-Pro	0.966	1.000	0.862	0.895
GPT-4o	0.931	0.931	0.966	0.895
GPT-5.1	0.897	0.966	0.931	0.842

A Supplementary Analyses on Earlier Model Rosters

The main paper reports every scenario on a single unified roster of six current frontier models. Two auxiliary analyses below were run on an earlier five-model roster and are reported here for completeness; their qualitative conclusions are unchanged by the roster and are cited from the main text only as supporting evidence, not as headline numbers.

A.1 S3 Supervisor Diagnosis (Earlier Roster)

The read-only *supervisor* task — locate the offending step, classify the fault kind, and propose the repair tool — was evaluated on five models (Claude-Opus-4.7, GPT-5.1, GPT-4o, Claude-Sonnet-4.5, Gemini-2.5-Pro), $n = 106$ traces/model, four injected fault types (`missing_step`, `out_of_order`, `wrong_arg`, `wrong_tool`).

Diagnosis is largely solved by the top models (Table 6; Claude-Opus-4.7 LocAcc 0.991), with the rest clustered at 0.92–0.97. Per-fault (Table 7), `missing_step` is hardest for every model: the supervisor must *infer* an absent step rather than diagnose an observable one — the agentic analogue of S2's adversarial blind spot. Because read-only diagnosis clusters tightly across models here, it fails to predict the end-to-end agent-capability gradient reported in the main text.

Gemini Format-Compliance Fix

Before patching, Gemini-2.5-Pro emitted free-form prose and occasionally a JSON *array*, yielding a 35.8% parse-fail rate (scored as wrong). Three minimal changes — a one-shot input/output exemplar, raising `max_tokens` to 4096, and accepting top-level arrays in the JSON stripper — bring parse-fail to zero across all four fault types (Table 8), lifting overall LocAcc to 0.934. Format compliance must be measured *jointly* with reasoning.

A.2 S2 Single-Model Self-Consistency (Earlier Roster)

To test whether single-model self-consistency (resampling at $T = 0.7$, five samples) recovers signal inside the consensus blind spot, we resampled five models (GPT-4o, GPT-5.1, Claude-Sonnet-4.5, Gemini-2.5-Pro, Qwen3-VL-235B) on the *stale* and *adversarial* slices. It does not: sample-agreement AUROC is 0.465/0.448 on stale/adversarial, indistinguishable from random, confirming that the blind spot is a stable property of the models rather than sampling noise.

Table 8: Gemini-2.5-Pro LocAcc by fault type, after the format/thinking-budget fix (§A.1). The pre-fix run scored **0.585 / 0.604 / 0.557 / 0.358** (LocAcc / KindAcc / ToolAcc / ParseFail) with the residual failures concentrated on `out_of_order`; raising `max_tokens` to 4096 and using a single-object format example brought ParseFail to zero across all four fault types.

Fault type	n	LocAcc	ParseFail
`missing_step`	29	0.966	0.000
`out_of_order`	29	1.000	0.000
`wrong_arg`	29	0.862	0.000
`wrong_tool`	19	0.895	0.000

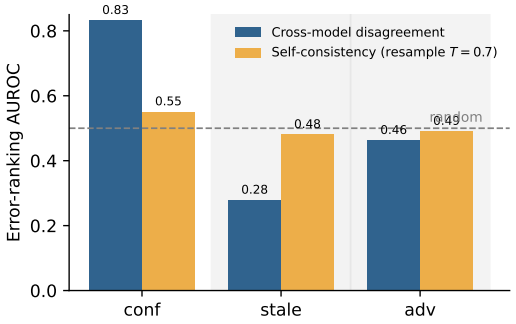


Figure 6: Single-model self-consistency (resampling at T=0.7, five samples) does not rescue the consensus blind spot. On the same five-model roster, cross-model disagreement carries signal on *conflicting* but both cross-model disagreement and self-consistency fall to chance on the shaded *stale* and *adversarial* slices, confirming the blind spot is a stable property of the models rather than sampling noise.

A.3 Gold-Label Reliability: Independent Judge Cross-Validation

Because our gold labels are constructed *programmatically* (S1 from COCO panoptic annotations, S2 from the pack-construction procedure), we audit their reliability with a panel of three judges drawn from *outside* the six-model evaluation roster (to avoid self-grading bias). Each sampled item is relabelled from scratch by all three judges; we take the majority vote and measure agreement with the programmatic gold. We sample 150 S1 probes and 120 S2 packs (seed 20260802).

On objective categories the gold is highly reliable. For the categories with an unambiguous ground truth—S1 `supported` / `contradicted`, S2 `sufficient` / `missing` / `irrelevant` / `conflicting`—independent-judge agreement with the programmatic gold is 0.940 (94/100) for S1 and 0.947 (71/75) for S2. Per-category rates are $\geq$ 0.80 throughout (S1 `supported` 0.975, `contradicted` 0.800; S2 `sufficient`/`missing` 1.000, `irrelevant` 0.962, `conflicting` 0.842), confirming the construction is sound where a defect is overt.

On the blind-spot categories the judges fail too—which validates the difficulty rather than the gold. Agreement collapses exactly on the categories the benchmark is designed to stress: S2 `adversarial` 0.000 and `stale` 0.364, S1 `insufficient` 0.415 and `non_visual` 0.000. The disagreements are not random noise but the *same* error the main findings report: on `adversarial` packs the independent judges are themselves lured away from `abstain` (toward `conflict-aware` 17 times and `answer` 6 times); on `stale` packs they answer from the plausible-but-stale passage (14 times) instead of flagging conflict; and on S1 they upgrade `insufficient` claims to `supported` (20 times), reproducing the hallucination bias. That a fresh, roster-independent panel walks into the same trap is third-party evidence that the blind-spot slices are genuinely hard, and that the programmatic gold—anchored to the construction procedure rather than to a fallible judge—is the *correct* label precisely where model consensus is not to be trusted. Validation records are released in `results/gold_validation_{s1,s2}.jsonl`.

A.4 Per-Slice Error-Ranking AUROC (ConflictRAG-Calib)

Figure 7 gives the per-slice breakdown of the two uncertainty signals referenced in Findings 1 and 3 of the main text.

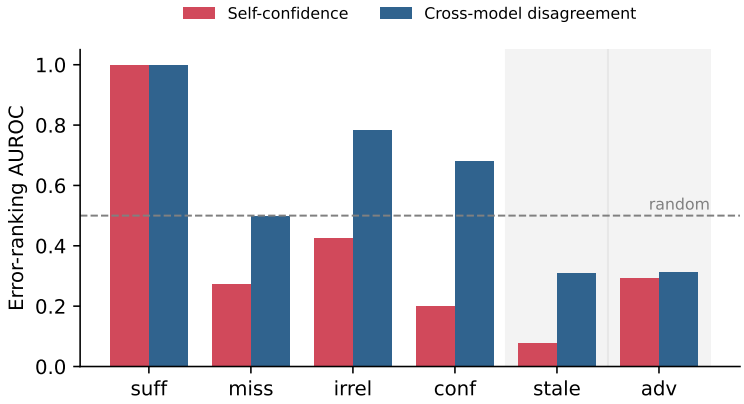


Figure 7: Per-slice error-ranking AUROC on ConflictRAG-Calib (six models pooled; full numbers tabulated above). Self-confidence (red) sits at or below the random 0.5 line on every contested slice — inverse calibration — while cross-model disagreement (blue) is informative on *conflicting/irrelevant* yet *both* signals collapse below chance on the shaded consensus blind-spot slices *stale* and *adversarial*, where models fail together.

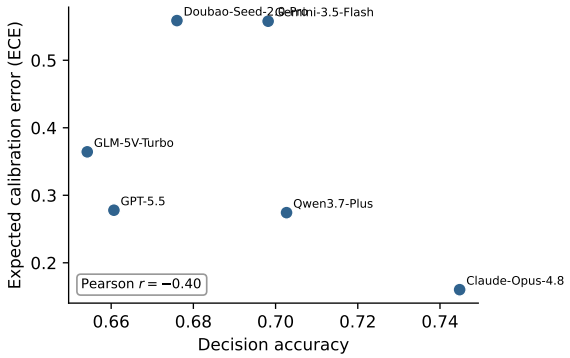


Figure 8: S2 decision accuracy versus expected calibration error (ECE), one point per model. Accuracy and calibration are only weakly coupled (r near zero): the most accurate model is not the best calibrated, so a single accuracy number hides the calibration axis this protocol isolates.

A.5 S2 Accuracy versus Calibration

Figure 8 tests whether stronger task performance also implies better calibrated decisions in S2. The model-level relationship is weak: decision accuracy and ECE vary largely independently, and the most accurate model is not the best calibrated. This separation supports treating calibration as an evaluation axis in its own right rather than using aggregate accuracy as a proxy for deployment reliability.

A.6 Extended Limitations

S1 CSR estimation. The S1 claim-support rate is a stratified-sample estimate ($\sim$ 9k claims, per image$\times$model); a full-census judge over every claim is left to future work.

Judge agreement. The three judges show high raw agreement (pairwise 0.74) but a modest Fleiss κ=0.21. This is an artefact of the high prevalence of *supported* claims ($P_e$$\approx$$0.69$) inflating chance agreement, not unstable judging; residual shared judge-model biases are mitigated, not eliminated, by using disjoint-family judges.

Defect coverage. The injected defects, though controlled and programmatically checkable, sample the possible defect space rather than exhaust it; the protocol extends easily, as new scenarios need only a defect generator, a programmatic checker, and the shared metric family.