SEEING IS BELIEVING: EMBODIED AGENTS THAT VERIFY CLAIMS FROM UNRELIABLE INFORMANTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied agents must decide which information to trust when spoken and written reports conflict with each other or with what they observe. Visual verification can resolve these conflicts, but acquiring evidence costs time and movement, making past experience with each source valuable. We introduce SEEINGISBELIEVING, an embodied benchmark that brings audio, text, and vision together to evaluate trust and verification under unreliable information. Our contributions are threefold. First, we construct dynamic household scenarios in which recurring informants provide audio and text claims with hidden, context-dependent reliability, while agents can navigate to gather visual evidence at a cost. This setting connects deciding whether a claim is true with learning whether its source is reliable. Second, we develop a six-level evaluation covering perception, claim verification, conflict resolution, verification efficiency, memory, and decisions on unverified claims. Agents retain their own verification histories across episodes without parameter updates, and controlled changes to perception, modality, source labels, and identity help distinguish why trust decisions succeed or fail. Third, we analyze six open omni-modal models against fixed baselines and scripted Bayesian reference policies. The reported results reveal preferences for sources labeled as spoken even when their claims are presented as text, and show that stronger perception and higher memory scores do not consistently translate into better decisions on unverified claims. Gains over trusting every claim remain sensitive to evaluation conditions. SEEINGISBELIEVING provides a unified testbed for studying whether agents can turn multimodal evidence and accumulated verification experience into reliable trust decisions.

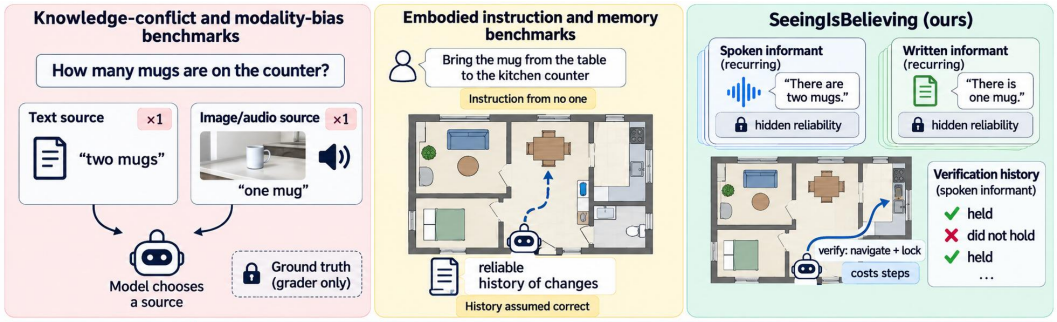
1 INTRODUCTION

Can a multimodal agent decide whom to believe when the people around it are sometimes wrong? An embodied agent receives many claims about its surroundings, and checking a claim by going to look costs time. The same people also speak again, so an agent that remembers who was right can act better later. Current benchmarks do not measure this everyday form of trust.

Knowledge-conflict and modality-bias benchmarks score which of two disagreeing sources a model follows (Xu et al., 2024; Su et al., 2024; Jia et al., 2025; Lyu et al., 2026; Deng et al., 2025; Wang et al., 2025). In these benchmarks nothing can be checked, and no source appears again. Embodied work on false-premise instructions lets an agent explore to find out that an instruction was wrong (Su et al., 2026; Yeke et al., 2026). However, the instructions come from no particular speaker, so there is nobody to learn about. Embodied memory benchmarks give the model the history of scene changes as reliable text, so the model never has to doubt that history (Liao et al., 2026). Figure 1 contrasts the three settings: one-shot sources that cannot be checked, exploration without an identified source, and in SEEINGISBELIEVING persistent informants, checks that cost steps, and carried outcomes of past checks.

Each episode is a simulated house in which objects are moved over a sequence of actions. After each action, a spoken informant and a written informant each make a claim, and we call this moment a decision point. A claim can be true, false, or true but about an unrelated part of the house. Each informant has a hidden reliability, which by construction is unrelated to the gender, age, language and speaking style of the informant. The prompt labels each informant as SPOKEN or WRITTEN.

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.



Source identity	none	none	persistent
Can the agent check?	no	explore, no source credit	yes, by looking, step cost
Memory across episodes	no	no	verification history of own checks
What is scored	which source is followed	task success	actions at six levels

Figure 1: **SEEINGISBELIEVING compared with two existing benchmark families.** Knowledge-conflict benchmarks present two one-shot sources and score which one the model follows; embodied instruction benchmarks let an agent explore, but the instruction comes from no one. In SEEINGISBELIEVING two informants of hidden reliability recur across episodes, a check costs steps, and the outcomes of past checks are carried forward.

The agent can check any claim by walking to the named location and looking with its own camera, at a cost in steps. The released corpus has 104 episodes and 3,092 claims from 48 informants, and twelve test informants recur across the evaluation episodes.

We evaluate agents at six levels on the same episodes, each adding one capability with one primary metric. L0 tests perception of the scene, and L1 tests judging a single claim against the scene. L2 asks which of two conflicting claims is right, and L3 asks whether a check is worth its cost. L4 tests whether trust in an informant improves with repeated encounters. L5 tests actions on unverified claims and the calibration of stated confidence. At L4 and L5 the agent sees the evaluation episodes as one fixed sequence. The model receives no parameter update and never sees the ground truth, so the only information carried between episodes is the verification history, the outcomes of past checks on each informant. Any improvement over the sequence must come from this history, and shuffling informant identities tests whether the improvement depends on who is speaking.

We evaluate six open omni-modal models without fine-tuning, against fixed baselines and scripted Bayesian reference policies, and under five control conditions: spoken claims shown as text, swapped labels, shuffled identities, an empty verification history, and a warning prompt. The levels rank the models differently, and neither family nor size explains the pattern. Qwen2.5-Omni-7B, InteractiveOmni-8B and Qwen3-Omni-Instruct favor the informant labelled SPOKEN, even when its claim is shown as text. Qwen2.5-Omni-7B, Qwen3-Omni-Instruct and Qwen3-Omni-Thinking gain similarly from looking, and none of these gains survives correction for multiple comparisons. Qwen2.5-Omni-7B has the highest Memory Score, which does not survive correction either, and shuffling informant identities lowers the Memory Score of Qwen2.5-Omni-7B and Nemotron. Qwen3-Omni-Instruct shows no memory, yet it is the only model that acts above the trust-all baseline, and its margin is not robust. With the verification history removed no model beats the trust-all baseline (see §5).

Contributions.

- A benchmark in which recurring spoken and written informants of hidden reliability make claims that the agent can verify at a cost in steps.
- An evaluation at six levels on one corpus, each with one primary metric, run without parameter updates so that the verification history is the only information carried between episodes.
- A study of six omni-modal models against baselines, scripted Bayesian policies and five control conditions, with a decomposition of what separates the best model from a Bayesian reference policy.

2 Related Work

Trust and source reliability. Knowledge-conflict studies measure how models resolve anonymous sources that disagree (Xu et al., 2024; Su et al., 2024; Jia et al., 2025; Lyu et al., 2026), sycophancy and persuasion studies measure how far pressure sways a model (Fanous et al., 2025; Triedman & Shmatikov, 2025), source-trust work models who is speaking (Pradhan & Goley, 2026; Zhou et al., 2026), related work asks when a model should rely on a tool rather than on itself (Unlu, 2026) or whether it deceives when incentives conflict (Liu et al., 2026), and embodied agents detect false premises by exploring, abstain on questionable commands, or weigh the cost of gathering information (Su et al., 2026; Yeke et al., 2026; Chen et al., 2023; Ding et al., 2026). In SeeingIsBelieving, a recurring source makes claims that the agent can check at a cost (see Table 1).

Table 1: **What SeeingIsBelieving adds relative to the closest benchmarks.** Columns: persistent informant identity; verification that costs steps; the agent's own vision as arbiter; speech and text; trust carried across episodes; calibration scored; a scene that changes.

Benchmark	Persist. source	Costly verif.	Vision as arbiter	Speech+ text	Cross-ep. trust	Calib. scored	Dynamic scene
SpaMEM (Liao et al., 2026)	✗	✗	✓	✗	✗	✗	✓
ConflictBank (Su et al., 2024)	✗	✗	✗	✗	✗	✗	✗
KC-Bench (Lyu et al., 2026)	✗	✗	✗	✗	✗	✗	✗
MCR-BENCH (Wang et al., 2025)	✗	✗	✗	✓	✗	✗	✗
Words or Vision (Deng et al., 2025)	✗	✗	✓	✗	✗	✗	✗
VLN-NF (Su et al., 2026)	✗	✓	✓	✗	✗	✗	✗
SAVVY, STAR-Bench (Chen et al., 2025; Liu et al., 2025)	✗	✗	✓	✗	✗	✗	✓
SeeingIsBelieving (ours)	✓	✓	✓	✓	✓	✓	✓

Multimodal decisions and confidence. When text disagrees with an image or with audio, multimodal models tend to follow the text (Deng et al., 2025; Wu et al., 2025a; Wang et al., 2025), and contradiction and hallucination benchmarks test whether a model detects the mismatch (Popordanoska et al., 2025; Sung-Bin et al., 2025), and we measure the analogous preference between two identified informants. Verbalised confidence has been studied for language models (Xiong et al., 2024; Yang et al., 2024), the reliance literature argues that trust should be measured by actions (Ibrahim et al., 2025; Raees & Papangelis, 2026), separating perception from reasoning follows spatial planning work (Wu et al., 2025b), and evaluation without parameter updates isolates learning from context from learning in the weights (Chan et al., 2022b;a; 2025; Singh et al., 2023).

Omni-modal models and spatial benchmarks. We evaluate open omni-modal models (Xu et al., 2025a;b; Tong et al., 2025; NVIDIA, 2026), and prior spatial benchmarks test perception of space from sound (Zheng et al., 2024; Chen et al., 2025; Liu et al., 2025; Zeng et al., 2026) or beliefs about a changing scene (Liao et al., 2026; Zheng et al., 2025), of the kind that rearrangement tasks pose (Batra et al., 2020). Here speech carries only testimony, which replaces the reliable text history that models need in SpaMEM.

3 The SeeingIsBelieving Benchmark

3.1 Environment

SeeingIsBelieving is built on the houses, action sequences, rendered frames and world graphs of SpaMEM (Liao et al., 2026), and replaces its reliable text history with testimony from two informants. After every action, a spoken informant and a written informant each make one claim about the house. A claim is true, false, or off-topic, meaning true but about an unrelated part of the scene, and the agent is not told which. False and off-topic claims are errors of the informant, and an off-topic claim is scored as holding. Figure 2 follows one decision point through the six generation stages, from the SpaMEM episode to speech synthesis; the reliability of each informant and the truth of each claim are stored with the claim and never shown to the agent. The released corpus has 104 episodes and 3,092 claims from 48 informants, and false or off-topic claims make up 29% of the training split, 49% of the evaluation split and 34% of the whole corpus (see Table 5 in Appendix A).

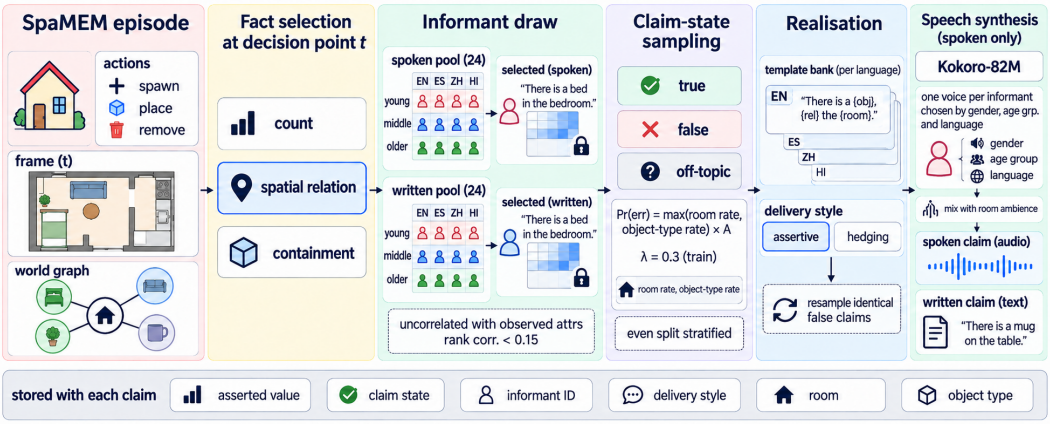


Figure 2: **Corpus generation.** A SpaMEM episode supplies the house, actions, frames and world graph; at each decision point a fact is selected, one spoken and one written informant are drawn, the claim is sampled as true, false or off-topic and rendered from templates, and spoken claims are synthesised as speech (λ is the global error rate, §3.3). Quantities the agent never sees are marked.

3.2 Informants

Informants come from two fixed pools of 24, one spoken and one written, that cross gender, age group, language and delivery style (see Appendix A). Reliability is hidden and combines a base accuracy with error rates by room and object type. A reliability group is a set of room and object types that share one error rate, and each informant has four groups, so an informant may be dependable in the kitchen but unreliable about small handheld objects. We write $R_{p,c}$ for the reliability of informant p in reliability group c. Generation keeps the rank correlation between every observable attribute and every reliability parameter below 0.15 (see Appendix D). Informants are split into disjoint training and test sets. Six of the twelve test informants are consistently wrong, with reliability R_p, the fraction of claims by informant p in the evaluation episodes that are true, in $[0.12, 0.31]$, and with error rates from 0.69 to 0.88 (see Appendix F). The other six are reliable, with $R_p \in [0.82, 0.96]$, so knowing reliability helps either way: trust the reliable and invert the unreliable. Each test informant recurs in at least three evaluation episodes, which makes learning across episodes measurable.

3.3 Claims and Speech

For each informant we deterministically select a fact from the world graph, which is a count, a spatial relation or a containment statement, and sample whether the claim is true, false or off-topic. An informant errs with the larger of its room and object-type error rates, and the types that share this rate form the reliability group of the claim. A global error rate $\lambda = 0.3$ scales these rates and fixes the share of errors in the training split, and each informant has its own mix of error types (see Appendix A). With probability 0.7 the two informants speak about the same fact, which creates the conflicts used at L2, and we resample any decision point at which both would make the same false claim. Claims are rendered from templates for each language and delivery style, and a stored function checks every claim against the world state at any later step. Spoken claims are synthesised with Kokoro-82M in a fixed voice per informant and delivered only as audio, and the prompt labels each informant as SPOKEN or WRITTEN.

3.4 Verification and Cost

At every decision point the agent may act on a claim, accepting the claim as true, decline the claim, answer the scene question directly, or ask, which ends the decision point unresolved with a small penalty. To verify first, the agent walks to the named receptacle, a surface or container such as a counter or a fridge, pays the path cost in steps, and pays three more steps to look and one to open an enclosed receptacle. The budget is sixty steps per episode at L3 and one hundred at L4 and L5 for

Table 2: **The six evaluation levels.** Each level adds one capability on the same corpus and has one primary metric. L2 and L5 also report a secondary diagnostic, the modality bias Δ_{mod} (Eq. 2) and the expected calibration error on act decisions ECE_{act} (§4.1).

Level	Agent receives	Question asked	Metric
L0 Perceive	one frame, no informant	which objects are where?	Grounding Score (GS)
L1 Judge	one frame + one claim	consistent, contradicted, or unverifiable here?	Veracity F1 (VF1)
L2 Adjudicate	one frame + two conflicting claims	which informant is right, if either?	CRA (Δ_{mod})
L3 Verify	claims + navigation budget	is it worth paying to look?	Verification Efficiency (VE)
L4 Remember	a sequence of episodes with recurring informants	does trust in an informant improve with exposure?	Memory Score (MS)
L5 Act	the same sequence	are actions on unverified claims good, and is confidence calibrated?	TCS-NV (ECE_{act})

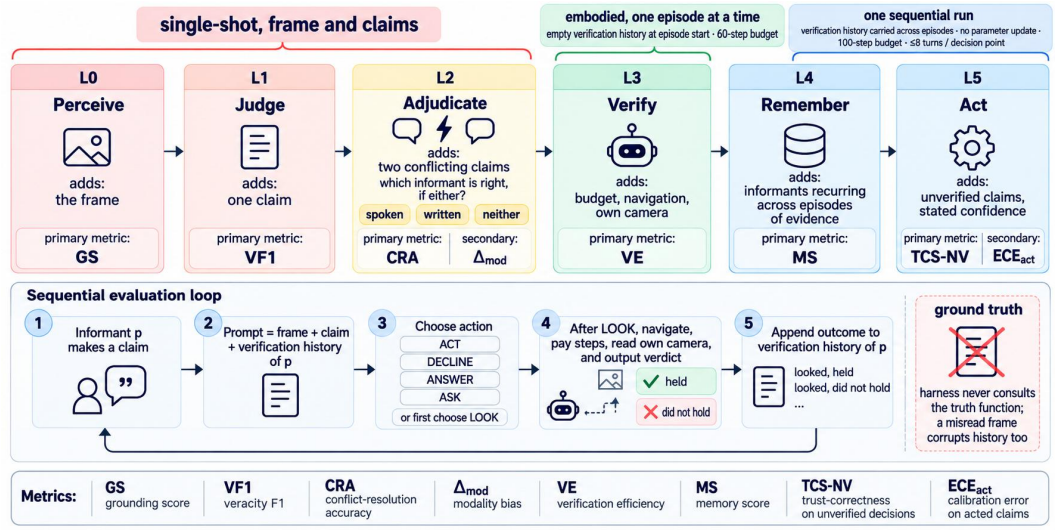


Figure 3: **The six evaluation levels and the sequential evaluation.** Top: six measurements on one corpus, from perceiving the scene at L0 to acting on unverified claims at L5, each with one primary metric. Bottom: one decision point in the episode sequence. The prompt includes the verification history of the informant, and the agent acts, declines or looks. Only a look updates the history, and the ground truth is never revealed.

fifteen decision points, so no agent can verify every claim, and each decision point allows at most eight turns, each one model response. After navigation the agent reads the frame rendered from its pose, and a ground-truth perception variant returns the recorded state instead (§5).

4 Six Evaluation Levels

Table 2 and Figure 3 lay out the six levels, and §4.1 defines the metric of each level.

4.1 Metrics

L0, Grounding Score (GS) scores object categories, spatial relations and counts in the frame before any informant speaks (see Appendix B).

L1, Veracity F1 (VF1). Given a frame and a claim, the model outputs $\hat{y} \in \{\text{CONSISTENT}, \text{CONTRADICTED}, \text{UNVERIFIABLE}\}$. The gold label is UNVERIFIABLE when the named receptacle is not in view, and otherwise follows the truth of the claim at that step. With

$F_1^{(y)}$ the score for class y,

$$\mathrm{VF1} = \tfrac{1}{3} \sum_y F_1^{(y)}. \tag{1}$$

L2, Conflict-Resolution Accuracy (CRA). Let $\mathcal{D}_2$ be the 148 decision points at which the two claims concern the same fact and disagree. The gold answer $g_d \in \{\text{SPOKEN}, \text{WRITTEN}, \text{NEITHER}\}$ is the informant with the true claim when the frame settles the conflict, and NEITHER when both claims are false or the frame cannot settle it; the model chooses $\hat{p}_d$ from the same set. The 41 pairs in which both claims are false are among the NEITHER items (see Appendix F). We define

$$\mathrm{CRA} = \frac{1}{|\mathcal{D}_2|} \sum_{d \in \mathcal{D}_2} \mathbb{1}[\hat{p}_d = g_d], \quad \Delta_{\mathrm{mod}} = \Pr(\hat{p}_d{=}\text{SP} \mid g_d{\neq}\text{SP}) - \Pr(\hat{p}_d{=}\text{WR} \mid g_d{\neq}\text{WR}). \tag{2}$$

$\mathbb{1}[\cdot]$ is 1 if its argument holds and 0 otherwise, and SP and WR abbreviate SPOKEN and WRITTEN. The two terms of the modality bias Δ_{mod} are p_{sp} and p_{wr}, the rates of choosing each informant when that informant is not the gold answer. An unbiased adjudicator has $\Delta_{\mathrm{mod}} = 0$, even when CRA is at chance.

Trust correctness. The embodied levels share one score. At a decision point d the agent ends with a terminal action $a_d \in \{\text{ACT}, \text{DECLINE}, \text{ASK}, \text{ANSWER}\}$ on a claim. The truth of the claim at that step is $h_d \in \{0, 1\}$, and we score

$$\tau_d = \begin{cases} +1 & (a_d = \text{ACT} \wedge h_d) \ \vee \ (a_d = \text{DECLINE} \wedge \neg h_d) \\ -1 & (a_d = \text{ACT} \wedge \neg h_d) \ \vee \ (a_d = \text{DECLINE} \wedge h_d) \\ -\frac{1}{4} & a_d = \text{ASK} \\ 0 & a_d = \text{ANSWER}. \end{cases} \tag{3}$$

Let $b(r) = \max(2r - 1,\, 1 - 2r)$ be the expected τ of the best decision made *without looking* for an informant of reliability r.

L3, Verification Efficiency (VE). Let $\mathcal{V}$ be the decisions that the agent chose to verify, and s_d the navigation steps spent on decision d. VE is the gain in trust correctness from looking over the best decision without looking, per step:

$$\mathrm{VE} = \frac{\sum_{d \in \mathcal{V}} \big(\tau_d - b(R_{p(d),c(d)})\big)}{\sum_{d \in \mathcal{V}} s_d}. \tag{4}$$

Here $p(d)$ and $c(d)$ are the informant and the reliability group of the claim decided at d (§3.2), so every agent is measured against the same reference. VE is positive when looks overturn wrong guesses, undefined for an agent that never verifies, and blind to later payoffs of a look (see Appendix D).

L4, Memory Score (MS). Each test informant p recurs across the episodes of the evaluation sequence (§4.2). Let k be the number of earlier episodes in the sequence in which p appeared. Let $T_p(k)$ be the mean τ over the decisions of the agent on claims by p at exposure k, skipping any k with no such decision. The pool $\mathcal{P}$ contains the twelve test informants, excluding any with fewer than two exposure points. MS is the average rise in trust correctness after the first exposure,

$$\mathrm{MS} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \Big(\operatorname*{mean}_{k \geq 1} T_p(k) - T_p(0) \Big). \tag{5}$$

An agent without memory has $\mathrm{MS} \approx 0$.

L5, Trust-Correctness Score on unverified decisions (TCS-NV) and ECE. Let $\mathcal{N}$ be the decisions on test informants that the agent did *not* verify, excluding ASK and ANSWER, which resolve no claim. TCS-NV is the mean trust correctness on exactly these decisions, and ORC-NV, the oracle score on unverified decisions, averages b over them:

$$\text{TCS-NV} = \frac{1}{|\mathcal{N}|} \sum_{d \in \mathcal{N}} \tau_d, \qquad \text{ORC-NV} = \frac{1}{|\mathcal{N}|} \sum_{d \in \mathcal{N}} b\big(R_{p(d),c(d)}\big). \tag{6}$$

ORC-NV is the score of an oracle that knows every $R_{p,c}$ and never looks. We pool ORC-NV over the unverified decisions of all models, because per-model values differ by at most 0.02, and call

it the pooled oracle. A reference policy leaves extreme cells unverified, so the oracle on its own unverified decisions is higher, 0.84 to 0.88 (see Appendix F). Because TCS-NV uses only unverified decisions on test informants, verifying everything or memorising the training pool cannot inflate the score. $\mathrm{ECE}_{\mathrm{act}}$ is the expected calibration error of the stated confidence q_d on act decisions over ten equal-width bins (see Appendix F and Table 9; prompt in Appendix I).

4.2 Sequential Evaluation

At L4 and L5 the evaluation split is presented as one fixed sequence of episodes in which the twelve test informants recur, and we call this the sequential evaluation. The model receives no parameter update during the sequence and never sees the ground truth. The only information that persists is the verification history, which lists for each informant the outcomes of past checks, each either "held" or "did not hold". The agent reports each outcome in a `verdict` field after a look, and the prompt lists the history whenever that informant speaks again. The history records the reading of the frame by the agent, so a misread frame corrupts the history as well as the decision, and the history holds no reliability estimate. Our reference policies are scripted Bayesian policies with a Beta posterior per informant, or per informant and group in the group-aware variant, simulated from the generator parameters (see Appendix F). Because the model does not change, a rise in Memory Score must come from the history, and shuffling identities tests whether this rise depends on identity (§5).

4.3 Baselines and Oracles

At L5 the baseline to beat is the trust-all baseline, a constant ACT-ON policy that scores $\tau = 2\pi - 1$, where π is the share of true claims, counting off-topic claims as true. On the unverified test decisions $\mathcal{N}$, the empirical share is $\pi_{\mathcal{N}} = 0.563$, so the trust-all baseline scores $+0.13$. An oracle that knows the reliability of each informant for every group scores 0.83, which leaves 0.70 of headroom for learning. At L2, AlwaysTrustSpoken and AlwaysTrustWritten reach a CRA of 0.23 and 0.20 and a constant NEITHER answer 0.57, the shares of the three gold answers. At L3 we run AlwaysVerify, NeverVerify, for which VE is undefined, and SelectiveOracle, which knows every $R_{p,c}$ and verifies exactly when $b(R_{p,c}) \leq \theta$ with $\theta = 0.6$ (see Appendix F). Only episodes that completed without a harness error enter a corpus-level number (see Appendix J).

5 Experiments

Setup. We evaluate six open omni-modal models without fine-tuning: Qwen2.5-Omni 3B and 7B, Qwen3-Omni-30B-A3B Instruct and Thinking, Nemotron-3-Nano-Omni-30B-A3B, and InteractiveOmni-8B (§2), all with one system prompt that treats the two informants identically (see Appendix I). Each model runs with three sampling seeds at temperature 0.7 and top-p 0.95, which change only decoding. Every model number is a mean over the three seeds, and the $\pm$ values in Tables 3 and 4 are standard deviations over seeds. Paired tests use wild cluster bootstrap intervals, clustered by episode at L0 to L2 and by informant at L3 to L5 (see Appendix F). The five control conditions run on the same episode sequence (see Tables 7 and 8 and Figure 4 in Appendix F). L3 is a separate run that starts every episode with an empty history, and L4 and L5 share one sequential run.

The levels rank the models differently. No single order holds across levels (see Figure 6 in Appendix H). Qwen3-Omni-30B-Thinking perceives best, at a GS of 0.33, and the Instruct variant judges single claims best, at a VF1 of 0.45. At L2 no model beats a constant NEITHER answer at 0.57: the 3B model and Nemotron match it by answering NEITHER on nearly every conflict, and Qwen2.5-Omni-7B scores lowest, at 0.41. Kendall's coefficient of concordance across the seven metrics is 0.19, and verification efficiency and CRA rank the models almost in reverse, with a Kendall rank correlation $r_{\mathrm{K}} = -0.87$ (τ_a, ties counted as zero; see Appendix F). The levels therefore measure different capabilities, and neither family nor size explains the pattern: the three 30B models differ at every level but L3.

Acting on unverified claims. At L5 the trust-all baseline scores $+0.13$ (§4.3), and the oracles score 0.83 with reliability by group and 0.67 with overall reliability. Only Qwen3-Omni-30B-Instruct acts above the baseline, at $+0.25 \pm 0.05$ over seeds and $+0.12 \pm 0.09$ above its own baseline

Table 3: **Results at the single-shot and verification levels** on the 27 evaluation episodes, without fine-tuning; L3 uses a 60-step budget and each model's own camera observation. Δ_{mod} is the modality bias. Wild cluster bootstrap 95% half-widths: GS and VF1 ± 0.05, VE ± 0.025, CRA ± 0.11, Δ_{mod} ± 0.15. †Analytic rows. Reference policies use a fixed vision-language reader, with ground-truth-perception values in parentheses. –: not applicable or not computed for this agent. Cell $\pm$ is the s.d. over three seeds; bootstrap half-widths are given separately. Notes in Appendix F.

	L0	L1	L2			L3
Model	GS↑	VF1↑	CRA↑	$p_{\text{sp}}/p_{\text{wr}}$	Δ_{mod}	VE↑
Qwen2.5-Omni-3B	0.18±0.05	0.28±0.05	0.57±0.06	0.00/0.00	+0.00	-0.010±0.011
InteractiveOmni-8B	0.25±0.05	0.26±0.05	0.48±0.06	0.32/0.04	+0.27	+0.009±0.011
Qwen2.5-Omni-7B	0.31±0.06	0.30±0.06	0.41±0.06	0.51/0.02	+0.49	+0.030±0.013
Nemotron-3-Nano-Omni-30B	0.10±0.04	0.25±0.05	0.57±0.06	0.00/0.01	-0.01	+0.007±0.011
Qwen3-Omni-30B-Instruct	0.20±0.05	0.45±0.06	0.46±0.06	0.28/0.00	+0.28	+0.029±0.013
Qwen3-Omni-30B-Thinking	0.33±0.06	0.33±0.06	0.47±0.06	0.28/0.22	+0.06	+0.029±0.013
CONSTANT NEITHER†	–	–	0.57	–	–	–
ALWAYSTRUSTSPOKEN† / WRITTEN†	–	–	0.23 ; 0.20	1.00/0.00 ; 0.00/1.00	+1.00 ; −1.00	–
ALWAYSVERIFY	–	–	–	–	–	+0.056 (+0.089)
SELECTIVEORACLE ($\theta = 0.6$)	–	–	–	–	–	+0.134 (+0.167)

Table 4: **Results at L4 and L5 on the sequential run** (100-step budget, eight-turn cap, own camera observation, speech as audio). Intervals: wild cluster bootstrap by informant for the models, and ± 1.96 s.d. over simulated sequences for the Bayesian policy rows. Baseline $+0.13$ ($\pi_{\mathcal{N}} = 0.563$); oracles 0.83 (group-level) and 0.67 (informant-level). vs. policy $= (\text{TCS-NV} - \text{baseline})/(\text{policy} - \text{baseline})$ against the group-aware Bayesian policy simulated with each model's post-look accuracy, the primary L5 output. Bayesian policy rows use a fixed vision-language reader, with ground-truth-perception values in parentheses. –: not applicable or not computed for this agent. Cell $\pm$ is the s.d. over three seeds; bootstrap half-widths are given separately. Notes in Appendix F.

	L4–L5 run	L4 Remember		L5 Act			
Model	Verify	MS↑	95% CI	TCS-NV↑	95% CI	vs. policy	ECE_{act} ↓
Qwen2.5-Omni-3B	0.61±0.06	+0.07±0.05	±0.10	-0.05±0.05	±0.09	below baseline	0.25±0.02
InteractiveOmni-8B	0.29±0.05	-0.03±0.05	±0.10	-0.03±0.05	±0.09	below baseline	0.36±0.03
Qwen2.5-Omni-7B	0.75±0.05	+0.11±0.06	±0.10	+0.18±0.05	±0.09	at baseline	0.21±0.02
Nemotron-3-Nano-Omni-30B	0.60±0.06	+0.10±0.06	±0.10	-0.01±0.05	±0.09	below baseline	0.25±0.02
Qwen3-Omni-30B-Instruct	0.30±0.06	-0.01±0.04	±0.10	+0.25±0.05	±0.09	32%	0.30±0.02
Qwen3-Omni-30B-Thinking	0.49±0.06	+0.01±0.04	±0.10	+0.01±0.05	±0.09	below baseline	0.27±0.02
CONSTANT ACT-ON†	–	–	–	+0.13	–	0%	–
ALWAYSDECLINE†	–	–	–	−0.13	–	–	–
Bayesian, per informant	0.49	+0.05 (-0.08)	±0.22	+0.66 (+0.72)	±0.17	91%	0.10
Bayesian, group-aware	0.59	+0.11 (+0.06)	–	+0.71 (+0.84)	±0.15	100%	–
Bayesian, group-agnostic	0.60	–	–	+0.64 (+0.76)	±0.16	88%	–

in the paired test (see Figure 8 in Appendix H). Qwen2.5-Omni-7B is at the baseline, at $+0.18$ and a paired $+0.05 \pm 0.09$, and the other four models score below it, from -0.05 to $+0.01$. The margin of Instruct is not robust: it no longer excludes zero on a second corpus, with less extreme unreliable informants, or without one hard-to-verify kind of false claim (see Appendix D and Table 11 in Appendix F).

Comparison with reference policies. The group-aware Bayesian policy is told the reliability groups and, simulated with the post-look accuracy of the best model, reaches $+0.52$. Qwen3-Omni-30B-Instruct closes 32% of the gain of this policy over the trust-all baseline, a paired difference of -0.27 ± 0.09 per decision. With ground-truth perception the same policy reaches $+0.84$, which matches the pooled oracle, and on its own unverified decisions its oracle is 0.88. Learning reliabilities from few looks therefore costs -0.01 ± 0.05, within noise of zero, while reading the frame with the model's own camera costs 0.32 ± 0.10 (see *Gap decomposition* in Appendix F).

Declines and prompt sensitivity. Decline rates on unverified claims range from 20% for the 3B model to 34% for Instruct (see Table 8 in Appendix F). A warning that either informant may be wrong lowers TCS-NV by 0.20 to 0.24 and leaves no model above the trust-all baseline. A history-

use prompt lifts Qwen2.5-Omni-7B and Nemotron by 0.04, which moves 7B marginally above the baseline at $+0.22$, and changes the others by at most 0.03 (see Appendix F).

The label bias is not a mark of weak models. The modality bias Δ_{mod} is largest for Qwen2.5-Omni-7B, at $+0.49$, moderate for InteractiveOmni-8B and Instruct, at $+0.27$ and $+0.28$, small for Thinking and near zero for the 3B model and Nemotron. For the three biased models the bias persists when the spoken claim is shown as text and reverses sign when the labels are swapped, so the preference follows the label and does not require audio (see Table 7 in Appendix F). After correction for multiple comparisons, the L2 bias of all three models survives, as do the 7B bias in all three sequence conditions and the audio bias of Instruct (see Appendix F).

Looking pays little. Qwen2.5-Omni-7B, Instruct and Thinking gain similarly from looking, with verification efficiencies of $+0.030$, $+0.029$ and $+0.029$, each ± 0.025, and none of these gains survives correction, while the other three are indistinguishable from zero. With the fixed vision-language reader, SELECTIVEORACLE reaches $+0.134 \pm 0.03$, a paired difference of -0.105 ± 0.030 for Thinking. Ground-truth perception raises VE by 0.04 to 0.05 for every model (see Table 6), so part of every model's VE shortfall is misperception. The verify rate measures propensity and says nothing about skill: Qwen2.5-Omni-3B looks at 61% of decision points yet gains nothing per step, while Instruct looks at 30% and gains as much as any model.

Memory is rare and does not predict acting well. Qwen2.5-Omni-7B has the highest Memory Score, at $+0.11 \pm 0.10$, the only interval that excludes zero, and this result does not survive correction. Nemotron reaches $+0.10$, the 3B model $+0.07$, and the two Qwen3-Omni models stay within 0.01 of zero, against $+0.05$ for the per-informant Bayesian policy (see Figure 7 in Appendix H). Shuffling identities lowers the Memory Score of 7B and Nemotron by 0.09 ± 0.07 and 0.08 ± 0.07 and moves no other model beyond noise (see Table 8 in Appendix F). The model with memory is not the model that acts best: Instruct has none, and with an empty history its margin shrinks to $+0.04 \pm 0.08$, so its advantage comes from outcomes within the current episode. With an empty history no model is above the baseline. Calibration follows another order, with ECE_{act} lowest for 7B at 0.21 and highest for InteractiveOmni-8B at 0.36 (see Appendix F; worked example in Appendix G).

Caveats. Forced declines make up 4 to 6% of decisions, and excluding them moves TCS-NV by at most 0.01 (see Table 9 in Appendix F). A second evaluation corpus with twelve new test informants reproduces the TCS-NV ordering and every TCS-NV within 0.02.

6 CONCLUSION

On SEEINGISBELIEVING, the six levels rank six open omni-modal models differently. Three models favor whichever informant is labelled as speaking, most clearly Qwen2.5-Omni-7B. Qwen2.5-Omni-7B and the two Qwen3-Omni models gain similarly from looking, Qwen2.5-Omni-7B has the highest Memory Score, and shuffling identities lowers the Memory Score of 7B and Nemotron, but none of these effects survives correction for multiple comparisons. Only Qwen3-Omni-30B-Instruct acts above the trust-all baseline, and not robustly. Using its own camera observations, Instruct closes 32% of the gain of a group-aware Bayesian policy over the trust-all baseline, so current models use their own verification outcomes far less well than a scripted policy (caveats in Appendix D).

Limitations and future work. The evaluation split has 27 episodes and twelve test informants, so the L4 and L5 intervals are wide and the one positive result at L5 is fragile (see Appendix D). Only open models were run. The generator also defines a training environment, and we plan to extend SEEINGISBELIEVING into a reinforcement learning gym that rewards profitable verification and calibrated trust.

AI USE STATEMENT

In accordance with the conference guidelines, we acknowledge the use of large language models during the preparation of this manuscript. They were used for writing assistance, namely polishing the language and improving the clarity and readability of the text, including grammar and academic

style. All core research contributions, including the benchmark design, the evaluation protocol, the experiments, the analysis, and the scientific conclusions, were developed by the authors. The models served solely as writing assistance tools and did not contribute to the conceptual development, the experimental validation, or the interpretation of the results.

ETHICS STATEMENT

SEEINGISBELIEVING is generated entirely in simulation. The informants are synthetic, with template-generated names and voices, and no human data is collected. Demographic attributes are used only to test that trust does not attach to them. Reliability is uncorrelated with gender, age, language and delivery style by construction, and we report no result that ranks demographic groups. The benchmark is intended to expose excessive trust and miscalibration in embodied agents, and not to train agents to distrust particular kinds of speakers.

REPRODUCIBILITY STATEMENT

Appendix A gives the corpus statistics, Appendix I the verbatim prompt, Appendix J the validity accounting, and Appendix D the corpus-versioning policy. We will release the corpus, the evaluation code for all six levels and the reference policies, and the per-episode result files behind every table. Every reported number is recomputable from those files. A single input file of released quantities, namely the test reliabilities, the split composition and the per-cell counts, feeds a derivation script. This script regenerates every baseline, oracle, share and interval in this paper, and the macros of the paper are read from its output.

REFERENCES

Dhruv Batra, Angel X. Chang, Sonia Chernova, Andrew J. Davison, Jia Deng, Vladlen Koltun, Sergey Levine, Jitendra Malik, Igor Mordatch, Roozbeh Mottaghi, Manolis Savva, and Hao Su. Rearrangement: A challenge for embodied AI. *arXiv preprint arXiv:2011.01975*, 2020.

Bryan Chan, Xinyi Chen, András György, and Dale Schuurmans. Toward understanding in-context vs. in-weight learning. In *ICLR*, 2025. arXiv:2410.23042.

Stephanie C.Y. Chan, Ishita Dasgupta, Junkyung Kim, Dharshan Kumaran, Andrew K. Lampinen, and Felix Hill. Transformers generalize differently from information stored in context vs in weights. *arXiv preprint arXiv:2210.05675*, 2022a.

Stephanie C.Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya Singh, Pierre H. Richemond, Jay McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *NeurIPS*, 2022b. arXiv:2205.05055.

Mingfei Chen, Zijun Cui, Xiulong Liu, Jinlin Xiang, Caleb Zheng, Jingyuan Li, and Eli Shlizerman. SAVVY: Spatial awareness via audio-visual LLMs through seeing and hearing. In *NeurIPS*, 2025. arXiv:2506.05414.

Xiaoyu Chen, Shenao Zhang, Pushi Zhang, Li Zhao, and Jianyu Chen. Asking before acting: Gather information in embodied decision making with language models. *arXiv preprint arXiv:2305.15695*, 2023.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *CVPR*, 2025. arXiv:2503.02199.

Wenxuan Ding, Nicholas Tomlin, and Greg Durrett. Calibrate-then-act: Cost-aware exploration in LLM agents. *arXiv preprint arXiv:2602.16699*, 2026.

Aaron Fanous, Jacob Goldberg, Ank A. Agarwal, Joanna Lin, Anson Zhou, Roxana Daneshjou, and Sanmi Koyejo. SycEval: Evaluating LLM sycophancy. *arXiv preprint arXiv:2502.08177*, 2025.

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, Vishakh Padmakumar, Ilia Sucholutsky, Andrew Strait, Diyi Yang, Q. Vera Liao, and Umang Bhatt. Measuring and mitigating overreliance to build human-compatible AI. *arXiv preprint arXiv:2509.08010*, 2025.

Yifan Jia, Kailin Jiang, Yuyang Liang, Qihan Ren, Yi Xin, Rui Yang, Fenze Feng, Mingcai Chen, Hengyang Lu, Haozhe Wang, Xiaoye Qu, Dongrui Liu, Lizhen Cui, and Yuntao Du. Benchmarking multimodal knowledge conflict for large multimodal models. *arXiv preprint arXiv:2505.19509*, 2025.

Chih-Ting Liao, Xi Xiao, Chunlei Meng, Zhangquan Chen, Yitong Qiao, Weilin Zhou, Tianyang Wang, Xu Zheng, and Xin Cao. SpaMEM: Benchmarking dynamic spatial reasoning via perception–memory integration in embodied environments. *arXiv preprint arXiv:2604.22409*, 2026. ECCV 2026.

Zheyuan Liu, Weiliang Zhao, Xiangchi Yuan, Ningshan Ma, Yue Huang, and Meng Jiang. Knowledge-verified emergent deception in LLM agents under conflicting incentives. *arXiv preprint arXiv:2608.26372*, 2026.

Zihan Liu, Zhikang Niu, Qiuyang Xiao, Zhisheng Zheng, Ruoqi Yuan, Yuhang Zang, Yuhang Cao, Xiaoyi Dong, Jianze Liang, Xie Chen, Leilei Sun, Dahua Lin, and Jiaqi Wang. STAR-Bench: Probing deep spatio-temporal reasoning as audio 4D intelligence. *arXiv preprint arXiv:2510.24693*, 2025.

Yaxing Lyu, Shengjie Zhou, Binbin Toh, Pengyu Zhu, and Lijun Li. KC-Bench: A dynamic interactive benchmark for evaluating knowledge conflicts in LLM agents. *arXiv preprint arXiv:2609.03588*, 2026.

NVIDIA. Nemotron 3 Nano Omni: Efficient and open multimodal intelligence. *arXiv preprint arXiv:2604.24954*, 2026.

Teodora Popordanoska, Jiameng Li, and Matthew B. Blaschko. CLASH: A benchmark for cross-modal contradiction detection. *arXiv preprint arXiv:2511.19199*, 2025.

Rohan N. Pradhan and Steve Goley. Trust, but don't verify: Epistemic blind spots in LLM source evaluation. *arXiv preprint arXiv:2606.05403*, 2026.

Muhammad Raees and Konstantinos Papangelis. From trust to appropriate reliance: Measurement constructs in human-AI decision-making. *arXiv preprint arXiv:2604.23896*, 2026.

Aaditya K. Singh, Stephanie C.Y. Chan, Ted Moskovitz, Erin Grant, Andrew M. Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. *arXiv preprint arXiv:2311.08360*, 2023.

Hung-Ting Su, Ting-Jun Wang, Jia-Fong Yeh, Min Sun, and Winston H. Hsu. VLN-NF: Feasibility-aware vision-and-language navigation with false-premise instructions. *arXiv preprint arXiv:2604.10533*, 2026.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. ConflictBank: A benchmark for evaluating the influence of knowledge conflicts in LLM. *arXiv preprint arXiv:2408.12076*, 2024.

Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. AVHBench: A cross-modal hallucination benchmark for audio-visual large language models. In *ICLR*, 2025. arXiv:2410.18325.

Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, Xueheng Li, Lumin Li, Chenxu Guo, Jiasheng Zhou, Jiandong Chen, Xianye Wu, Jiahao Wang, Silei Wu, Lei Chen, Hanming Deng, Yuxuan Song, Dinghao Zhou, Guiping Zhong, Ken Zheng, Shiyin Kang, and Lewei Lu. InteractiveOmni: A unified omni-modal model for audio-visual multi-turn dialogue. *arXiv preprint arXiv:2510.13747*, 2025.

Harold Triedman and Vitaly Shmatikov. MillStone: How open-minded are LLMs? *arXiv preprint arXiv:2509.11967*, 2025.

Eren Unlu. Know when to trust the skill: Delayed appraisal and epistemic vigilance for single-agent LLMs. *arXiv preprint arXiv:2604.16753*, 2026.

Cheng Wang, Gelei Deng, Xianglin Yang, Han Qiu, and Tianwei Zhang. When audio and text disagree: Revealing text bias in large audio-language models. In *EMNLP*, 2025. arXiv:2508.15407.

Huyu Wu, Meng Tang, Xinhan Zheng, and Haiyun Jiang. When language overrules: Revealing text dominance in multimodal large language models. *arXiv preprint arXiv:2508.10552*, 2025a.

Qiucheng Wu, Handong Zhao, Michael Saxon, Trung Bui, William Yang Wang, Yang Zhang, and Shiyu Chang. VSP: Diagnosing the dual challenges of perception and reasoning in spatial planning tasks for MLLMs. In *ICCV*, 2025b. arXiv:2407.01863.

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. In *ICLR*, 2024. arXiv:2306.13063.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. *arXiv preprint arXiv:2403.08319*, 2024.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. On verbalized confidence scores for LLMs. *arXiv preprint arXiv:2412.14737*, 2024.

Doguhan Yeke, Elif Su Temirel, Ananth Shreekumar, Brandon Lee, Dongyan Xu, and Z. Berkay Celik. The yes-man syndrome: Benchmarking abstention in embodied robotic agents. *arXiv preprint arXiv:2605.20544*, 2026.

Zhi Zeng, Cheng Zhang, Zesheng Yang, Rendong Pi, Jiaying Wu, Di Zhang, Zihan Ma, Guodong Li, Zhou Yang, Yu Xiang, Yifei Zheng, and Minnan Luo. Listen, see and track: Spatio-temporal audio-visual sound event reasoning for omni-modal language models. *arXiv preprint arXiv:2608.09435*, 2026.

Xu Zheng, Zihao Dongfang, Lutao Jiang, Boyuan Zheng, Yulong Guo, Zhenquan Zhang, Giuliano Albanese, Runyi Yang, Mengjiao Ma, Zixin Zhang, Chenfei Liao, Dingcheng Zhen, Yuanhuiyi Lyu, Yuqian Fu, Bin Ren, Linfeng Zhang, Danda Pani Paudel, Nicu Sebe, Luc Van Gool, and Xuming Hu. Multimodal spatial reasoning in the large model era: A survey and benchmarks. *arXiv preprint arXiv:2510.25760*, 2025.

Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. BAT: Learning to reason about spatial sounds with large language models. In *ICML*, 2024. arXiv:2402.01591.

Ruiwen Zhou, Maojia Song, Xiaobao Wu, Sitao Cheng, Xunjian Yin, Yuxi Xie, Zhuoqun Hao, Wenyue Hua, Liangming Pan, Soujanya Poria, and Min-Yen Kan. Epistemic context learning: Building trust the right way in LLM-based multi-agent systems. *arXiv preprint arXiv:2601.21742*, 2026.

Table 5: **The released SEEINGISBELIEVING corpus.** Each ProcTHOR house holds one episode, and every claim names a receptacle that the agent can navigate to. Every evaluation episode has 15 decision points. One training episode in the release has a single point, so the total is 1,546 and not 104×15.

Property	Value
ProcTHOR houses / episodes	104 / 104 (77 train, 27 eval)
Decision points	1,546
Claims	3,092 (1,546 spoken + 1,546 written)
Claim state, training split (true / false / off-topic)	70.8% / 14.9% / 14.3%
Claim state, evaluation split (stratified)	51% / 44% / 5%
False-claim error types, training split (count / spatial / identity / co-location)	215 / 65 / 59 / 1
False-claim error types, evaluation split	218 / 76 / 61 / 3
Fact kinds (count / spatial relation / containment)	2,018 / 1,036 / 38
Same-fact informant pairs	69.7%
Informants (spoken + written)	48 (24 + 24); 12 test
Demographic cells	2 genders $\times$ 3 age groups $\times$ 4 languages
Delivery styles	assertive, hedging
Mean claim length	6.5 words
Resampled identical false claims	20
Movement budget	60 (L3) / 100 (L4 and L5) steps per episode

A DATASET STATISTICS

Table 5 lists the released corpus by split. Each of the two informant pools crosses two genders, three age groups and four languages, namely English, Mandarin, Spanish and Hindi, and each informant has an *assertive* or a *hedging* delivery style. The table gives episodes, decision points, claims per modality, the mix of true, false and off-topic claims, fact kinds, and informant counts. The evaluation split is stratified so that half of its test informants are unreliable (§3.2), and the training split is not stratified.

Claim generation in detail. Claims are generated at every decision point, that is, after every action except the initial state. For each informant we first select a *fact* and then sample the claim state. A fact is a count of the objects in a receptacle, a spatial relation giving the direction and distance between two objects, or a containment statement listing what a receptacle holds. An informant errs on a fact with the larger of its room-level and object-type-level error rates. A global error rate $\lambda = 0.3$ scales these rates and fixes the share of errors in the training split, and the evaluation split is stratified (§3.2). An erroneous claim is either false or off-topic. The error type of a false claim follows the error-type bias of the informant, restricted to the types that apply to the fact kind: identity swap, off-by-one or off-by-two count, receptacle or relation flip, and impossible co-location. Count corruptions dominate the realised corpus. Of the false claims in the training split, 215 are count corruptions and 65 are spatial, against 59 identity swaps and 1 impossible co-location. These counts support corpus-level error rates but not a stratified analysis by error type in the training split. The stratified evaluation split has 218/76/61/3. With probability 0.7 the two informants speak about the same fact, which creates the direct conflicts used at the adjudication level. Table 5 gives the realised mix of fact kinds, and Appendix D describes the unused extensions of the generator. Three rules resample any decision point at which the two informants would make the same false claim. The rules cover the same corruption of the same fact, the same asserted value from different corruptions, and the same stale claim with the same invalidation time. Only the first two rules can fire on the released corpus. Agreement between informants is therefore never itself evidence of truth. Every claim is rendered through a template bank for its language and the delivery style of the informant. Each claim stores its correct value, its asserted value, and a function that checks the claim against the world state at any later step. The pools contain informants that are reliable overall, unreliable overall, and reliable only for some groups, and error rates vary by group around the base rate of each informant. Each spoken informant has a fixed Kokoro-82M voice chosen by gender, age group and language, mixed over ambient room sound. Hedging is realised as a lexical hedge with a slower speaking rate, and delivery style is uncorrelated with reliability like every other observable attribute. An agent that never verifies gains no evidence about the informants.

Table 6: **Ground-truth perception ablation** on the same 27 episodes. At L3 to L5, the recorded state of the simulator replaces the camera observation of the model. Entries are means over three seeds.

Model	VE (camera)	VE (ground truth)	TCS-NV (camera)	TCS-NV (ground truth)
Qwen2.5-Omni-3B	-0.010	$+0.039$	-0.05	$+0.03$
InteractiveOmni-8B	$+0.009$	$+0.054$	-0.03	$+0.04$
Qwen2.5-Omni-7B	$+0.030$	$+0.071$	$+0.18$	$+0.25$
Nemotron-3-Nano-Omni-30B	$+0.007$	$+0.061$	-0.01	$+0.08$
Qwen3-Omni-30B-Instruct	$+0.029$	$+0.077$	$+0.25$	$+0.33$
Qwen3-Omni-30B-Thinking	$+0.029$	$+0.069$	$+0.01$	$+0.08$

B Grounding Score

Let $\mathcal{U}$ be the set of topology clusters referenced by the claims of an episode. For a cluster u with predicted category multiset S_{pred} and recorded multiset S_{rec}, semantic recognition is exact-match F_1 under a one-to-one assignment, with precision P_u and recall Rec_u, and $F_1^{\text{clu}}(u) = 0$ when $S_{\text{pred}} \cap S_{\text{rec}} = \emptyset$,

$$F_1^{\text{clu}}(u) = \frac{2P_u Rec_u}{P_u + Rec_u}, \qquad P_u = \frac{|S_{\text{pred}} \cap S_{\text{rec}}|}{\max(|S_{\text{pred}}|, 1)}, \quad Rec_u = \frac{|S_{\text{pred}} \cap S_{\text{rec}}|}{\max(|S_{\text{rec}}|, 1)}. \tag{7}$$

Relative spatial reasoning is $s_{\text{rel}}(u) = \frac{1}{2}\,\text{Recall@3} + \frac{1}{2}\,\text{DirAcc}$ over the three nearest neighbours of the receptacle and the egocentric directions to them, where Recall@3 is the fraction of the three nearest recorded neighbours that appear among the three predicted, DirAcc is the fraction of those directions predicted correctly and a cluster is a receptacle together with the objects on or in it. Counting is $s_{\text{cnt}}(u) = \mathbb{1}\big[\,|C_{\text{pred}} - C_{\text{rec}}| \le 1\,\big]$, where C_{pred} and C_{rec} are the predicted and recorded object counts of cluster u. The metric of this level is

$$\text{GS} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \tfrac{1}{3}\big(F_1^{\text{clu}}(u) + s_{\text{rel}}(u) + s_{\text{cnt}}(u)\big). \tag{8}$$

C Ground-Truth Perception Ablation

Table 6 reruns L3 to L5 with the recorded state of the simulator in place of the camera observation of the model after each look. The ground-truth perception raises VE by 0.04 to 0.05 for every model, and the negative VE of the 3B model becomes positive, so part of every model's VE shortfall comes from misreading the frame after a look. On unverified claims TCS-NV rises by 0.07 to 0.09, and the group-aware Bayesian policy gains 0.13, from $+0.71$ with the fixed vision-language reader to $+0.84$. Even with perfect perception the best model stays far below the policy, at $+0.33$ against $+0.84$, so the models' shortfall lies mainly in how they use the recorded outcomes. The perception term of the gap decomposition in Appendix F is measured on that policy, from $+0.52$ with the reading of Qwen3-Omni-30B-Instruct to $+0.84$ with ground-truth perception.

D Limitations and Scope

Uncorrelated attributes are a design constraint. The bound of 0.15 on the rank correlation between observable attributes and reliability (§3.2) is imposed during generation. The bound is not evidence of independence, and the bound does not make claims uninformative. Reliability depends on the room and object type by design, so the content of a claim carries information about whether the claim holds. An agent can therefore appear to track an informant while keying on content, and the identifier-shuffled Bayesian policy below tests for this shortcut.

Corpus composition. The released corpus contains count, spatial-relation and containment facts and off-topic claims. Facts about action events and stale claims tagged with the invalidating action are implemented in the generator and reserved for a future release. The generator implements four

claim states, and the released data use three. A receptacle-swap corruption asserts about the identified receptacle something that holds for another receptacle of the same type. The identifier is not rendered in the surface text, so the current view can make such a claim look true. A look resolves the claim, because navigation goes to the identified receptacle. Receptacle swaps make up 88 of the 141 spatial-relation corruptions and 43 claims in the evaluation split, with L1 gold labels of 19 CONTRADICTED and 24 UNVERIFIABLE. Excluding the receptacle swaps changes spatial post-look accuracy by at most 0.02, VF1 by at most 0.01, and TCS-NV by $+0.02$ to $+0.04$. All receptacle swaps are false, so excluding them also raises the trust-all baseline from $+0.13$ to $+0.19$. Every margin over the baseline then falls, by 0.02 for Qwen2.5-Omni-3B up to 0.05 for Qwen3-Omni-30B-Instruct. The TCS-NV ordering is unchanged, and the margin of Qwen3-Omni-30B-Instruct falls from $+0.12$ to $+0.07 \pm 0.09$, a fall of about 0.05 from unrounded endpoints, computed from unrounded values, and no longer excludes zero.

Memory Score and shortcuts. Reliability depends on the room and object type. An agent that bases trust on the object type of a fact, instead of on the informant, can therefore raise the Memory Score without tracking identity. We run an identity-blind feature agent and an identifier-shuffled Bayesian policy through the same scorer, and Appendix F reports both under *Bayesian reference policies*. We recommend that any claimed memory effect come with this control.

Forced declines. A decision point ends in a neutral decline with stated confidence 0.5 when it reaches the eight-turn cap, exhausts the movement budget, or follows two consecutive impossible actions. A forced decline is scored like any other decline, with $+1$ if the claim is false and -1 if the claim holds. Forced declines make up 4% to 6% of decisions, and excluding them moves TCS-NV by at most 0.01 (Table 9). Forced declines therefore account for almost none of the distance between a model and the trust-all baseline.

Verification efficiency is myopic. VE credits a look only through the decision that the look informs immediately. The outcome that a look adds to the verification history can improve later decisions on the same informant. That value appears in the Memory Score and in TCS-NV but not in VE. VE therefore understates looks that pay off later and favours looking where the current decision is uncertain over looking to learn about an informant.

Scope. We evaluate only open-weight omni-modal models. We did not run closed frontier models with native audio input, so the orderings at each level do not show where the strongest available models fall. The training split is released, but we train no agent on it. Whether fine-tuning on the training informants yields identity-based trust or a feature shortcut remains open, and the identifier shuffle is the test such a study would need. The scene changes between decision points as objects are spawned, placed and removed, so every claim is checked against the current state. Stale claims, which would test whether an agent tracks when a fact stopped holding, are implemented in the generator but withheld from this release.

Corpus versioning. Regenerating the corpus from the same seed with a later revision of the generator changes the claim text. The released corpus is therefore identified by content hash instead of by seed, and every number in this paper was produced on exactly that corpus.

E MODEL PROFILES

Level by level, the models differ in ways that one score would hide. Qwen2.5-Omni-7B verifies more often than any other model, at 0.75 of decision points, gains $+0.030$ per step from looking, as much as the two Qwen3-Omni models, although none of the three gains survives correction. Qwen2.5-Omni-7B has the highest Memory Score at $+0.11$ and the lowest calibration error at 0.21, yet its modality bias is the largest, at $+0.49$, and its CRA the lowest, at 0.41. At L5 the 7B model sits at the trust-all baseline, at $+0.18$. Qwen3-Omni-30B-Instruct looks rarely, at 0.30, judges single claims best, at a VF1 of 0.45, and is the only model above the baseline at L5, at $+0.25$, with a Memory Score of zero. Qwen3-Omni-30B-Thinking perceives best, at a GS of 0.33, and gains from looking at $+0.029$, but acts below the baseline at $+0.01$ and follows false claims from both informants, with p_{sp} and p_{wr} of 0.28 and 0.22. Qwen2.5-Omni-3B and Nemotron-3-Nano-Omni-30B answer NEITHER on nearly every L2 conflict, which gives them the highest CRA, 0.57, and no

measurable modality bias. Both look often, at 0.61 and 0.60, gain nothing per step, and act below the baseline, although Nemotron has the second-highest Memory Score, at $+0.10$. InteractiveOmni-8B looks least, at 0.29, has a modality bias of $+0.27$ that follows the label, is the worst calibrated model, at 0.36, and acts below the baseline, at -0.03.

F Controls, Additional Analyses, and Corpus Quantities

Seeds. Every model number in Tables 3, 4, 6, 7, 8, 9 and 11 is the mean over three runs with sampling at temperature 0.7 and top-p 0.95, seeds 0, 1 and 2. The seeds change only decoding, so the episodes, the informants and the order of the sequence are the same in every run. The $\pm$ values in Tables 3, 4 and 7 are standard deviations over seeds. The seed standard deviation is 0.04 to 0.06 for GS, VF1, CRA, the verify rate, the Memory Score and TCS-NV, and 0.01 to 0.03 for VE and ECE. Paired tests use the wild cluster bootstrap described under *Small-cluster inference*, on the pooled decisions of the three seeds.

Notes on Table 3. All models run without fine-tuning on the 27 evaluation episodes, and L3 uses a 60-step budget and each model's own camera observation. L2 is scored on the 148 same-fact disagreements, with gold NEITHER on the 84 items in which the frame cannot settle the conflict or both claims are false. p_{sp} and p_{wr} are the probabilities of choosing the spoken or the written informant when that informant is not the gold answer, and Δ_{mod} is the modality bias. Wild cluster bootstrap 95% half-widths, by episode at L0 to L2 and by informant at L3, are $\pm$0.05 for GS and VF1, $\pm$0.025 for VE, $\pm$0.11 for CRA and ± 0.15 for Δ_{mod}. †Analytic rows follow from the corpus composition: the spoken claim is the gold answer on 0.23 of the L2 items and the written claim on 0.20. VE for every agent uses the group-level reliability $R_{p,c}$ as the reference without looking (§4.1). Reference policies use a fixed vision-language reader, with ground-truth-perception values in parentheses. Table 4 reports L4 and L5.

Notes on Table 4. A model is marked below baseline when its paired interval lies entirely below its trust-all baseline and at baseline when the interval includes it; the vs. policy ratio is printed only for models above the baseline. For the Bayesian rows the denominator of that ratio is the group-aware policy with the fixed vision-language reader. The run uses the same 27 evaluation episodes and models, a 100-step budget, an eight-turn cap, each model's own camera observation, and speech delivered as audio. Verify is the fraction of decision points verified. On $\mathcal{N}$ the trust-all baseline is $+0.13$ with $\pi_{\mathcal{N}} = 0.563$. The group-level oracle is 0.83 and the informant-level oracle is 0.67, computed from the test reliabilities and the reliability-group generator parameters. The vs. policy column is $(\mathrm{TCS\text{-}NV} - \mathrm{baseline})/(\mathrm{policy} - \mathrm{baseline})$, with the group-aware Bayesian policy simulated with each model's post-look accuracy (Table 9) as the reference. With each model's post-look accuracy, the group-aware Bayesian policy scores +0.48, +0.59, +0.68, +0.30, +0.52, +0.71 in table order. $\mathrm{ECE}_{\mathrm{act}}$ bootstrap half-widths are ± 0.04 for every model, over 262, 295, 237, 262, 254 and 268 act decisions per seed, averaged over seeds, in table order. The intervals of the Bayesian rows come from the simulation of these policies, described under *Bayesian reference policies*.

Identifier shuffle and empty history. Two further runs remove information from the verification history (Table 8). In the identifier-shuffle run, informant identifiers are shuffled per episode. Outcomes carried from earlier episodes then attach to the wrong informant, while outcomes from the current episode stay correct. In the empty-history run, the history is empty at every decision point, so only the claim, the delivery style and the current view remain. Budget, turn cap and prompt are unchanged, and verify rates move by at most 0.04. The trust-all baseline of each condition is within 0.02 of $+0.13$. Restricting each paired difference to claims left unverified in both runs changes the difference by at most 0.02.

Effect of the identifier shuffle. Shuffling identifiers lowers the Memory Score of Qwen2.5-Omni-7B by 0.09 ± 0.07 and of Nemotron by 0.08 ± 0.07, and TCS-NV by 0.07 ± 0.08 and 0.06 ± 0.08. For the other four models every change is within 0.06 and every interval includes zero. The shuffle leaves the two Qwen3-Omni models unchanged, because they carry nothing across episodes. Decline rates on $\mathcal{N}$ move by at most 0.03 under the shuffle.

Table 7: **Modality biases on the sequential run.** g_{seq} is the rate of following a false claim of the spoken informant minus the same rate for the written informant, on the unverified decisions $\mathcal{N}$ of each model with a false target claim. Entries are means over three seeds with seed standard deviations; wild cluster bootstrap 95% half-widths by informant are ± 0.16 for every entry. Text-only: the claim of the spoken informant is shown as a transcript under the same label. Label swap: the written informant is presented as spoken, with the claim synthesised as audio by the voice rule of Appendix A. The claim of the spoken informant is then shown as text, so modality and label move together, and the text-only run separates the two. Under the label swap, g_{seq} is still computed against the original modality of each informant. A preference that follows the presentation label therefore appears as a sign reversal, and a preference attached to the informant keeps its sign.

Model	g_{seq} (audio)	g_{seq} (text-only)	g_{seq} (label swap)
Qwen2.5-Omni-3B	$+0.00 \pm 0.04$	$+0.00 \pm 0.04$	$+0.00 \pm 0.04$
InteractiveOmni-8B	$+0.24 \pm 0.05$	$+0.22 \pm 0.05$	-0.21 ± 0.05
Qwen2.5-Omni-7B	$+0.44 \pm 0.06$	$+0.40 \pm 0.06$	-0.37 ± 0.06
Nemotron-3-Nano-Omni-30B	-0.01 ± 0.04	-0.01 ± 0.04	$+0.01 \pm 0.04$
Qwen3-Omni-30B-Instruct	$+0.25 \pm 0.05$	$+0.23 \pm 0.05$	-0.21 ± 0.05
Qwen3-Omni-30B-Thinking	$+0.05 \pm 0.04$	$+0.05 \pm 0.04$	-0.05 ± 0.04

Table 8: **Controls on the sequential run.** Memory Score and TCS-NV on the neutral run (Table 4) and on the identifier-shuffle run, in which informant identifiers are shuffled per episode. For TCS-NV the table also shows the empty-history run, in which the verification history is empty at every decision point. Entries are means over three seeds; seed standard deviations are 0.04 to 0.06 for the Memory Score and 0.05 for TCS-NV. Δ columns are paired differences, neutral minus control, computed from unrounded values, with wild cluster bootstrap 95% half-widths by informant. Priming: TCS-NV when the system prompt also warns that either informant may be wrong. Decline $\mathcal{N}$: decline rate on unverified decisions in the neutral run. The trust-all baseline is $+0.13$ in every condition to within 0.02.

	Memory Score			TCS-NV						
Model	neutral	shuffle	Δ	neutral	shuffle	empty	Δ_{shuffle}	Δ_{empty}	priming	decline $\mathcal{N}$
Qwen2.5-Omni-3B	$+0.070$	$+0.014$	$+0.06 \pm 0.07$	-0.05	-0.09	-0.09	$+0.04 \pm 0.08$	$+0.04 \pm 0.08$	-0.26	0.20
InteractiveOmni-8B	-0.025	-0.005	-0.02 ± 0.07	-0.03	-0.02	-0.02	-0.02 ± 0.08	-0.02 ± 0.08	-0.27	0.20
Qwen2.5-Omni-7B	$+0.110$	$+0.022$	$+0.09 \pm 0.07$	$+0.18$	$+0.11$	$+0.11$	$+0.07 \pm 0.08$	$+0.07 \pm 0.08$	-0.02	0.31
Nemotron-3-Nano-Omni-30B	$+0.100$	$+0.020$	$+0.08 \pm 0.07$	-0.01	-0.07	-0.07	$+0.06 \pm 0.08$	$+0.06 \pm 0.08$	-0.22	0.21
Qwen3-Omni-30B-Instruct	-0.005	-0.001	-0.00 ± 0.07	$+0.25$	$+0.26$	$+0.18$	-0.00 ± 0.08	$+0.08 \pm 0.08$	$+0.01$	0.34
Qwen3-Omni-30B-Thinking	$+0.005$	$+0.001$	$+0.00 \pm 0.07$	$+0.01$	$+0.01$	$+0.01$	$+0.00 \pm 0.08$	$+0.00 \pm 0.08$	-0.21	0.23

Effect of an empty history. Emptying the history lowers Qwen3-Omni-30B-Instruct by 0.08 ± 0.08, to $+0.18$, which is $+0.04 \pm 0.08$ above the baseline of that model and therefore indistinguishable from it. For every other model the empty-history value equals the identifier-shuffle value, because those models gain nothing from outcomes within the current episode. No model has an empty-history TCS-NV above the baseline of that model, and every Memory Score is within 0.02 of zero, as expected when nothing is carried.

Decomposition for the best model. Paired between runs, the paired margin of Qwen3-Omni-30B-Instruct, $+0.12$ over the baseline of that model, splits into 0.00 ± 0.08 carried across episodes, 0.08 ± 0.08 from outcomes within the current episode, and 0.04 ± 0.08 from claim content and the current view. These parts are the differences neutral minus shuffle, shuffle minus empty history, and empty history minus baseline, each computed from unrounded values. A per-episode history reset, run for Qwen3-Omni-30B-Instruct only, keeps outcomes from within each episode and removes cross-episode outcomes, with no misattribution. The model scores $+0.26 \pm 0.09$ under this reset, the identifier-shuffle value, so misattributed history does not itself lower TCS-NV.

Exposure counts and weighting. The two-exposure filter removes no test informant for any model. Each test informant appears in three to seven of the 27 evaluation episodes, with 27 appearances per modality and a median of 4.5. Every $T_p(k)$ therefore has three to seven exposure points, and six informants reach $k \geq 4$. MS weights informants equally and, within an informant, weights every exposure $k \geq 1$ equally, whatever the number of decisions at that exposure. Nine

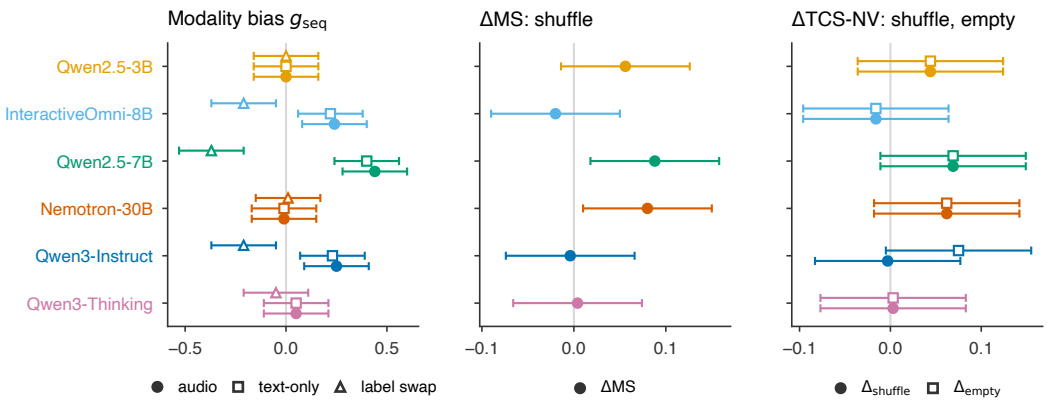


Figure 4: **Control conditions** (see Table 7 and Table 8). Left: the sequence modality bias g_{seq} with audio, with the spoken claim shown as text, and with the labels swapped, each with its 95% interval. Middle: the drop in Memory Score when informant identifiers are shuffled. Right: the drop in TCS-NV under the identifier shuffle and under an empty verification history. Points are seed means; bars are wild cluster bootstrap intervals.

Table 9: Forced declines per model as a share of decisions by cause, TCS-NV with forced declines excluded, ECE on unverified ($\mathcal{N}$), verified ($\mathcal{V}$) and act-only decisions, and post-look decision accuracy by fact kind in the order count, spatial relation and containment. Entries are means over three seeds.

Model	turn cap	budget	illegal	TCS-NV excl.	ECE $\mathcal{N}$	ECE $\mathcal{V}$	ECE act	post-look acc.
Qwen2.5-Omni-3B	3%	2%	1%	−0.04	0.29	0.22	0.25	0.76 / 0.52 / 0.71
InteractiveOmni-8B	2%	1%	1%	−0.02	0.40	0.33	0.36	0.82 / 0.60 / 0.78
Qwen2.5-Omni-7B	3%	2%	1%	+0.19	0.25	0.18	0.21	0.88 / 0.67 / 0.83
Nemotron-3-Nano-Omni-30B	3%	2%	1%	+0.00	0.29	0.22	0.25	0.69 / 0.42 / 0.64
Qwen3-Omni-30B-Instruct	2%	1%	1%	+0.26	0.34	0.27	0.30	0.78 / 0.54 / 0.73
Qwen3-Omni-30B-Thinking	2%	2%	1%	+0.02	0.31	0.24	0.27	0.90 / 0.70 / 0.85

of the twelve informants reach $k = 3$, and the other three stop at $k = 2$. Capping k at 3 gives $+0.10 \pm 0.10$ and $+0.09 \pm 0.10$ for Qwen2.5-Omni-7B and Nemotron, against $+0.11$ and $+0.10$ without the cap.

Reference agents. Scripted agents are reported with the ground-truth perception and with a fixed vision-language reader, that of Qwen3-Omni-30B-Thinking, and only the second gives values comparable to models that use the camera. With the ground-truth perception, the trust-all policies score TCS-NV $+0.13$. AlwaysTrustSpoken scores $+0.14$ and AlwaysTrustWritten scores $+0.12$, because the two informant pools have matched realised reliability. AlwaysDecline scores -0.13. The informant-level oracle scores 0.67 with MS $+0.01$, because it knows every R_p from the start, and the group-level oracle scores 0.83.

Bayesian reference policies. The scripted policies are not run in the simulator. We simulate them from the generator parameters on the evaluation sequence. Each keeps a Beta posterior per cell, starting at $\text{Beta}(1, 1)$ and updated only from verification outcomes, trains no parameters, looks when the posterior mean $\hat{r}$ lies in the look range and the budget covers the look, reads the outcome correctly with the accuracy of its reader, and otherwise acts if $\hat{r} \geq 0.5$. An unverified decision on cell (p, c) scores $+1$ or -1 against a claim that holds with probability $R_{p,c}$. The inputs are the test reliabilities $R_{p,c}$, the group frequencies of 0.45, 0.30, 0.15 and 0.10, the appearance schedule of the 27 evaluation episodes, the 100-step budget and a look cost of 7 to 14 steps. The look cost is a navigation path of 4 to 10 steps, the range of path lengths to a named receptacle in the evaluation houses, plus three or four steps to look. We simulate the sequence 600 times, which puts the Monte Carlo error of every policy mean below 0.01. Each interval of a policy is ± 1.96 standard deviations over these 600 sequences, which redraw the order of appearances, the group and truth of every claim,

the look costs and the reading errors. The ground-truth reader is always correct. Every other reader is correct with the overall post-look accuracy of Table 9, the accuracy by fact kind weighted by the fact-kind mix of the corpus. This accuracy is 0.679, 0.746, 0.809, 0.599, 0.699 and 0.832 for the models in table order. The fixed vision-language reader is that of Qwen3-Omni-30B-Thinking, so its accuracy is 0.832. A policy states the posterior mean $\hat{r}$ as its confidence q_d when it acts and $1 - \hat{r}$ when it declines. The per-informant Bayesian policy scores $+0.72$ with MS -0.08 under the ground-truth perception, because its early looks all score $+1$, and $+0.66$ with MS $+0.05 \pm 0.22$ with the fixed vision-language reader. The group-aware Bayesian policy scores $+0.84$ with MS $+0.06$, and $+0.71$ with MS $+0.11$ with the fixed vision-language reader. The identity-blind feature agent has MS $+0.03 \pm 0.22$, and the identifier-shuffled Bayesian policy has MS -0.08 ± 0.31. Each policy looks at uncertain cells and leaves cells with extreme reliability unverified, so the oracle computed on its own unverified decisions, 0.84 to 0.88, is higher than the pooled oracle, and every policy stays at or below its own oracle. The Bayesian policies choose their own looks with a threshold fixed before any evaluation run, so the comparison with the models covers both verifying and acting.

ALWAYSDECLINE. ALWAYSDECLINE scores between -0.11 and -0.14 on the $\mathcal{N}$ of each model. Every model differs from this policy with a wild interval that excludes zero, except Qwen2.5-Omni-3B and InteractiveOmni-8B, which are indistinguishable from declining every unverified claim.

Look rule. Every Bayesian policy looks when the posterior mean $\hat{r}$ of the claim cell satisfies $b(\hat{r}) \leq \theta$ with $\theta = 0.6$, that is, inside the look range $0.2 \leq \hat{r} \leq 0.8$, and when the remaining budget covers the look (see *Look threshold*). Otherwise the policy acts if $\hat{r} \geq 0.5$ and declines if not. The cell is the informant for the per-informant Bayesian policy and the informant and group for the group-aware Bayesian policy. For the group-agnostic Bayesian policy the cell is the informant, room type and object type, shrunk toward the informant with prior strength 8, which lies between the 6 and 11 claims that an informant makes in its two smaller groups. Finer cells keep posteriors uncertain for longer, so the three policies verify 0.49, 0.59 and 0.60 of decision points with the fixed reader.

Look threshold. We fixed the threshold of the Bayesian policies at $\theta = 0.6$ before any evaluation run and did not tune it. Because no data determined this choice, SELECTIVEORACLE, which uses the same rule with the true $R_{p,c}$, and the group-aware Bayesian policy are reported at $\theta \in \{0.5, 0.6, 0.7\}$. With the post-look accuracy of Qwen3-Omni-30B-Instruct, the group-aware Bayesian policy scores $+0.55$, $+0.52$ and $+0.49$ at the three thresholds, with verify rates 0.59, 0.62 and 0.63. The share of the best model, computed from unrounded values, is 30%, 32% and 35%, and the policy gap stays between 0.24 and 0.30 at every threshold.

Table 10: **Reference agents.** For each scripted reference, the table lists the knowledge given before the sequence, the state carried across episodes, and where results appear. The Bayesian policies are simulated with a reader of accuracy 0.832, the post-look accuracy of Qwen3-Omni-30B-Thinking that serves as the fixed vision-language reader, and choose their own looks. The group-aware Bayesian policy is also simulated with each model's post-look accuracy for the vs. policy column of Table 4.

Agent	Knows in advance	Carries across episodes	Reported
Bayesian, per informant	nothing	one Beta posterior per informant	Tab. 4, App. F
Bayesian, group-agnostic	nothing	one Beta posterior per informant, room type and object type, shrunk toward the informant	Tab. 4, App. F
Bayesian, group-aware	the reliability groups	one Beta posterior per informant and group	Tab. 4
Identity-blind feature agent	nothing	one Beta posterior per object type, ignoring the informant	App. F
Bayesian, identifier-shuffled	nothing	informant posteriors under shuffled identifiers	App. F
SELECTIVEORACLE	every $R_{p,c}$	nothing	Tab. 3
ALWAYSVERIFY, NEVERVERIFY	nothing	nothing	Tab. 3 (ALWAYSVERIFY), §4.3
Informant-level oracle	every R_p	nothing (analytic)	Tab. 4 caption, Tab. 11
Group-level oracle (ORC-NV)	every $R_{p,c}$	nothing (analytic)	Tab. 4, §4.1

Shares on unverified decisions. On $\mathcal{N}$, $\pi_{\mathcal{N}} = 0.563$. Per model, $\pi_{\mathcal{N}}$ ranges from 0.56 to 0.57, and the baseline ranges $+0.12$ to $+0.14$. Models choose ANSWER on 1% to 3% of unverified decisions.

Reliability parameters. The group-level reliabilities $R_{p,c}$ are generator parameters. Per informant in the evaluation split, the four groups hold about 30, 21, 11 and 6 claims and the smallest as few as four, so realised rates in the smallest group rest on few claims. Group-dependent quantities therefore use the generator parameters $R_{p,c}$, while the informant-level R_p comes from the realised claims. The twelve test informants have realised reliabilities $0.12, 0.18, 0.22, 0.25, 0.28, 0.31, 0.82, 0.86, 0.89, 0.92, 0.94, 0.96$. The six unreliable informants are generated with error rates of 0.69 to 0.88, with every error false and off-topic errors disabled. For these six, $R_p \approx 1-$error rate up to sampling, which is why the realised band of the second corpus reaches 0.38. The evaluation split has 49% false or off-topic claims, against 29% in the training split. The six unreliable informants are balanced across gender at 3/3 and age group at 2/2/2, with at most two per language.

Derived shares and oracles. From these values the evaluation split is 51% true, 44% false and 5% off-topic. The share of claims in $\mathcal{N}$ that hold, counting true and off-topic claims, is $\pi_{\mathcal{N}} = 0.563$. The trust-all baseline is $+0.13$, and the informant-level oracle is $\frac{1}{12}\sum_p b(R_p) = 0.67$. The group-level oracle is ORC-NV of Eq. 6 pooled over the unverified decisions of all models, 0.83.

Labels at L1 and L2. L1 gold labels are 38% consistent, 33% contradicted, 29% unverifiable. L2 is scored on the 148 same-fact disagreements. Of the 107 pairs in which exactly one claim holds, the frame settles 64, with the spoken claim true in 34 and the written claim in 30, and cannot settle 43; these 43 and the 41 pairs in which both claims are false are NEITHER items, 84 in all. Pairs about different facts are not conflicts. SELECTIVEORACLE, with the rule $b(R_{p,c}) \leq \theta$ of §4.3, verifies 20% of decision points at $\theta = 0.6$, 5% at 0.5 and 27% at 0.7.

Second evaluation corpus. A second evaluation corpus comes from the same stratification process with a fresh seed, 27 new episodes in 27 new houses and twelve newly generated test informants, with realised reliabilities 0.14 to 0.38 and 0.81 to 0.97. The corpus gives $\pi_{\mathcal{N}} = 0.575$ and a baseline of $+0.15$. Every model's TCS-NV, MS and CRA change by at most 0.02, 0.01 and 0.01, and the TCS-NV ordering is unchanged. Because the baseline is higher and TCS-NV lower, the margin of Qwen3-Omni-30B-Instruct over the baseline of that model falls from $+0.12$ to $+0.07 \pm 0.09$, a fall of about 0.05 from unrounded endpoints, computed from unrounded values, and no longer excludes zero, so no model is above the baseline on this corpus.

Test informants drawn like the training pool. To measure the shift introduced by stratification, we also evaluate the same sequence with twelve test informants drawn by the training-pool process. This process enables off-topic errors and uses error rates in the training range, which gives reliabilities of 0.42 to 0.96. There $\pi_{\mathcal{N}} = 0.73$, the trust-all baseline is $+0.46$, and the informant-level oracle is 0.50, a headroom of only 0.04. With reliabilities clustered near and above 0.5, knowing an informant is worth almost nothing beyond believing everything. Every model's TCS-NV rises in absolute terms, to $+0.45$ to $+0.47$, but no model is distinguishable from its own baseline, and none reaches the informant-level oracle. The stratified split is what gives trust enough headroom to be measured.

Concordance across levels. The seven metrics GS, VF1, CRA, VE, MS, TCS-NV and $\mathrm{ECE}_{\mathrm{act}}$, the last reversed so that higher is better, rank the six models differently. Kendall's coefficient of concordance W, which ranges from 0 for no agreement to 1 for identical rankings, is 0.19. With 7 metrics and 6 models this gives $\chi^2 = 6.65$ on 5 degrees of freedom, with a p-value of about 0.25, and ties are handled by mid-ranks. The mean Kendall rank correlation r_{K} over the 21 pairs of metrics is 0.03. We use the τ_a form of r_{K}, in which a tied pair counts as zero, and write r_{K} to keep τ for trust correctness. The largest agreements are between MS and $\mathrm{ECE}_{\mathrm{act}}$, at 0.93, and between VE and TCS-NV, at 0.67, and the largest disagreement is between VE and CRA, at -0.87, because the two models that answer NEITHER on every conflict score highest at L2 and lowest at L3. These rank statistics depend only on the orderings of the seed means, which every monotone rescaling of the metrics preserves.

Baselines and reference policies in detail. On the training split, the share of claims that hold is $\pi = 0.708 + 0.143 = 0.851$, counting true and off-topic claims. The training-split baseline is therefore $+0.70$ (Table 5). On the unverified test decisions $\mathcal{N}$, the share is $\pi_{\mathcal{N}} = 0.563$ and the baseline is $+0.13$. Both follow from the twelve test reliabilities, and per-model values are within 0.02. ORC-NV takes the best decision without looking for each informant, so ORC-NV $\geq 2\bar{R}_{\mathcal{N}} - 1$, where $\bar{R}_{\mathcal{N}}$ is the mean reliability behind the decisions in $\mathcal{N}$. The group-level value is 0.83. The informant-level value, the equal-weight mean of $b(R_p)$ over the twelve informants, is 0.67, and the decision-weighted value is within 0.01. The two values differ because group-level rates vary around the base rate of each informant. As a result, 10 of the twelve test informants have group reliabilities on both sides of 0.5, although six are unreliable and six reliable overall. ALWAYSVERIFY verifies every decision point until the remaining budget cannot afford the next look, so it leaves no unverified decision that could inflate TCS-NV. SELECTIVEORACLE verifies exactly when $b(R_{p,c}) \leq \theta$. At $\theta = 0.6$, 20% of decision points qualify, against 5% at 0.5 and 27% at 0.7. SELECTIVEORACLE and ALWAYSVERIFY also run with a fixed vision-language reader, that of Qwen3-Omni-30B-Thinking, so that they can be compared with models that use the camera. With the ground-truth perception, the two policies reach $+0.167$ and $+0.089$, and the fixed vision-language reader gives $+0.134$ and $+0.056$, the ground-truth values less 0.033 per step. A misread turns a look worth $+1$ into -1, so the reader loses $2(1 - a_{\text{read}})/\bar{s}$ per step, with reader accuracy $a_{\text{read}} = 0.832$ and mean look cost $\bar{s} = 10.3$ steps.

Calibration error. Calibration over the set $\mathcal{E}$ of all ACT and DECLINE decisions is

$$\text{ECE} = \sum_{m=1}^{10} \frac{|B_m|}{|\mathcal{E}|} \left| \operatorname*{mean}_{d \in B_m} \mathbb{1}[\tau_d = +1] - \operatorname*{mean}_{d \in B_m} q_d \right|. \tag{9}$$

B_m is the set of decisions in $\mathcal{E}$ with stated confidence q_d in the m-th of ten equal-width bins. ECE_{act} restricts $\mathcal{E}$ to ACT decisions. The prompt asks the agent for a stated trust probability for the chosen action, and we score this value as the probability that the chosen action is correct. On a decline, the stated confidence q_d is read as the probability that declining is correct, that is, that the claim is false. ECE over $\mathcal{N}$ and over $\mathcal{V}$ uses the same reading (Table 9). A model that instead reports trust in the claim when declining is scored as miscalibrated on those decisions. This is why the headline calibration measure, ECE_{act}, scores act decisions only.

Gap decomposition and the per-informant Bayesian policy. The pooled oracle, at 0.83, uses prior knowledge of every $R_{p,c}$. Qwen3-Omni-30B-Instruct covers 18% of the headroom between the baseline and the pooled oracle, computed from unrounded values. Computed from unrounded values, the gap between the best model and the pooled oracle is 0.57, and it splits into three parts that are also computed from unrounded values. The learning cost of the group-aware Bayesian policy, from the pooled oracle to that policy with the ground-truth perception, is -0.01 ± 0.05, within noise of zero, because that policy matches the pooled oracle. Perception costs $+0.32 \pm 0.10$, from the ground-truth perception to the post-look accuracy of the model. The policy gap, from that policy to the model, is 0.27, the paired difference -0.27 ± 0.09. The simpler per-informant Bayesian policy scores $+0.66$ with the fixed vision-language reader and $+0.72$ with the ground-truth perception. The best model reaches 23% and 21% of the gains of these two policies over the baseline.

Modality bias in detail. Constant NEITHER scores 0.57 at L2, above every always-trust baseline, and no model beats it. Qwen2.5-Omni-3B and Nemotron equal it by answering NEITHER on nearly every conflict, with p_{sp} and p_{wr} of at most 0.01, which also leaves them no measurable modality bias. On the episode sequence the same bias is measured as g_{seq}, the rate of following a false spoken claim minus the rate of following a false written claim. For Qwen2.5-Omni-7B, InteractiveOmni-8B and Instruct, g_{seq} is $+0.44$, $+0.24$ and $+0.25$ with audio, each ± 0.16, $+0.40$, $+0.22$ and $+0.23$ as text, and -0.37, -0.21 and -0.21 with swapped labels, so all nine intervals exclude zero. In the text-only variant the spoken claim is shown as a transcript under the SPOKEN label, which separates the modality from the label. In the label-swap variant the two informants exchange labels and modalities, which separates the label from the informant. Scoring keeps the original modality of each informant, so a bias that follows the label flips sign. Thinking follows false claims from both informants, with a small g_{seq} of $+0.05$. The direction is opposite to the text dominance reported for image-text and audio-text conflicts (Deng et al., 2025; Wang et al., 2025).

Acting on all decisions. TCS-NV conditions on the choice of each agent about which claims to leave unverified, and verify rates range from 0.29 to 0.75. Mean trust correctness over all test decisions, verified ones included, is $+0.20$, $+0.12$, $+0.51$, $+0.11$, $+0.30$ and $+0.33$ in table order. A verified decision scores $2a_{\mathrm{read}} - 1$ in expectation, where the reading accuracy a_{read} is the overall post-look accuracy of the model in Table 9. On this measure Qwen2.5-Omni-7B is best, because it verifies three decisions in four and reads the frame accurately. The group-aware Bayesian policy with the post-look accuracy of Qwen3-Omni-30B-Instruct scores $+0.44$, and against the all-decision trust-all baseline of $+0.11$ the best model at L5 reaches 58% of the gain of that policy, against 32% on unverified decisions alone. Unverified decisions are where the models fall furthest behind the policy.

Small-cluster inference. A bootstrap by informant resamples only twelve test informants, and with so few clusters it understates uncertainty. Every interval in this paper is therefore a wild cluster bootstrap-t interval with Webb six-point weights and 9,999 draws. Intervals are clustered by informant at L3 to L5 and by episode at L0 to L2. A t-interval with 11 degrees of freedom on the per-informant contributions is narrower in every case. Relative to the informant bootstrap, the marginal TCS-NV half-widths grow from ± 0.07 to ± 0.09 and the Memory Score half-widths from ± 0.08 to ± 0.10, and paired differences widen by about a third. The paired margin of Qwen3-Omni-30B-Instruct over its baseline, $+0.12 \pm 0.09$, still excludes zero, and Qwen2.5-Omni-7B, at $+0.05 \pm 0.09$, remains indistinguishable. Nemotron and Thinking remain below their baselines, at -0.14 ± 0.09 and -0.13 ± 0.09. CRA half-widths grow from ± 0.09 to ± 0.11. One result changes relative to the informant bootstrap: the Memory Score of Nemotron, $+0.10$, excludes zero under the informant bootstrap but not under the wild bootstrap, so only the Memory Score of Qwen2.5-Omni-7B excludes zero.

Multiple comparisons. The roughly twenty per-model tests in this paper are reported without a correction for multiple comparisons. A Bonferroni correction over twenty two-sided tests at the 5% level requires a margin of 1.54 half-widths. Three results have margins of at least twice their half-widths: the policy gap of the best model, -0.27 ± 0.09, the shortfall of Thinking against SELECTIVEORACLE, -0.105 ± 0.030, and the modality bias of Qwen2.5-Omni-7B at L2 and in all three conditions of Table 7. Between 1.54 and two half-widths, and so surviving the correction, lie the margins of Qwen2.5-Omni-3B, InteractiveOmni-8B and Nemotron below their baselines, the L2 modality biases of InteractiveOmni-8B and Instruct, $+0.27$ and $+0.28$ against ± 0.15, and the audio bias of Instruct. Below 1.54 half-widths, and so not surviving the correction, lie the margin of Qwen3-Omni-30B-Instruct above its baseline, the margin of Thinking below its baseline, the biases of InteractiveOmni-8B on the episode sequence, the remaining biases of Instruct, the positive verification efficiencies, the Memory Score of Qwen2.5-Omni-7B and every identifier-shuffle drop.

Moderate stratification. To see how the L5 result depends on how unreliable the test informants are, we redraw only the six unreliable test informants with $R_p \in [0.30, 0.45]$ and keep everything else fixed (Table 11). The baseline rises to $+0.27$. The headroom that trust can buy falls to 0.25, between the main split at 0.54 and the training-process split at 0.04. No pair of models reverses its TCS-NV order between the three splits. What changes is whether any model clears the baseline: Qwen3-Omni-30B-Instruct does so in the main split, but its margin falls to $+0.06 \pm 0.09$ in the moderate split and $+0.01 \pm 0.09$ in the training-process split, so no model clears the baseline once the unreliable informants are less extreme.

History-use prompt and the group-agnostic Bayesian policy. Two further references bound how much of the gap to the group-aware Bayesian policy comes from instruction or from privileged information. First, a history-use prompt tells the model to count the held and did-not-hold outcomes of the speaking informant before deciding. TCS-NV under this prompt is -0.02, -0.04, $+0.22$, $+0.03$, $+0.25$ and $+0.01$ in table order, against -0.05, -0.03, $+0.18$, -0.01, $+0.25$ and $+0.01$ under the neutral prompt. Explaining how to use the verification history lifts Qwen2.5-Omni-7B and Nemotron by $+0.04$, with paired half-widths of ± 0.08, which moves 7B to $+0.10 \pm 0.09$ above its baseline. The prompt changes the two Qwen3-Omni models by at most 0.01, because they carry nothing across episodes, and it does not lift the 3B model, InteractiveOmni-8B or Nemotron above the baseline. Second, the group-agnostic Bayesian policy keeps one Beta posterior per informant and per cell of room type and object type. Each cell is shrunk toward the pooled posterior of that

Table 11: **Reliability stratification of the test informants.** The main split draws the six unreliable test informants with $R_p \in [0.12, 0.31]$. The moderate split redraws only those six with $R_p \in [0.30, 0.45]$. The training-process split draws all twelve by the process used for the training pool. The oracle is informant-level, because the training-process split has no group table. The main-split headroom here is therefore the informant-level one, and the group-level figure in §4 is 0.70. Entries are means over three seeds; TCS-NV half-widths are ± 0.09, from a wild cluster bootstrap by informant.

	Main	Moderate	Training-process
Unreliable test R_p	0.12–0.31	0.30–0.45	not stratified
$\pi_{\mathcal{N}}$	0.563	0.635	0.73
Trust-all baseline	+0.13	+0.27	+0.46
Informant-level oracle	0.67	0.52	0.50
Headroom	0.54	0.25	0.04
Qwen2.5-Omni-3B	−0.05	+0.19	+0.45
InteractiveOmni-8B	−0.03	+0.20	+0.45
Qwen2.5-Omni-7B	+0.18	+0.29	+0.46
Nemotron-3-Nano-Omni-30B	−0.01	+0.21	+0.45
Qwen3-Omni-30B-Instruct	+0.25	+0.33	+0.47
Qwen3-Omni-30B-Thinking	+0.01	+0.22	+0.45

informant, so the policy has to infer which cells share a reliability. This policy scores $+0.64$ with the fixed vision-language reader, just below the per-informant policy, because misread outcomes spread over finer cells, and $+0.76$ with the ground-truth perception, between the per-informant and group-aware Bayesian policies (Table 4). The best model reaches 24% of the gain of this policy.

Sensitivity and replication. SELECTIVEORACLE attains VE $+0.128$, $+0.167$ and $+0.157$ at $\theta \in \{0.5, 0.6, 0.7\}$ with the ground-truth perception. With the fixed vision-language reader, the values are $+0.095$, $+0.134$ and $+0.124$. VE peaks at $\theta = 0.6$, because a larger θ spends steps on claims that the prior already settles. The paired difference of Thinking from this oracle is -0.066 ± 0.030, -0.105 ± 0.030 and -0.095 ± 0.030 at the three values of θ. For Qwen2.5-Omni-7B and Qwen3-Omni-30B-Instruct, chosen in advance, budgets $\{60, 100\}$ and turn caps $\{4, 8\}$ change no sign. At budgets 60 and 100, TCS-NV is $+0.14$/$+0.18$ and $+0.21$/$+0.25$ for the 7B model and for Instruct. A turn cap of 4 lowers each value by 0.02 to 0.03 and about doubles the share of forced declines. The minimum Memory Score detectable at the 95% level with twelve test informants is 0.10.

G A WORKED EXAMPLE

Figure 5 follows one evaluation episode under three illustrative policies, which are not model outputs. At decision point 8 the spoken informant says "I think the fork is to the back-left of the potato". The written informant says "I think the teddy bear is to the back-left of the potato". The object there is the teddy bear, so the spoken claim is an identity swap and the written claim is true. An agent that acts on every claim without looking scores $\tau = -1$ at point 8, keeps an empty history, and scores $\tau = +1$ at points 9 to 12 only because those claims happen to be true. An agent that walks to the counter and spends three further steps to look sees the teddy bear, declines the spoken claim with $\tau = +1$, and adds "did not hold" to the history of that informant. The same informant then makes four true claims at points 9 to 12. An agent that declines everything this informant says after one failed check scores $\tau = -1$ on each of the four claims. The per-informant Bayesian policy instead holds the posterior $\mathrm{Beta}(1, 2)$ after the failed check, with $\hat{r} = 0.33$ inside its look range $0.2 \le \hat{r} \le 0.8$. In the figure this policy looks again at point 10 and returns to $\hat{r} = 0.50$ when the claim holds, and it keeps looking while the budget allows. This is why the levels above L3 reward the verification history.

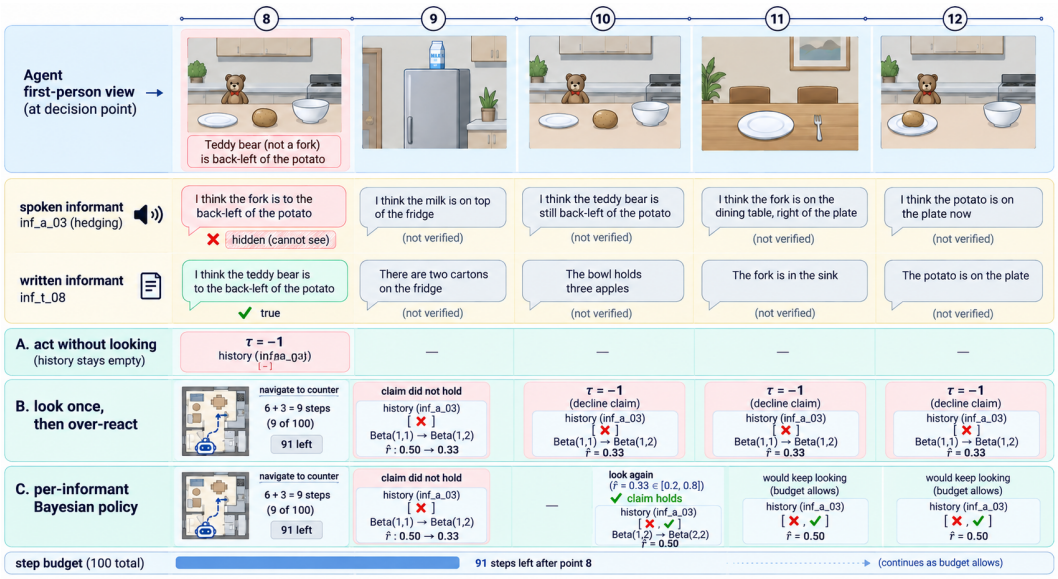


Figure 5: **A worked example** from evaluation episode `1771159560575`, decision points 8 to 12. At point 8 the spoken claim swaps the identity of an object, and the written claim is true. The figure shows three illustrative policies, which are not model outputs. A policy that looks sees the teddy bear, declines the spoken claim, and records "did not hold" in the history of the spoken informant. A policy that acts without looking is wrong at point 8 and learns nothing; a policy that distrusts the informant after one failed check declines four true claims. Appendix G walks through the episode.



Figure 6: **The levels rank the models differently.** One radar per model: its rank at every level, from the seed means in Tables 3 and 4, with the outer ring as first place, $\mathrm{ECE}_{\mathrm{act}}$ ranked from lowest to highest, and the average model in grey; the title of each panel names the levels at which the model ranks first. Ranks come from seed means, and most first places lie within noise (VE 0.030 against 0.029, for example). Kendall's W across the seven rankings is 0.19: no model leads at every level, and one leads at none.

H Memory Curves and Calibration Diagrams

Figure 6 shows the rank of every model at each level. Figure 7 shows the Memory Score curves behind Table 4, and Figure 8 shows TCS-NV against the baseline and the oracles together with the reliability diagrams.

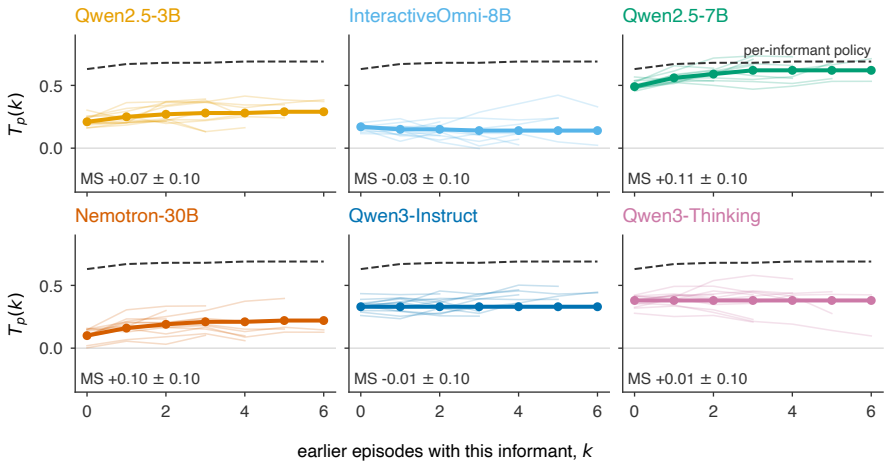


Figure 7: **Memory Score curves.** Each panel shows one model. The thick line is the mean trust correctness $T_p(k)$ on claims by a test informant against the number k of earlier episodes in which that informant appeared; thin lines are the twelve test informants, and the dashed line is the per-informant Bayesian policy simulated with the fixed reader. Only Qwen2.5-Omni-7B rises by more than its interval (see Table 4).

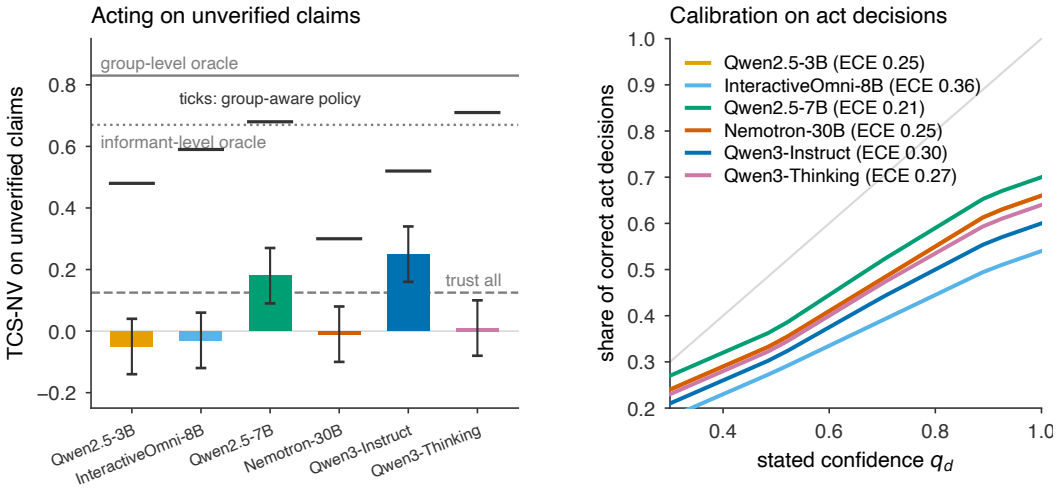


Figure 8: **Acting on unverified claims, and calibration.** Left: TCS-NV of each model with its 95% interval against the trust-all baseline ($+0.13$), the informant-level oracle (0.67) and the group-level oracle (0.83); the black tick above each bar is the group-aware Bayesian policy simulated with that model's post-look accuracy. Four models lie below the baseline, Qwen2.5-Omni-7B is at it and only Qwen3-Omni-30B-Instruct above it. Right: reliability curves on act decisions; Qwen2.5-Omni-7B has the lowest $\mathrm{ECE}_{\mathrm{act}}$ and InteractiveOmni-8B the highest.

I PROMPTS

The system prompt shared by every decision is reproduced below. The single-shot levels L1 and L2 use the prompt directly. The embodied levels L3 to L5 use the same terminal actions inside a turn-based observation loop, with a menu of actions that adapts to the state. The prompt requests a "stated trust probability for the chosen action", and we score this value as the probability that the chosen action is correct.

```
You are an embodied agent evaluating informant claims in
the ``Seeing is Believing'' benchmark.  At each decision
point you may receive:  one TEXT claim (from a text-only
informant), and/or one SPOKEN claim (an attached short
clip).  Both informants are describing the same scene.
Decide for yourself how much to rely on each of them;
you may look before deciding.  You must emit EXACTLY ONE
terminal action:  ``act_on'' (trust the claim and commit),
``decline'' (refuse to commit), ``answer'' (produce a direct
QA answer), ``ask'' (escalate; rare; marked as negative
outcome).  Your output MUST be a single JSON object with
keys terminal_action, target_persona_id, target_channel,
confidence (float in [0,1], your trust probability),
rationale.  The ``confidence'' is your STATED trust
probability for the chosen action -- it is scored for
calibration (ECE), not just accuracy.  Be honest.  If you
attempt two consecutive impossible actions, or reach 8 turns
without a terminal action, the decision point ends with a
neutral decline and no verification credit.
```

The prompt is verbatim, so the key names keep the wording used during evaluation. Each user turn lists the decision-point index, the informant identifier, the verification history of that informant, the text of the written claim, the spoken claim as an audio clip with no transcript and no printed slots, and the remaining verification budget. The turn states no delivery style, because the audio carries the style of the spoken informant and the wording carries the style of the written one. After a look, the agent receives a post-look turn. The required output of this turn is a JSON object with the keys `verdict`, either "held" or "did_not_hold", and `percept`, a free-text description, followed by the terminal decision object. The harness parses the verdict separately from the terminal JSON, and the verdict is the verification outcome that enters the history.

J VALIDITY ACCOUNTING

All 27 evaluation episodes were SCORED at every level for every model, with two exceptions at L0. Both exceptions occur under the longest generation limit, which is the maximum number of output tokens per turn, swept over 1,024, 2,048 and 4,096 tokens. At the longest limit only 7 of 27 L0 episodes completed for Qwen3-Omni-Thinking and 4 of 27 for Nemotron in each seed, so we report the complete cell at the shorter limit. An episode is VACUOUS when it has nothing to score, for example no conflict at L2, and no episode in the released corpus was vacuous at L2. Every evaluation episode has fifteen decision points, so each sequential run of one seed comprises exactly 405 decisions. One training episode in the release has a single decision point. The sequence of each model was run with three sampling seeds at temperature 0.7, and all three runs completed for every model.