ELICITING LLM PROPOSITIONAL BELIEFS IN AGENTIC TASKS: CONSISTENCY, ACCURACY, AND RESPONSE TENDENCIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly deployed as autonomous agents in interactive environments that require multi-step planning, reasoning, and state tracking. Evaluating an agent's beliefs about environment states and task progress can therefore help diagnose behavioral failures. However, reference belief labels are often unavailable or difficult to obtain. In this paper, we introduce a general, LLM-assisted evaluation pipeline that combines proposition generation with reference labels. The pipeline grounds LLM-authored proposition templates in agent trajectories and obtains labels from an LLM-assisted reference observer using formal logical rules and Bayesian inference. Using this pipeline, we systematically evaluate models' self-reported belief consistency and accuracy across World, Epistemic, Affordance, and Decision propositions in ALFWorld, WebShop, and Noisy Cabinets and examine their relationship with task success. T/F-only Affordance and Decision accuracy are positively associated with task success after centering within environments. Furthermore, we find that response tendencies help explain both observed consistency and accuracy variation: the tendency to answer Unknown contributes to cross-model differences in consistency, and a True/False tendency model explains degradation in accuracy and negation consistency as context grows.

1 INTRODUCTION

Large language models (LLMs) are increasingly deployed as autonomous agents that act (Yao et al., 2023) in partially observable environments, from household simulators (Shridhar et al., 2021) to online shops (Yao et al., 2022). In these agentic environments, agents' internal beliefs or elicited confidences (Pal et al., 2026) are frequently used as summaries for their action behavior (Yamin et al., 2026) and to diagnose task failure modes. For example, Wang et al. (2026) find that agents can exhibit belief inertia, where they maintain outdated beliefs despite new evidence, which will lead agents to perform ineffective actions. Therefore, systematically evaluating agents' beliefs is crucial for understanding their decision-making processes.

Conceptually, a faithful belief state must satisfy two criteria: **accuracy** and **consistency**. Accuracy reflects whether an agent's beliefs match the physical world, which is foundational to self-assessment (Kadavath et al., 2022) and world modeling (Chae et al., 2025). In contrast, belief consistency evaluates internal logical cohesion across queries (Elazar et al., 2021), semantic entailments (Jang et al., 2022), and temporal updates (Betz & Richardson, 2023; Tsujimura et al., 2026; Kassner et al., 2021). Despite their importance, simultaneously evaluating general belief accuracy and consistency in agentic settings introduces a fundamental challenge: *how to define belief ground truth under partial observability?*

To bridge this gap, we propose an LLM-assisted evaluation pipeline (Figure 1; Section 3). An authoring LLM synthesizes proposition templates and formal evaluation rules, grounding them turn-by-turn along recorded trajectories and environment information. To obtain the reference label, we introduce a *reference rational observer* that computes belief labels through Bayesian and logical rules. We systematically generate propositions across four dimensions—*World*, *Epistemic*, *Affor-*

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

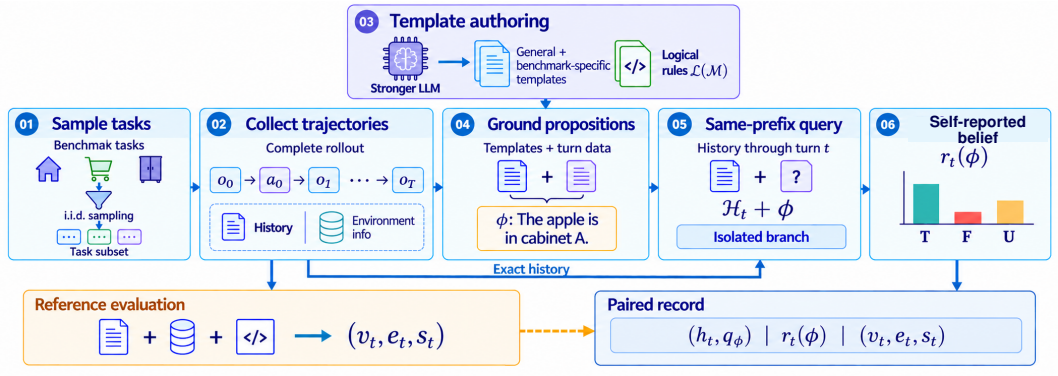

Figure 1: Data pipeline. Policies solve stratified tasks (01–02) from different environments; LLM-authored templates and logical rules (03) are grounded into propositions at each turn with trajectory context and environment information (04); Same-prefix queries (05) elicit self-reported beliefs (06). The reference evaluation module generates labels v_t and probabilities e_t from the trajectory context and environment information.

dance, and *Decision*—in three representative benchmarks: ALFWorld, WebShop, and Noisy Cabinets.

Using this pipeline, we benchmark models across normative accuracy and multiple axes of consistency (permutation, negation, sequential, and rephrasing; Section 4). Beyond aggregate performance, our analysis reveals that internal belief accuracy provides strong diagnostic signals for task success. Furthermore, we find models' form-specific biases (affirming ϕ while denying $\neg\phi$; Section 5) that systematically amplify through context length.

In summary, our key contributions are:

- **A Normative Belief Evaluation Framework:** We develop an automated pipeline coupled with a Bayesian rational observer that systematically benchmarks LLM agents' propositional beliefs in partially observable environments.
- **Behavioral Evaluation and Diagnosis:** By evaluating self-reported beliefs across four proposition types and three environments, we show that belief accuracy in *Affordance* and *Decision* states is strongly predictive of downstream task success, empirically validating the diagnostic value of internal beliefs.
- **Characterizing Compounding Response Tendencies:** We uncover that models' varying tendencies for "Unknown" answers explain discrepancies in measured consistency. Furthermore, by formally modeling form-specific tendencies (affirming ϕ while denying $\neg\phi$), we demonstrate that this polarity bias links accuracy with consistency and explains belief degradation through growing context.

2 Related Work

Model beliefs and probing. Recovering latent beliefs from neural representations can be broadly categorized into internal probing and verbalized elicitation. White-box methods train linear probes to detect entity properties (Li et al., 2021; Feng et al., 2025a), discover truth-related subspace directions (Burns et al., 2023), map belief geometry (Shai et al., 2024), or decode internal states into natural language (Pan et al., 2026; Karvonen et al., 2026). However, internal probes require access to model activations. Conversely, verbal elicitation directly probes explicit self-assessments and expressed uncertainties (Lin et al., 2022; Kadavath et al., 2022) in a model-agnostic manner. In this work, we focus on semantically explicit self-reports, pairing grounded propositions with verification rules to get beliefs directly.

Connecting elicited beliefs to agent behavior. A burgeoning line of research investigates whether elicited confidences faithfully reflect an agent's downstream decision-making. Recent studies evaluate this link by testing compatibility with utility maximization (Yamin et al., 2026), inferring latent

beliefs across parallel reasoning continuations (Smolin & Wilder, 2026), simulating social trust dynamics (Mannekote et al., 2025), or measuring strategic deference to user interventions (Pal et al., 2026). While these works establish the correlation between internal beliefs and immediate choices, they did not test whether the agent's beliefs are objectively correct. In partially observable environments, an agent may act coherently according to a severely ungrounded or hallucinated world model. We bridge this gap by benchmarking beliefs against a mathematically normative rational observer, demonstrating that proposition-level accuracy in affordances and decision states can predict overall task success.

Consistency across formulations and trajectories. Evaluating logical consistency in language models traditionally relies on static perturbations, including semantic paraphrasing (Elazar et al., 2021), negation symmetry (Jang et al., 2022), probabilistic coherence (Betz & Richardson, 2023), and constraint satisfaction (Kassner et al., 2021; Tsujimura et al., 2026). In interactive domains, consistency evaluations have largely focused on behavioral reproducibility across repeated rollouts (Mehta, 2026; Yagubyan, 2026) or target persistence in constrained dialogues (Luo et al., 2026). However, these paradigms neglect how epistemic uncertainty (reporting "Unknown") and sequential evidence accumulation interact. We provide a unified evaluation protocol that evaluates four distinct consistency relations alongside observer-referenced ground truth at exact interaction prefixes, uniquely uncovering how underlying response tendencies and compounding polarity biases systematically erode consistency over extended trajectories.

3 Data Collection and Labeling Pipeline

We first establish the notation and describe proposition generation and trajectory collection (Sections 3.1–3.2). We then explain how reference labels and same-prefix self-reports are obtained (Sections 3.3–3.4).

3.1 Environment and model notation

We model each agent task environment as a partially observable Markov decision process (POMDP; Kaelbling et al., 1998). At turn t, the environment has hidden state $s_t \in \mathcal{S}$, the agent takes action a_t, and then receives observation o_{t+1}. The transition and observation kernels are $T(s' \mid s, a)$ and $O(o \mid s', a)$. With task instruction g, the available interaction history is $H_t = (g, o_0, a_0, \ldots, a_{t-1}, o_t)$. A frozen model with parameters θ induces the policy $a_t \sim \pi_\theta(\cdot \mid H_t)$.

Let $X_{t,1:c}$ be the exact recorded prefix at token cutpoint c. It contains the policy-visible history; in the thinking condition, it also includes the current turn's reasoning and action text (Appendix B.1). Appendix A.3 specifies the associated representations.

3.2 Proposition generation and trajectory collection

We use an authoring LLM to generate proposition templates $\mathcal{M}$ and corresponding generic evaluation rules $\mathcal{L}(\mathcal{M})$ (block 3 in Figure 1). Combining these templates with environment information yields grounded propositions ϕ and their evaluation programs. We ask the authoring LLM to generate four types of propositions: *World* describes physical or session facts; *Epistemic* describes observed or known information; *Affordance* specifies immediate action preconditions; and *Decision* assesses task progress or heuristic priorities.

We instantiate the pipeline in three environments: ALFWorld (Shridhar et al., 2021), WebShop (Yao et al., 2022), and Noisy Cabinets (NC).[1] We draw stratified task samples and collect trajectories from six model conditions: Qwen3-8B, Qwen3-8B (think), and Qwen3-14B (Qwen Team, 2025); Llama-3.1-8B-Instruct (Llama-3.1-8B) (Llama Team, 2024); and WS-RL and AF-RL, two Qwen3-8B checkpoints fine-tuned on WebShop and ALFWorld, respectively.

3.3 Reference labeling

For a grounded proposition ϕ, world truth $Y_t^\phi = \phi(s_{0:t}) \in \{0, 1\}$ is determined by the state history $s_{0:t} = (s_0, \ldots, s_t)$. Because this history is hidden, obtaining world truth from the interaction history

[1] NC is an environment we created with noisy observations; see Appendix C.

alone is not always possible. We therefore use the coding capabilities of the authoring LLM to construct a **reference rational observer** based on the recorded history H_t, token prefix $X_{t,1:c}$, and benchmark metadata $\mathrm{BasicInfo}$.[2] The information sets for the observer and world truth are

$$\mathcal{F}_t^{\mathrm{ref}} = \sigma(H_t, X_{t,1:c}, \mathrm{BasicInfo}), \qquad \mathcal{F}_t^{\mathrm{world}} = \mathcal{F}_t^{\mathrm{ref}} \vee \sigma(s_{0:t}). \tag{1}$$

Stochastic evaluation. The observer follows Bayesian inference in stochastic environments such as NC, updating its state belief $b_t(s) = P(s_t = s \mid \mathcal{F}_t^{\mathrm{ref}})$ by

$$\begin{aligned} \widetilde{b}_t(s') &= \sum_{s \in \mathcal{S}} T(s' \mid s, a_{t-1}) b_{t-1}(s), \\ b_t(s') &= \frac{O(o_t \mid s', a_{t-1}) \widetilde{b}_t(s')}{\sum_{\bar{s} \in \mathcal{S}} O(o_t \mid \bar{s}, a_{t-1}) \widetilde{b}_t(\bar{s})}. \end{aligned} \tag{2}$$

For a current-state proposition, this gives the reference probability

$$e_t(\phi) = P(Y_t^{\phi} = 1 \mid \mathcal{F}_t^{\mathrm{ref}}) = \sum_{s \in \mathcal{S}} b_t(s) \mathbf{1}[\phi(s) = 1]. \tag{3}$$

Deterministic evaluation. In deterministic environments such as ALFWorld and WebShop, the observer computes labels using the deterministic rules generated by the authoring LLM.

For example, the ALFWorld template "you still need to search for κ" is evaluated by

$$d_{\mathrm{search}}(t, \kappa) = \neg \mathrm{Located}_t(\kappa) \wedge \neg \mathrm{Held}_t(\kappa) \wedge \mathbf{1}[|\mathcal{C}_t^{\mathrm{unexp}}| > 0]\,. \tag{4}$$

Here $\mathrm{Located}_t(\kappa)$ indicates that the observer knows an object's location, $\mathrm{Held}_t(\kappa)$ that the agent holds one, and $\mathcal{C}_t^{\mathrm{unexp}}$ is the set of unexplored receptacles.

Reference labels. When a scalar reference probability $e_t(\phi)$ is available, we define the label $v_t(\phi) \in \{\mathrm{T}, \mathrm{F}, \mathrm{U}\}$ using thresholds $\tau_{\mathrm{F}} = 0.05$ and $\tau_{\mathrm{T}} = 0.95$: $v_t(\phi) = \mathrm{T}$ if $e_t(\phi) \geq \tau_{\mathrm{T}}$, $v_t(\phi) = \mathrm{F}$ if $e_t(\phi) \leq \tau_{\mathrm{F}}$, and $v_t(\phi) = \mathrm{U}$ otherwise. When no scalar reference probability is available, the corresponding evaluation rule directly assigns a T/F/U label. We add an Unknown label which means that the available information does not permit a sufficiently certain true/false judgment. The same observer rules and thresholds apply to every model. Appendix C.4 compares reference labels with recorded world truth and checks the stochastic observer.

3.4 Internal beliefs and same-prefix self-reported beliefs

Let $b_{\theta,t}^{\mathrm{int}}(\phi)$ denote the model's internal belief in ϕ. This belief is not directly observable, and the concept may not apply well to LLMs (Smolin & Wilder, 2026). Even if it exists, it may not be Bayesian or logically coherent. Therefore, we measure a proxy by appending the proposition to the recorded token prefix and reading its next-token probabilities over the answer letters.

$$r_t(\phi) = \big(r_t(\mathrm{T} \mid \phi), r_t(\mathrm{F} \mid \phi), r_t(\mathrm{U} \mid \phi)\big) \in \Delta^2. \tag{5}$$

We call this a self-reported belief. The corresponding hard label is $y_t(\phi) = \arg\max_{k \in \{\mathrm{T},\mathrm{F},\mathrm{U}\}} r_t(k \mid \phi)$. Causal attention prevents positions from attending to later tokens (Vaswani et al., 2017), so appending a query preserves the exact prefix's hidden representations.

For fixed model parameters and query, the model-accessible information is $\mathcal{F}_t^{\mathrm{model}} = \sigma(X_{t,1:c})$. Obviously we have $\mathcal{F}_t^{\mathrm{model}} \subseteq \mathcal{F}_t^{\mathrm{ref}} \subseteq \mathcal{F}_t^{\mathrm{world}}$. The observer's information set lies between the model's and the world's, allowing it to provide reference labels that are easier to obtain than world truth for evaluating self-reported belief accuracy. Appendices B and C give the elicitation and reference-labeling details.

4 Consistency and Accuracy of Self-Reported Beliefs

Using these self-reports, we define the evaluation measures (Section 4.1), compare model performance (Section 4.2), and examine how accuracy relates to task success (Section 4.3).

[2] This may include state-related information that is not generally observable from the interaction history but is available from the benchmark, such as the numerical probabilities of stochastic events and the set of possible objects. See Appendix C.2 for Noisy Cabinets.

Table 1: Self-reported belief agreement (Agg, %) on the reference-complete subsets. **Left:** full query/pair sets, including U reports. **Right:** pairs with both reports in T/F; UU and all single-U pairs are excluded. Permutation retains all six-order comparisons. Bold: highest value per column.

	Full set				T/F-only pairs		
Model	**Perm.**	**Neg.**	**Seq.**	**Rephr.**	**Neg.**	**Seq.**	**Rephr.**
Qwen3-8B	67.12	72.25	86.70	84.24	79.66	92.99	91.01
Qwen3-8B (think)	70.86	59.94	79.75	82.85	75.43	93.86	94.14
Qwen3-14B	**80.34**	**75.06**	88.57	77.46	**83.01**	93.68	89.90
Llama-3.1-8B	65.97	55.84	88.74	78.53	62.05	**96.08**	88.35
WS-RL	69.53	72.06	88.79	83.36	80.11	94.05	90.83
AF-RL	77.19	73.39	**95.35**	**89.44**	77.13	95.00	**95.09**

4.1 Measurement

Consistency. We evaluate four forms of self-reported belief consistency.[3] *Permutation consistency* measures the fraction of pairs of answer-option orderings that yield the same T/F/U label. *Rephrasing consistency* tests whether the model's belief remains unchanged when a proposition is expressed in different, semantically equivalent ways at the same interaction prefix (Elazar et al., 2021).

We also construct the negated proposition $\neg\phi$ to test *negation consistency* (Jang et al., 2022). This extends the rephrasing test to a logically complementary formulation: T and F should be exchanged, while U should remain U. Finally, *sequential consistency* tests whether the model gives the same answer to an unchanged proposition across queried turns of a trajectory.[4] Consistency is summarized by agreement (Agg), the fraction of answer pairs satisfying the corresponding relation.

Accuracy. Accuracy is the proportion of queries for which the self-reported label matches the reference label. Two-letter outcomes list the reference first and report second: TU/FU denotes Unknown reports against T/F references. T/F-only accuracy requires both labels to be definite; accuracy w/o TU/FU excludes only those Unknown reports. Appendices D.1 and E.1 give the pairing rules, evaluation populations, and detailed definitions of consistency and accuracy, respectively.

4.2 Evaluations

Consistency and accuracy comparison. Table 1 reports overall agreement under the four relations; sequential agreement pools WebShop and NC, since ALFWorld has no qualifying pairs. The larger Qwen3-14B performs particularly well on permutation and negation; its negation result is consistent with evidence that larger models can handle negation better in multilingual textual entailment (Vrabcová et al., 2025). It also achieves the highest overall belief accuracy in the Overall column of Table 2.

The observer supplies a common scoring rule, while TU/FU identify Unknown reports against definite references, including possible information-gap cases. Overall accuracy is a conservative lower bound under an adjustment that excludes any subset of these mismatches (Appendix E.1). Excluding all TU/FU gives the final column of Table 2.

Model tendency to answer Unknown. Because the tendency to answer Unknown affects consistency and accuracy, we conduct two analyses.

First, we examine how the model's Unknown responses change with the prompt and how this affects agreement with a commonsense answer key. Using the elicitation pipeline in Section 3, we query task-irrelevant commonsense propositions, such as "Paris is the capital of France," under identical histories with interaction-only and general-knowledge prompts. Figure 2 shows frequent Unknown

[3] Unless stated otherwise, queries use the interaction-only prompt (Appendix B.2). Analyses are reference-complete: every query and both pair endpoints have T/F/U reference labels; U is valid.

[4] We restrict this test to propositions whose referent, world truth, reference label, and available reference probability remain unchanged over the recorded trajectory.

Table 2: Accuracy of self-reported beliefs (%) across environments. **Left:** T/F-only accuracy for World, Epistemic (Epi.), Affordance (Aff.), and Decision (Dec.) propositions. **Right:** accuracy pooled across categories. Overall uses all nine reference–report outcomes; T/F-only retains TT, FF, TF, and FT; w/o TU/FU removes only reference-T/F, report-U errors and retains UU. Bold: highest value per column.

	Category (T/F-only)				All categories		
Model	**World**	**Epi.**	**Aff.**	**Dec.**	**Overall**	**T/F-only**	**w/o TU/FU**
Qwen3-8B	83.51	78.87	66.89	68.62	43.39	76.18	78.26
Qwen3-8B (think)	78.72	82.36	**70.70**	70.82	48.82	76.42	74.09
Qwen3-14B	**88.36**	**94.55**	55.97	**78.85**	**67.10**	**81.06**	75.48
Llama-3.1-8B	72.21	83.89	55.07	58.66	53.61	69.59	69.96
WS-RL	84.35	78.53	58.93	70.33	43.56	75.48	77.00
AF-RL	87.05	79.39	63.46	66.55	34.82	77.19	**83.07**

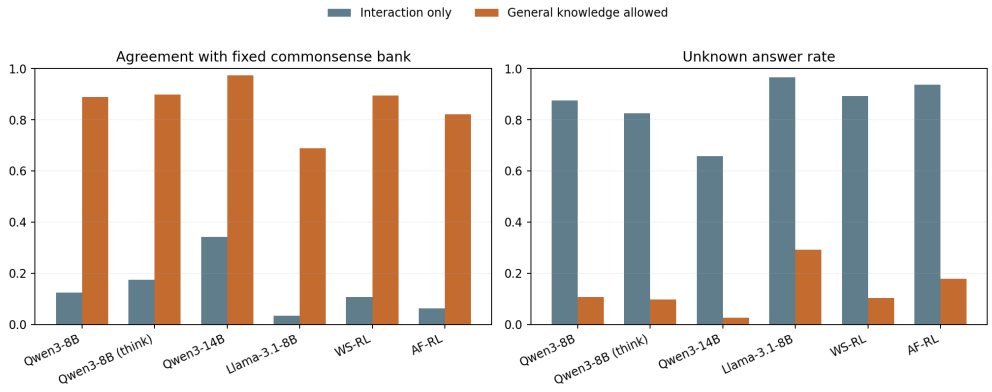

Figure 2: Self-reported beliefs on fixed commonsense assertions under interaction-only and general-knowledge prompts. **Left:** agreement with the commonsense answer key. **Right:** frequency of U self-reported beliefs. Paired bars compare the two prompts for each model; vertical axes show proportions.

responses and low agreement with the answer key under the interaction-only prompt; allowing general knowledge reduces Unknown responses and increases agreement across all models.

Next, we analyze consistency and accuracy after excluding Unknown outcomes. The T/F-only columns of Tables 1 and 2 report the results for definite answers. Sequential and rephrasing consistency are similar across models, while negation consistency still varies. The reduced spread in the first two relations suggests that Unknown-response tendencies contribute to their overall differences. T/F-only accuracy rises for every model, and Qwen3-14B retains the highest pooled accuracy and an advantage in most proposition categories. Model rankings depend on the retained outcomes: AF-RL has the highest accuracy when only TU/FU are excluded. Appendix Figure 9 gives the corresponding retention fractions. Appendix Figure 8 reports each model's Unknown share, and Tables 7, 8, and 10 provide the consistency and accuracy breakdowns by environment and proposition category.

4.3 Belief accuracy and task success

Agent evaluation relates behavioral consistency to correctness (Mehta, 2026; Yagubyan, 2026), while state predictions, action-quality feedback, and reward models support planning (Chae et al., 2025; Miao et al., 2025; Xia et al., 2025). We examine a complementary association: whether more accurate proposition-level judgments accompany higher task-success rates.

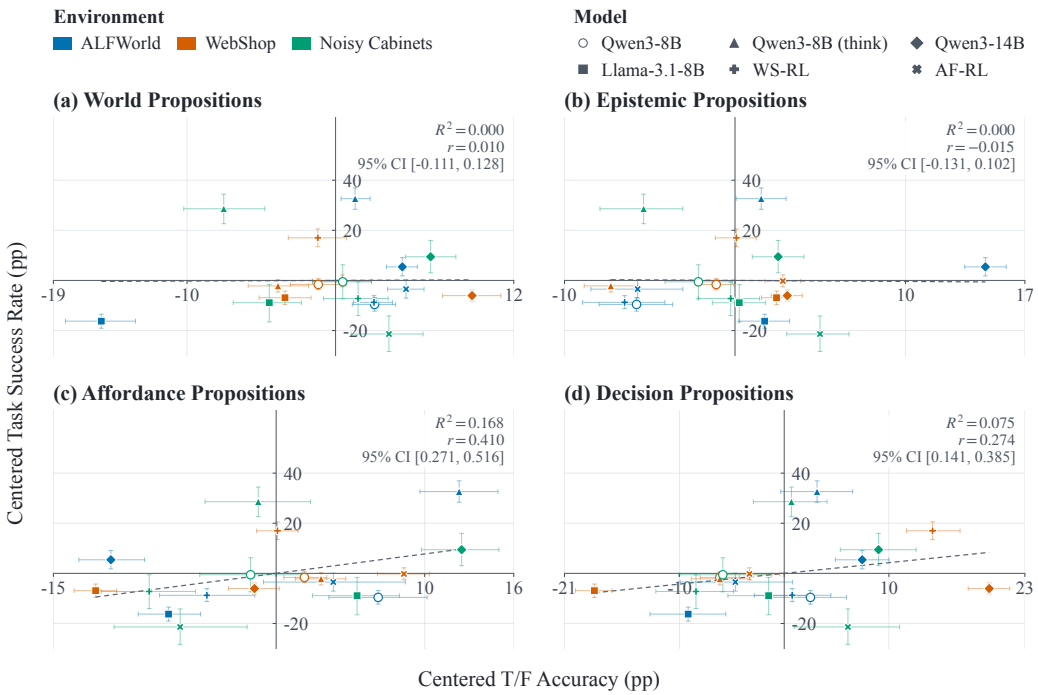

Figure 3: T/F-only category accuracy versus task success. **(a–d)** World, Epistemic, Affordance, and Decision. Both reference and self-report must be T/F; task success uses the original task population. Both axes are centered within each environment. Colors denote environments, markers denote models, and dashed lines show OLS fits; error bars indicate 95% task-cluster bootstrap intervals.

Figure 3 retains queries for which both reference and self-report are T/F, using the same query union as Table 2. Accuracy is computed separately for each category and model–environment condition. Task success uses the original task set, including tasks without a retained query. Both axes subtract their six-model environment means, and each category receives an equal-condition ordinary least squares (OLS) fit.

Affordance accuracy is positively associated with task success, as is Decision accuracy. World and Epistemic associations are near zero, with intervals spanning zero. Thus, within the evaluated conditions, judgments about immediate action feasibility and task progress provide task-relevant diagnostics on a common observer scale. The Affordance result complements feasibility-based planning (Ahn et al., 2023). Both associations remain positive after omitting individual models or environments (Appendix E.4).

5 Model Response Tendencies

We next examine how consistency and accuracy vary with context length (Section 5.1), then develop a response-tendency model to connect the two measures (Section 5.2).

5.1 Consistency and Accuracy Across Context Length

We profile reports along each agent's own trajectory, complementing controlled input-length and logical-extension tests (Levy et al., 2024; Tsujimura et al., 2026). We apply the measures of Section 4 by recorded prefix length at the measurement cutpoint, excluding belief questions. We adapt the trajectory-length bins to each environment. For each model and bin, we compute T/F-only negation consistency and accuracy, with confidence intervals for both metrics. We also evaluate permutation and rephrasing consistency in Appendix F.1.

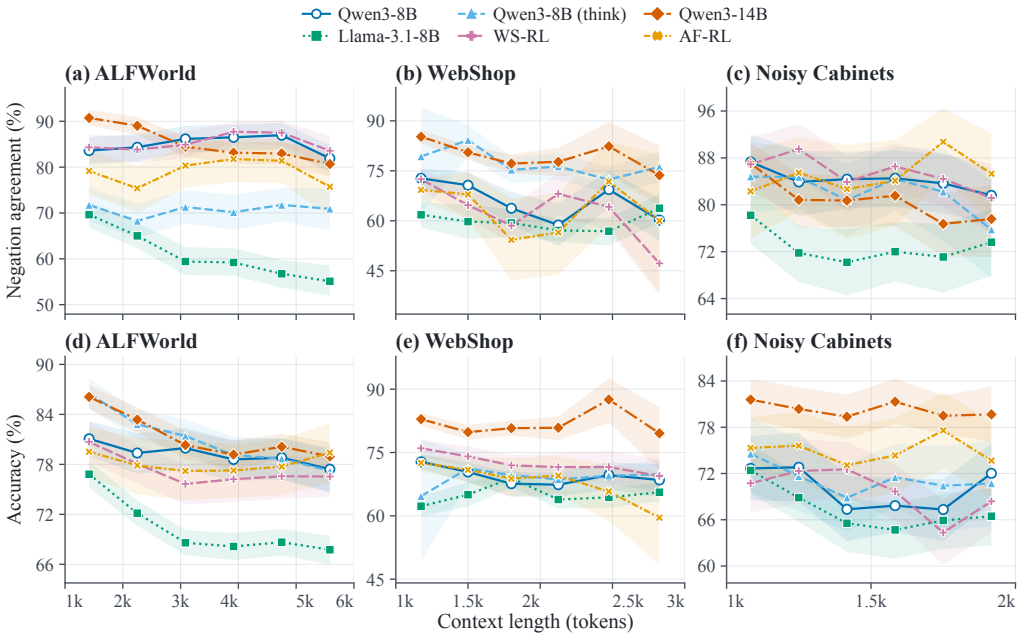

Figure 4: T/F-only negation agreement and accuracy across context lengths. **Top (a–c):** negation agreement. **Bottom (d–f):** accuracy on the Base, Negation, and Rephrasing union. Both reference labels and self-reported beliefs are T/F. Colors and markers denote models; vertical scales differ. Shading shows 95% joint task-cluster bootstrap intervals.

Figure 4 shows an overall tendency toward lower negation agreement and accuracy at longer contexts, most clearly in ALFWorld. Length indexes the accumulated interaction, including the task states reached by the policy. The permutation and rephrasing consistency profiles in Appendix Figure 10 do not show the same declining pattern as negation consistency and accuracy.

We further examine two error types separately: endorsing false-reference ϕ and rejecting true-reference $\neg\phi$. Both error rates tend to increase with context length in Appendix Figure 11. These trends suggest tendencies to endorse ϕ and reject its negated form and motivate the model below.

5.2 Response Tendency Model

We model form-specific tendencies among definite answers as a mixture of fixed responses and reference-following judgments. For ϕ, the report is fixed T with probability α_T, fixed F with α_F, and follows the reference v_ϕ with probability $1 - \alpha_T - \alpha_F$. The negated form uses β_T, β_F. We estimate all four parameters on the same T/F pairs used to evaluate negation.

Under this mixture, only fixed T errs against an F reference, and only fixed F against T. The mixture applies when the two fixed-response weights sum to at most one per form, as they do in every Figure 5 bin. The corresponding class-conditional error rates are $\alpha_T = P(y_\phi = T \mid v_\phi = F)$ and $\alpha_F = P(y_\phi = F \mid v_\phi = T)$ for ϕ, and $\beta_T = P(y_{\neg\phi} = T \mid v_{\neg\phi} = F)$ and $\beta_F = P(y_{\neg\phi} = F \mid v_{\neg\phi} = T)$ for $\neg\phi$.

TT and FF pairs are contradictions; E_{TT} and E_{FF} denote their joint probabilities. With $\pi_T = P(v_\phi = T)$ and $\pi_F = 1 - \pi_T$, let κ be the reference-weighted covariance of the two T-report indicators within reference classes.[5] Joint negation errors become

$$\begin{aligned} E_{TT} &= \pi_T(1-\alpha_F)\beta_T + \pi_F\alpha_T(1-\beta_F) + \kappa, \\ E_{FF} &= \pi_T\alpha_F(1-\beta_T) + \pi_F(1-\alpha_T)\beta_F + \kappa. \end{aligned} \tag{6}$$

[5]Conditional independence gives $\kappa = 0$. For the observed $\kappa < 0$ and positive denominators, this baseline upper-bounds contradiction rates and lower-bounds negation agreement. The sign alone does not establish a reasoning mechanism.

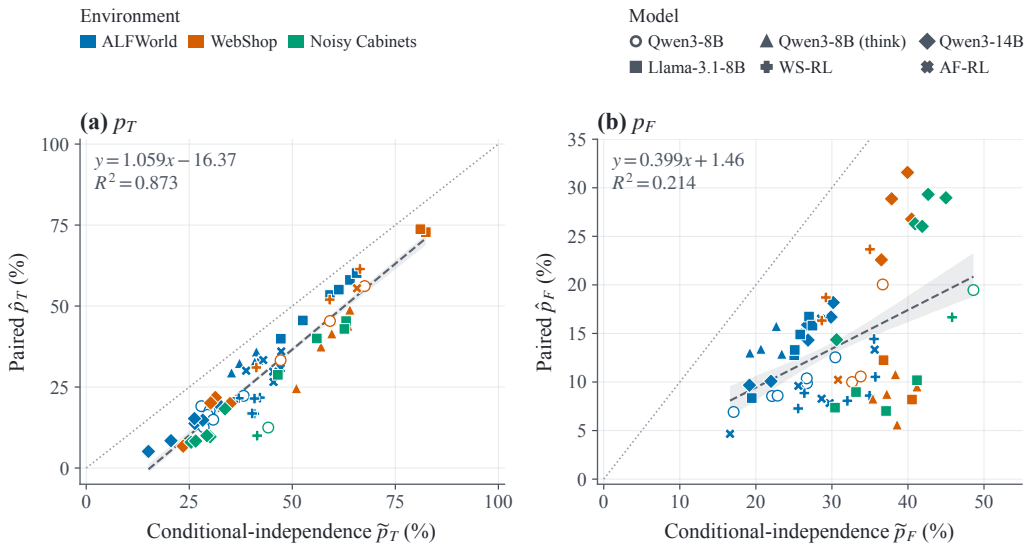

Figure 5: Observed versus accuracy-derived conditional contradiction rates (%): **(a)** p_T and **(b)** p_F (Section 5.2). Points are model–environment–length bins. Colors: environments; markers: models; dashed lines: OLS fits; gray bands: 95% task-cluster bootstrap intervals for the fitted mean; dotted lines: equality. Panel scales differ.

Tildes denote the $\kappa = 0$ baseline. Let $m_T = P(y_{\neg\phi} = T)$ and $m_F = P(y_\phi = F)$. Then $p_T = E_{TT}/m_T$ is the probability of affirming ϕ given affirmation of $\neg\phi$, and $p_F = E_{FF}/m_F$ of rejecting $\neg\phi$ given rejection of ϕ. For positive denominators, they satisfy $p_T = \widetilde{p}_T + \kappa/m_T$ and $p_F = \widetilde{p}_F + \kappa/m_F$, while negation agreement satisfies $\mathrm{Agg}_{\mathrm{neg}} = \widetilde{\mathrm{Agg}}_{\mathrm{neg}} - 2\kappa$.

The decomposition separates marginal errors from paired dependence; its identities also hold when the parameters are interpreted directly as conditional error rates.

Results. Figure 5 links accuracy-derived errors to observed negation errors, strongly for TT contradictions and more weakly for FF contradictions, supporting shared response tendencies as an account of their co-variation. Observations predominantly below $y = x$ indicate negative κ: models maintain negation consistency more often than the independent-response baseline predicts. The marginal-and-dependence framework also extends to permutation and rephrasing (Appendix Figure 14), although a shared-marginal predictor fits these relations less well and transfer requires relation-specific assumptions.

Within-condition and adjacent-bin comparisons support the relationship beyond pooled model differences (Appendix Figure 12). Parameters estimated on training tasks predict conditional contradictions and negation agreement on disjoint test tasks, with substantially lower error when training dependence is included (Appendix Figure 13; Table 16). Together, these results support a shared role for models' response tendencies in explaining changes in both consistency and accuracy.

6 Conclusion

We introduced a framework to benchmark LLM agent beliefs against a rational Bayesian observer in partially observable environments. Accounting for "Unknown" responses is essential for fairly measuring consistency, and that belief accuracy in affordances and decisions directly predicts an agent's downstream task success. Models develop strong form-specific response tendencies over extended trajectories—habitually affirming a proposition ϕ while rejecting its negation $\neg\phi$. Our model jointly explains declining accuracy and increasing logical contradictions through these tendencies, generalizing reliably to unseen tasks. Building dependable agents therefore requires jointly evaluating their beliefs, logical consistency, and uncertainty expression.

AI USE STATEMENT

Generative AI was used to generate proposition templates and the corresponding evaluation-rule code for dataset synthesis (Section 3). AI tools also assisted with GPU performance tuning during data collection, dataset partitioning, writing data-analysis code, and adjusting figure formatting. They were further used to assist with literature searches and to polish and improve the manuscript. All references and their formatting were manually checked by the authors. The authors developed the experimental design, conducted the experiments and analyses, computed all reported numerical results, and wrote the initial manuscript draft. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appendices A–C describe the data collection and labeling pipeline, prompts, and reference-label validation; Appendices D–E provide detailed consistency and accuracy results and task-success analyses; Appendix F presents the context profiles, response-tendency derivations, and within-condition and task-disjoint validation. To support reproducibility, we retain the task and model identifiers, trajectories, exact token prefixes and measurement cutpoints, reference labels, and self-reports, together with the configurations and analysis scripts used to generate the reported tables and figures. The code and data will be uploaded to GitHub.

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 287–318. PMLR, 2023. URL https://proceedings.mlr.press/v205/ichter23a.html.

Gregor Betz and Kyle Richardson. Probabilistic coherence, logical consistency, and bayesian learning: Neural language models as epistemic agents. *PLOS ONE*, 18(2):1–29, 02 2023. doi: 10.1371/journal.pone.0281372. URL https://doi.org/10.1371/journal.pone.0281372.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Hyungjoo Chae, Namyoung Kim, Kai Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee, and Jinyoung Yeo. Web agents with world models: Learning and leveraging environment dynamics in web navigation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 63707–63738, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a00548031e4647b13042c97c922fadf1-Paper-Conference.pdf.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410. URL https://aclanthology.org/2021.tacl-1.60/.

Edgar Erdfelder, Carolina E. Küpper-Tetzel, and Sandra D. Mattern. Threshold Models of Recognition and the Recognition Heuristic. *Judgment and Decision Making*, 6(1):7–22, 2011. doi: 10.1017/S1930297500002060. URL https://doi.org/10.1017/S1930297500002060.

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. Monitoring latent world states in language models with propositional probes. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International*

Conference on Learning Representations, volume 2025, pp. 19337–19359, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3132d0405fabe24b2a7b6cd7ba9de6b5-Paper-Conference.pdf.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 46375–46408. Curran Associates, Inc., 2025b. doi: 10.52202/085713-1544. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/420c9f777c0b4f78d515e53cf74d58b2-Paper-Conference.pdf.

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. BECEL: Benchmark for consistency evaluation of language models. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3680–3696, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.324/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134, 1998. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(98)00023-X. URL https://www.sciencedirect.com/science/article/pii/S000437029800023X.

Adam Karvonen, James Chua, Clément Dumas, Kit Fraser-Taliente, Subhash Kantamneni, Julian Minder, Euan Ong, Arnab Sen Sharma, Daniel Wen, Owain Evans, and Samuel Marks. Activation oracles: Training and evaluating LLMs as general-purpose activation explainers. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=DbZjxkZrZm.

Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. BeliefBank: Adding memory to a pre-trained language model for a systematic notion of belief. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8849–8861, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.697. URL https://aclanthology.org/2021.emnlp-main.697/.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same Task, More Tokens: the Impact of Input Length on the Reasoning Performance of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15339–15353, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.818. URL https://aclanthology.org/2024.acl-long.818/.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. Implicit representations of meaning in neural language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1813–1827, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.143. URL https://aclanthology.org/2021.acl-long.143/.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Llama Team. The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783, 2024. URL https://arxiv.org/abs/2407.21783.

Yifan Luo, Kangping Xu, Yanzhen Lu, Yang Yuan, and Andrew Chi-Chih Yao. Probing the lack of stable internal beliefs in llms, 2026. URL https://arxiv.org/abs/2603.25187.

Amogh Mannekote, Adam Davies, Guohao Li, Kristy Elizabeth Boyer, ChengXiang Zhai, Bonnie J. Dorr, and Francesco Pinto. Do Role-Playing Agents Practice What They Preach? Belief-Behavior Consistency in LLM-Based Simulations of Human Trust. arXiv preprint arXiv:2507.02197v1, 2025. URL https://arxiv.org/abs/2507.02197v1.

Aman Mehta. When agents disagree with themselves: Behavioral consistency as an uncertainty signal for llm agents, 2026. URL https://arxiv.org/abs/2602.11619.

Bingchen Miao, Yang Wu, Minghe Gao, Qifan Yu, Wendong Bu, Wenqiao Zhang, Yunfei Li, Siliang Tang, Tat-Seng Chua, and Juncheng Li. Boosting Virtual Agent Learning and Reasoning: A Step-Wise, Multi-Dimensional, and Generalist Reward Model with Benchmark. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 44054–44075. PMLR, 2025. URL https://proceedings.mlr.press/v267/miao25b.html.

Arka Pal, Teo Kitanovski, Arthur Liang, Akilesh Potti, and Micah Goldblum. Knowing what you know is not enough: Large language model confidences don't align with their actions, 2026. URL https://arxiv.org/abs/2511.13240.

Alexander Pan, Lijie Chen, and Jacob Steinhardt. Latentqa: Teaching llms to decode activations into natural language. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 78726–78759, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/7f19b99e63762d20e9df91144056f1ee-Paper-Conference.pdf.

Qwen Team. Qwen3 Technical Report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Adam S. Shai, Sarah E. Marzen, Lucas Teixeira, Alexander Gietelink Oldenziel, and Paul M. Riechers. Transformers represent belief state geometry in their residual stream. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 75012–75034. Curran Associates, Inc., 2024. doi: 10.52202/079017-2387. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/8936fa1691764912d9519e1b5673ea66-Paper-Conference.pdf.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Alex Smolin and Bryan Wilder. Beliefs and behavior in language models, 2026. URL https://arxiv.org/abs/2609.07943.

Tomoki Tsujimura, Matīss Rikters, Masaki Asada, Shusaku Egami, Tatsuya Ishigaki, Ken Yano, and Hiroya Takamura. Assessing the belief consistency of large language models on the logical conversation process. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40032–40055, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1860. URL https://aclanthology.org/2026.acl-long.1860/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Tereza Vrabcová, Marek Kadlčík, Petr Sojka, Michal Štefánik, and Michal Spiegel. Towards the Roots of the Negation Problem: A Multilingual NLI Dataset and Model Scaling Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 25537–25551, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1391. URL https://aclanthology.org/2025.findings-emnlp.1391/.

Hanlin Wang, Chak Tou Leong, Jian Wang, and Wenjie Li. Seeing isn't believing: Mitigating belief inertia via active intervention in embodied agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 37796–37815, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1884. URL https://aclanthology.org/2026.findings-acl.1884/.

Yu Xia, Jingru Fan, Weize Chen, Siyu Yan, Xin Cong, Zhong Zhang, Yaxi Lu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. AgentRM: Enhancing Agent Generalization with Reward Modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19277–19290, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.945. URL https://aclanthology.org/2025.acl-long.945/.

Abel Yagubyan. How consistent are llm agents? measuring behavioral reproducibility in multi-step tool-calling pipelines, 2026. URL https://arxiv.org/abs/2605.28840.

Khurram Yamin, Jingjing Tang, Santiago Cortes-Gomez, Amit Sharma, Eric Horvitz, and Bryan Wilder. When agents say one thing and do another: Validating elicited beliefs from llms, 2026. URL https://arxiv.org/abs/2602.06286.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 20744–20757. Curran Associates, Inc., 2022. doi: 10.52202/068431-1508. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/82ad13ec01f9fe44c01cb91814fd7b8c-Paper-Conference.pdf.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Table 3: Grounded Noisy Cabinets propositions, one per category, with their reference rules. Statements illustrate template instantiations, not additional observations.

Category	Grounded statement	Reference rule
World	The red key is in cabinet 1.	Joint-posterior location marginal, then threshold.
Epistemic	You know where the gem is.	Held gem or maximum location posterior at least 0.95.
Affordance	You can unlock the main door now.	Door adjacency, lock state, and matching held key.
Decision	You still need to search the study.	Gem location not known/held, with an unsearched container in that room.

A Dataset Construction and Coverage

A.1 Tasks, policies, and trajectories

All six model conditions use the same tasks and initial configurations, while each policy generates its own trajectory. Sampling is stratified: ALFWorld balances task types and scenes with one trial per task directory; WebShop selects one goal per target product, stratified by category and option count; and Noisy Cabinets stratifies world seeds by room and lock counts. WebShop uses synthetic shopping goals.

The thinking and non-thinking conditions of Qwen3-8B share weights; all other conditions use the non-thinking protocol, and Llama-3.1-8B denotes Llama-3.1-8B-Instruct. Both RL checkpoints are Qwen3-8B models: WS-RL is the public checkpoint `webshop-qwen3-8b-outcome-gs80`, whose name indicates outcome-reward training on WebShop, and AF-RL is the step-400 checkpoint trained on ALFWorld with GiGPO (Feng et al., 2025b). Appendix B.1 describes the policy prompts and history recording.

A.2 Template generation and proposition taxonomy

We used GPT Astra, GPT-5.6 Sol, and Opus 5 to generate proposition templates and the corresponding code. To generate templates, we provide the authoring LLM with the environment's existing proposition families, typed arguments, example statements, and declared evaluation rules. These inputs constrain the semantics; the model proposes clearer formulations or new families. The prompt requires a semantic category, typed placeholders, a statement and its logical negation, separate rules for world truth and reference labels, and a rationale, while preserving temporal anchors and the distinction between physical facts, knowledge, and action conditions.

Proposals require semantic review and an environment-specific grounding implementation before registration in the executable bank. Automated checks verify that registered families have matching implementations, that statement and negation placeholders use only declared arguments, and that both world-truth and reference-label rules are present. The retained bank makes statements, argument bindings, and evaluation rules inspectable. Analyses retain grounded queries with well-defined reference and response labels; U remains a valid label, while missing labels and undefined referents are excluded. Representative instances across the four categories were inspected against recorded observation–action histories using the prompt semantics in Appendix B.2; Appendix C.4 gives examples and quantitative checks.[6]

For example, "You could unlock door d right now" has a door-valued argument d. Its rule requires the agent to be adjacent to a locked door and hold a matching key, fixing "could" to mean immediate

[6]The current proposal interface saves the authoring model identifier and revision, exact prompt, input/output token IDs, and generated text. Historical per-template authoring traces are incomplete, so these details cannot be recovered for every retained template.

Table 4: Proposition taxonomy by semantic category. A/W/N denote ALFWorld/WebShop/Noisy Cabinets.

Type	Definition & example	Proposition families
World		
Current state	Current physical or session facts. *e.g., the apple is in the fridge.*	**A:** Agent at, At, Holding, In Open Receptacle, Is Clean, Is Cool, Is Hot, Is on, Is Open, Same Receptacle **W:** All Goal Options Selected, On Page, Option Count Ge, Option Selected, Product Has Option, Product Price Ok, Query Contains, Results Contain Price Ok, Viewing **N:** Agent in, Holding, Is Locked, Is Open, Light on
History	Past actions or displayed content. *e.g., the fridge was opened.*	**A:** Was Open before **W:** Attribute Visible Match, Has Searched
Hidden location	Uncertain item or key location. *e.g., the red key is in the box.*	**N:** Color, In
World / expert	Hidden item facts and outcomes. *e.g., buying it meets the goal.*	**W:** Attribute World Match, Would Succeed if Bought Now
Epistemic		
Knowledge	Information known to the agent. *e.g., the key's location is known.*	**A:** Inventory Known, Known Location **W:** Known Price **N:** Known Location
History	Past observations and inspections. *e.g., the Reviews tab was opened.*	**A:** Examined, Explored, Seen, Visited before **W:** Viewed Description, Viewed Features, Viewed Reviews **N:** Searched
Affordance		
Action precondition	Whether an action is feasible now. *e.g., the door can be unlocked.*	**A:** Can Clean, Can Cool, Can Heat, Can Open, Can Put, Can Take **W:** Buy Eligible, Can Go Back **N:** Can Open, Can Unlock
Decision		
Decision	Task checks and search heuristics. *e.g., the room still needs searching.*	**A:** Last Action Succeeded, Needs Search, Subgoal Done, Target Found **W:** Needs Search **N:** Gem Found, Needs Search, Subgoal Done
World / expert	Reference action or plan length. *e.g., the expert opens the fridge.*	**A:** Expert Would Do, Remaining Expert Steps Le **W:** Expert Would Do

feasibility. Instantiating d as the main door gives "You could unlock the main door right now." Table 3 gives examples and their reference rules across the four categories.

The taxonomy groups templates into World, Epistemic, Affordance, and Decision, then distinguishes more specific semantic groups within each category (Table 4). For example, Epistemic contains *Knowledge*, such as "You know which container the gem is in," and *History*, such as "You

have already opened or examined cabinet 1." These subcategories distinguish knowing a fact from having performed an inspection; the latter need not resolve the former under noisy observations.

A.3 Proposition–label alignment

The recorded layer-ℓ residual at the final prefix position is

$$h_{t,c}^{(\ell)} = f_\theta^{(\ell)}(X_{t,1:c})_c \in \mathbb{R}^{d_\theta}, \tag{7}$$

where $f_\theta^{(\ell)}$ is the decoder-block output and d_θ the hidden dimension. At each turn t and token cutpoint c, a proposition ϕ is associated with its self-reported belief $r_{t,c}(\phi)$, reference label $v_t(\phi)$, and, when available, reference probability $e_t(\phi)$ and world truth Y_t^ϕ. Base propositions and their variants share the same model, task, and exact recorded prefix. Reference labels and world truth are used only for evaluation and never enter the query.

A.4 Coverage and analysis populations

The dataset includes propositions about entities in the current observation or mentioned earlier, as well as queries tagged as unobserved. These coverage descriptors can overlap and are separate from the semantic categories. Unobserved is not equivalent to Unknown: a cabinet can be mentioned while its contents remain unobserved, and negative evidence can establish an object's location without direct observation.

Consistency and accuracy use task-related propositions with reference labels in T/F/U; commonsense assertions form a separate control. The corresponding query and pair populations are defined in Appendices D.1 and E.1.

B Self-Reported Belief Elicitation

This appendix specifies how a grounded proposition becomes a self-reported belief. Each query branches from the recorded history without changing the policy's subsequent trajectory.

B.1 Policy prompts and token boundaries

The policy prompt combines the task goal and environment rules with the observation–action history and currently admissible actions, then asks for the next action. We use each model's native chat format. Earlier turns enter this history as action lines only; their reasoning is not retained. In NC, the prompt describes observation noise qualitatively without supplying the observer's numerical noise parameters.

We append each belief question to the original token prefix at a fixed boundary. For non-thinking models, this is immediately after observation o_t and before the next action is generated. For the thinking condition, it is after the model's reasoning and action text but before observation o_{t+1} arrives. Both measurements use reference labels anchored to s_t, before the action's resulting observation. The comparison evaluates the two protocols at their respective cutpoints. Each query forms a separate continuation and does not enter the subsequent policy history.

B.2 Self-reported belief elicitation

The belief-elicitation prompt uses the following template, with a grounded statement inserted:

> Using only the interaction so far, is this statement true right now?
> "The red key is in cabinet 1."
> A. True
> B. False
> C. Unknown: it cannot be determined from the interaction so far
> D. The question is unclear or cannot be parsed.
> Reply with one letter (A, B, C, or D).

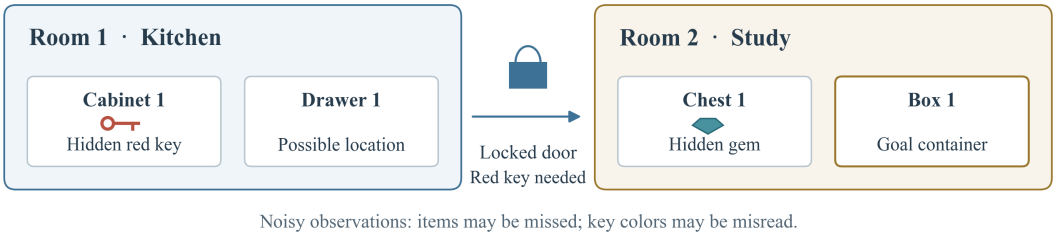

Figure 6: Illustrative hidden state in Noisy Cabinets, with a key, a locked door, the target gem, and the goal container.

Table 5: World-generation and observation parameters of Noisy Cabinets. Probability entries specify the simulator and observer rather than empirical error frequencies.

Parameter	Default	High noise
World generation		
Rooms in a chain	2–3	2–3
Containers (at least one per room)	4–8	4–8
Maximum colored keys	3	3
Door initially locked, given a generated key	0.70	0.70
Observation and dynamics		
Object detection drop	0.01	0.02
Examine detection drop	0.02	0.03
Key-color flip	0.05	0.05
Exogenous movement	0	0
Initial room darkness	0.30	0.60

Only the grounded statement is illustrative. Each query is measured under all six assignments of T/F/U to A/B/C, while D remains fixed. Unknown refers to information available in the interaction, whereas D denotes an unclear or unparsable query. For every Qwen3 condition, including the thinking condition, the answer letter is read after an empty reasoning block, so no additional reasoning trace is generated for the query.

Let a_l^σ be the next-token probability mass assigned to answer letter l under option ordering σ. The semantic distribution and the aggregate self-reported belief are

$$p_t^\sigma(k \mid \phi) = \frac{a^\sigma_{\sigma^{-1}(k)}}{a^\sigma_A + a^\sigma_B + a^\sigma_C}, \qquad r_t(\phi) = \frac{1}{6}\sum_\sigma p_t^\sigma(\cdot \mid \phi), \quad k \in \{\mathrm{T}, \mathrm{F}, \mathrm{U}\}. \tag{8}$$

D is excluded from this mixture, and the aggregate answer is $y = \arg\max_{k\in\{\mathrm{T},\mathrm{F},\mathrm{U}\}} r(k)$. Averaging after semantic mapping reduces dependence on answer position. The resulting distribution is conditional on the T/F/U options and is not a calibrated probability of world truth. The records retain the median across orders of $D/(A+B+C+D)$ as a format diagnostic. Its mean is below 0.001% for each Qwen condition and 0.447% for Llama-3.1-8B; the latter's 95th percentile is 1.09%. No retained query exceeds 5% on this median diagnostic, so excluding queries above that threshold leaves aggregate accuracies unchanged. This concerns the saved order-median share, not a reconstructed four-way argmax.

B.3 Unknown answers

Because the query restricts answers to the interaction so far, U is a legitimate answer whenever the history is inconclusive. It expresses uncertainty about a well-defined proposition, whereas D indicates difficulty interpreting the question. A reference label of U and a self-reported belief of U describe the observer's and the model's judgments, respectively.

Table 6: Reference validation against available world truth. **Top:** definite World references; coverage is relative to definite World queries. **Bottom:** NC physical-probability calibration on recorded queries and on independent history-only simulations. ECE is in percentage points; the final columns report empirical truth frequencies (%).

Environment	Truth coverage (%)	Agreement (%)
ALFWorld	28.85	99.95
WebShop	87.47	100.00

NC population	ECE (pp)	$P(Y = 1 \mid v = \mathrm{T})$	$P(Y = 1 \mid v = \mathrm{F})$
Recorded queries	0.4524	99.882	0.155
Default simulation	0.0023	99.954	0.019
High-noise simulation	0.0036	99.948	0.029

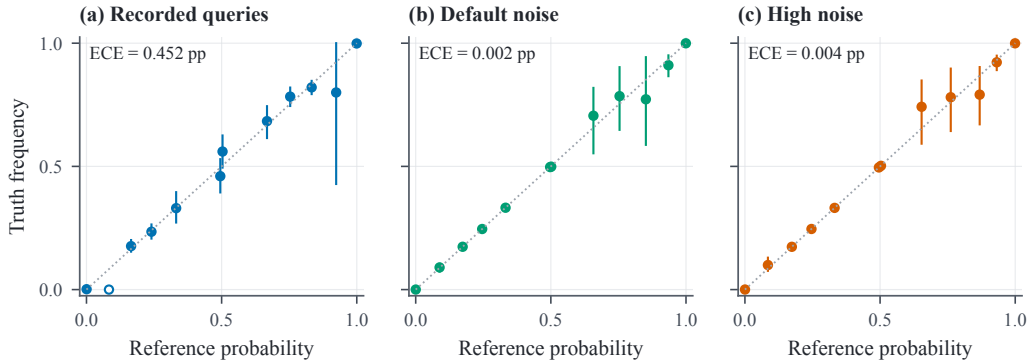

Figure 7: NC reference probabilities and physical truth. **(a)** Recorded physical queries. **(b–c)** History-only simulations under the default and high-noise profiles. Vertical bars show 95% task/world-cluster intervals for truth frequencies; open markers denote unavailable intervals. Dotted lines indicate equality. ECE is in percentage points.

C Reference Labels and Noisy Cabinets

C.1 Environment-specific reference labels

ALFWorld and WebShop use deterministic observation-history trackers to assign the reference label $v_t(\phi)$ independently of model answers. They assign T or F when the available evidence establishes a proposition or its negation, and U when it remains unresolved. Evaluation requires well-defined temporal and entity references; physical states, action preconditions, and observed outcomes are treated separately.

NC instead integrates propositions against an exact joint posterior over latent item configurations. For physical predicates, $e_t(\phi)$ is the posterior probability of truth, thresholded as in Section 3.3. Knowledge predicates differ: `known_location` is true when an item is held or its maximum location posterior reaches τ_{T}. Its reference probability is therefore the Boolean indicator of this condition. Physical presence, knowledge of a location, and a past inspection are distinct propositions.

A reference label of U marks uncertainty relative to the available observations. For physical location propositions, it typically reflects posterior mass between the decision thresholds; for knowledge predicates, a lack of knowledge makes the proposition false. Some reference rules assign U directly without a scalar probability, so three-valued labels and continuous probabilities are not interchangeable.

C.2 World generation and observation model

NC separates world truth from reference labels through noisy observations. The agent navigates connected rooms, finds keys, unlocks doors, and places a target gem in a designated container (Figure 6). Each required key is placed in a preceding reachable room, ensuring solvability; other item placements follow the permitted-container constraints. Table 5 summarizes the world-generation and observation parameters.

The prior respects the same placement constraints. Inventory, agent location, lights, door locks, and container apertures update deterministically after valid actions, while unobserved item locations remain uncertain; there is no exogenous item movement. The observer knows the noise parameters and the generated world structure, including every container, the item set, and the key color for each locked door, whereas the agent receives only the qualitative noise description in Appendix B.1. Reference labels therefore depend on information that is not fully available to the agent.

C.3 Exact Bayesian inference

The observer maintains $b_t(z) = P(z \mid \mathcal{F}_t^{\mathrm{ref}})$ over joint latent configurations and applies the prediction and conditioning steps of Equation 2. Failed actions provide evidence through their preconditions; an uninformative "nothing happens" response does not establish that an item is absent.

A missed-detection example. Consider a stationary item, a uniform prior over $N \geq 2$ containers, independent detection drops with probability $d \in [0, 1]$, no false alarms, and no other observations. After $k \geq 1$ unsuccessful inspections of container 1,

$$P(\text{item in container 1} \mid k \text{ misses}) = \frac{d^k/N}{d^k/N + (N-1)/N} = \frac{d^k}{N - 1 + d^k}. \tag{9}$$

As $d \to 0$, the posterior approaches zero; as $d \to 1$, it returns to the prior $1/N$. Repeated unsuccessful inspections can therefore yield an F reference label even when an item is physically present but missed. The full observer also incorporates placement constraints and action history. This distinction illustrates why reference labels and world truth need not coincide.

C.4 Reference-label validation

We validate references against world truth where the quantities have matching semantics. In deterministic environments, we compare definite World references with recorded Boolean truth for exactly matched propositions and cutpoints. Table 6 reports agreement and the share of definite World queries with an available truth field. Residual ALFWorld disagreements concern locations of carried objects: recorded inventory, successful take observations, and admissible move commands support the reference that a carried object is no longer at the queried receptacle. These discrepancies remain in the raw comparison.

For NC, we compare the retained physical-proposition probabilities with recorded world truth and separately evaluate simulations under both noise profiles. The simulation policy uses displayed commands, observation history, and independent randomness. Every item–location event, including the held location, is scored without selecting events by truth. Figure 7 and Table 6 report the resulting calibration and threshold behavior; the recorded-query and simulation populations remain distinct.

An independently enumerated joint-state observer matches the implemented filter to numerical precision (maximum cell error 1.11×10^{-16}). The missed-detection boundary cases follow the posterior above, and impossible evidence raises an error. These checks establish agreement of the independent probability calculations under the declared environment model.

Grounded examples and rule checks. The prompt's temporal and informational qualifications determine each label. In one recorded ALFWorld trajectory, the agent visits and opens fridge 1 while searching for a mug. Its listed contents contain no mug, supporting "You still need to search for a mug" (Decision, T). Before any visit to coffeemachine 1, "You have already looked inside/at the coffeemachine 1" is F (Epistemic). While at the fridge, "You could open cabinet 11 right now" is F (Affordance), because the command is not immediately available; the unobserved open status of cabinet 7 remains U (World). These examples instantiate the rules in Table 3 and the prompt in Appendix B.2.

Table 7: Self-reported belief agreement (Agg, %) by environment on the reference-complete subsets. **Left:** full query/pair sets, including U reports. **Right:** pairs with both reports in T/F; UU and all single-U pairs are excluded. Permutation retains all six-order comparisons. Bold: highest value per column within each panel. Proposition categories are pooled.

	Full set				T/F-only pairs		
Model	**Perm.**	**Neg.**	**Seq.**	**Rephr.**	**Neg.**	**Seq.**	**Rephr.**
ALFWorld							
Qwen3-8B	67.35	74.73	—	85.17	84.36	—	93.62
Qwen3-8B (think)	70.63	56.22	—	82.77	72.97	—	95.31
Qwen3-14B	78.46	75.44	—	74.10	**84.82**	—	89.44
Llama-3.1-8B	63.63	53.03	—	77.15	61.24	—	87.27
WS-RL	70.27	74.29	—	83.36	84.10	—	92.44
AF-RL	**78.88**	**76.68**	—	**89.73**	77.74	—	**98.23**
WebShop							
Qwen3-8B	64.47	61.03	83.94	73.70	67.08	86.79	87.70
Qwen3-8B (think)	71.49	63.96	**91.18**	80.55	76.37	**97.17**	**93.80**
Qwen3-14B	**85.87**	**75.44**	88.46	**85.09**	**77.61**	92.28	89.73
Llama-3.1-8B	71.25	53.72	79.09	82.11	56.02	89.47	92.31
WS-RL	63.85	57.77	53.03	77.97	66.28	72.73	89.30
AF-RL	69.58	55.76	75.41	83.99	66.84	83.87	93.27
Noisy Cabinets							
Qwen3-8B	68.84	72.72	87.43	91.98	86.21	96.97	89.71
Qwen3-8B (think)	70.92	70.77	70.07	87.16	81.88	88.19	91.04
Qwen3-14B	**79.73**	72.88	88.68	79.76	84.91	94.76	**92.15**
Llama-3.1-8B	68.40	70.83	90.31	79.47	74.34	96.78	86.79
WS-RL	71.56	**75.45**	91.23	89.04	87.08	98.56	88.43
AF-RL	75.39	71.54	**97.26**	**92.27**	**87.44**	**100.00**	89.80

Independent comparisons with recorded admissible commands give agreement between 99.38% and 99.95% for the six checked ALFWorld action families. WebShop's checked back-navigation and purchase-eligibility rules agree with recorded navigation, price, and instruction-option conditions. On the checked ALFWorld action queries with T/F references and reports, accuracy is 54.34% against the original references and 54.60% against command availability. The remaining differences include spatial accessibility and placement preconditions. WebShop purchase eligibility is evaluated by price and instruction options; its stored truth field additionally requires target-item identity and is therefore kept separate. Epistemic and Decision labels retain their declared history and task-progress semantics.

D Consistency: Methods and Detailed Results

This appendix defines the consistency measures in Table 1, reports agreement by environment and proposition category, and decomposes the paired outcomes. Appendix F presents the context-length analyses.

D.1 Pairs and aggregation

All base queries have reference labels in T/F/U, as do both endpoints of each relation. The shared base set excludes explicit negation wording and unrelated controls, and every retained negated variant contains an explicit negation token. U remains a valid reference label and self-reported belief.

Negation and rephrasing compare a base proposition with its variant at the same task, turn, and recorded prefix. Sequential relations compare base propositions across adjacent queried turns, re-

Table 8: Self-reported belief agreement (Agg, %) by proposition category on the reference-complete subsets. **Left:** full query/pair sets, including U reports. **Right:** pairs with both reports in T/F; UU and all single-U pairs are excluded. Permutation retains all six-order comparisons. Bold: highest value per column within each panel. Environments are pooled.

	Full set				T/F-only pairs		
Model	**Perm.**	**Neg.**	**Seq.**	**Rephr.**	**Neg.**	**Seq.**	**Rephr.**
World							
Qwen3-8B	73.90	81.86	88.58	90.17	88.24	92.78	92.52
Qwen3-8B (think)	75.18	66.15	85.35	87.77	80.51	**96.76**	94.58
Qwen3-14B	80.64	**83.89**	88.64	83.60	**93.07**	94.82	94.58
Llama-3.1-8B	67.15	67.91	83.07	84.00	76.48	96.17	**96.29**
WS-RL	76.03	79.71	86.70	89.37	84.88	88.99	94.01
AF-RL	**81.60**	78.98	**92.18**	**93.92**	83.36	93.15	94.74
Epistemic							
Qwen3-8B	65.43	57.29	81.62	80.42	79.58	90.91	93.84
Qwen3-8B (think)	71.68	39.38	82.98	78.09	60.95	**100.00**	95.18
Qwen3-14B	**85.97**	**90.89**	**98.21**	**89.33**	**94.69**	**100.00**	94.21
Llama-3.1-8B	79.70	66.22	97.55	85.88	70.91	98.41	93.45
WS-RL	68.10	56.75	87.63	81.06	79.71	93.33	93.25
AF-RL	73.58	62.36	98.14	83.05	69.42	**100.00**	**100.00**
Affordance							
Qwen3-8B	66.10	77.80	95.00	92.56	63.91	**100.00**	90.71
Qwen3-8B (think)	68.63	75.12	73.68	89.65	70.25	90.91	**93.04**
Qwen3-14B	72.86	54.83	89.74	70.10	59.52	93.75	84.84
Llama-3.1-8B	60.98	50.60	82.87	71.16	45.72	91.60	78.77
WS-RL	70.10	78.99	96.94	91.35	65.11	**100.00**	89.33
AF-RL	**85.38**	**90.05**	**99.39**	**96.85**	**74.75**	—	91.67
Decision							
Qwen3-8B	59.13	67.75	82.69	68.09	76.05	93.44	85.23
Qwen3-8B (think)	65.13	53.84	52.25	69.89	**85.68**	74.51	93.53
Qwen3-14B	**82.52**	66.69	84.76	64.50	73.18	88.89	82.56
Llama-3.1-8B	54.58	30.97	94.79	72.63	42.54	96.98	81.33
WS-RL	60.01	**70.20**	89.75	65.94	82.40	98.41	81.10
AF-RL	64.52	58.03	**95.43**	**79.49**	75.17	**100.00**	**94.12**

quiring stable referents, world truth, reference labels, and available reference probabilities throughout the episode, including unqueried steps. For U references, this requires a defined reference probability.

For the paired relations, let $\mathcal{P}_h$ contain pairs of aggregate hard labels from Section 3.4, with T and F swapped in the second label for negation. For permutation, let n_{qc} count the mapped votes for label $c \in \{\mathrm{T}, \mathrm{F}, \mathrm{U}\}$ among the six answer orderings of base query q, across N queries. Agreement is

$$\mathrm{Agg} = \frac{1}{|\mathcal{P}_h|} \sum_{i \in \mathcal{P}_h} \mathbf{1}[y_{i1} = y_{i2}], \qquad \mathrm{Agg}_{\mathrm{perm}} = \frac{1}{30N} \sum_{q,c} n_{qc}(n_{qc} - 1). \tag{10}$$

Permutation averages agreement over all pairs of option-order labels within each query. The aggregate pools these comparisons across queries.

The T/F-only columns retain TT, FF, TF, and FT, excluding outcomes that contain U. The permutation column includes T/F/U votes. Rates with empty denominators are undefined.

Table 9: Nine-outcome decomposition of paired consistency (%). Shares use all pairs within each model and relation. Letter order is base–variant for negation/rephrasing and earlier–later for sequential. Correct outcomes are TF/FT/UU for negation and TT/FF/UU otherwise; their sum is Agg. Bold marks the highest Agg within each relation.

Neg.

		Correct			Incorrect					
Model	**Agg**	**TF**	**FT**	**UU**	**TT**	**FF**	**UT**	**UF**	**TU**	**FU**
Qwen3-8B	72.25	12.86	15.07	44.32	4.58	2.55	16.04	1.09	1.77	1.71
Qwen3-8B (think)	59.94	14.07	18.22	27.65	7.78	2.74	20.99	1.14	5.01	2.40
Qwen3-14B	**75.06**	27.01	35.21	12.84	6.23	6.51	2.87	0.46	3.94	4.93
Llama-3.1-8B	55.84	18.38	17.08	20.39	19.70	1.98	15.70	3.35	1.62	1.80
WS-RL	72.06	14.79	15.27	42.00	5.03	2.44	15.96	1.12	1.85	1.55
AF-RL	73.39	8.75	5.87	58.77	3.29	1.05	16.83	0.93	3.68	0.84

Seq.

		Correct			Incorrect					
Model	**Agg**	**TT**	**FF**	**UU**	**TF**	**FT**	**UT**	**UF**	**TU**	**FU**
Qwen3-8B	86.70	2.09	21.84	62.77	0.57	1.23	1.14	6.55	0.85	2.94
Qwen3-8B (think)	79.75	6.75	38.28	34.72	0.98	1.96	1.35	8.71	1.60	5.64
Qwen3-14B	88.57	8.16	61.15	19.26	2.72	1.96	0.33	4.13	0.00	2.29
Llama-3.1-8B	88.74	8.33	39.95	40.46	0.70	1.27	0.83	4.07	0.57	3.82
WS-RL	88.79	1.26	21.64	65.89	0.58	0.87	0.97	5.89	0.68	2.22
AF-RL	**95.35**	1.25	3.31	90.79	0.10	0.14	0.24	2.30	0.29	1.58

Rephr.

		Correct			Incorrect					
Model	**Agg**	**TT**	**FF**	**UU**	**TF**	**FT**	**UT**	**UF**	**TU**	**FU**
Qwen3-8B	84.24	17.46	9.78	57.01	0.75	1.94	2.53	1.86	2.14	6.53
Qwen3-8B (think)	82.85	23.15	14.60	45.10	0.70	1.65	1.40	2.80	2.80	7.80
Qwen3-14B	77.46	30.48	34.51	12.48	2.84	4.46	0.42	2.93	4.50	7.39
Llama-3.1-8B	78.53	32.38	14.63	31.52	3.53	2.67	2.88	5.66	3.60	3.13
WS-RL	83.36	17.76	10.36	55.24	0.84	2.00	3.36	1.40	2.36	6.68
AF-RL	**89.44**	13.03	4.89	71.52	0.35	0.57	3.30	0.66	2.73	2.95

D.2 Agreement by environment and proposition category

Table 7 pools proposition categories within each environment; Table 8 pools environments within each category. Both use the full and T/F-only definitions of Table 1. ALFWorld has no sequential pairs satisfying the reference conditions; dashes denote undefined estimates. Model comparisons describe agreement on the queries and pairs associated with each policy.

D.3 Unknown reports and nine-outcome decomposition

Figure 8 uses the union of retained base, negated, and rephrased statements, counting each query once. The pair-level decomposition in Table 9 instead uses each relation's own pair support. Letters are ordered as base then variant for negation/rephrasing and earlier then later for sequential comparisons.

Negation is correct for TF, FT, and UU; sequential and rephrasing agreement is correct for TT, FF, and UU. The table reports each outcome as a percentage of all pairs for that model and relation. The shares cover all paired outcomes, and the consistent outcomes sum to Agg.

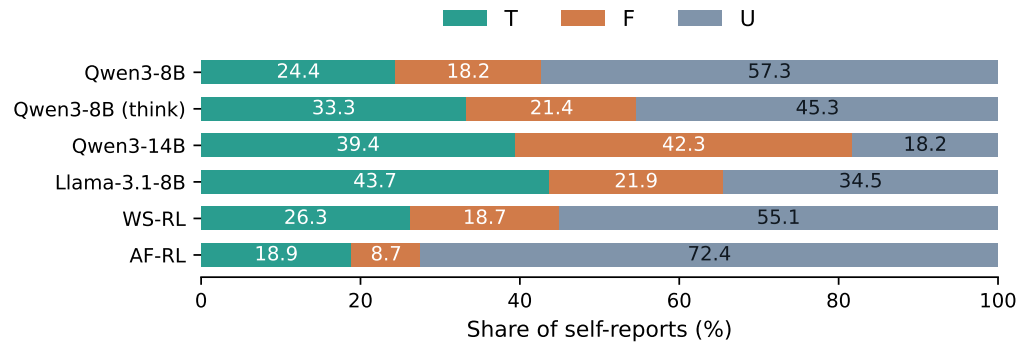

Figure 8: T/F/U self-reported belief composition (%) on the reference-complete query union. Each base, negated, or rephrased query is counted once.

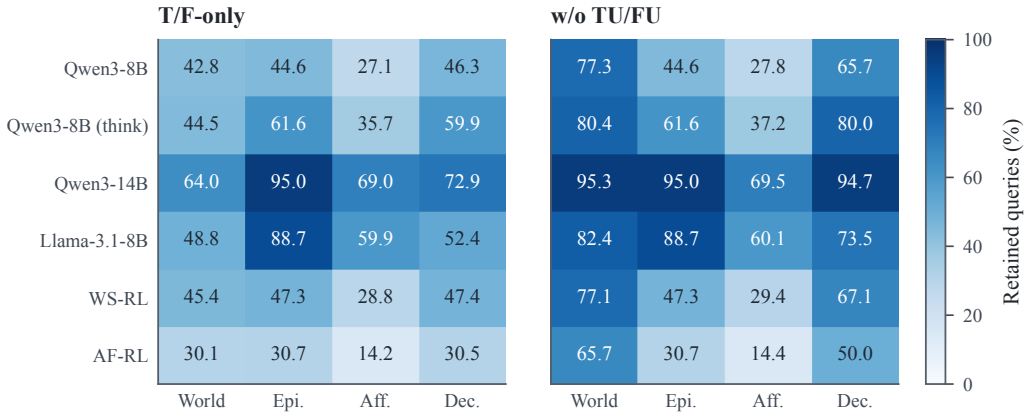

Figure 9: Retained-query fractions under the conditional accuracy definitions (%). **Left:** T/F-only. **Right:** w/o TU/FU. Rows denote models and columns proposition categories; Epi., Aff., and Dec. abbreviate Epistemic, Affordance, and Decision. Both panels use the original category query union as denominator and the same color scale.

Single-U outcomes dominate disagreements in most model–relation comparisons, while UU accounts for much of AF-RL's agreement. T/F-only measures omit both UU agreements and single-U disagreements, so their relation to full-population agreement depends on the balance between these outcomes. Agreement alone does not establish reference accuracy.

E Accuracy: Populations and Detailed Results

E.1 Evaluation population and conditional measures

The accuracy population is the deduplicated union of base, negated, and rephrased queries. Sequential relations reuse Base and do not add accuracy observations. Every record has a reference label and an aggregate self-reported belief in T/F/U. Reference U is a valid target, not a missing value.

For the evaluated query set G, each query contributes once. Let N_{ab} count $(v_i, y_i) = (a, b)$, with $N = |G| = \sum_{a,b\in\{\mathrm{T,F,U}\}} N_{ab}$. The overall and conditional accuracies are

$$A_{\text{all}} = \text{Accuracy}_{\text{ref}}(G) = \frac{1}{N}\sum_{i\in G} \mathbf{1}[y_i = v_i] = \frac{N_{TT} + N_{FF} + N_{UU}}{N}, \tag{11}$$

$$A_{\text{TF}} = \frac{N_{TT} + N_{FF}}{N_{TT} + N_{FF} + N_{TF} + N_{FT}}, \tag{12}$$

Table 10: Accuracy of self-reported beliefs (%) by environment and proposition category. Environment panels pool categories; category panels pool environments. Overall uses all outcomes; T/F-only retains TT, FF, TF, and FT; w/o TU/FU excludes reference-T/F, report-U errors. Bold: highest value per column within each panel.

By environment

Model	Overall	T/F-only	w/o TU/FU
ALFWorld			
Qwen3-8B	39.00	79.03	82.45
Qwen3-8B (think)	46.44	80.86	78.76
Qwen3-14B	**66.45**	**81.00**	76.60
Llama-3.1-8B	52.53	70.38	70.84
WS-RL	39.69	76.64	79.43
AF-RL	30.11	78.63	**85.39**
WebShop			
Qwen3-8B	57.41	71.03	67.34
Qwen3-8B (think)	52.60	68.68	63.41
Qwen3-14B	**65.34**	**80.60**	69.67
Llama-3.1-8B	54.48	64.99	64.89
WS-RL	57.30	74.16	68.23
AF-RL	52.50	71.78	**73.34**
Noisy Cabinets			
Qwen3-8B	48.26	73.64	79.74
Qwen3-8B (think)	53.85	71.55	75.85
Qwen3-14B	**72.60**	**81.98**	80.25
Llama-3.1-8B	56.93	72.28	74.17
WS-RL	47.48	72.91	79.59
AF-RL	42.32	78.12	**85.65**

By proposition category

Model	Overall	T/F-only	w/o TU/FU
World			
Qwen3-8B	67.94	83.51	87.88
Qwen3-8B (think)	64.74	78.72	80.56
Qwen3-14B	**75.93**	**88.36**	79.69
Llama-3.1-8B	62.12	72.21	75.39
WS-RL	67.98	84.35	88.21
AF-RL	61.23	87.05	**93.21**
Epistemic			
Qwen3-8B	35.20	78.87	78.87
Qwen3-8B (think)	50.76	82.36	82.36
Qwen3-14B	**89.78**	**94.55**	**94.55**
Llama-3.1-8B	74.43	83.89	83.89
WS-RL	37.17	78.53	78.53
AF-RL	24.37	79.39	79.39
Affordance			
Qwen3-8B	18.47	66.89	66.47
Qwen3-8B (think)	25.47	**70.70**	**68.48**
Qwen3-14B	**38.64**	55.97	55.56
Llama-3.1-8B	32.99	55.07	54.89
WS-RL	17.26	58.93	58.69
AF-RL	9.14	63.46	63.52
Decision			
Qwen3-8B	42.87	68.62	65.27
Qwen3-8B (think)	47.27	70.82	59.07
Qwen3-14B	**60.51**	**78.85**	63.87
Llama-3.1-8B	41.14	58.66	55.94
WS-RL	43.48	70.33	64.79
AF-RL	35.54	66.55	**71.09**

$$A_{-TU,-FU} = \frac{N_{TT} + N_{FF} + N_{UU}}{N - N_{TU} - N_{FU}}. \tag{13}$$

T/F-only requires both the reference and report to be T or F. Excluding TU/FU removes only reference-T-or-F/report-U errors, retains UU, UT, and UF, and leaves the correct numerator unchanged. This diagnostic sets aside mismatches that may reflect an information gap between the reference and model. These are conditional descriptions; a higher rate after deleting outcomes is not an improvement on the original population. Undefined denominators remain undefined.

A conservative bound for Unknown-error adjustment. Let $C = N_{TT} + N_{FF} + N_{UU}$ and $U_{\text{gap}} = N_{TU} + N_{FU}$. Suppose S is any subset of these TU/FU mismatches attributed to an information gap and set aside, without altering other labels or removing correct reports. The adjusted score is $A_S = C/(N - |S|)$. Since $0 \le |S| \le U_{\text{gap}}$, for nonempty retained populations,

$$A_{\text{all}} = \frac{C}{N} \;\le\; A_S \;\le\; \frac{C}{N - U_{\text{gap}}} = A_{-TU,-FU}. \tag{14}$$

Table 11: Nine-outcome decomposition of accuracy (%). Letters denote reference then report. Each cell gives its share of all evaluated queries; TT, FF, and UU sum to overall accuracy (Acc). Bold marks the highest overall accuracy.

Accuracy

		Correct			Incorrect					
Model	**Acc**	**TT**	**FF**	**UU**	**TF**	**FT**	**UT**	**UF**	**TU**	**FU**
Qwen3-8B	43.39	17.59	13.02	12.78	3.70	5.87	0.96	1.53	16.22	28.34
Qwen3-8B (think)	48.82	22.84	14.78	11.20	4.12	7.49	2.96	2.50	9.79	24.32
Qwen3-14B	**67.10**	28.18	31.78	7.14	6.07	7.94	3.29	4.50	4.32	6.79
Llama-3.1-8B	53.61	26.06	16.46	11.09	4.36	14.22	3.38	1.06	7.05	16.32
WS-RL	43.56	18.78	13.14	11.64	4.01	6.36	1.14	1.50	15.42	28.01
AF-RL	34.82	14.01	6.47	14.34	1.77	4.28	0.58	0.46	21.20	36.89

Table 12: T/F-only accuracy–success sensitivity. Entries are correlations; ranges summarize omission-specific point estimates, not confidence intervals. Environment means are recomputed after each omission.

Category	All	Omit one model	Omit one environment	Omit both RL conditions
Affordance	0.410	0.328–0.479	0.316–0.452	0.305
Decision	0.274	0.139–0.379	0.213–0.365	0.228

This defines the lower-bound claim relative to an explicit information-gap adjustment. Equality holds with no exclusions; if $C = 0$, all defined scores are zero. Removing the entire population leaves an undefined score. T/F-only separately evaluates definite judgments and need not preserve this ordering because it also removes UU. Across models, common observer rules and the same Unknown treatment provide a common scoring scale; the nine-cell table shows how each model's errors and retained population contribute to the conditional scores.

E.2 Accuracy by environment and proposition category

Table 10 reports accuracy by environment and proposition category. Environment panels pool categories, while category panels pool environments. Values are query-weighted proportions; bold marks the highest observed accuracy within each panel and definition.

E.3 Correct and incorrect outcome composition

Table 11 gives the nine outcomes underlying the Overall column. The correct group is TT/FF/UU; the error group is TF/FT/UT/UF/TU/FU. Each cell gives its share of all evaluated queries; TT, FF, and UU sum to overall accuracy. The directions matter: TU/FU are model-U responses to T/F references, whereas UT/UF are definite reports against a U reference.

AF-RL's errors are dominated by U reports against definite references, whereas these outcomes account for a smaller share of Qwen3-14B's errors. Consequently, the conditional measures emphasize different aspects of each model's performance. Comparisons across checkpoints also involve policy-dependent query populations and do not isolate the effect of post-training.

Figure 9 makes the two conditional denominators explicit. For AF-RL Affordance queries, T/F-only and w/o TU/FU retain 14.18% and 14.39%, respectively. A model that always reports U attains w/o TU/FU accuracy of one whenever reference-U queries remain; the statistic therefore characterizes its retained outcomes rather than overall reporting quality.

Table 13: Context-length analysis settings. All profiles and response-tendency analyses use the same ranges; lengths exclude the appended query.

Environment	Length range (tokens)	Binning
ALFWorld	1,000–6,000	Equal width
WebShop	1,000–3,000	Near-equal width
Noisy Cabinets	1,000–2,000	Equal width

Each range has six bins shared across models. Bins are left-closed and right-open, with the final upper bound included. WebShop boundaries are 1,000, 1,350, 1,650, 1,950, 2,300, 2,650, and 3,000. Parameters, error shares, and observed contradiction rates use the same negation pairs. Intervals use joint task-cluster resampling and 95% empirical percentile bounds. Lines retain defined estimates; intervals require support across tasks and defined estimates in at least 95% of resamples. The scatter plots require at least 100 pairs and effective task support of 20 in each conditioning group: reference T/F, base report F, and negated report T. All eight rates—α_T, α_F, β_T, β_F, p_T, p_F, $\widetilde{p}_T$, and $\widetilde{p}_F$—must be defined in at least 99% of resamples and have interval width at most 0.20. Both scatter panels use the same fixed selection; OLS gives each displayed bin equal weight.

E.4 T/F-ONLY ACCURACY–SUCCESS ASSOCIATIONS

Figure 3 uses the same query union with both reference and report restricted to T/F. Thus the numerator is TT+FF and the denominator is TT+TF+FT+FF. Task success uses the original task set, including tasks without a retained query.

For category k, model m, and environment e, Figure 3 centers T/F-only category accuracy, A_{mek}, and task success by their six-model environment means:

$$\widetilde{A}_{mek} = A_{mek} - \frac{1}{6}\sum_{m'} A_{m'ek}, \qquad \widetilde{S}_{me} = S_{me} - \frac{1}{6}\sum_{m'} S_{m'e}. \tag{15}$$

Each panel fits $\widetilde{S}_{me} = a_k + b_k \widetilde{A}_{mek} + \epsilon_{mek}$ with equal weight per condition. Centering makes the fitted intercept zero up to numerical precision and gives $R^2 = r^2$; it removes environment means, not every source of benchmark difficulty.

Affordance and Decision accuracy have positive associations with task success, with task-bootstrap intervals above zero (Figure 3). World and Epistemic point estimates are near zero and their intervals span zero. The ordering of point estimates does not by itself establish a statistically significant difference between categories.

Confidence intervals use percentile task-cluster bootstrap resampling, with common task draws across models and categories within each environment. Each draw recomputes query-weighted accuracy, task success, and environment means; models and environments remain fixed. These associations are descriptive because the conditional subsets depend on model responses and the comparisons are between conditions rather than individual actions.

Sensitivity to evaluated conditions. Table 12 recomputes environment means after omitting each model condition, each environment, or both RL checkpoints. The T/F-only Affordance and Decision point estimates remain positive throughout these comparisons. The ranges describe sensitivity of the observed associations; their uncertainty remains distinct from the shared-task intervals for the full set of conditions.

F CONTEXT LENGTH AND RESPONSE TENDENCIES

The analyses below share the main-text length ranges and task resampling. We present the supplementary profiles, derive the paired-report decomposition, validate its within-condition and task-disjoint behavior, and extend the baseline to meaning-preserving relations.

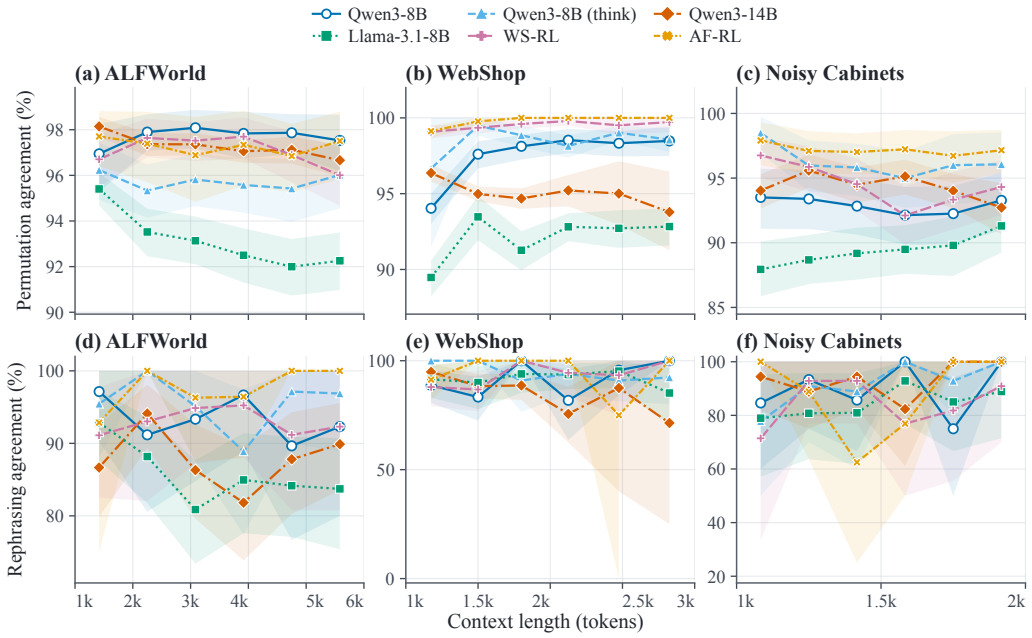

Figure 10: T/F-only agreement across context lengths (%). **Top:** permutation. **Bottom:** rephrasing. Colors and markers denote models; panel scales differ. Shading shows 95% joint task-cluster bootstrap intervals, where supported.

F.1 Data preparation, length bins, and uncertainty

Profiles use recorded prefix length at the measurement cutpoint, excluding the appended query. Models share the environment-specific intervals in Table 13. Accuracy uses the deduplicated Base, Negation, and Rephrasing union, retaining T/F references and reports independently for each query. Negation uses matched same-prefix pairs with T/F reports and complementary T/F references. Its error decomposition and response parameters therefore share one paired population.

Rates pool outcomes within each bin. ALFWorld and NC use equal widths; WebShop uses near-equal widths on a token grid with nonincreasing pooled support, chosen without inspecting accuracy or agreement. Boundaries remain fixed across profiles. Sequential pairs retain their original adjacency and reference stability and are assigned to the later turn's bin; filtering never creates new neighbors.

The joint task bootstrap resamples complete tasks within environments, sharing weights across models, bins, and metrics. Each draw recomputes pooled rates; 95% pointwise intervals use inverted-CDF percentiles. Undefined conditional rates remain unavailable. Length profiles describe accumulated interactions and the states reached by each policy.

F.2 Additional agreement profiles

Figure 10 complements the main negation profile. Permutation retains Base queries with T/F references and T/F reports under every option order, averages mapped agreement over order pairs, and then pools queries. Rephrasing retains T/F labels at both endpoints and counts TT/FF as agreement. Both use the same length boundaries and task draws.

F.3 Errors on matched proposition forms

Let N_{ab} count paired reports $(y_\phi, y_{\neg\phi}) = (a, b)$ and $N = \sum_{a,b\in\{T,F\}} N_{ab}$ within a bin. Double affirmation and double rejection contribute

$$E_{TT} = \frac{N_{TT}}{N}, \qquad E_{FF} = \frac{N_{FF}}{N}, \qquad \mathrm{Agg}_{\mathrm{neg}} = 1 - E_{TT} - E_{FF}. \tag{16}$$

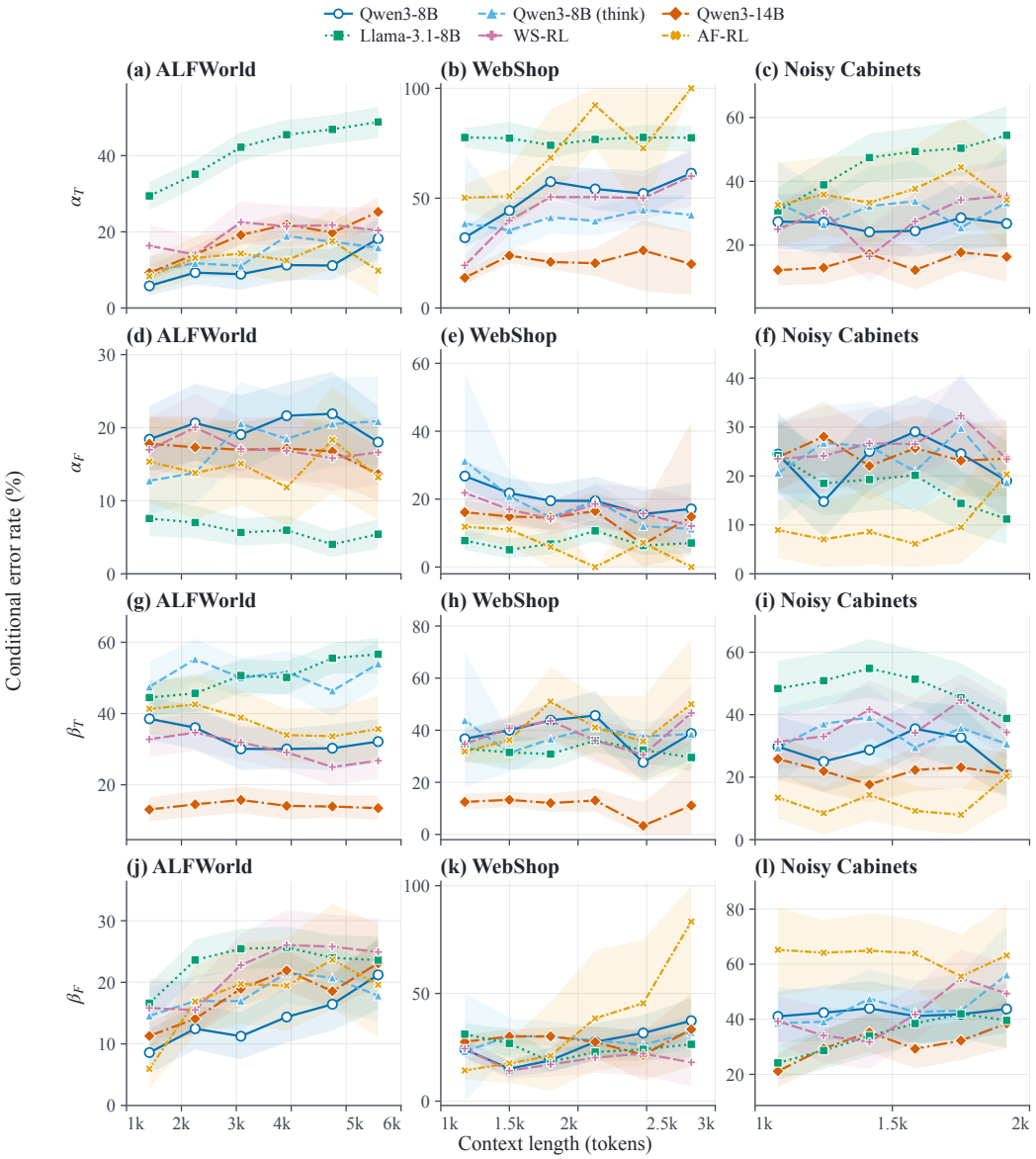

Figure 11: Conditional error rates across context lengths (%). Rows: α_T, α_F, β_T, β_F; columns: environments. Estimates use the same T/F-only pairs and bins as Figure 4(a–c). Bands show 95% joint task-cluster bootstrap intervals.

Figure 11 resolves endpoint accuracy into the four conditional error rates defined in Section 5.2. Its first and last rows measure endorsement of false ϕ and rejection of true $\neg\phi$, respectively, motivating the response-tendency model in Section 5.2. All rates use the pairs and bins of Figure 4(a–c), so their variation and the paired contradictions share the same evaluated population.

F.4 From endpoint errors to paired contradictions

All probabilities below refer to one model–environment–length bin of matched T/F pairs. The main-text mixture parallels a two-high-threshold recognition model with equal detection probabilities for true and false references (Erdfelder et al., 2011). Its reference-independent weights require $\alpha_T + \alpha_F \leq 1$ and $\beta_T + \beta_F \leq 1$. Both constraints hold in every Figure 5 bin. The additional longest-context AF-RL/WebShop bin has sums 1 and $4/3$, respectively; its unmodified values retain the conditional-error interpretation in the full profiles.

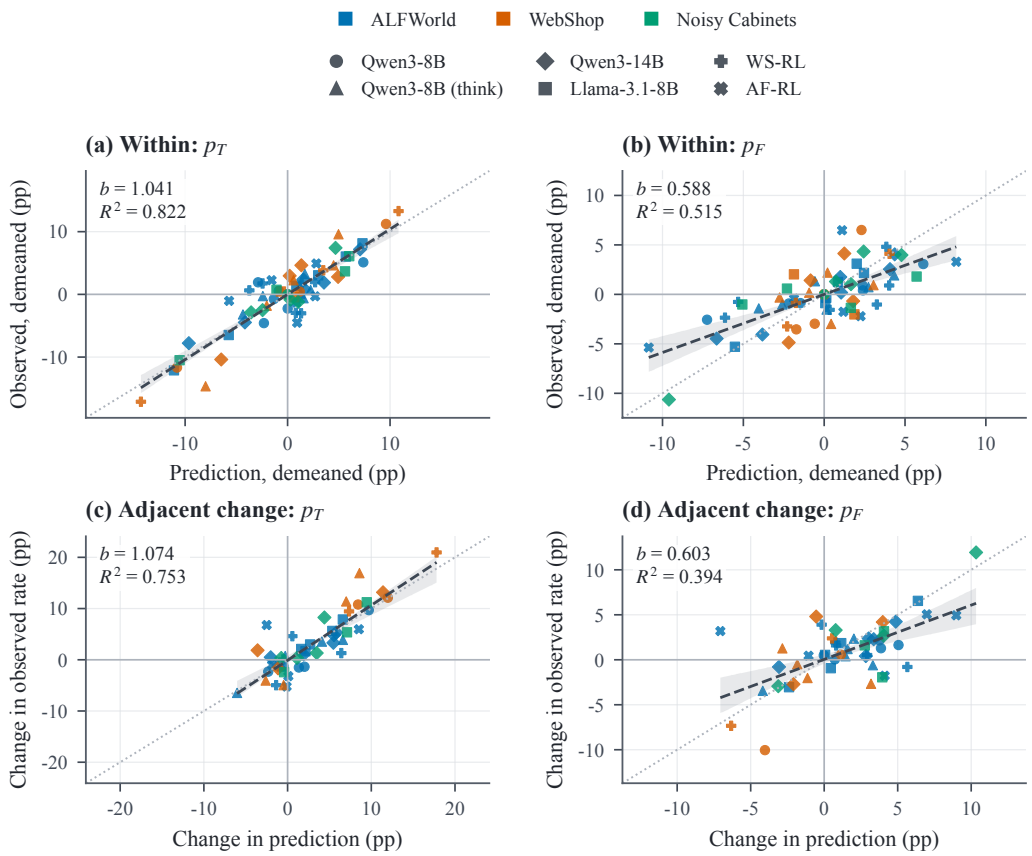

Figure 12: Within-condition and adjacent-bin relationships. **Top:** predictions and observed rates after subtracting model–environment means. **Bottom:** changes between original adjacent bins. Colors denote environments and markers denote models; b is the fitted slope. Dashed lines show fits, gray bands show 95% shared task-bootstrap intervals for fitted means, and dotted lines indicate equality. Panel scales differ.

Table 14: Regression estimates for Figure 5. Brackets give 95% joint task-bootstrap intervals. Intercepts are in percentage points, matching the figure's axes.

Rate	Slope	Intercept (pp)	R^2
p_T	1.059 [1.009, 1.089]	-16.37 [-17.81, -14.04]	0.873 [0.830, 0.892]
p_F	0.399 [0.295, 0.525]	1.46 [-2.24, 4.65]	0.214 [0.124, 0.337]

Table 15: Within-condition and adjacent-bin regressions. Brackets give 95% shared task-bootstrap intervals. Within fits remove model–environment means; adjacent fits include an intercept and use original neighboring bins.

Rate	Comparison	Slope	R^2
p_T	Within condition	1.041 [0.900, 1.100]	0.822 [0.690, 0.850]
p_T	Adjacent bins	1.074 [0.813, 1.151]	0.753 [0.534, 0.809]
p_F	Within condition	0.588 [0.424, 0.722]	0.515 [0.290, 0.615]
p_F	Adjacent bins	0.603 [0.343, 0.792]	0.394 [0.157, 0.569]

Marginals and dependence. Let $V = v_\phi$, $X = \mathbf{1}[y_\phi = T]$, $Z = \mathbf{1}[y_{\neg\phi} = T]$, $\mu_v = E[X \mid V = v]$, and $\nu_v = E[Z \mid V = v]$. Complementary references imply

$$(\mu_T, \nu_T) = (1 - \alpha_F, \beta_T), \qquad (\mu_F, \nu_F) = (\alpha_T, 1 - \beta_F). \tag{17}$$

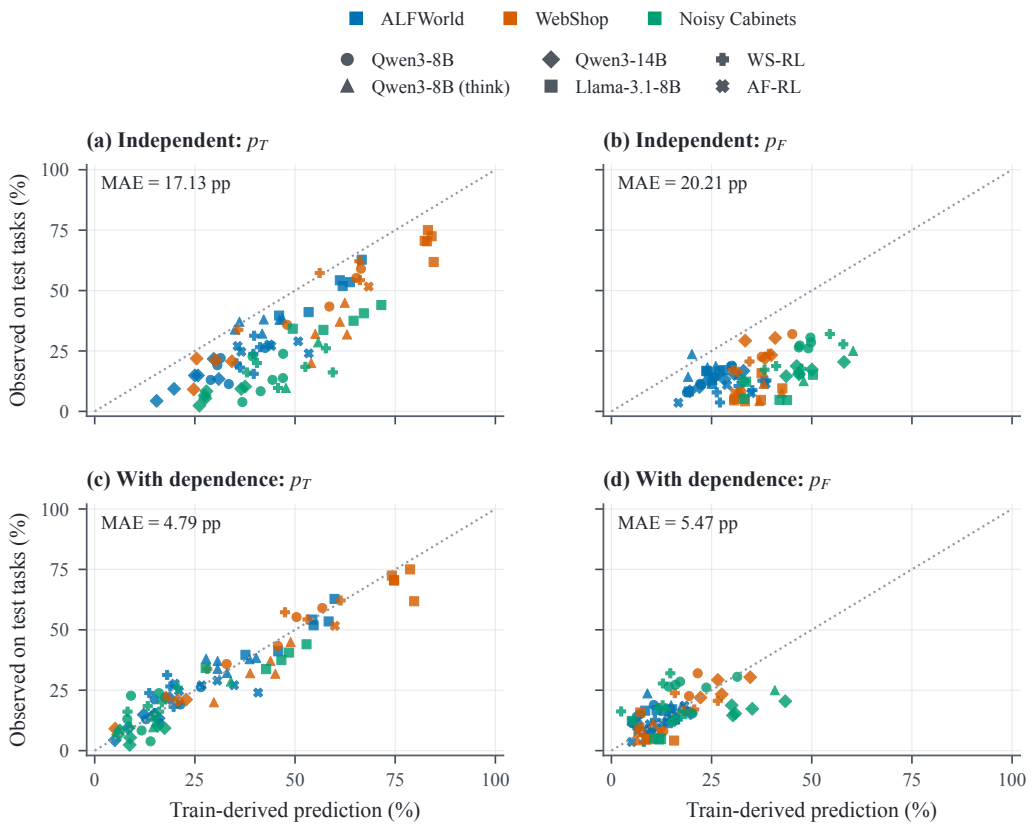

Figure 13: Train-derived predictions and contradictions observed on test tasks. **Top:** independent products of training endpoint marginals. **Bottom:** training marginals with training conditional dependence. Both use test reference proportions and the same support. Dotted lines indicate equality; MAE is in percentage points. No line is fitted to test outcomes.

Independence multiplies these marginals within each reference class. More generally, with $c_v = \mathrm{Cov}(X, Z \mid V = v)$,

$$\begin{aligned} P(X = 1, Z = 1 \mid V = v) &= \mu_v \nu_v + c_v, \\ P(X = 0, Z = 0 \mid V = v) &= (1 - \mu_v)(1 - \nu_v) + c_v, \\ \max(0, \mu_v + \nu_v - 1) - \mu_v \nu_v \le c_v &\le \min(\mu_v, \nu_v) - \mu_v \nu_v. \end{aligned} \tag{18}$$

The bounds ensure nonnegative joint probabilities and force $c_v = 0$ when either marginal is deterministic. Mixing the classes with $\pi_v = P(V = v)$ gives $\kappa = \pi_T c_T + \pi_F c_F$ and the corrections in Equation 6.

The conditioning probabilities are

$$m_T = \pi_T \beta_T + \pi_F (1 - \beta_F), \qquad m_F = \pi_T \alpha_F + \pi_F (1 - \alpha_T). \tag{19}$$

Dividing the TT/FF probabilities by these denominators yields the conditional-rate identities in Section 5.2. Empirically, $\widehat{p}_T = N_{TT}/(N_{TT} + N_{FT})$ and $\widehat{p}_F = N_{FF}/(N_{FT} + N_{FF})$. Zero denominators leave conditional rates undefined, while agreement is defined from joint shares whenever the pair population is nonempty. An absent reference class has zero weight, with no imputation of its conditional rates.

Offsets and slopes. Write $x_j = \widetilde{p}_{s,j}$ and $\delta_j = \widehat{\kappa}_j / \widehat{m}_{s,j}$ for $s \in \{T, F\}$, so $\widehat{p}_{s,j} = x_j + \delta_j$. With an intercept and nonzero predictor variance, OLS gives

$$\widehat{b}_s = 1 + \frac{\mathrm{Cov}_j(x_j, \delta_j)}{\mathrm{Var}_j(x_j)}, \qquad \widehat{a}_s = \overline{\delta} - (\widehat{b}_s - 1)\overline{x}. \tag{20}$$

Table 16: Test-task mean absolute error (pp). Predictions use training response parameters and test reference proportions, with no test-set regression calibration. Intervals summarize the paired error reduction under shared task resampling.

Rate	Independent	With dependence	Reduction [95% CI]
p_T	17.13	4.79	12.34 [8.47, 12.52]
p_F	20.21	5.47	14.75 [10.44, 15.50]
Negation agreement	16.32	4.05	12.27 [8.98, 12.53]

A negative correction lowers contradiction levels, while its covariation with the baseline determines the slope. The shared κ is divided by different conditioning probabilities in the two panels. Changes along the profiles satisfy $\Delta\mathrm{Agg}_{\mathrm{neg}} = \Delta\widetilde{\mathrm{Agg}}_{\mathrm{neg}} - 2\Delta\kappa$; dependence can offset or reinforce the marginal change.

Figure 5 uses the common criteria in Table 13, fixed before fitting. OLS weights bins equally, and each shared task-bootstrap draw recomputes both coordinates and the fit. Table 14 reports coefficient intervals; the figure bands describe the fitted mean.

F.5 Validation across length bins and disjoint tasks

Within-condition and adjacent-bin comparisons. To separate trajectory variation from differences between model–environment conditions, we regress observed rates on baseline predictions after removing each condition's mean. A second comparison uses differences between original neighboring retained bins, without bridging gaps. Both axes, group means, and fits are recomputed in the shared task-bootstrap draws, with the original Figure 5 selection held fixed.

The TT relationship remains strong within conditions and across adjacent changes; the FF relationship is weaker but clearer within conditions than in the pooled fit (Table 15, Figure 12). The all-defined-bin and per-token-difference calculations are retained with the analysis outputs. These comparisons concern changes along the observed trajectories.

Prediction on disjoint tasks. We reuse the original train/test task assignment and estimate endpoint response marginals using training reports only. Test reference-class proportions determine the mixture weights. The independence predictor multiplies the training marginals; a second predictor retains their training class-conditional covariance. Every report-dependent parameter comes from training tasks, with no regression calibration on test rates.

The common support requires at least 20 training reports in each reference class and at least 20 conditioning events for each rate in both splits, fixed before prediction errors are evaluated. Both predictors use this support, with equal weight per bin. Independent train/test task resamples share weights across models within each environment and give the paired error-reduction intervals.

Training-derived dependence reduces error for both conditional contradictions and negation agreement (Table 16; Figure 13). The gain supports transfer of the joint response structure across the recorded task split.

F.6 Shared tendencies for permutation and rephrasing

Both relations preserve the reference label. If the two reports share conditional T-report probabilities $q_T = 1 - \alpha_F$, $q_F = \alpha_T$ and are independent given reference V, their disagreement probability is

$$\widetilde{D} = 2\pi_T(1-\alpha_F)\alpha_F + 2\pi_F\alpha_T(1-\alpha_T), \qquad \widetilde{\mathrm{Agg}} = 1 - \widetilde{D}. \tag{21}$$

The prediction is at most one half. Both perfect reference-following and a fixed identical response give zero disagreement, illustrating why agreement and accuracy differ.

Figure 14 estimates the common predictor from T/F-only Base queries. Permutation observations require all six actual option-order reports to be T/F; rephrasing uses aggregate T/F reports at both endpoints. Transferring the Base predictor assumes common reference proportions and conditional

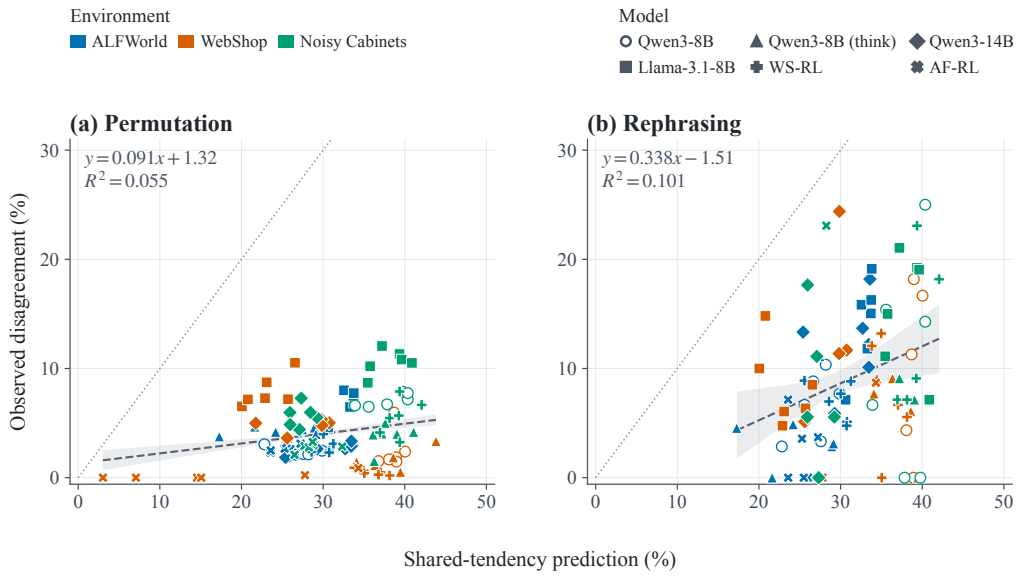

Figure 14: T/F-only invariance disagreement versus a shared-marginal prediction. **(a)** Permutation. **(b)** Rephrasing. Predictions use Base accuracy in Equation 21. Colors denote environments and markers denote models. Dashed lines show OLS fits, gray bands show 95% shared task-bootstrap intervals for fitted means, and dotted lines indicate equality.

error rates across these relation-specific populations and reporting units. Fits give equal weight to bins with effective task support $(\sum_t n_t)^2 / \sum_t n_t^2 \geq 10$ in each contributing population; shared task draws recompute both axes on the fixed support.

For endpoint T-report probabilities q_{iv}, q_{jv} and within-reference covariance $c_{ij,v}$, paired disagreement more generally satisfies

$$D_{ij} = \sum_{v \in \{T,F\}} \pi_v \left[q_{iv} + q_{jv} - 2q_{iv}q_{jv} - 2c_{ij,v} \right]. \tag{22}$$

Positive covariance lowers disagreement relative to the matched independent baseline. Re-estimating separate marginals on matched rephrasing pairs strengthens its association with observed disagreement, which remains below the baseline in most bins. The shared Base predictor fits both relations weakly, and the direct association between their observed rates is also weak. This extension identifies the assumptions needed to transfer a response-tendency account across relations.

G LIMITATIONS

Model coverage. The evaluated conditions cover models up to 14B parameters and do not include mixture-of-experts models. Broader model and architecture coverage would test how widely the observed patterns extend.

Reference labels and query populations. Accuracy uses the observer's specified information, including NC noise and world-structure parameters. Models answer from their own policy contexts, so comparisons include differences in visited states, query composition, and permitted evidence. Checkpoint comparisons do not isolate post-training effects.

Mechanisms behind response tendencies. The decomposition and task-disjoint comparisons connect endpoint errors with paired reports. Controlled changes to context and evidence would further distinguish why response tendencies vary along trajectories from the accompanying changes in task state and proposition composition.

From evaluation to reinforcement learning. Proposition-level accuracy and consistency could provide step-level feedback for process reward models. Testing whether such feedback improves agent judgments and task performance is a direction for future work.