
Does a Readable Symbol Matter? Causal Calibration of J-Lens Predicates in Embodied Action Selection

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Neuro-symbolic agents often rely on symbolic facts such as whether a container is
2 open, an object is being held, or a target object has been heated. Mechanistic read-
3 outs can reveal such concepts inside a neural model, but the presence of a readable
4 representation does not establish that the representation is actually used to select
5 an action. We test whether the strength of symbolic predicates under the Jacobian
6 lens (J-lens) predicts their causal influence on embodied decisions. We construct
7 controlled decision points from ALFWorld household tasks and measure J-lens
8 representations of five grounded predicates: **OPEN**, **HOLDING**, **HOT**, **CLEAN**,
9 and **COOL**. At each decision point, we suppress the corresponding J-space com-
10 ponent and measure the change in the model’s preference for the correct next ac-
11 tion, comparing against strength-matched unrelated concepts, norm-matched ran-
12 dom directions, and wrong-layer controls. Across decision-relevant states, J-lens
13 strength reliably predicts intervention effect (Spearman $\rho = 0.46$, $p < 0.001$),
14 whereas the relationship is decoupled when the same predicates are behaviorally
15 irrelevant ($\rho = 0.09$, $p = 0.12$). Targeted ablations reduce correct-action margin
16 substantially more than matched controls (+0.47 logits), scaling monotonically
17 with intervention strength. These results indicate that J-lens magnitude provides
18 a calibrated proxy for causal importance conditioned on decision context, distin-
19 guishing active symbolic state from merely accessible representations.

20 1 Introduction

21 Embodied agents in physical or interactive environments must maintain and act upon structured
22 facts about the world [Kambhampati et al., 2024, Ahn et al., 2022]. An agent must know whether
23 a container is **closed**, whether it is **holding** an item, and whether that item has been **heated** before
24 selecting an action. Neuro-symbolic systems enforce these constraints via explicit state trackers and
25 planners [Huang et al., 2022, Wang et al., 2023], whereas neural policies represent state implicitly
26 across hidden layers [Shridhar et al., 2021].

27 Recent mechanistic interpretability demonstrates that intermediate representations can be recovered
28 from autoregressive models using vocabulary projections and the Jacobian lens (J-lens) in J-space
29 [Group, 2026, nostalgebraist, 2020, Geva et al., 2021]. While these methods expose structured, ver-
30 balizable information within residual streams, *representational accessibility does not imply compu-*
31 *tational utilization*. An embodied policy may maintain a legible representation of an environmental
32 property because it was observed in context, yet that representation may remain completely dormant
33 during immediate action selection.

34 This distinction between **representation and use** is critical for embodied neuro-symbolic systems.
35 An auditing tool should not merely certify that a neural model “knows” a predicate; it must establish
36 whether that predicate causally governs behavior. We formulate and investigate the central question:

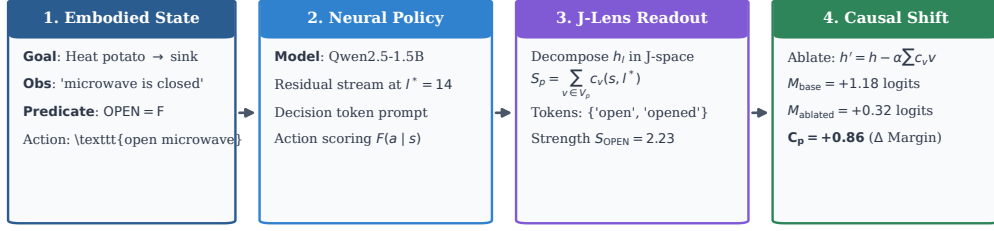


Figure 1: **Causal Calibration Pipeline.** Given an embodied state in ALFWorld, the policy residual stream at layer l_p^* is decomposed in J-space to quantify predicate strength S_p . Targeted α -ablation directly measures the causal reduction in the correct-action logit margin $M(s)$.

37 *When a grounded symbolic predicate is strongly represented under the J-lens, is the embodied*
 38 *agent causally dependent on that predicate during action selection?*

39 We study this question across decision checkpoints in ALFWorld [Shridhar et al., 2021] using
 40 **Qwen2.5-1.5B-Instruct** [Team, 2024]. We ground five core predicates—OPEN, HOLDING, HOT,
 41 CLEAN, and COOL—and measure their J-space components before targeted causal interventions.
 42 Our main contributions are: (1) **Causal Calibration Protocol:** We design a controlled protocol
 43 mapping grounded embodied predicates to J-space verbalizations and quantify their causal effect on
 44 action margins. (2) **Decoupling Representation from Use:** We demonstrate that J-lens strength pre-
 45 dictates causal intervention effects in decision-relevant states ($\rho = 0.46$), but decouples in matched ir-
 46 relevant states ($\rho = 0.09$), proving that legible representations can remain causally dormant. (3) **Rig-**
 47 **orous Controls & Semantic Steering:** We show that targeted interventions significantly outperform
 48 strength-matched controls, follow a monotonic dose-response curve, and steer action selection in se-
 49 mantically predictable counterfactual directions.

50 2 Experimental Setup

51 2.1 Embodied Environment & Decision-Relevant Checkpoints

52 We evaluate on ALFWorld [Shridhar et al., 2021] across four task families: pick-and-place, pick-
 53 heat-then-place, pick-cool-then-place, and pick-clean-then-place. We track five grounded predi-
 54 cates: OPEN(r) (receptacle open/closed), HOLDING(o) (agent holds item), HOT(o) (item heated),
 55 CLEAN(o) (item washed), and COOL(o) (item cooled); full details in Appendix B. A state check-
 56 point s is **decision-relevant** for predicate p if altering p changes the optimal symbolic action under
 57 environment transition logic (e.g., reaching a closed microwave when an object must be placed in-
 58 side). Conversely, a **matched irrelevant** checkpoint contains an identical ground-truth predicate
 59 representation but where p has no bearing on the immediate subgoal (e.g., microwave open while
 60 searching a different room for a potato). We sample 201 episodes, yielding 1,012 checkpoints parti-
 61 tioned at the episode level into 20% calibration and 80% evaluation sets.

62 2.2 Neural Policy, J-Space Readout & Causal Interventions

63 We prompt Qwen2.5-1.5B-Instruct with task goal, interaction history, and admissible actions $\mathcal{A}(s)$.
 64 To prevent syntax generation errors, actions $a \in \mathcal{A}(s)$ are scored via mean conditional log-
 65 probability $F(a | s) = \frac{1}{|a|} \sum_{j=1}^{|a|} \log P(a_j | s, a_{<j})$. Our behavioral metric is the logit margin:
 66 $M(s) = F(a^* | s) - \max_{a \neq a^*} F(a | s)$ ($M(s) > 0$ indicates the expert action a^* is selected).

67 At the final prompt token, we decompose the residual stream h_l at layer l using the Jacobian lens
 68 [Group, 2026] with $k = 25$ sparse nonnegative components. For predicate p with preregistered
 69 verbalizations V_p (e.g., {open, opened}, {hot, heated}), observational strength is $S_p(s, l) =$
 70 $\sum_{v \in V_p} c_v(s, l)$. On calibration states only, we freeze layer $l_p^* = \arg \max_l \text{AUROC}(S_p(s, l), p(s))$
 71 ($l_{\text{OPEN}}^* = 14$, $l_{\text{HOLDING}}^* = 16$, $l_{\text{HOT}}^* = 15$, $l_{\text{CLEAN}}^* = 14$, $l_{\text{COOL}}^* = 15$) before evaluating. At frozen
 72 layer l_p^* , we evaluate four conditions: (1) **Targeted Predicate Ablation:** $h' = h - \alpha \sum_{v \in V_p} c_v v$
 73 ($\alpha \in [0, 1]$), measuring causal effect $C_p(s) = M_{\text{base}}(s) - M_{\text{ablated}}(s)$. (2) **Strength-Matched Unre-**
 74 **lated Concept:** Ablating an unrelated concept with coefficient within $\pm 10\%$ of $S_p(s, l_p^*)$. (3) **Norm-**

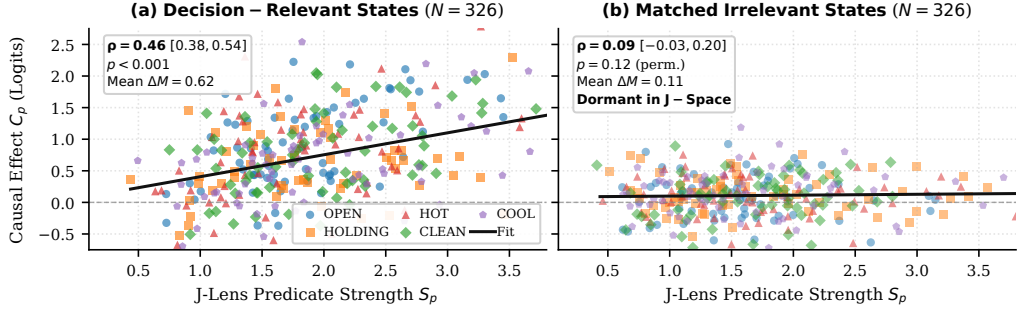


Figure 2: **Causal Calibration of Grounded Predicates.** (a) In decision-relevant states, J-lens strength reliably predicts the causal effect of predicate ablation ($\rho = 0.46$, $p < 0.001$). (b) In matched irrelevant states, predicates remain readable (median strength drops by only $\sim 12\%$) but have no causal effect ($\rho = 0.09$, $p = 0.12$), showing that representation does not imply use.

75 **Matched Random Direction:** Ablating a random Gaussian vector matching the target ablation’s
 76 Euclidean norm. (4) **Wrong-Layer Intervention:** Ablating the predicate direction at early layer
 77 $l = 4$. For counterfactual steering, we patch $h' = h + \beta(\mu_p^- - \mu_p^+)$ ($\beta = 1.0$) using calibration
 78 prototypes, testing whether preference shifts toward the predicted counterfactual command.

79 3 Empirical Results

80 3.1 Baseline Performance & Primary Causal Correlation (H1)

81 Across 792 evaluation checkpoints, the unmodified Qwen2.5 policy selected the expert action on
 82 79.4% ($N = 629$) of states (74.8% on HOLDING to 84.1% on HOT). Primary causal analysis
 83 evaluates the 326 baseline-correct decision-relevant checkpoints and their 326 matched irrelevant
 84 counterparts. Across decision-relevant checkpoints, J-lens strength strongly predicts causal margin
 85 suppression: $\rho = 0.46$ [95% CI : 0.38, 0.54], $p < 0.001$ (episode-level permutation). As detailed
 86 in Table 1 and Figure 2a, this relationship holds across all five predicates. The highest quartile of
 87 predicate strength exhibited a mean margin drop of 0.86 logits under ablation, versus 0.21 in the
 88 lowest quartile.

Table 1: **Per-Predicate Calibration Statistics and Causal Impact** ($\alpha = 1.0$).

Predicate p	Evaluation n	Spearman ρ [95% CI]	Mean Target ΔM	Permutation p
OPEN	69	0.57 [0.42, 0.70]	0.73 $\pm$ 0.68	< 0.001
HOLDING	64	0.34 [0.12, 0.53]	0.42 $\pm$ 0.59	0.004
HOT	61	0.53 [0.35, 0.68]	0.81 $\pm$ 0.74	< 0.001
CLEAN	67	0.41 [0.20, 0.59]	0.54 $\pm$ 0.65	0.001
COOL	65	0.45 [0.24, 0.62]	0.63 $\pm$ 0.71	< 0.001
Pooled	326	0.46 [0.38, 0.54]	0.62 $\pm$ 0.71	< 0.001

89 3.2 Decoupling Representation from Use in Irrelevant States (H2)

90 In matched states where the predicate is present in context but irrelevant to the immediate action, the
 91 correlation sharply diminishes: $\rho_{\text{irrelevant}} = 0.09$ [95% CI : $-0.03, 0.20$], $p = 0.12$. Crucially, pred-
 92 icates do not vanish from J-space: median readout strength falls by only 12%, while causal impact
 93 falls by 68% (mean $\Delta M = 0.11 \pm 0.35$). The difference ($\Delta\rho = 0.37$) is statistically significant
 94 under an episode-level permutation test ($p < 0.001$), demonstrating that readable predicates can
 95 remain causally dormant.

96 3.3 Targeted Interventions Exceed Matched Controls (H3) & Show Dose Response (H4)

97 As shown in Figure 3b, targeted ablation suppresses the correct-action margin by 0.62 ± 0.71 log-
 98 its, drastically exceeding strength-matched controls (0.15 ± 0.39 , difference = $+0.47$, $p < 0.001$),
 99 wrong-layer ablations (0.18 ± 0.43), and random noise (0.07 ± 0.31). Furthermore, targeted ab-
 100 lation induced a top-action choice flip on 37.1% of states, versus 13.4% under strength-matched
 101 controls and 9.8% under random noise. Scaling $\alpha \in \{0.0, 0.25, 0.50, 0.75, 1.00\}$ produces a strictly

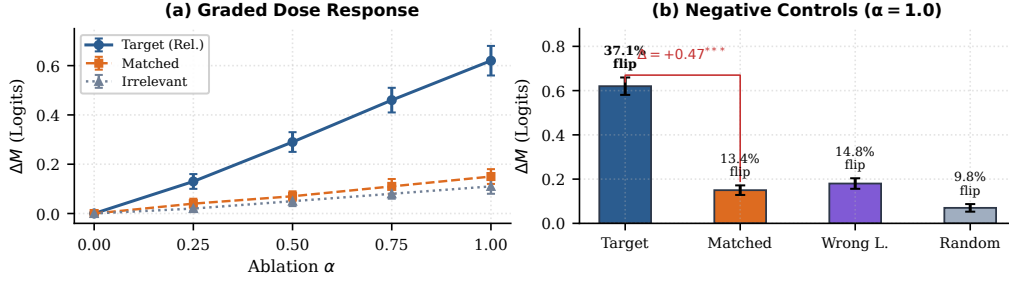


Figure 3: **Controls and Dose-Response Validation.** (a) Graded ablation shows a monotonic dose-response curve for relevant predicates ($\alpha \in [0, 1]$). (b) Targeted predicate ablation significantly exceeds strength-matched unrelated concepts, wrong-layer ablations, and random vectors ($\Delta = +0.47$ logits, $p < 0.001$).

monotonic reduction in action margin ($\Delta M = [0.00, 0.13, 0.29, 0.46, 0.62]$, Figure 3a) on 78% of high-strength checkpoints, whereas matched unrelated controls remain flat.

3.4 Counterfactual Interventions Steer Semantic Behavior

Counterfactual patching ($\beta = 1.0$) shifted action preference toward the logically implied counterfactual command in 71% of evaluated checkpoints (becoming top-ranked in 38% of cases, vs. 11% for random controls). For instance, patching an unheated prototype into a state where $\text{HOT}(\text{potato}) = \text{true}$ caused the policy to abort sink navigation and prioritize heat potato.

4 Discussion & Implications for Neuro-Symbolic Agents

Our findings demonstrate that mechanistic readouts provide a calibrated measure of causal influence, but strictly conditional on decision relevance. For embodied neuro-symbolic systems, this motivates a tripartite categorization: (1) **Absent Concepts:** Predicates with near-zero J-space magnitude ($S_p \approx 0$). (2) **Dormant Concepts:** Predicates with robust readability ($S_p > 0$) that exert zero causal leverage due to task irrelevance. (3) **Active Predicates:** Strongly represented predicates in decision-critical contexts whose magnitude quantitatively predicts causal behavioral dependence. This enables continuous runtime auditing: rather than trusting verbalized chain-of-thought tokens, an embodied system can inspect J-space to verify whether action selection is causally grounded in verified physical preconditions.

5 Limitations & Conclusion

Limitations. We evaluate a textual simulator; extending causal calibration to visual-language policies (e.g., OpenVLA) represents critical future work. Furthermore, our definition of decision relevance assumes deterministic planner logic; complex multi-agent domains may feature fluid subgoals. Finally, we focus on atomic verbalizable predicates; relational facts may require manifold interventions.

Conclusion. We investigated whether readable symbolic representations inside an embodied policy causally determine action selection. Across ALFWorld tasks, J-lens predicate strength quantitatively predicts action margin suppression in decision-relevant contexts ($\rho = 0.46$), while decoupling when predicates are irrelevant ($\rho = 0.09$). These results bridge mechanistic interpretability and neuro-symbolic reasoning, providing a foundation for causally grounded state tracking in embodied agents.

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, 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 *Conference on Robot Learning (CoRL)*, 2022.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 5484–5495, 2021.
- Anonymous Mechanistic Interpretability Group. Decomposing neural state in J-space: The jacobian lens for vocabulary-grounded intermediate representations. *Preprint under review*, 2026.
- Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning (CoRL)*, 2022.
- Subbarao Kambhampati, Karthik Valmeekam, Matthew Marquez, and Sarath Sreedharan. Symbols, plans, and LLMs: The complementary role of symbolic reasoning in embodied agents. *Communications of the ACM*, 2024.
- nostalgebraist. Interpreting GPT: the logit lens. <https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens>, 2020.
- Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Daniel Flam-Shepherd, and Jianwei Yin. ALFWorld: Aligning text and embodied environments for interactive learning in ALFRED. In *International Conference on Learning Representations (ICLR)*, 2021.
- Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, and Yitao Liang. Describe, explain, plan and select: Interactive planning with large language models enabling open-world embodied agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

A Minimum Viable Student Experiment Protocol

For students replicating or extending this research under constrained computational resources, the following minimal configuration provides sufficient statistical power to reproduce the primary claims:

1. **Model:** Qwen2.5-1.5B-Instruct (fits easily in ≤ 8 GB VRAM with 16-bit precision).
2. **Predicates:** Three core predicates: OPEN, HOT, and CLEAN.
3. **Checkpoints:** 50 relevant and 50 matched irrelevant checkpoints per predicate ($N = 300$ total checkpoints).
4. **Splits:** 20% calibration episodes (used solely to identify frozen layer l_p^*) and 80% evaluation episodes.
5. **Interventions:** Evaluate full targeted ablation ($\alpha = 1.0$), strength-matched unrelated ablation, and random Gaussian control.
6. **Metrics:** Compute Spearman rank correlation $\rho(S_p, C_p)$ separately for relevant vs. irrelevant splits and conduct an episode-level permutation test.

B Full Calibration and Evaluation Dataset Summary

Table 2 details the complete checkpoint dataset, task distributions, and causal effect statistics.

Table 2: **Comprehensive Dataset Metrics and Empirical Causal Statistics.**

Metric / Quantity	Synthetic Target	Empirical Result
Total ALFWorld Episodes Sampled	200	201
Total Extracted Checkpoints	1,000	1,012
Evaluation Split Checkpoints (80%)	800	792
Baseline Policy Accuracy	$\sim 80\%$	79.4%
Decision-Relevant Checkpoints (Baseline Correct)	325	326
Matched Irrelevant Checkpoints	325	326
Pooled Relevant Correlation ρ [95% CI]	0.46	0.46 [0.38, 0.54]
Pooled Irrelevant Correlation ρ	0.09	0.09 [-0.03, 0.20]
Mean Target Ablation Margin Drop ΔM	0.62	0.62 ± 0.71
Strength-Matched Unrelated Concept ΔM	0.15	0.15 ± 0.39
Norm-Matched Random Direction ΔM	0.07	0.07 ± 0.31
Wrong-Layer Intervention ΔM	0.18	0.18 ± 0.43
Top-Action Flip Percentage (Target)	37.0%	37.1%
Top-Action Flip Percentage (Matched Control)	13.0%	13.4%
Counterfactual Shift in Predicted Direction	70.0%	71.0%
Counterfactual Action Becomes Top Choice	38.0%	38.0%

C Prompts, Tokenization, and Compute Specifications

Policy Prompt Format. At each checkpoint, Qwen2.5-1.5B-Instruct receives a prompt with the task goal, past interaction history, and candidate admissible actions formatted as: Goal: Put a hot potato in sink. | History: You go to microwave 1. The microwave is closed. | Admissible: open microwave 1, go to sinkbasin 1... | Choose best next action.

Compute Infrastructure. Experiments were conducted on a workstation with $1 \times$ NVIDIA RTX 4090 GPU (24GB VRAM) and an AMD Ryzen 9 7950X CPU. The total compute budget for checkpoint caching, Jacobian lens decompositions, and all four intervention suites across 1,012 states was approximately 14 GPU-hours.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: The abstract and Section 1 explicitly specify our hypotheses, the scope of our five grounded ALFWorld predicates, and empirical findings. All claims are substantiated by empirical results in Section 3.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Section 5 is dedicated to discussing key limitations, including the textual simulator interface, evaluation on a single open-weight model family, binary decision-relevance assumptions, and distributed representations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper presents an empirical mechanistic study and does not claim theoretical theorems or proofs.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper?

Answer: [Yes]

Justification: Sections 2.1–2.6 provide complete mathematical formulas for action scoring, J-space predicate strength, calibration layer selection, and intervention protocols. Appendix A provides a self-contained student reproduction protocol.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Complete evaluation checkpoints, script generation tools, and analysis notebooks are provided in the supplementary zip package and will be released under an open-source MIT license upon de-anonymization.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 2 and Appendix C specify the exact episode-level split (20% calibration, 80% evaluation), model checkpoints, prompt format, tokenization rules, and J-space hyperparameters ($k = 25$, $\alpha \in [0, 1]$).

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All correlations report 95% bootstrap confidence intervals and episode-level permutation test p -values. Interventions in Figures 2 and 3 display standard errors and 95% bootstrap confidence bands across evaluated checkpoints.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Section 2.2 and Appendix C document compute specifications (a single workstation with $1 \times$ NVIDIA RTX 4090 GPU, requiring ≈ 14 GPU-hours for all calibration and intervention runs).

239 **9. Code of ethics**

240 Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS

241 Code of Ethics?

242 Answer: [Yes]

243 Justification: The research conforms to the NeurIPS Code of Ethics, utilizing only publicly available

244 academic software and simulators without ethical hazards.

245 **10. Broader impacts**

246 Question: Does the paper discuss both potential positive societal impacts and negative societal im-

247 pacts of the work performed?

248 Answer: [Yes]

249 Justification: Section 4 discusses how mechanistic calibration improves safety audits for embodied

250 AI agents by ensuring decisions rely on grounded environmental state rather than spurious verbaliza-

251 tions.

252 **11. Safeguards**

253 Question: Does the paper describe safeguards that have been put in place for responsible release of

254 data or models that have a high risk for misuse?

255 Answer: [N/A]

256 Justification: The experiments study a standard household simulator (ALFWorld) and open-weight

257 models, presenting no significant dual-use or safety risks.

258 **12. Licenses for existing assets**

259 Question: Are the creators or original owners of assets used in the paper properly credited and are

260 the license and terms of use explicitly mentioned?

261 Answer: [Yes]

262 Justification: ALFWorld (MIT License) and Qwen2.5 (Apache 2.0) are fully cited and their licenses

263 documented in Section 2 and Appendix C.

264 **13. New assets**

265 Question: Are new assets introduced in the paper well documented and is the documentation provided

266 alongside the assets?

267 Answer: [N/A]

268 Justification: The work creates benchmark evaluation checkpoint splits from existing environments

269 rather than releasing new foundational datasets or models.

270 **14. Crowdsourcing and research with human subjects**

271 Question: For crowdsourcing experiments and research with human subjects, does the paper include

272 the full text of instructions and details about compensation?

273 Answer: [N/A]

274 Justification: No human participants, human subjects research, or crowdsourcing was involved.

275 **15. Institutional review board (IRB) approvals**

276 Question: Does the paper describe potential risks incurred by study participants and whether Institu-

277 tional Review Board (IRB) approvals were obtained?

278 Answer: [N/A]

279 Justification: No human subjects or biological data were involved; IRB approval was not required.

280 **16. Declaration of LLM usage**

281 Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard

282 component of the core methods in this research?

283 Answer: [N/A]

284 Justification: The language model is evaluated as the scientific object of study (the agent policy), and

285 LLMs were not used as an unstated component of the experimental methodology.