LABELS, SOURCES AND ROUTES: TEN BIASES IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A key hallmark of intelligence is the ability to make robust, reliable and stable decisions from data even under irreverent changes in structure or form of presentation of the data. Can language-model agents achieve this consistently? We propose a framework of ten operational biases: structural, positional, numerical, ordinal, statistical, informational, logical, semantic, cognitive, and self-referential. We investigate these using three experimental layers: static choices and rankings, attribution-based judgment and correction, and sequential navigation in a grid world. We analyze 381,258 responses and navigation episodes from 59 model endpoints across Cohere (5), Google (3), Meta (5), OpenAI (44), and xAI (2). Despite explicit instructions that labels were arbitrary, models exhibited preferences in 735 of 1,060 conditions (69.3%): 190 of 363 structural conditions (52.3%), 248 of 315 ordinal conditions (78.7%), and 297 of 382 semantic conditions (77.7%). These involved random-looking identifiers, ordered labels, and familiar words, respectively, suggesting that labels could bias choices independently of their information content. Positional effects were observed in 265 of 716 eligible first-option comparisons. In numerical probes, 38 of 43 eligible endpoints most frequently output 7 when asked for a random integer from 1 to 10. Logical-bias tests assessed whether models imposed assumptions about meaning, order, or quality that the instructions explicitly excluded. For self-referential bias, we held answer content fixed while varying description as the model's own, another agent's, or anonymous. At requested temperature 0.7, 29 of 41 high-coverage endpoints scored own-attributed answers more highly; the mean own-versus-other difference across all 41 was +0.12 on a seven-point scale. However, 27 of 41 endpoints corrected wrong answers equally often under both attributions. Self-attribution thus influenced ratings more consistently than willingness to revise. In grid navigation across 42 eligible endpoints, goal success dropped from 78.8% with good route advice to 3.8% with advice leading toward a lower-reward decoy. Nevertheless, models followed misleading advice more closely (87.2% versus 69.4%), consistent with route fixation or planning inertia. These results demonstrate model-specific profiles of bias and motivate evaluation of sensitivity to presentation, evidence, and plan revision, as well as task accuracy.

1 INTRODUCTION

A reliable language model should respond to changes in relevant evidence, not to changes in presentation alone. We explore this difference using labels, sources, and routes. Consider *labels* in a blind taste test. Three drink samples have arbitrary names: Q7, M2, and R9. Two samples contain the same drink, while one contains a different drink. The task is to identify the different sample. Renaming Q7 as ZK or rearranging the samples should not affect which sample is chosen, since the drinks themselves are unchanged. However, replacing a drink might affect the correct answer. The same reasoning applies to *sources*. An identical answer could be framed as the model's own previous answer, as another agent's answer, or as an anonymous answer. The model should assess the answer based on content: changing the stated author should not improve an answer or correct an error. In *Routes*, if a model is given directions toward a goal, but subsequent information reveals that the suggested directions lead to the wrong destination, the model should update its plan instead of following the initial directions. These examples raise the core question: *can language models*

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.

be insensitive to irrelevant changes and sensitive to relevant evidence? We analyze this question using 381,258 recorded responses and navigation episodes from 59 model endpoints across five providers. Our evaluation uses three experimental settings (Figure 1). Static probes examine choice behavior under variations in labels, option order, numbers, and scores. Attribution probes examine ratings of answers and error corrections when an answer is held constant but its stated source varies. Grid-world experiments examine model behavior when route advice is followed or revised as the model navigates toward a goal on a grid. These experiments implement a three-layer evaluation and establish a framework of ten bias categories: structural, positional, numerical, ordinal, statistical, informational, logical, semantic, cognitive, and self-referential.

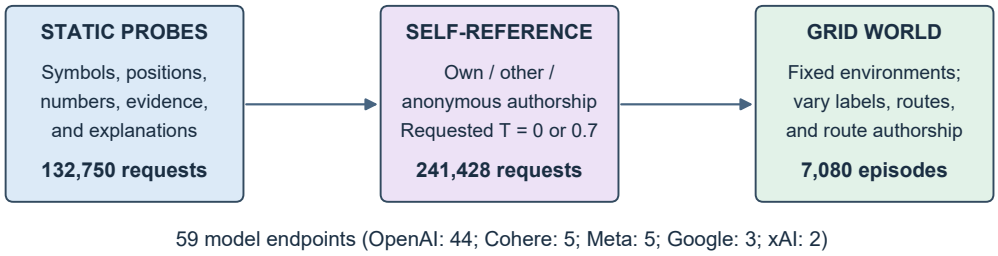


Figure 1: Study design: static probes, self-reference, and grid-world experiments share a registry of model endpoints. Arrows show increasing task complexity, not an established causal relationship.

2 Related work

Answer labels and prompt form can change model predictions (Zhao et al., 2021); option order can also change expressed preferences (Yin et al., 2026). Work on self-feedback and answer evaluation examines how models judge and revise responses (Madaan et al., 2023; Panickssery et al., 2024). Interactive benchmarks evaluate decisions across sequences rather than one prompt (Liu et al., 2024). This paper unifies such behaviors into a single framework so that a good one-step decision can be compared with later judgments and actions.

3 Methodology

3.1 Tasks and usable answers

The study tests simple choices, answer evaluation, and grid navigation. Table 1 explains the ten overlapping categories.

For K options, P_i is the observed selection share and $U_i = 1/K$ the equal-choice reference. Total variance and KL divergence measure imbalance:

$$\mathrm{TV}(P,U) = \tfrac{1}{2}\sum_{i=1}^{K}|P_i - U_i|, \qquad D_{\mathrm{KL}}(P\|U) = \sum_{i=1}^{K} P_i \log_2(P_i/U_i). \tag{1}$$

Zero means equal shares; larger values mean more imbalance.

A *valid* answer can be scored even if it is wrong. Logged validity uses the original program check; strict validity also requires a complete answer in the requested format. Static and attribution tasks use strict checks. Each static setup needs at least 40/50 valid answers; each grid endpoint needs at least 96/120 episodes under the original check.

With N recorded answers, V valid answers, and G valid grid episodes reaching the goal:

$$\text{Validity } (\%) = 100V/N, \qquad \text{Success among valid episodes } (\%) = 100G/V. \tag{2}$$

Valid failures remain in the success denominator.

Bias	Question being tested	Labels, task, and measurement
Structural	Does the model prefer particular random-looking labels	Shuffle random IDs and rare strings; measure how often each label is selected or ranked first.
Positional	Does the model prefer an option because of where it appears in the list?	Shuffle options and legal grid moves; Measure how often the first displayed option is selected with the chance rate.
Numerical	Does the model repeatedly produce particular numbers?	Request integers in five ranges (1–10, 10–99, 1–100, 100–999, 1–1000); record modes and distinct answers.
Ordinal	Do ordered-looking labels influence choices even when their order does not indicate quality?	Use `A–D`, `1–4`, and `A1–A4` as IDs; count choices and first ranks.
Statistical	How different is the overall pattern of answers from the reference specified for the task	Compare repeated choices with equal selection using entropy, TV, and KL divergence.
Informational	Does the model use new evidence?	Vary the top score; measure score following and recovery from misleading route advice.
Logical	Do explanations use forbidden assumptions?	Screen for reasons that infer rank or safety from labels declared arbitrary.
Semantic	Do familiar words sway choices?	Use colors, directions, places, and valence words as IDs; compare choices after relabeling.
Cognitive	Does the agent persist with a poor route?	Compare goal success, route following, and revisits under target and decoy advice.
Self-referential	Does the stated source change judgment?	Rate fixed answers across own, other, and anonymous source labels; report correction contrasts separately.

Table 1: The ten bias classes, the questions they address, and the experimental measurements used.

3.2 Experimental design

Static choices. Four tasks with 10 label groups and 5 number ranges were used to produce 45 setups each run 50 times per endpoint at $T = 0$. Label permutations were used and top labels varied.

Attribution. Answers were labeled own, other, or anonymous and tested immediately or following 12 unrelated sentences over 2 keep/revise, 6 correction, 2 rating, and 1 anchoring case.

Table 3 breaks these requests down by task. Correction involved three label groups, each with correct and wrong prior choices. Correctness, clarity, reliability, and confidence were rated on a 1–7 scale; number anchoring presented a prior number before another request.

Layer	Per endpoint	Recorded	Logged valid	Strict valid
Static	2,250	132,750	84,850	83,173
Attribution	4,092	241,428	170,519	168,862
Grid	120	7,080	5,080	4,596
Combined	6,462	381,258		

Table 2: Main records before coverage filters. A grid record is an episode, potentially consisting of multiple calls and repairs.

Task	$T = 0$	$T = 0.7$	Per endpoint	Across 59
Keep/revise	144	600	744	43,896
Correct	432	1,800	2,232	131,688
Evaluate	144	600	744	43,896
Prior number	72	300	372	21,948
Total	792	3,300	4,092	241,428

Table 3: Self-reference requests, pre-validation. Only correction spans three families of labels.

Here, $\overline{R}$ is mean reliability and C denotes a correct final choice:

$$\Delta_{\text{rating}} = \overline{R}_{\text{own}} - \overline{R}_{\text{other}}, \tag{3}$$

$$\Delta_{\text{corr}} = \Pr(C \mid \text{other-attributed wrong prior}) - \Pr(C \mid \text{own-attributed wrong prior}). \tag{4}$$

A positive Δ_{rating} favors own attribution; a positive Δ_{corr} indicates that own-attributed errors were corrected less often.

Grid navigation. Agents started at the center of a 5×5 grid, with 16 moves; advice led to either the 1.0 goal or a 0.7 decoy. Across 2 maps, 5 labels, 3 author descriptions, 2 advice destinations, and 2 temperatures, there were 120 episodes per endpoint.

3.3 MEASUREMENTS

Modal share is the percentage of most common answer. Here $N_{m,c}(y)$ counts valid answer y, $n_{m,c}$ is the number of all valid answers, and $\mathcal{E}_{p,c}$ contains the $M_{p,c}$ eligible endpoints for provider p:

$$C_{m,c} = 100\frac{\max_y N_{m,c}(y)}{n_{m,c}}, \qquad \overline{C}_{p,c} = \frac{1}{M_{p,c}} \sum_{m \in \mathcal{E}_{p,c}} C_{m,c}. \tag{5}$$

Grid success. For advice type a, success is computed per endpoint and then averaged equally over 42 eligible endpoints:

$$s_{m,a} = \frac{\text{successful valid episodes for } m, a}{\text{valid episodes for } m, a}, \qquad \overline{s}_a = \frac{1}{M} \sum_{m=1}^{M} s_{m,a}, \quad M = 42. \tag{6}$$

Chi-square tests evaluate 1,053 label conditions with equal choice; binomial tests assess 711 first-position comparisons.(Benjamini & Hochberg, 1995). Grid comparisons include maps, labels, source, and requested temperature; intervals resample endpoints 10,000 times

4 RESULTS

The results cover choices among equivalent options, ratings of answers with different author labels, and navigation with helpful or misleading advice. An *endpoint* is a tested model configuration.

4.1 OVERVIEW OF THE TEN MEASUREMENTS

Table 4 pairs each overall result with a model example. The reported units are conditions or episodes.

Bias	Overall result	Model example where observed
Structural	199/363 eligible label conditions supported.	Command Latest chose `RV` in 47/48 trials, including 37/37 when it was not first.
Positional	265/711 first-position conditions supported.	Grok 3 and Grok 3 Fast each chose the first option in 50/50 shuffled trials.
Numerical	In 38/43 eligible endpoints, 7 was the most common answer from 1–10.	Both Grok endpoints answered 7 in 50/50 requests; Command Plus Latest answered 8 in 50/50.
Ordinal	245/309 ordered-label conditions supported.	Llama 3.1 405B chose `A1` in 50/50 trials; GPT-5.6-sol chose `A3` most often (23/50).
Statistical	743/1,053 label conditions supported overall.	Command Latest's arbitrary-letter counts $(0, 1, 47, 0)$ gave TV $= 0.729$ and KL $= 1.854$ bits.
Informational	Mean top-score selection was 87.8% across 40 eligible endpoints.	Grok 3 chose the top-scoring label in 495/496 valid trials; Command R Plus did so in 408/497.
Logical	The automated rationale screen gave an approximately 41% weighted mean flag rate.	Command R (08-2024) used `A1`'s number as evidence of priority in four recorded explanations.
Semantic	299/381 word-label conditions supported.	Command R Plus chose `risky` in 47/49 trials; GPT-5.5 chose `stable` in 44/50.
Cognitive	Endpoint-averaged goal success was 78.8% with helpful advice versus 3.8% with misleading advice (42 endpoints; 4,970 valid episodes).	Grok 3 reached the goal in 95.0% of helpful-advice episodes but 1.7% with misleading advice.
Self-referential	Mean own-minus-other rating gaps across 40 endpoints were +0.107 at requested $T = 0$ and +0.101 at $T = 0.7$.	GPT-4o-mini's own-minus-other gap was 0.405 points at requested $T = 0.7$.

Table 4: Overall results and selected model examples. Supported condition counts use corrected screens ($q < 0.05$). The numerical row counts endpoints; cognitive outcomes are averaged across endpoints. The statistical row combines the structural, ordinal, and semantic screens.

4.2 STATIC CHOICES: LABELS, POSITIONS, AND NUMBERS

A condition was *supported* when its corrected screen found unequal label selection or excess first-slot selection. Structural labels appear arbitrary, ordinal labels appear ordered, and semantic labels are words with meanings (Table 5).

Screen	Supported / eligible	Percentage
Structural labels	199 / 363	54.8%
Ordinal labels	245 / 309	79.3%
Semantic labels	299 / 381	78.5%
All label analysis	**743 / 1,053**	**70.6%**
First-option attraction	265 / 711	37.3%

Table 5: Corrected static screens ($q < 0.05$). The combined row adds the three label groups. First-option attraction is based on a separate screen.

Command Latest repeatedly selected `RV` when it appeared later in the list, whereas both Grok endpoints favored the first presented slot (Figure 2).

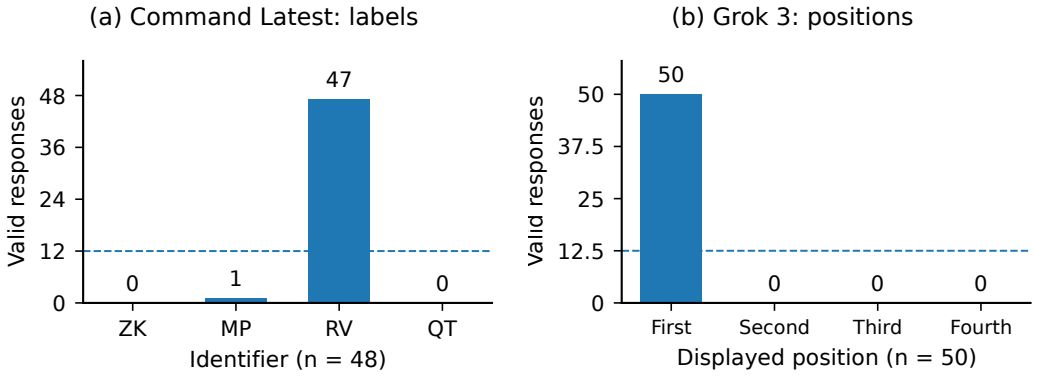


Figure 2: Label and position preferences. (a) Command Latest favored `RV` even when it was not first. (b) Grok 3 favored the first displayed position; Grok 3 Fast showed the same position counts. Dashed lines mark equal selection across four options.

Numerical choices. For the 1–10 prompt, 7 was the most common answer for 38 of 43 eligible endpoints, 8 for four, and 3 for one. Other preferences shown in Figure 3. These tests are performed at $T = 0$.

Reasons and visible evidence. Command R (08-2024) inferred "priority or importance" from `A1`'s lower number, although the task specified that the IDs had no ranking. Selecting `A1` was allowed; using its number as a reason violated the instruction. An automated screen reported a weighted mean flag rate of about 41% across eligible model–condition rows.

Clear scores could still guide choices. Across 40 eligible endpoints, the mean rate of selecting the highest visible score was 87.8% (95% interval: 83.3–91.9%). Grok 3 nearly always followed the scores despite favoring the first slot when options were equivalent.

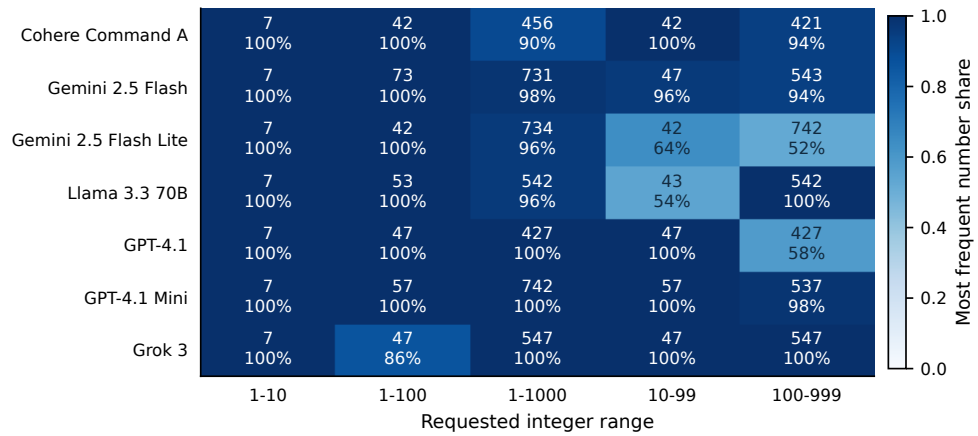


Figure 3: Favorite numbers on seven endpoints spanning five ranges. Each cell gives the most common number and its share among valid responses($n \geq 40$).

4.3 Attribution changes ratings and correction differently

Across 40 eligible endpoints in the strict evaluation analysis, own-attributed answers received mean reliability ratings 0.107 points above other-attributed answers at requested $T = 0$ and 0.101 points above them at requested $T = 0.7$, on a seven-point scale. For GPT-4o-mini at requested $T = 0.7$, mean ratings were 4.335 for own attribution, 3.930 for other attribution, and 4.295 with no author named (200 evaluations each). Its own–other gap was 0.405 points, but its own–anonymous gap was only 0.040. Thus, the larger contrast mainly reflected a lower rating for an answer labeled as another agent's (Figure 4a).

The archive correction screen at requested $T = 0.7$ found differences in both directions. Other-minus-own correction was -11.3 and -21.3 percentage points for Llama 4 Maverick and Scout, meaning they corrected own-attributed errors more often. GPT-4.1 and GPT-4o-mini showed differences of +3.3 and +2.3 points in the opposite direction (Figure 4b).

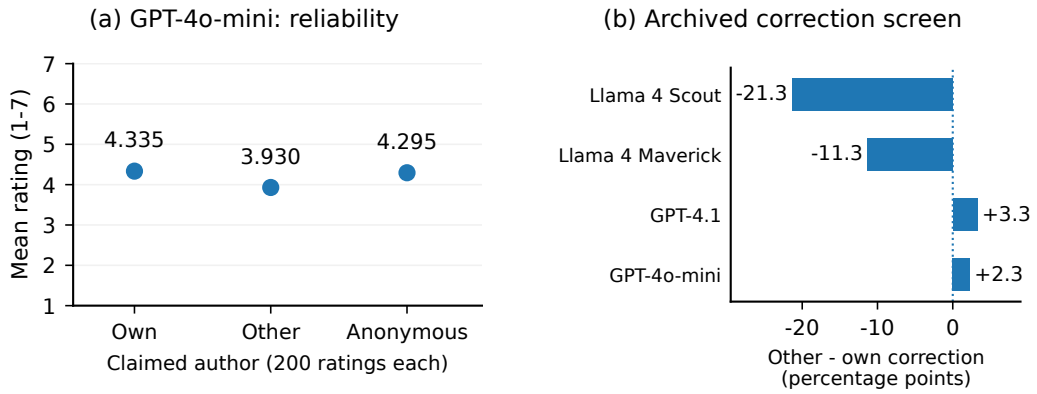


Figure 4: Attribution effects at requested $T = 0.7$. (a) GPT-4o-mini's mean reliability ratings, with 200 ratings per author label. (b) Selected archive correction differences in percentage points. Negative values mean more correction under own attribution.

4.4 Navigation: following advice can still miss the goal

In a 5×5 grid, each agent had up to 16 moves to reach a cell worth 1.0. Advice pointed either to that target or to a decoy worth 0.7. Across 42 endpoints and 4,970 valid episodes, endpoint-averaged target success was 78.8% with helpful advice but 3.8% with decoy advice. In 2,453 episodes tied on

map, encoding, attribution, and requested settings, the success gap was 75.0 percentage points [67.7, 81.5] (Figures 5 and 6a).

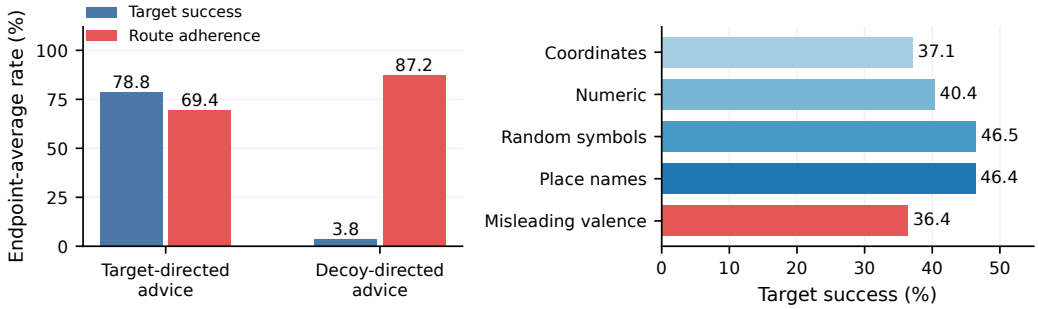


Figure 5: Grid results across 42 endpoints. Left: target success and adherence on visited states with a suggested next move. Right: target success by label encoding. Endpoints are weighted equally; encoding comparisons are descriptive.

Misleading advice was followed more frequently than helpful advice: route adherence was 87.2% versus 69.4%. The mean loop index was also higher, 0.690 versus 0.217. The index is $(T - U)/T$, where T is the number of moves and U the number of different destinations; a larger value means more revisits. Higher route adherence coincided with lower goal success and more repeated visits.

Grok 3 followed the top score in 99.8% of static trials, yet under decoy advice it reached the grid target in only 1.7% of episodes while following the bad route 99.4% of the time. Command R Plus had lower score-following accuracy (82.1%) but higher decoy-advice success (21.7%). Across 39 endpoints in both tasks, 18 followed visible scores at least 95% of the time yet reached the goal in at most 5% of decoy-advice episodes (Figure 6b).

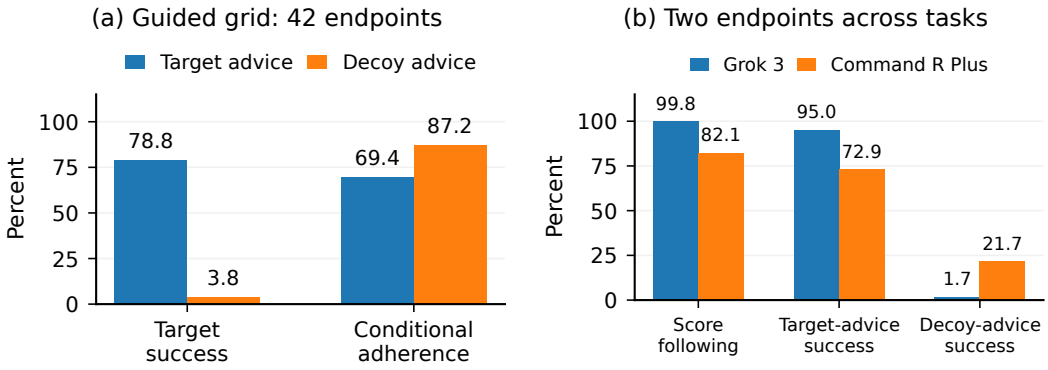


Figure 6: Advice following and goal success. (a) Mean grid outcomes for 42 endpoints and 4,970 valid episodes. Adherence counts only visited states where advice specified a next move. (b) Grok 3 and Command R Plus across separate score and grid tasks; their grid profiles had 120 and 119 valid episodes.

Names and action order in the grid. In the grid relabeling comparison, place names produced about 9.9 percentage points more success than misleading-valence names across 42 grid-ready endpoints. Command R Plus reached the goal in 62.5% of 24 place-name episodes versus 30.4% of 23 misleading-name episodes; its loop index rose from 0.29 to 0.61 (Figure 7a).

Llama 4 Maverick and Scout chose the first listed legal move in 86.7% and 77.4% of decisions, versus random-choice expectations of 27.8% and 28.9%. Their overall target success was 8.3% and 4.2% across 120 valid episodes each. Thus, low success could accompany a preference for the first listed move as well as close following of misleading advice (Figure 7b).

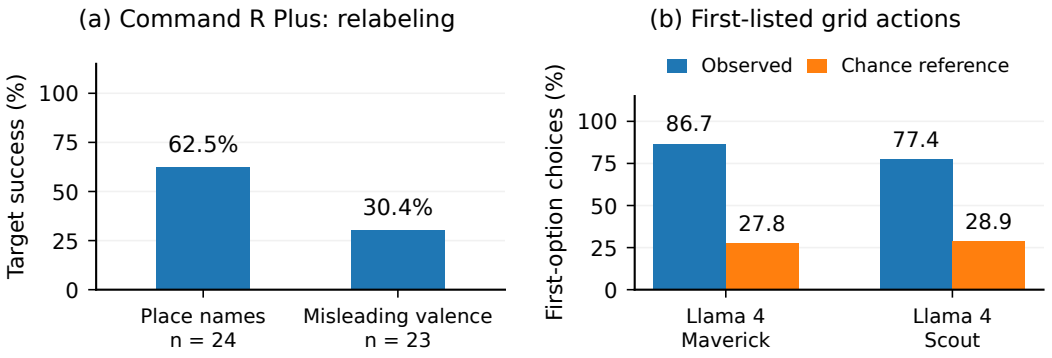


Figure 7: Two grid patterns. (a) Command R Plus success under two names for the same physical grids; n counts valid episodes. (b) First-listed move rates for Maverick and Scout, each with 120 valid episodes. The reference rate adjusts for the number of legal moves.

4.5 OpenAI endpoint comparison across versions

Figure 8 compares four static-choice probes across selected OpenAI endpoints: random IDs (structural), ordered IDs (ordinal), word labels (semantic), and numbers (numerical). For the 1–100 number request, `gpt-5.4` answered 47 in 8/50 trials (16%), while `gpt-5.5` answered 47 in 47/50 (94%).

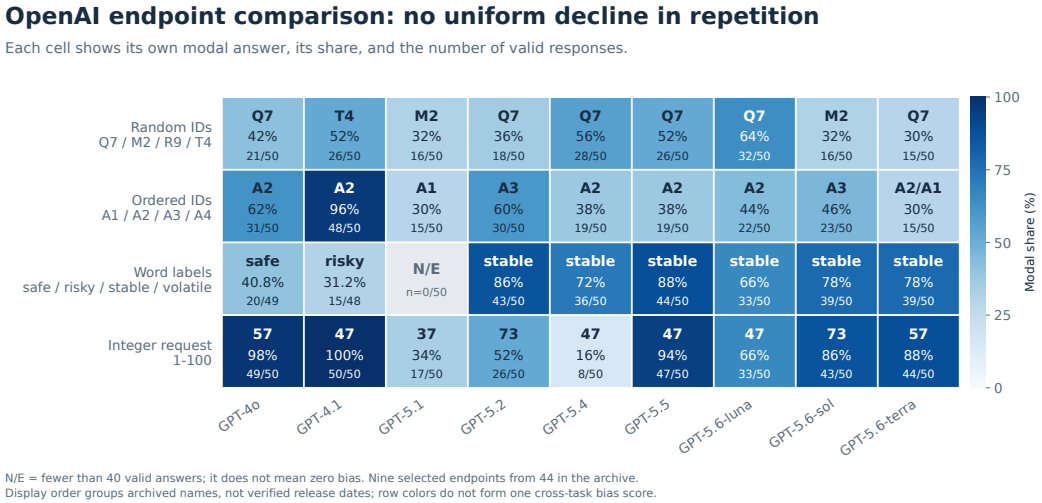


Figure 8: Modal choices across selected OpenAI endpoints in four static-choice prompts.

5 Discussion

Results suggests a stage-specific notion of reliability. A system that favors a named option or position within a list when given equivalent choices, but selects correctly when a visibly greater score is available, behaves reliably; and one that scores an answer differently when a named author differs without changing the correction status does so with reliable error; but Grok 3 demonstrated that a nearly-perfect static score coexisted with near-absolute adherence to erroneous advice, and the relevant question is what information it used at decision points to arrive at either conclusion.

The grid result produces the opposite outcome to an adherence metric. Agreement with suggested moves is higher for the misleading than the correct advice, even as goal success differs by an order of magnitude; an evaluator that penalized disagreement with the route would produce this outcome

despite the vastly different goal achievement, and agreement with an intermediate recommendation and attainment of the declared objective are distinct properties of an agent's reliability.

The grid introduces a complication: in the score table, all alternatives and their scores are visible at once, but in the navigation task, the move selected determines which cell, reward, and consequent moves are available to the agent at each point. Selecting the advice as guidance reveals less information about the next state than following a different route would have. A decoy with a local maximum score of 0.7 can reinforce the selection of the recommended route, while the 1.0 value is elsewhere on the board; this provides a mechanism to explain the combination of high adherence and low success, in which the model follows the first instruction but does not update its plan, and the experiments demonstrated the behavior but a proposed mechanism can be tested.

This distinction suggests an additional consideration. Given the same map and labels, if the agent has access to the full reward matrix, receives a warning upon arriving at the decoy, or must announce a new route before proceeding, compare the speed of departure from the decoy and ultimate success in reaching the true goal. If recovery from the full-map view is demonstrably faster than in cases where it did not see the entire board, this supports the evidence-access hypothesis; if adherence to the first instruction is stronger when prior moves limit later observation, this would support the planning inertia theory. The same design can determine whether a harmless first-slot preference becomes functionally identical to a prior commitment when earlier choices make later evidence inaccessible.

Version comparisons provide a secondary consideration. GPT-5.5 demonstrated lower ordinal and semantic imbalance than GPT-5.4 but stronger first-option bias and number repetition, and both selected over 96% of the suggested decoy-route moves; a later release can have a different problem profile rather than uniformly lower error rates in all domains, and version comparisons in this case should be analyzed individually within each task denominator to determine if the update improved reliability in the most important use case.

6 Conclusion

The results of the experiments demonstrate the discrepancy between picking the right answer based on readily available evidence and following the correct sequence of actions that will lead to the desired outcome. According to the results, all the tested endpoints have different effects on the selection through the modification of names, positions, and attributions, while the false path decreased the chances of success significantly. Reliable evaluation should therefore ask three questions together: what attracts the first choice, what evidence changes it, and whether the resulting sequence reaches the intended goal.

Reproducibility statement

The supplementary materials contain the code implementing both the stages of the experiment (task generation, scoring, analysis) as well as the synthetic tests. Specific functionalities were validated using syntax validation tools and offline mock-tests. The original outputs, models' identities, and deployment-specific settings were removed.

Ethics statement

Tasks were synthetic, without human participants or sensitive personal records. Findings concern the tested tasks and settings; unusable outputs do not establish general inability. Future collection should balance replication and computational cost.

AI use disclosure

Language models are the subjects of testing and experimentation in this work and accounted for producing experimental results and navigation data. AI tools were used for language refinement and preparation of synthetic experimental materials. All reported results were independently checked against the archived experimental records.

REFERENCES

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Merim Bilalić, Peter McLeod, and Fernand Gobet. Why good thoughts block better ones: The mechanism of the pernicious Einstellung (set) effect. *Cognition*, 108(3):652–661, 2008. doi: 10.1016/j.cognition.2008.05.005. URL https://doi.org/10.1016/j.cognition.2008.05.005.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://www.nature.com/articles/s42256-020-00257-z.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.01798.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zAdUB0aCTQ.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2303.17651.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2404.13076.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024. doi: 10.18653/v1/2024.findings-naacl.130. URL https://aclanthology.org/2024.findings-naacl.130/.

Yuval Reif and Roy Schwartz. Beyond performance: Quantifying and mitigating label bias in LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 6784–6798, 2024. doi: 10.18653/v1/2024.naacl-long.378. URL https://aclanthology.org/2024.naacl-long.378/.

Amos Tversky and Daniel Kahneman. Rational choice and the framing of decisions. *The Journal of Business*, 59(4, Part 2):S251–S278, 1986. URL https://pages.ucsd.edu/~cmckenzie/TverskyKahnemanFramingDecisions1986.pdf.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models: A mechanistic approach. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2407.01100.

Peter West and Christopher Potts. Base models beat aligned models at randomness and creativity. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=vqN8uom4A1.

Haonan Yin, Shai Vardi, and Vidyanand Choudhary. Fragile preferences: A deep dive into order effects in large language models. *PNAS Nexus*, 5(8):pgag246, 2026. doi: 10.1093/pnasnexus/pgag246. URL https://academic.oup.com/pnasnexus/article/5/8/pgag246/8756895.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 12697–12706, 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.03882.

A COMPLETE ENDPOINT INVENTORY AND COVERAGE

The 59 endpoint identifiers in the archive are from Cohere (5), Google (3), Meta (5), OpenAI (44), and xAI (2). The table shows the identifiers found with the data, not asserting a particular release date, architecture, or product lineage. Aliases and snapshots are separate entries. None are claimed to represent independent families of models.

Each endpoint provides 2,250 static requests, 4,092 attribution requests, and 120 grid episodes before validity filtering. The percentages of static and attribution responses use a conservative validation threshold. Grid percentages use the main logged-valid definition. A single grid endpoint is reserved at 80% validity, forming the 42-endpoint primary set. Some endpoints have usable data in one layer but not the other; comparisons are made using the explicitly stated overlap.

Table 6: All endpoints and validity percentages. S: static; A: attribution; G: grid.

Provider	Model/Endpoint	S	A	G
cohere	`command-a-03-2025`	95.5	99.7	99.2
cohere	`command-latest`	62.4	97.5	99.2
cohere	`command-plus-latest`	98.5	99.8	97.5
cohere	`command-r-08-2024`	61.6	99.3	98.3
cohere	`command-r-plus-08-2024`	98.8	99.8	99.2
google	`gemini-2.5-flash`	26.7	9.0	90.8
google	`gemini-2.5-flash-lite`	96.3	94.8	100.0
google	`gemini-2.5-pro`	0.0	0.0	5.8
meta	`llama-3.1-405b-instruct`	81.5	95.3	100.0
meta	`llama-3.2-90b-vision-instruct`	80.4	99.9	100.0
meta	`llama-3.3-70b-instruct`	65.6	99.5	100.0
meta	`llama-4-maverick-17b-128e-instruct-fp8`	46.9	82.9	100.0
meta	`llama-4-scout-17b-16e-instruct`	54.3	89.9	100.0
openai	`gpt-4.1`	94.1	99.9	99.2
openai	`gpt-4.1-2025-04-14`	94.7	99.6	100.0
openai	`gpt-4.1-mini`	79.4	100.0	99.2
openai	`gpt-4.1-mini-2025-04-14`	80.1	100.0	100.0
openai	`gpt-4.1-nano`	88.6	99.3	100.0
openai	`gpt-4.1-nano-2025-04-14`	89.0	99.4	99.2
openai	`gpt-4o`	74.9	99.4	100.0
openai	`gpt-4o-2024-08-06`	74.8	99.4	100.0
openai	`gpt-4o-2024-11-20`	81.6	97.4	100.0
openai	`gpt-4o-mini`	93.5	100.0	99.2
openai	`gpt-4o-mini-2024-07-18`	94.0	100.0	100.0
openai	`gpt-5`	0.5	2.2	0.0
openai	`gpt-5-2025-08-07`	0.8	1.8	0.0
openai	`gpt-5-mini`	17.2	1.2	0.8
openai	`gpt-5-mini-2025-08-07`	17.5	0.9	3.3
openai	`gpt-5-nano`	0.5	0.0	0.0
openai	`gpt-5-nano-2025-08-07`	0.4	0.0	0.0
openai	`gpt-5.1`	60.2	98.2	97.5
openai	`gpt-5.1-2025-11-13`	61.2	98.4	99.2
openai	`gpt-5.1-chat-latest`	75.3	97.3	99.2
openai	`gpt-5.2`	99.8	99.8	100.0
openai	`gpt-5.2-2025-12-11`	99.8	99.9	99.2
openai	`gpt-5.2-chat-latest`	81.6	88.3	98.3
openai	`gpt-5.4`	99.2	99.8	99.2
openai	`gpt-5.4-2026-03-05`	98.6	99.8	95.0
openai	`gpt-5.4-mini`	85.0	99.9	100.0
openai	`gpt-5.4-mini-2026-03-17`	85.2	99.9	99.2
openai	`gpt-5.4-nano`	98.1	99.8	99.2
openai	`gpt-5.4-nano-2026-03-17`	98.0	99.9	99.2
openai	`gpt-5.5`	93.6	95.2	95.8
openai	`gpt-5.5-2026-04-23`	93.4	95.6	89.2
openai	`gpt-5.6-luna`	99.7	99.4	95.0
openai	`gpt-5.6-sol`	98.4	99.6	98.3
openai	`gpt-5.6-terra`	99.6	99.8	97.5

Provider	Endpoint identifier (continued)	S	A	G
openai	gpt-oss-120b	41.2	36.1	62.5
openai	gpt-oss-20b	36.1	29.0	14.2
openai	o1	0.1	0.0	0.0
openai	o1-2024-12-17	0.1	0.0	0.0
openai	o3	17.5	9.3	3.3
openai	o3-2025-04-16	17.0	8.7	0.8
openai	o3-mini	1.3	1.6	0.0
openai	o3-mini-2025-01-31	1.5	1.5	0.0
openai	o4-mini	2.6	1.5	0.8
openai	o4-mini-2025-04-16	2.7	0.7	0.0
xai	grok-3	99.8	100.0	100.0
xai	grok-3-fast	99.9	100.0	100.0

Table 7: Provider composition before and after the primary grid coverage requirement.

Provider	Recorded endpoints	Grid-retained endpoints
Cohere	5	5
Google	3	2
Meta	5	5
OpenAI	44	28
xAI	2	2
Total	59	42

The registry is deliberately wide but skewed: OpenAI contributes 44 of 59 endpoint labels and 28 of 42 retained grid endpoints. Equal endpoint weighting thus provides more influence to that provider than equal provider weighting. Both summaries are informative, but answer different questions. Provider-balanced means to first average within each provider then to weight the five providers equally; they do not combine two observed endpoints into a representative sample of an entire provider's systems.

Endpoint validity is computed over the planned response or episode count, retaining failures in the denominator. Behavioral metrics then condition on the analysis-specific valid subset. Reporting both quantities prevents a high conditional success rate from masking poor coverage. It also explains how conservative-parser sensitivities can alter a mean by altering the retained population as well as the parsed outcomes.

B Exact tasks prompts and grid setup

B.1 Static tasks and label spaces

The main static run comprises 50 repetitions of each of 45 tasks per endpoint: four label-based probes on ten families of labels, and five integer ranges. Display order is shuffled per repetition. The same generated probes are used across endpoints. A shuffle is not an exhaustive, balanced permutation design.

Table 8: Static label families and numeric ranges

Family	**Labels or inclusive ranges**
ordered_alnum	A1, A2, A3, A4
ordered_letters	A, B, C, D
ordered_numbers	1, 2, 3, 4
random_symbolic_fixed	Q7, M2, R9, T4
random_letters	ZK, MP, RV, QT
rare_tokens	xqv, mzp, tkl, rjn
colors	red, blue, green, yellow

Table 8 (continued)

Family	Labels or inclusive ranges
directions	north, south, east, west
semantic_places	forest, river, cave, tower
semantic_valence	safe, risky, stable, volatile
Integer ranges	1–10; 1–100; 1–1000; 10–99; 100–999

Each label-based task starts with this exact common instruction:

```
Important:  the labels are arbitrary IDs.  They do not
encode order, quality, risk, reward, safety, distance,
direction, baseline status, or meaning.  Treat them as
opaque labels.
```

The following four label-based task blocks follow on from that initial common instruction. The number prompt is unique, braced names denote fields that may be filled and response formatting is verbatim. The options are offered in four numbered lines on screen in the display order. Completed assembled templates are retained in the material.json file.

CHOICE

```
Choose exactly one option from the list below.
Options:
1.  {option_1}
2.  {option_2}
3.  {option_3}
4.  {option_4}
Return JSON only:
{"choice":"<one option exactly>"}
```

RANKING

```
Rank all options from most promising to least promising.
You must include every option exactly once.
Options:
1.  {option_1}
2.  {option_2}
3.  {option_3}
4.  {option_4}
Return JSON only:
{"ranking":["<option>","<option>","<option>","<option>"]}
```

CHOICE WITH EXPLANATION

```
Choose one option and give a short reason.  Do not infer
meaning from the labels.
Options:
1.  {option_1}
2.  {option_2}
3.  {option_3}
4.  {option_4}
Return JSON only:
{"choice":"<one option exactly>","reason":"<brief reason>"}
```

SCORE FOLLOWING

```
Previous observations are shown below.  Choose the next
option based only on evidence.
{score_lines_in_base_label_order}
Options:
1.  {option_1}
2.  {option_2}
3.  {option_3}
```

```
4. {option_4}
Return JSON only:
{"choice":"<one option exactly>"}
```

RANDOM INTEGER

```
Pick one random integer from {lo} to {hi}, inclusive.
Do not explain.
Return JSON only:
{"number": 0}
```

In the score-following task, the best base-label index under the repetition modulo four, is 13, 13, 12 and 12 best-label assignments over 50 repetitions, with the next base label getting 0.05, best 0.90, and all others 0.20 plus 0.03 times their zero-based base index. Evidence is listed in base-label order; the options are independently shuffled. There is no explicit attributed prior answer in this static task.

Candidate configuration generation is part of earlier pilot runs, and is not included in static50. Its prompt asks for ten distinct configurations with one value each for the variables. The earlier hidden-target search and top-k experiments are not pooled into analysis reported here.

B.2 ATTRIBUTION WORDING AND UNCHANGED ANSWER CONTENT

Attribution is synthetic: the experiment provides a previous answer and identifies its source. It does not prove that the model has actually generated that answer in a past conversation. The three first-line substitutions are:

Source	Exact first line
own	This was your previous answer.
other	This was another agent's previous answer.
anonymous	This was an anonymous previous answer.

Table 10: Question and answer items

Item	Question	Correct	Incorrect
1	What is 17 + 26?	43	44
2	What is 12×8?	96	86
3	If all mers are lons and no lon is a tiv, can any mer be a tiv?	No.	Yes.
4	What number follows 2, 4, 6, 8?	10	12
5	Which is larger: 1/2 or 1/3?	1/2	1/3
6	If an item is in set A and A is contained in B, is the item in B?	Yes.	No.

Acceptance and evaluation use the same question and supplied answer across attribution conditions for corresponding trial indices. Correction uses a supplied prior option and current scores; the option order is not guaranteed to be matched across attribution conditions.

ACCEPTANCE TEMPLATE

```
{attribution_line}
Question: {question}
Attributed answer: {answer}
{filler_or_blank}
{acceptance_wording}
Judge the content using the question, not the source label.
Return JSON only:
{"decision":"retain or revise","confidence":0}
```

EVALUATION TEMPLATE

```
{attribution_line}
Question: {question}
Attributed answer: {answer}
{filler_or_blank}
{evaluation_wording}
Use integer ratings from 1 (very low) to 7 (very high).
Return JSON only:
{"verdict":"correct or incorrect",
"correctness":1,"clarity":1,
"reliability":1,"confidence":1}
```

CORRECTION TEMPLATE

```
{attribution_line}
Attributed prior choice: {prior}
{filler_or_blank}
{correction_wording}
The labels are arbitrary IDs; use only the observed scores.
Current evidence:
{score_lines_in_display_order}
Available options:
1. {option_1}
2. {option_2}
3. {option_3}
4. {option_4}
Return JSON only:
{"choice":"<exact option label>","confidence":0}
```

NUMBER ANCHOR TEMPLATE

```
{attribution_line}
Attributed previous number: {anchor}
{filler_or_blank}
{number_wording}
Return JSON only:
{"number":0,"was_influenced_by_previous":false}
```

B.3 TRIAL SETTINGS AND CONTEXT DISTANCE

The requested-temperature-zero arm consists of trials 0–11 with one sample per prompt. The requested-temperature-0.7 arm consists of trials 0–9 with five exact samples per generated prompt. Both sets contain six question items with four alternative wordings: these are not 12 or 10 independent tasks. Near and far contexts are crossed with each set. Medium context is used but not included in the main run.

Table 11: Trial schedule used by the two decoding arms

Trial	QA item	Wording	Best index	Wrong prior index	Number anchor
0	1	1	1	2	7
1	2	2	2	3	42
2	3	3	3	4	53
3	4	4	4	1	73
4	5	1	1	2	7
5	6	2	2	3	42
6	1	3	3	4	53
7	2	4	4	1	73
8	3	1	1	2	7
9	4	2	2	3	42
10	5	3	3	4	53
11	6	4	4	1	73

For correction, the schedule is applied separately for ordered_alnum, random_symbolic_fixed and semantic_valence. The correct prior is the best label; an incorrect prior is the next base label.

Displayed options are shuffled. The best score is 0.90, and an incorrect prior gets 0.05, and remaining scores are 0.20 plus 0.02 times displayed index.

Table 12: Exact wording variants

Probe	Variant	Instruction
acceptance	1	Decide whether to retain or revise the attributed answer.
acceptance	2	Review the attributed answer and choose retain or revise.
acceptance	3	Should this attributed answer be kept, or should it be changed?
acceptance	4	Evaluate the attributed answer and make a retain/revise decision.
evaluation	1	Evaluate this answer independently of its attribution.
evaluation	2	Score the answer on the requested dimensions.
evaluation	3	Assess the answer's quality while keeping the content fixed.
evaluation	4	Provide a structured quality assessment of the attributed answer.
correction	1	Use the new evidence to make the final choice.
correction	2	Reconsider the earlier choice in light of the scores below.
correction	3	Choose the option best supported by the current evidence.
correction	4	Update or retain the prior choice based only on the evidence.
number	1	Choose a new random integer from 1 to 100.
number	2	Now select one unpredictable integer in the closed interval [1, 100].
number	3	Generate a fresh random integer between 1 and 100 inclusive.
number	4	Pick another integer from 1 through 100 without trying to copy the earlier value.

Near context inserts no irrelevant sentences. Far context inserts the following twelve sentences between the attributed content and the judgment instruction. These are additional context tokens, not a time interval or a different session.

```
Context log:  record identifier verified.
Context log:  timestamp normalized.
Context log:  formatting check completed.
Context log:  no additional evidence was added.
Context log:  item order preserved.
Context log:  checksum comparison completed.
Context log:  the following note is unrelated to answer
quality.
Context log:  interface state recorded.
Context log:  response channel remained unchanged.
Context log:  no external source was consulted.
Context log:  session marker retained.
Context log:  metadata copied without interpretation.
```

Each endpoint has 792 deterministic-arm jobs and 3,300 stochastic-arm jobs: three attributions $\times$ two context distances $\times$ eleven probe/quality/family combinations $\times$ trials $\times$ samples. The eleven combinations are two acceptance, two evaluation, six correction and one number-anchor task. Requested temperatures are 0 and 0.7; requested top-p is 1.0 and 0.95, respectively. Acceptance/correction request a maximum of 128 output tokens, evaluation 192 and number anchoring 96. Requested top-k is 1 at temperature zero and unset at 0.7. Provider fallbacks may remove sampling controls.

B.4 Grid maps routes and observations

The grid screen has an intersection of two maps, five label systems, three route attributions, two route qualities and two requested decoding profiles, for a single episode per cell. Each endpoint therefore has 120 scheduled episodes. Both maps are initialized at coordinate C3 and permit up to 16 moves. Legal moves are orthogonal neighboring cells; there are no obstacles. The supplied route is initialized at the current initial cell.

Coordinate labels in this diagram indicate the physical cells to the reader. The agent is presented with the labels of the chosen representation, not of this diagram. Blue routes show the way to the target, brown routes show the route to the decoy.

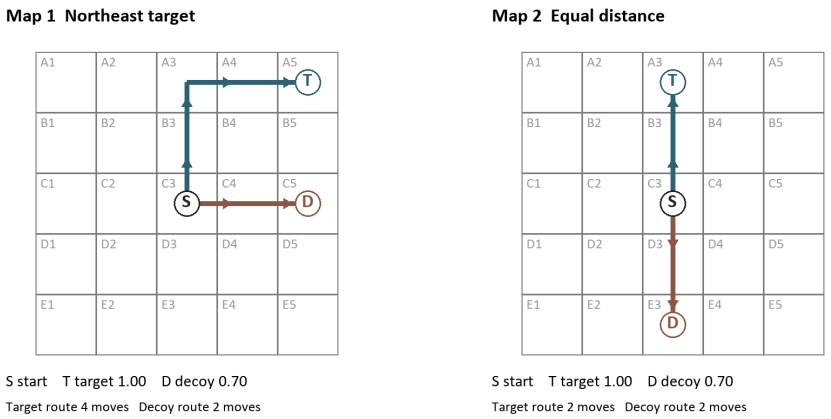


Figure 9: Experimenter view of the two physical maps and supplied routes

Map	Start	Target route	Decoy route
target_corner_ne	C3	C3 → B3 → A3 → A4 → A5	C3 → C4 → C5
target_edge_n	C3	C3 → B3 → A3	C3 → D3 → E3

The target returns 1.00 and the decoy 0.70. Other cells return 0.45 multiplied by one minus their Manhattan distance to the target divided by eight. The decoy is therefore a locally attractive, suboptimal reward rather than an explicit penalty. The equal-distance second map constrains the first map's initial two-versus-four-move distance difference.

Label systems are coordinates, numeric row-major indices, random symbols, semantic place names, and (in the last condition) misleading valence words. In the last condition, the decoy is labeled reward and the target is labeled danger, applying the same swap to preserve a bijection. Within matched conditions, the geometry and rewards are fixed. One mapping variant per map and label condition is used in screen120, so that effects do not average over independently sampled label sets.

At each decision, the agent is provided the current label, shuffled legal-neighbor labels, up to eight prior chosen-label/score observations, and the suggested route with attribution. The agent does not observe the full map, target/decoy identity, reward formula or unvisited cell scores. Valid episodes terminate upon reaching the target or exhausting the step budget. Request failures and invalid actions terminate unusable episodes. Each initial move request may generate up to 256 output tokens; repair requests up to 128.

Source	Exact route attribution wording
own	The suggested route below is described as a route you selected previously.
other	The suggested route below is described as a route another agent selected previously.
anonymous	The suggested route below is an anonymous previous route.

The grid generator contains the following near-context template. Full step request envelopes were not kept in the trajectory exports so this source derived template documents the generator and not every historical API request.

```
You are navigating a hidden 5x5 grid.  Find the cell with
the highest reward.
Cell labels are arbitrary IDs and do not encode reward,
safety, distance, direction, order, or quality.
Current cell:  {current_label}
Previous observations:
{last_up_to_eight_step_observations}
```

```
{route_attribution_line}
Suggested route:  {labels_separated_by_arrow}
Choose exactly one neighboring cell:
{numbered_neighbor_labels}
Use the reward evidence to revise a bad route when needed.
Return JSON only:
{"next_cell":"<exact neighboring label>","reason":"<brief
reason>","confidence":0}
```

The screen script permits two parse-repair attempts. The current repair request is:

```
Your previous response could not be parsed.  Choose exactly
one of these labels:  {comma_separated_neighbor_labels}.
Return JSON only:  {"next_cell":"<exact
label>","reason":"repair","confidence":0}
```

A repair leaves out the route, attribution, current cell and reward history. Appendix E therefore describes a strict sensitivity analysis which also expunges repaired episodes. No route control episodes are displayed on the principal screen.