REPRESENTATION AND MOTIVATION ARE SUPERADDITIVE: UNLOCKING EXPLORATION FOR LLM AGENTS IN STRUCTURED SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM agents are increasingly expected to discover solutions beyond what is already known, yet prior work shows that they search locally, optimizing among known options rather than pursuing unexplored alternatives. We identify an upstream representational bottleneck in structured search: when alternatives must be recovered from relational history, unexplored regions must first be represented as candidate destinations before motivation can redirect choice. In a task adapted from Wu et al. (2018), baseline agents across DeepSeek V4.1 Flash, Qwen3.8-flash, and GPT-4.1-mini select new positions with mean uncertainty percentiles of only 1.6–3.6, compared with 22.8 for matched humans; the deficit persists when provider-native reasoning is enabled and in a GPT-5.6 terra diagnostic. Factorially combining a coverage representation with topology-free motivation produces superadditive gains in search extent: the joint effect exceeds the sum of the individual effects by 4.2–12.6 uncertainty-percentile points. The combination raises the uncertainty percentile to 21.9–23.2, closing the roughly order-of-magnitude gap to the human benchmark. Same-information controls isolate coverage status from generic history summarization, while rendering controls show that computing the same coverage fields does not guarantee their use in choice. In a restless bandit with named arms and 857 human participants, the same motivation increases switching without a positive interaction. Together, the task contrast identifies a structural boundary: motivation acts directly among named alternatives but gains leverage from representation when destinations must be recovered from history.

1 INTRODUCTION

Scientific discovery, design, software development, and adaptation to unfamiliar environments require agents to test alternatives outside familiar trajectories (Antoniades et al., 2026). Yet existing agents often remain near known-good options (Krishnamurthy et al., 2024; Schmied et al., 2026). Krishnamurthy et al. (2024) obtain robust bandit exploration only by pairing chain-of-thought with an externally constructed summary of per-arm statistics, leaving open how to organize history beyond arm-wise state. Subsequent methods supply algorithmic guidance, online feedback, or explicit reasoning (Nie et al., 2025; Monea et al., 2025; Zhang et al., 2025b); Arumugam & Griffiths (2026) use LLMs to implement posterior sampling for reinforcement learning. Work beyond bandits optimizes checkpoint coverage (Ye et al., 2026), jointly trains memory and exploration with code-coverage novelty (Tian et al., 2026), or diversifies candidate actions within tree search (Tang et al., 2026). We ask how to enable LLM agents to move beyond local search. In structured search, unexplored regions must be recovered from relations among past choices and outcomes. We identify a bottleneck in bringing that information into choice and test a training-free intervention that addresses it.

Our intervention begins from two requirements for moving beyond local search. The agent must identify where it has not searched and value learning what those areas contain. A coverage representation addresses the first requirement by reorganizing past choices into visited and unvisited positions. Its sentence rendering is part of the intervention: it presents coverage information as candidate destinations without predicting their rewards or recommending an action. Topology-free motivation addresses the second requirement by giving the agent a reason to leave a known-good

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.

option without locating uncertain alternatives. We call coverage information *actionable* when it is available for use in choosing where to search. We predict that representation and motivation have superadditive effects on non-local search: making recoverable coverage information actionable increases motivation's effect. When alternatives are independently named, the task already supplies candidate destinations.

We test this argument across two tasks that differ in how unexplored alternatives become available for choice. In the primary structured-search task, adapted from Wu et al. (2018), agents search 16 smooth but hidden payoff landscapes, each containing 30 ordered positions. After one position is revealed, they make five or ten choices and observe the reward only at each selected position. The finite action set lets us vary whether identical histories are represented as candidate destinations while holding actions and outcomes fixed. Across DeepSeek V4.1 Flash, Qwen3.8-flash, and GPT-4.1-mini ($n = 50$ per factorial cell), we cross the coverage representation with topology-free motivation. Our primary measure ranks each newly selected position by its uncertainty among positions not yet observed, with matched human trajectories providing the benchmark. We then cross the same motivation with an arm-level summary in a four-arm restless bandit with named arms and 857 human participants. Because its arms are named, this task tests whether motivation can act without first recovering candidate destinations from history.

Baseline agents in structured search remain almost entirely around known locations. Across the three model families, newly selected positions have mean uncertainty percentiles of only 1.6–3.6, compared with 22.8 for matched humans. Motivation alone has little effect on the uncertainty percentile, while representation alone produces a partial shift away from local search. Their joint effect exceeds the sum of their individual effects by 4.2–12.6 uncertainty-percentile points. The combined intervention raises the percentile to 21.9–23.2, close to the human benchmark. In the restless bandit with named arms, motivation raises switching and has a small negative interaction with the arm-level summary. These results separate departure from destination choice: motivation can increase departure while coverage representation changes where structured search goes.

We then separate information availability from use in choice. A same-information control tests whether reorganizing observed rewards, in the spirit of per-arm summaries, extends to structured search: it leaves unvisited positions unidentified and produces a much smaller shift than the coverage representation. Models can recover substantial coverage information when queried, but whether that information changes choice depends on rendering: sentence rendering produces much larger shifts than a four-line list. Token-level probes show a corresponding shift toward distant actions, amplified after the generated rationale. Positive interactions persist for the logit transform of uncertainty percentile, its unconditional top-quartile share, and its product with new-position probability, as well as across prompt specifications, rough and two-dimensional landscapes, and DeepSeek V4.1 Flash provider-native reasoning.

Contributions. **(1)** We establish superadditive effects of coverage representation and motivation on search extent across three model families. **(2)** We provide a training-free intervention that pairs sentence-rendered coverage information with topology-free motivation, bringing non-local search close to the human benchmark. **(3)** Same-information and rendering controls, together with a cross-task comparison, show how coverage representation increases motivation's effect in structured search, while motivation increases departure among already named alternatives.

2 Experimental Design and Measurement

2.1 Tasks and human benchmarks

Our design varies whether exploration requires recovering relations among actions. Structured search is the primary setting; the restless bandit provides independently named alternatives. The cross-task analysis asks whether representation adds leverage within each task, not whether task-specific interactions have equal magnitudes. Human trajectories provide post hoc benchmarks; agents never receive them, and they are not used to fit agent policies.

The structured task reproduces the one-dimensional condition of Wu et al. (2018) from the original code. Each environment contains 30 ordered positions drawn from a radial-basis-function Gaussian process ($\ell = 2$) (Rasmussen & Williams, 2006), producing separated high-value regions. One po-

sition is revealed initially; the agent then makes five or ten choices across 16 environments, earning slightly noisy rewards on revisits. Our benchmark is the cumulative-reward, smooth-landscape cell ($n = 19$) from the 81-participant study.

A dimensional replication runs the complete DeepSeek V4.1 Flash factorial on the matched 11 $\times$ 11 grid condition from Wu et al. (2018); Appendix A gives the representation, horizons, and human benchmark.

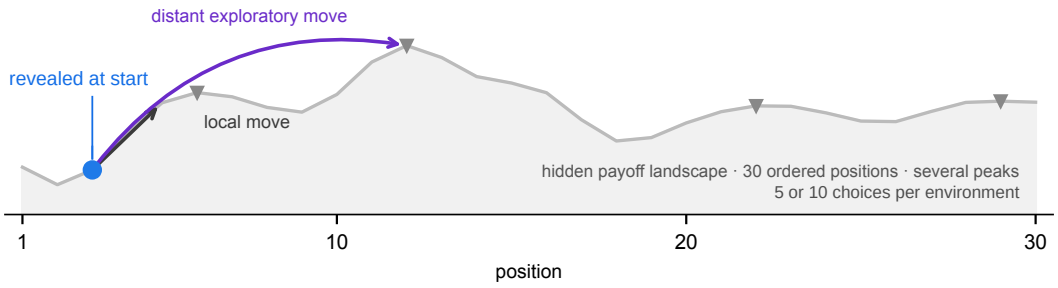


Figure 1: **One-dimensional structured search.** Each environment contains 30 ordered positions on a hidden, smooth, multi-peaked payoff landscape. One position is revealed initially, after which the decision maker makes five or ten choices. Effective exploration requires linking the observed history to unvisited regions that can be reached by subsequent actions.

The second task uses the four-armed restless bandit dataset included in Psych-101 (Binz et al., 2025). Four letter-named arms generate independently drifting rewards over 150 choices. Their names make every alternative a directly named candidate destination, so switching measures whether the agent leaves its current arm. We compare agents with all 857 source participants. Table 1 summarizes the task contrast, representations, and measures.

2.2 Agents, Interventions, and Evaluation Protocol

We evaluate DeepSeek V4.1 Flash, Qwen3.8-flash, and GPT-4.1-mini with 50 independent agents per primary factorial cell. Each receives the original instructions and its complete choice–reward history, then produces a generated rationale and delimited action. Appendix A gives prompts, identifiers, decoding settings, seeds, parser diagnostics, provider-native reasoning, and additional-model tests.

The experiment crosses representation, $R \in \{0, 1\}$, with topology-free motivation, $M \in \{0, 1\}$. The structured-search representation deterministically renders coverage information as sentences: visited and unvisited positions, the longest unvisited interval, and the best observation. For example, after observing positions 5 and 14–16, the renderer includes: "26 of the 30 options have never been touched in this environment: options 1-4, options 6-13, options 17-30." The two-dimensional version describes touched regions, untouched rows and columns, the largest untouched rectangle, and the best observation. In the bandit, an arm-level summary reports each arm's count, mean and latest reward, and time since selection. Appendix A gives full examples and reduced-representation ablations.

The motivation prompt couples aspiration with an intrinsic value for learning (Cyert & March, 1963; March, 1991). It contains the same two sentences in both tasks: "You are hard to satisfy: even when an option pays well you keep wondering whether another pays more, and you go and check fairly often." "Figuring something out is worth something to you in its own right; what you come to understand counts as part of what you came away with." Neither sentence supplies spatial or reward-structure information.

The full-interface condition renders the human board and feedback in text. A same-information control reorganizes observed readings in the same format and at similar length as the coverage representation while leaving untouched positions unnamed. The semantic frame (motivation + topology) interprets the task as a segmented pricing market; it and a reduced-motivation variant are detailed in Appendix A.

We also test whether agents can recover the required coverage information themselves. At 40 mid-environment decision points per model, reconstruction questions ask for the unvisited set, longest unvisited run, and best observed position using only the raw history. In separate DeepSeek V4.1 Flash experiments, the model writes free-form notes or completes a four-field coverage template before every choice. A complete $2 \times 2 \times 2$ then holds motivation present and crosses coverage content (model-written or ground truth), attribution (an own-notes label or none), and rendering (the four-line list or the coverage description rendered as sentences). Model-written content is parsed and formatted without correcting its errors, and every coverage block appears between the history and the choice question. The design separates computation of coverage information from its use in choice across renderings.

Task	Action structure	Representation	Primary exploration measure	Human reference
Structured search (Wu et al., 2018)	30 ordered positions; spatially correlated rewards (RBF, $\ell = 2$); 16 environments with 5 or 10 choices	Visited and unvisited positions, longest unvisited interval, and current best observation	Uncertainty percentile among unvisited positions	22.8 ($n = 19$)
Two-dimensional replication (Wu et al., 2018)	121 grid tiles; spatially correlated rewards; eight environments with 20 or 40 choices	Touched regions, untouched rows and columns, largest untouched block, and current best observation	Uncertainty percentile among unvisited tiles	14.9 ($n = 22$)
Restless bandit with named arms (Binz et al., 2025)	Four independently named arms with independently drifting payoffs; 150 choices	Per-arm counts, mean and latest reward, and time since last choice	Switch rate	0.391 ($n = 857$)

Table 1: **Task structure, representation, and primary exploration measures.** The topology-free motivation prompt is reproduced verbatim across tasks and uses task-general aspiration and learning language while holding spatial, topological, and reward-structure content absent. Each main factorial cell contains 50 independent agents per model.

2.3 Measures and estimands

Our primary structured-search measure is the uncertainty percentile of each newly selected position, conditional on the choice being new. Let $\mathcal{U}_t$ denote the positions unobserved before choice t, and let $r_t(j)$ be the ascending rank of posterior standard deviation $\sigma_t(j)$ among $\{\sigma_t(k) : k \in \mathcal{U}_t\}$, using average ranks for ties. We compute the posterior with the task's Gaussian-process kernel from observed locations only; posterior uncertainty is confined to evaluation and excluded from agent contexts. For a newly selected position a_t,

$$q_t(a_t) = 100 \, \frac{r_t(a_t) - 1}{|\mathcal{U}_t| - 1}. \tag{1}$$

Thus, $q_t = 0$ selects the most certain unobserved position, $q_t = 100$ the most uncertain, and uncertainty-blind choice has expectation 50. The percentile measures the extent of search away from well-observed regions. Best-found value measures discovery quality, while per-step reward measures payoff under the cumulative-reward instructions. These outcomes can diverge: uniform-random choice reaches percentile 46.2 but best-found value 87.4%, compared with 93.9% for humans (Appendix Table A4).

In two dimensions, the posterior uses Euclidean grid distance. We average q_t within individual over newly selected positions, then weight individuals equally. Companion measures are new-position probability, the unconditional share of all choices in the top uncertainty quartile among unobserved positions, and the product of new-position probability and conditional uncertainty percentile. The primary bandit measure is switch rate, the share of choices that change arms. A Kalman-filter evaluator using the known restless dynamics ranks switch destinations by posterior uncertainty and mean, separating departure from destination choice.

A *far jump* moves at least five positions from the preceding choice, while a *re-click* selects a position already observed in the current environment. Both best-found value and per-step reward are divided by that environment's attainable maximum. For any outcome y, the representation-by-motivation interaction is

$$\Delta_{\text{int}} = (\bar{y}_{11} - \bar{y}_{01}) - (\bar{y}_{10} - \bar{y}_{00}), \tag{2}$$

where the first subscript indexes representation and the second topology-free motivation. Positive Δ_{int} means that their joint effect exceeds the sum of their separate effects on the additive scale of outcome y. It is our primary factorial test of the prediction that motivation becomes more effective after relational history has been converted into candidate destinations. Because factorial cells share environment seeds 1000–1049, primary confidence intervals resample seeds as blocks, carrying all four cell observations for a seed together. Human intervals resample participants; other analyses resample their stated observational unit (Efron & Tibshirani, 1993).

Situation responsiveness is analyzed separately through a per-individual logistic model of far jumps (Appendix D).

2.4 Reference policies and mechanism probes

Gaussian-process upper-confidence-bound policies provide computational references on the same landscapes; Appendix E defines them and Table A4 reports a reference policy alongside uniform-random choice.

Token-level probes separate spatial locality from numeric-label effects by comparing action probabilities under the original labels, random permutations, and two-digit codes. At matched DeepSeek V4.1 Flash and GPT-4.1-mini decision states, we compare distant-action mass at the bare action head and after the generated rationale under baseline instructions, the joint intervention, and the semantic frame (Appendix E).

3 Results

3.1 Baseline agents remain local despite richer interfaces and native reasoning

Baseline agents concentrate choices near known rewards. Across DeepSeek V4.1 Flash, Qwen3.8-flash, and GPT-4.1-mini, newly selected positions have mean uncertainty percentiles of only 1.6–3.6, compared with 22.8 for matched humans; best-found values reach 78.1–83.3% of the attainable maximum, compared with 93.9% for humans. Rendering the complete human-facing board changes the uncertainty percentile by at most one point (Appendix Table A4).

Provider-native reasoning intensifies rather than resolves locality. With identical DeepSeek V4.1 Flash prompts and environment seeds, enabling it lowers the baseline uncertainty percentile from 1.6 to 0.9; under the joint intervention the percentile reaches 10.8, with a positive representation-by-motivation interaction of +6.6 [+5.4,+7.9] points. Far-jump behavior shifts in the same direction. A separate GPT-5.6 terra diagnostic shows the same qualitative pattern, with the joint intervention raising the uncertainty percentile from 0.8 to 6.3. In the GPT-5.6 terra diagnostic, best-found value rises from 79.3% to 85.2% as per-step reward falls from 70.9% to 63.3% (Table 2; Appendix Tables A11 and A18).

Provider-native reasoning	Unc. pct.	Far jump	Re-click	Best-found
DeepSeek V4.1 Flash (off)	1.6 / 23.2	.026 / .408	.428 / .151	78.1 / 86.6
DeepSeek V4.1 Flash (on)	0.9 / 10.8	.001 / .180	.516 / .153	79.4 / 88.7
GPT-5.6 terra (on)	0.8 / 6.3	.001 / .155	.506 / .066	79.3 / 85.2

Table 2: **Provider-native-reasoning diagnostics.** One-dimensional search, $n = 50$ agents per condition. Entries report baseline/joint-intervention pairs.

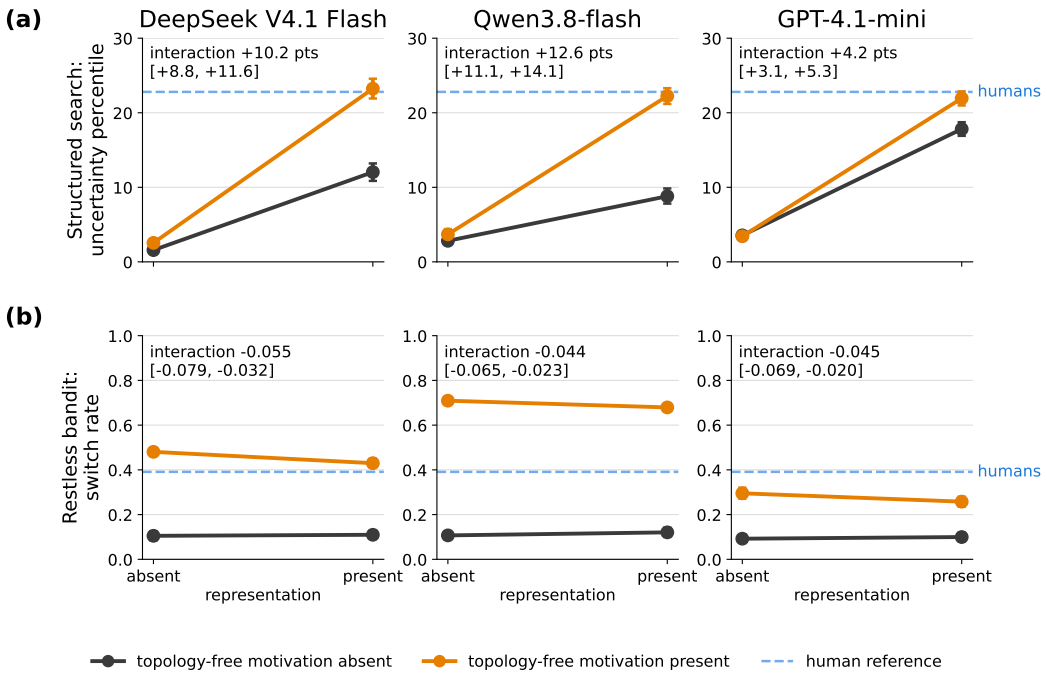


Figure 2: **Representation and motivation have superadditive effects on search extent.** In structured search, the joint intervention exceeds the sum of the individual effects (a). In the restless bandit with named arms, motivation directly raises switching and interactions are small and negative (b). Interaction intervals are seed-paired bootstrap 95% CIs; dashed lines show human references.

3.2 Representation and motivation superadditively extend search

Coverage representation and motivation have superadditive effects on search extent: their combined effect exceeds the sum of their individual effects. Motivation alone changes the uncertainty percentile by only -0.1 to +1.0 points across the three models, whereas representation alone increases it by 6.0–14.3 points. Combining coverage representation with motivation raises the uncertainty percentile to 21.9–23.2, close to the human benchmark of 22.8. The seed-paired representation-by-motivation interactions are +10.2 [+8.8,+11.6], +12.6 [+11.1,+14.1], and +4.2 [+3.1,+5.3] points for DeepSeek V4.1 Flash, Qwen3.8-flash, and GPT-4.1-mini, respectively (Figure 2a).

Uncertainty-percentile interactions remain positive on a per-agent logit scale, in the unconditional top-uncertainty-quartile share (+5.7, +5.9, and +2.7 points), and after deleting the direct "go and check" clause (+8.1, +8.5, and +3.3 points; Appendix Tables A7, A20, and A19).

The joint intervention raises one-dimensional best-found value by 4.2–9.0 points, to 86.6%, 89.8%, and 87.5% for DeepSeek V4.1 Flash, Qwen3.8-flash, and GPT-4.1-mini. All remain below the human 93.9%; human-scale search extent therefore does not imply human-level discovery quality. The factorial result concerns uncertainty percentile; Appendix Table A7 reports best-found differences from humans. On an 11×11 grid, the uncertainty-percentile interaction is +11.3 [+10.2,+12.4] points and joint-condition best-found value reaches 95.7%, compared with 95.6% for matched humans. The interaction also remains positive in rough landscapes (Appendix Figure A2 and Table A15). The semantic frame raises the uncertainty percentile further, to 27.4–38.6, but adds structural information about how rewards are arranged. We therefore treat it as a proof-of-concept condition rather than the factorial motivation factor (Appendix Table A10).

3.3 Coverage identifies destinations; rendering affects their use

The controls isolate what the coverage representation contributes. The full human interface exposes every position but leaves exploration near baseline. With motivation present, the same-information control reaches uncertainty percentiles of 4.0–7.7, whereas the coverage representation reaches

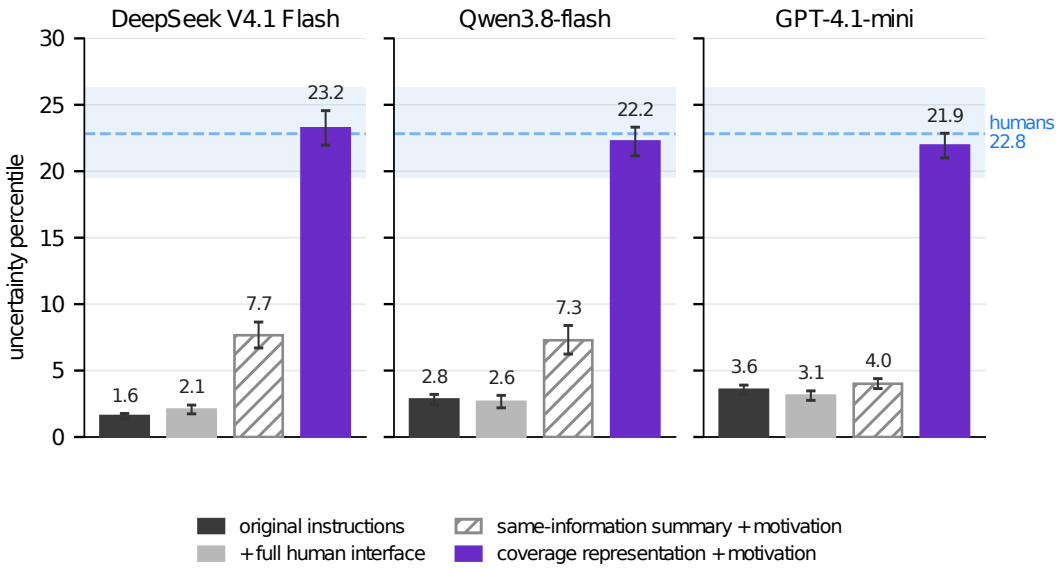


Figure 3: **A coverage representation identifies candidate destinations.** Uncertainty percentile among newly selected positions for three model families ($n = 50$ per cell). The full-human-interface condition uses the original instructions without motivation; the same-information and coverage conditions both include topology-free motivation. Restoring the full human interface leaves exploration near baseline. The same-information control reorganizes observed readings but does not identify untouched positions; the coverage representation raises the uncertainty percentile by 15.0–17.9 points over that control across models (seed-paired 95% CIs). The dashed line and band show the human mean and bootstrap 95% CI ($n = 19$).

21.9–23.2, gains of 15.0–17.9 points across models (Figure 3). In this control, organizing observed rewards does not reproduce the effect of identifying unvisited destinations.

The gain does not depend on identifying the longest unvisited interval or the best observation. A reduced representation retaining only the visited–unvisited partition preserves interactions of +6.1, +11.1, and +1.6 points (Appendix Table A9). Nor is the effect reproduced by directly displaying the quantity used to evaluate the uncertainty percentile. In DeepSeek V4.1 Flash development runs ($n = 30$ per condition), printing the posterior variance of every position at each step raises the uncertainty percentile only from 1.5 to 5.0, compared with 11.5 for the coverage representation alone (Appendix Table A6). Exposing uncertainty scores does not by itself organize the action space into candidate destinations.

A separate factorial distinguishes computing coverage information from using it in choice. Models recover substantial coverage information when queried (Appendix Table A21), but whether that information changes choice depends on rendering. Sentence rendering raises the uncertainty percentile by 12.9 [+12.2,+13.5] points relative to the four-line list; the same qualitative pattern appears even with imperfect model-written coverage (Appendix Table A17). Information availability alone does not ensure use in choice.

3.4 Named alternatives let motivation directly induce departure

Departure and destination choice can also be separated within structured search. Motivation alone increases new-position probability from 56.9% to 71.8% for DeepSeek V4.1 Flash and from 58.9% to 79.0% for Qwen3.8-flash, while their uncertainty percentiles stay at 2.5–3.7; GPT-4.1-mini already selects new positions on 97.1% of baseline choices (Appendix Table A20). Motivation can therefore increase new choices without substantially extending their reach.

In the restless bandit, every alternative is independently named. Motivation raises switching by 0.20–0.60 without requiring an arm-level summary, while representation alone adds only 0.005–

0.014. The seed-paired interactions are small and negative: -0.055 [-0.079,-0.032], -0.044 [-0.065,-0.023], and -0.045 [-0.069,-0.020] (Figure 2b).

The destination analysis identifies these switches as value-seeking. Under motivation, destination posterior-mean ranks range from 0.69 to 0.79, whereas destination uncertainty ranks remain below the chance level of 0.5. The bandit therefore isolates motivation's direct effect on leaving the current arm. These switches favor arms with higher estimated values, consistent with the task already supplying named candidate destinations (Appendix Tables A5 and A23).

3.5 Generated rationales amplify the shift toward distant actions

Finally, token-level probes locate locality within the action distribution and test whether it is only an artifact of numeric labels. Permuting the labels or replacing them with random two-digit codes weakens but does not eliminate the spatial gradient in choice probability (Figure 4a; Appendix Table A12).

The intervention shifts probability toward distant actions before the generated rationale and more strongly afterward. For DeepSeek V4.1 Flash, distant-action mass rises from 2.9% to 6.3% at the bare action head and from 3.8% to 19.8% after the rationale; for GPT-4.1-mini, the corresponding changes are 14.6% to 36.2% and 30.9% to 56.4% (Figure 4b,c). The coverage representation changes the initial action distribution, while the generated rationale amplifies that shift. This probe concerns the generated rationale elicited by the choice prompt, not provider-native reasoning.

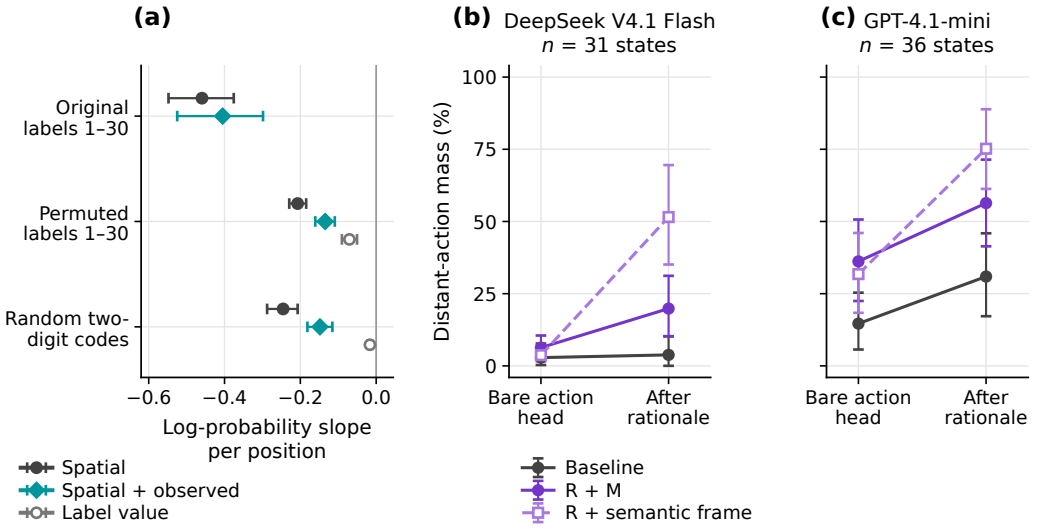


Figure 4: **Generated rationales amplify the shift toward distant actions.** Panel (a) measures the local-action prior under alternative label schemes, including a control for observed status. Panels (b,c) compare distant-action probability at the bare action head and after the generated rationale. R is coverage representation; M is topology-free motivation. The semantic frame adds topology. Whiskers are bootstrap 95% CIs.

4 Discussion and Related Work

Local search as a representational bottleneck. LLM exploration methods use history summaries, learned policies, and explicit search. Krishnamurthy et al. (2024) obtain robust exploration only by pairing chain-of-thought with an externally computed per-arm summary, whose extension beyond arm-wise state they leave open. Related methods use algorithmic guidance, posterior sampling, demonstrations, fine-tuning, online rewards, or explicit reasoning (Nie et al., 2025; Arumugam & Griffiths, 2026; Monea et al., 2025; Zhang et al., 2025b); behavioral studies find greedy choice, premature decisions in combination search, and sensitivity to semantic action labels (Schmied et al., 2026; Pan et al., 2025; Austin et al., 2026). Recent approaches optimize checkpoint

coverage (Ye et al., 2026) or query diversity (Chen et al., 2027); Agent Alpha diversifies candidate actions within Monte Carlo tree search (Tang et al., 2026). Intelligent Go-Explore maintains an archive of discovered states and uses foundation models to choose where to return and how to explore (Lu et al., 2025). This is a close precedent for explicitly organizing destinations. Our factorial instead isolates how coverage representation and its rendering interact with motivation within a fixed model and task. The controls refine the role of external summaries: they gain leverage when they identify unexplored alternatives as candidate destinations, while generic reorganization of observed rewards is insufficient.

Representation upstream of motivation. Search requires candidate moves and a reason to pursue them. Aspiration and intrinsic-motivation accounts explain why uncertain alternatives become valuable (Cyert & March, 1963; March, 1991; Pathak et al., 2017; Burda et al., 2019); classical methods operationalize that value through optimism, uncertainty, visitation, or novelty (Auer et al., 2002; Srinivas et al., 2010; Osband et al., 2016; Houthooft et al., 2016; Bellemare et al., 2016; Tang et al., 2017; Ostrovski et al., 2017; Badia et al., 2020). Visitation-based methods already use coverage to direct exploration. Our contribution concerns whether an LLM uses recoverable coverage information in context and how that use changes the effect of motivation. Organizational search similarly makes cognitive representation central to which alternatives can be considered (Simon, 1955; Cyert & March, 1963; March, 1991; Levinthal & March, 1993; Levinthal, 1997; Gavetti & Levinthal, 2000). Our cross-task factorial gives this distinction an empirical signature: motivation directly increases switching among named arms, whereas coverage representation greatly increases motivation's effect on search extent in relational search. The contrast is consistent with task structure changing the role of representation; the two tasks also differ in horizon, dynamics, and number of alternatives.

Memory must organize history for choice. Agent-memory architectures store, retrieve, summarize, and reflect on experience (Park et al., 2023; Packer et al., 2023). OPRO arranges past solutions and their scores into an optimization trajectory that guides new proposals (Yang et al., 2024). Tian et al. (2026) jointly train latent memory and exploration policies using code-coverage novelty signals. Our controls distinguish organizing observed outcomes from identifying unvisited alternatives. Models recover substantial coverage information when queried, but its effect on choice depends on rendering. Generic summaries leave unexplored regions unnamed, whereas the coverage representation identifies candidate destinations. Sentence rendering changes several linguistic properties at once, so the experiment does not isolate which property drives the difference (Appendix Table A17). The deterministic implementation holds coverage information fixed across renderings. Open-ended environments will require memory systems that organize retrieved history into candidate destinations as action spaces evolve.

Human benchmarks and discovery. Human exploration research distinguishes directed choice, stochastic choice, uncertainty-regulated switching, and generalization across related alternatives (Daw et al., 2006; Wilson et al., 2014; Gershman, 2018; Speekenbrink & Konstantinidis, 2015; Wu et al., 2018). Research using LLMs as models of human behavior further shows that aggregate similarity need not imply the same decision policy or consequences (Argyle et al., 2023; Aher et al., 2023; Binz & Schulz, 2023; Binz et al., 2025; Hu et al., 2026; Kirk et al., 2024; Murthy et al., 2025; Xiao et al., 2024; Zhang et al., 2025a). We use human trajectories as a behaviorally meaningful scale for the extent of non-local search, while treating spatial reach, situation responsiveness, and discovery outcomes as distinct properties. The central question is therefore whether unexplored regions are represented as candidate destinations toward which motivation can be directed.

Limitations. The evidence comes from enumerable landscapes and a four-arm bandit, leaving open how coverage representations can be constructed for evolving action spaces in code, web, or scientific agents. The coverage representation is deterministic and task-specific; the factorial does not isolate why sentence rendering affects choice. The primary human comparison has $n = 19$ and a visual rather than textual interface, while the two-dimensional and provider-native-reasoning factorials use DeepSeek V4.1 Flash only. Because Psych-101 is public, pretraining exposure cannot be excluded; the bandit therefore provides a structural rather than out-of-distribution test.

REPRODUCIBILITY STATEMENT

Appendix A provides verbatim interventions and prompts, model identifiers, decoding settings, environment seeds, and parsing diagnostics. Appendix tables report cell means, sample sizes, and uncertainty intervals; primary factorial intervals resample shared environment seeds as blocks. Computational reference policies and the token-probe model are specified in Appendix E. The supplementary material contains the environment and agent code, analysis scripts, and saved agent trajectories recording choices and rewards. The accompanying README maps analysis scripts to the tables and figures they reproduce offline.

AI USE STATEMENT

In this work, we used generative AI tools (OpenAI ChatGPT and Codex; Anthropic Claude and Claude Code) to provide feedback on the research question, conceptual framing, research execution, intervention wording, measurement, robustness analyses, and interpretation; to assist with data processing and the implementation, debugging, and review of experiment and analysis code; and to support literature search, manuscript structuring, drafting, editing, and LaTeX preparation. LLMs also generated the agent trajectories analyzed in the study; that experimental use is described in Sections 2–3 and Appendix A. The authors designed and directed the study, made all substantive decisions, inspected the experimental outputs, verified reported numbers against the analysis records, checked cited sources, and reviewed all AI-assisted text and code. The authors take responsibility for the final content, including claims, code, and artifacts produced with AI assistance.

REFERENCES

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 337–371, 2023.

Antonis Antoniades, Deepak Nathani, Ritam Saha, Alfonso Amayuelas, Ivan Bercovich, Zhaotian Weng, Vignesh Baskaran, Kunal Bhatia, and William Yang Wang. Heuresis: Search strategies for autonomous AI research agents across quality, diversity and novelty. *arXiv preprint arXiv:2606.25198*, 2026. URL https://arxiv.org/abs/2606.25198.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

Dilip Arumugam and Thomas L. Griffiths. Toward efficient exploration by large language model agents. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4431224d3762aa655f0aee4eaf04ff16-Abstract-Conference.html.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002. doi: 10.1023/A:1013689704352.

David Eric Austin, Kaheer Suleman, and Jackie Chi Kit Cheung. In-context exploration-exploitation is biased by semantic priors. In *Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=PbNQoIllIM.

Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martín Arjovsky, Alexander Pritzel, Andrew Bolt, and Charles Blundell. Never give up: Learning directed exploration strategies. In *International Conference on Learning Representations*, 2020.

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023. doi: 10.1073/pnas.2218523120.

Marcel Binz et al. A foundation model to predict and capture human cognition. *Nature*, 644(8078): 1002–1009, 2025. doi: 10.1038/s41586-025-09215-4.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019.

Xinyue Chen, Pengyu Gao, Jiangjiang Song, Liqian Chen, Kehang Zeng, Xin Yang, Xinjian Chen, and Xiaoyang Tan. Restoring exploration via Diversity-Driven Policy Optimization in search-based LLM agents. *Information Processing & Management*, 64(2):105104, 2027. doi: 10.1016/j.ipm.2026.105104.

Richard M. Cyert and James G. March. *A Behavioral Theory of the Firm*. Prentice-Hall, Englewood Cliffs, NJ, 1963.

Nathaniel D. Daw, John P. O'Doherty, Peter Dayan, Ben Seymour, and Raymond J. Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441:876–879, 2006. doi: 10.1038/nature04766.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York, 1993.

Giovanni Gavetti and Daniel A. Levinthal. Looking forward and looking backward: Cognitive and experiential search. *Administrative Science Quarterly*, 45(1):113–137, 2000. doi: 10.2307/2666981.

Samuel J. Gershman. Deconstructing the human algorithms for exploration. *Cognition*, 173:34–42, 2018. doi: 10.1016/j.cognition.2017.12.014.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. SimBench: Benchmarking the ability of large language models to simulate human behaviors. In *International Conference on Learning Representations*, 2026.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5a68d05006d5b05dd9463dd9c0219db0-Abstract-Conference.html.

Akshay Krishnamurthy, Keegan Harris, Dylan J. Foster, Cyril Zhang, and Aleksandrs Slivkins. Can large language models explore in-context? In *Advances in Neural Information Processing Systems*, 2024.

Daniel A. Levinthal. Adaptation on rugged landscapes. *Management Science*, 43(7):934–950, 1997. doi: 10.1287/mnsc.43.7.934.

Daniel A. Levinthal and James G. March. The myopia of learning. *Strategic Management Journal*, 14(S2):95–112, 1993.

Cong Lu, Shengran Hu, and Jeff Clune. Intelligent go-explore: Standing on the shoulders of giant foundation models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/369a30aac2765950865efd318cef7f76-Abstract-Conference.html.

James G. March. Exploration and exploitation in organizational learning. *Organization Science*, 2 (1):71–87, 1991.

Giovanni Monea, Antoine Bosselut, Kianté Brantley, and Yoav Artzi. LLMs are in-context bandit reinforcement learners. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=c0RsezY2D1.

Sonia Krishna Murthy, Tomer Ullman, and Jennifer Hu. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 11241–11258. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.561.

Allen Nie, Yi Su, Bo Chang, Jonathan Lee, Ed H. Chi, Quoc V. Le, and Minmin Chen. EVOLvE: Evaluating and optimizing LLMs for in-context exploration. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 46346–46376. PMLR, 2025. URL https://proceedings.mlr.press/v267/nie25b.html.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Georg Ostrovski, Marc G. Bellemare, Aäron van den Oord, and Rémi Munos. Count-based exploration with neural density models. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2721–2730, 2017.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv:2310.08560*, 2023.

Lan Pan, Hanbo Xie, and Robert C. Wilson. Large language models think too fast to explore effectively. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3428.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023. doi: 10.1145/3586183.3606763.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *ICML*, pp. 2778–2787, 2017.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

Thomas Schmied, Jörg Bornschein, Jordi Grau-Moya, Markus Wulfmeier, and Razvan Pascanu. LLMs are greedy agents: Effects of RL fine-tuning on decision-making abilities. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/cf2fcda40de8520027af2eadcafdecd6-Paper-Conference.pdf.

Herbert A. Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69 (1):99–118, 1955.

Maarten Speekenbrink and Emmanouil Konstantinidis. Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science*, 7(2):351–367, 2015. doi: 10.1111/tops.12145.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 1015–1022, 2010.

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. #exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Sizhe Tang, Rongqian Chen, and Tian Lan. Agent alpha: Tree search unifying generation, exploration and evaluation for computer-use agents. *arXiv preprint arXiv:2602.02995*, 2026. URL https://arxiv.org/abs/2602.02995.

Shizuo Tian, Xiaohong Weng, Rui Kong, Yuxuan Chen, Guohong Liu, Yuebing Song, Jiacheng Liu, Yuchen Li, Dawei Yin, Ting Cao, Yunxin Liu, and Yuanchun Li. Joint agent memory and exploration learning via novelty signals. *arXiv preprint arXiv:2606.01528*, 2026. URL https://arxiv.org/abs/2606.01528.

Robert C. Wilson, Andra Geana, John M. White, Elliot A. Ludvig, and Jonathan D. Cohen. Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, 143(6):2074–2081, 2014. doi: 10.1037/a0038199.

Charley M. Wu, Eric Schulz, Maarten Speekenbrink, Jonathan D. Nelson, and Björn Meder. Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, 2(12): 915–924, 2018.

Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J. Su. On the algorithmic bias of aligning large language models with RLHF: Preference collapse and matching regularization. *arXiv preprint arXiv:2405.16455*, 2024.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bb4VGOWELI.

Ziang Ye, Wentao Shi, Yuxin Liu, Yu Wang, Zhengzhou Cai, Yaorui Shi, Qi Gu, Xunliang Cai, and Fuli Feng. Look before you leap: Autonomous exploration for LLM agents. *arXiv preprint arXiv:2605.16143*, 2026. URL https://arxiv.org/abs/2605.16143.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, and Weiyan Shi. Verbalized sampling: How to mitigate mode collapse and unlock LLM diversity. *arXiv preprint arXiv:2510.01171*, 2025a.

Ziyuan Zhang, Darcy Wang, Ningyuan Chen, Rodrigo Mansur, and Vahid Sarhangian. Comparing exploration–exploitation strategies of LLMs and humans: Insights from standard multi-armed bandit tasks. *arXiv preprint arXiv:2505.09901*, 2025b.

A Reproducibility details

A.1 Verbatim interventions

Topology-free motivation. The following text is identical in both tasks and is appended to the system prompt:

> *How you tend to deal with choices like this:*
> *- You are hard to satisfy: even when an option pays well you keep wondering whether another pays more, and you go and check fairly often.*
> *- Figuring something out is worth something to you in its own right; what you come to understand counts as part of what you came away with.*

The reduced-motivation ablation deletes only the clause "and you go and check fairly often" from the first sentence; the aspiration and intrinsic-learning statements remain.

Coverage representation. The structured-search coverage representation deterministically converts the accumulated history into a description such as:

> Looking at the whole line of 30: you have taken readings on option 5 (value 68), options 14-16 (values 16-38).
> 26 of the 30 options have never been touched in this environment: options 1-4, options 6-13, options 17-30. The longest untouched stretch is 14 options long, at the right end.
> The highest reading you have taken here is 68, at option 5.

The reduced one-dimensional representation uses the first two sentences only: it lists the positions already sampled and partitions all remaining positions as unvisited, while omitting the longest-untouched-stretch and highest-reading sentences. The full two-dimensional coverage representation follows the same structure: it describes touched connected regions and their value ranges, counts untouched tiles, lists untouched rows and columns, identifies the largest untouched rectangular block and its location, and reports the highest reading. The reduced two-dimensional representation removes the largest-block and highest-reading sentences while retaining the preceding relational coverage summary. All reported two-dimensional trajectories use these fixed representations and seeds 1000–1049. The restless-bandit representation uses the same principle but treats each named arm as an independently named candidate destination:

> Summary so far — A: pressed 2 times, average 29.5, last 31 (2 steps ago); O: pressed 1 time, average 43.0, last 43 (0 steps ago); U: pressed 1 time, average 69.0, last 69 (1 step ago); Y: never pressed. 146 presses left.

Both representations deterministically organize the transcript into explicit coverage and performance variables while leaving payoff inference and action selection to the model.

Self-written coverage notes. In the self-representation experiment, DeepSeek V4.1 Flash receives the raw transcript and writes notes about what has and has not been sampled before each choice. The notes are appended to the subsequent decision context, replacing the supplied coverage representation. The model receives the same decision prompt and, where applicable, the same topology-free motivation text as in the main factorial. A subsequent complete $2 \times 2 \times 2$ holds motivation present and crosses coverage content (the model's filled four-field template or ground-truth coverage), attribution (preceded by "Your own notes on the current environment, written just now:" or by no label), and rendering (the four-line template or the coverage representation's sentence rendering). For model-written content, a parser formats the fields while retaining the model's errors; rendering succeeds on 99% of steps. All eight coverage blocks appear between the history and the choice question.

Semantic frame (motivation + topology). The proof-of-concept condition replaces the abstract line with a segmented-market interpretation:

> You are setting the price for a consumer product, one new market after another; there are 16 markets in total. In each market you can pick a price point from 1 to 30, where 1 is the cheapest positioning and 30 the most premium. Picking a price point tells you the profit per unit you earn at that price in this market. Buyers in a market are split into several distinct segments spread along the price line, and neighbouring price points reach the same segment, so they earn similar profit. Different segments sit far apart on the line, and any one of them may turn out to be the profitable one in a given market; between segments there are stretches where few people buy. In each market you will be able to set the price either 5 or 10 times. When you have used up your attempts in a market, you move on to the next, unfamiliar market. Which segments are worth serving differs from market to market, so it has to be learned anew in each one. Each market starts with the profit at one price point already known to you. You may return to a price point you have already used; the profit may shift slightly. Your aim is to earn as much profit as possible across all 16 markets.

The single-peak semantic control retains the pricing cover story and task mechanics while removing descriptions of distinct market segments, their separation, and intervening low-demand stretches.

Decision prompt. All conditions use the same response format:

> You are a participant in a psychology experiment. Read the experiment so far and decide your next choice. First write one or two sentences of your private thoughts about what to do next (as inner monologue, first person). Then, on the last line, give only the key you press, formatted exactly as `<<X>>`.

The user message contains the complete task transcript followed by "What do you do next?" The parser takes the last delimited action.

A.2 Models, seeds, and parsing diagnostics

Table A1 records the model identifiers returned by the providers. Evaluations reported in the final batches were run from September 15–19, 2026. DeepSeek V4.1 Flash requests used `deepseek-chat` and later `deepseek-flash`; response metadata returned `deepseek-flash` with one system fingerprint across reported runs. The restless-bandit baseline was rerun on September 19 under that served version. Provider-native reasoning was disabled for DeepSeek V4.1 Flash and Qwen3.8-flash in the main factorial; GPT-4.1-mini is used in its non-reasoning configuration. Every call uses temperature 1.0, provider-default top-p, and no API sampling seed. Maximum output is 400 tokens for choices, 8 for key-only fallbacks, and 300 for reconstruction questions and self-written notes. The generated rationale requested by the decision prompt is ordinary generated text, distinct from provider-native reasoning output. The provider-native-reasoning robustness uses the same DeepSeek V4.1 Flash prompts and seeds with thinking enabled and an 8,000-token budget; the generated-rationale request remains in the prompt. The core cells use seeds 1000–1049, with one independent agent trajectory per seed. The shared seed range holds sampled environments fixed across cells while model calls remain independent.

Paper label	API/served identifier	Evaluated cells	Agents per cell
DeepSeek V4.1 Flash	`deepseek-chat`/`deepseek-flash`; returned `deepseek-flash`	1D and bandit factorials; 2D replication; interface; provider-native reasoning	50
Qwen3.8-flash	`qwen3.8-flash`	both $R \times M$ factorials; interface	50
GPT-4.1-mini	`gpt-4.1-mini-2025-04-14`	both $R \times M$ factorials; interface	50
GPT-4.1-nano	`gpt-4.1-nano-2025-04-14`	auxiliary structured-search cells	50

Table A1: Model identifiers and coverage. DeepSeek V4.1 Flash supplies the two-dimensional replication. GPT-4.1-nano provides an auxiliary scale comparison outside the clean factorial test.

A response that contains a valid action without delimiters is recovered as a format fallback. If no valid action can be recovered, the agent repeats its preceding action. Table A2 reports event counts in the structured-search audit cells used for the baseline, coverage-representation, semantic-frame, and full-interface comparisons. These events concentrate in the original-instruction conditions and are retained in the reported outcomes. A repeated fallback action does not enter the uncertainty-percentile average directly because Equation 1 is defined only for newly selected positions; it can

affect far jumps, re-clicks, performance, and the subsequent trajectory. Across clean factorial cells, unrecoverable outputs account for 0–0.62% of steps (1.15% in the provider-native-reasoning baseline). The interaction estimates are stable after excluding every affected agent (Table A8). The clean DeepSeek V4.1 Flash factorial additionally contains 14 position-zero fallback events in the baseline cell and eight in the motivation-only cell. Position zero lies outside the action set and therefore does not enter the uncertainty-percentile calculation; no joint-intervention cell contains such an event.

Across the four full-representation two-dimensional cells, 69/49/64/27 responses required format recovery for neither/representation/motivation/both; all 48,000 choices were recoverable, and both reduced-representation cells likewise had no unrecoverable choices.

Model	Baseline	Representation	Semantic frame	Representation + frame	Full interface
DeepSeek V4.1 Flash	33/16	5/3	2/0	0/0	0/0
Qwen3.8-flash	84/20	28/17	5/0	0/0	9/0
GPT-4.1-mini	94/35	11/0	2/0	0/0	0/0
GPT-4.1-nano	21/45	—	—	0/0	3/0

Table A2: Format fallback / unrecoverable-output events across 50 agent trajectories per available cell. An unrecoverable output repeats the preceding action. Counts are events, not affected agents.

B Complete numerical results

Condition	n	Unc. pct.	Far jump	Re-click	Per-step (%)	Best-found (%)	$\beta_{\text{situation}}$
Cumulative × smooth (main comparison)	19	22.8	0.275	0.086	70.0	93.9	-7.37
Cumulative × rough	17	20.8	0.219	0.271	67.2	88.0	-8.86
Max-value × smooth	23	26.3	0.315	0.042	68.4	93.3	-4.48
Max-value × rough	22	28.7	0.323	0.068	61.8	89.8	-4.62

Table A3: Human behavior across the four conditions of the structured-search experiment. The cumulative-reward rough-condition uncertainty percentile uses the matched $\ell = 1$ posterior, as in Table A15.

Model	Condition	n	q	Far jump	Re-click	Per-step (%)	Best-found (%)	β_{sit}	β_{pos}
DeepSeek V4.1 Flash	Baseline	50	1.6	0.026	0.428	64.9	78.1	-0.89	+1.91
	R	50	12.0	0.233	0.198	63.8	85.4	-2.82	+2.01
	SF	50	6.5	0.105	0.286	68.6	85.0	-2.74	+1.98
	R + SF	50	33.8	0.549	0.095	65.9	92.0	-2.22	+0.21
	Full interface	50	2.1	0.066	0.274	65.0	81.6	-1.90	+2.42
Qwen3.8-flash	Baseline	50	2.8	0.038	0.408	69.3	80.8	-1.00	+4.29
	R	50	8.8	0.161	0.242	68.7	86.2	-3.10	+2.64
	SF	50	8.8	0.102	0.386	73.0	85.6	-6.07	+0.18
	R + SF	50	27.4	0.427	0.156	69.2	90.8	-2.38	-0.05
	Full interface	50	2.6	0.042	0.356	69.8	83.0	-3.98	+2.33
GPT-4.1-mini	Baseline	50	3.6	0.131	0.023	58.7	83.3	-0.16	+2.19
	R	50	17.8	0.411	0.023	59.1	87.4	-1.49	+1.18
	SF	50	4.5	0.078	0.291	68.0	83.3	-4.35	+0.65
	R + SF	50	38.6	0.658	0.018	61.0	91.3	-1.82	-0.92
	Full interface	50	3.1	0.162	0.032	57.9	82.7	+0.63	+2.61
GPT-4.1-nano	Baseline	50	12.7	0.238	0.407	56.1	78.6	-0.20	+0.62
	R + SF	50	26.4	0.412	0.280	64.1	86.5	-1.84	-1.44
	Full interface	50	13.7	0.258	0.375	58.3	79.9	-0.09	+1.30
Humans	—	19	22.8	0.275	0.086	70.0	93.9	-7.37	+0.50
GP-UCB, $\kappa = 0.5$	—	30	35.1	0.322	0.392	75.7	93.0	-3.32	-4.57
Uniform random	—	30	46.2	0.719	0.146	52.2	87.4	+0.18	-0.10

Table A4: Structured-search comparison conditions and reference policies. R denotes coverage representation; SF denotes the semantic frame (motivation + topology), distinct from the topology-free motivation factor in Table A5. q is uncertainty percentile; the human row is the cumulative-reward smooth condition.

Model	Baseline	R	M	$R{+}M$	Interaction [95% CI]
Structured search: uncertainty percentile					
DeepSeek V4.1 Flash	1.6	12.0	2.5	23.2	+10.2 [+8.8,+11.6]
Qwen3.8-flash	2.8	8.8	3.7	22.2	+12.6 [+11.1,+14.1]
GPT-4.1-mini	3.6	17.8	3.4	21.9	+4.2 [+3.1,+5.3]
Structured search: situation coefficient, β_{sit}					
DeepSeek V4.1 Flash	-0.89	-2.82	-1.02	-1.13	+1.81 [-0.06,+3.63]
Qwen3.8-flash	-1.00	-3.10	-0.17	-2.20	+0.08 [-1.42,+1.58]
GPT-4.1-mini	-0.16	-1.49	-0.29	-1.24	+0.38 [-0.48,+1.25]
Humans ($n = 19$)	-7.37 [-9.04,-5.97]				
Restless bandit: switch rate					
DeepSeek V4.1 Flash	0.105	0.110	0.481	0.430	-0.055 [-0.079,-0.032]
Qwen3.8-flash	0.107	0.121	0.709	0.679	-0.044 [-0.065,-0.023]
GPT-4.1-mini	0.092	0.100	0.295	0.258	-0.045 [-0.069,-0.020]

Table A5: Representation $\times$ topology-free motivation factorial results, with $n = 50$ agents per cell. Interactions use seed-paired bootstrap intervals. More negative β_{sit} indicates stronger situation responsiveness; the human coefficient interval resamples participants.

Condition	n	Unc. pct.	Best-found (%)	Re-click
Baseline (abstract instructions)	30	1.5	82.2	0.368
Posterior variance printed each step	30	5.0	84.7	0.350
Coverage representation only	30	11.5	88.2	0.155

Table A6: Selected development-stage controls comparing explicit uncertainty scores with coverage representation in DeepSeek V4.1 Flash. Each condition uses $n = 30$ and predates the main $n = 50$ factorial experiment.

C Robustness and mechanism analyses

Model	Additive interaction (points)	Per-agent logit interaction	Far-jump logit interaction	$R+M-$human, unc. pct. (90% CI)	$R+M-$human, best-found (90% CI)
DeepSeek V4.1 Flash	+10.2 [+8.8,+11.6]	+0.38 [+0.15,+0.60]	−0.26 [−0.57,+0.05]	+0.4 [−2.7,+3.5]	−7.2 [−8.9,−5.5]
Qwen3.8-flash	+12.6 [+11.1,+14.1]	+0.93 [+0.65,+1.20]	+0.14 [−0.17,+0.44]	−0.6 [−3.7,+2.4]	−4.1 [−5.7,−2.4]
GPT-4.1-mini	+4.2 [+3.1,+5.3]	+0.28 [+0.12,+0.44]	+0.20 [+0.04,+0.35]	−0.9 [−4.0,+2.1]	−6.3 [−7.9,−4.7]

Table A7: **Scale sensitivity and comparison with humans.** The primary uncertainty-percentile interaction remains positive under a per-agent logit transformation in all three models. Far-jump interactions vary with scale. Human comparisons report agent-minus-human differences: uncertainty-percentile point differences range from -0.9 to +0.4 points, with 90% CIs extending roughly four points in either direction; best-found differences range from -7.2 to -4.1 points.

Model	Unrecoverable-output steps, % (neither / R / M / $R+M$)	Interaction after excluding affected agents (points)
DeepSeek V4.1 Flash	0.27 / 0.05 / 0.13 / 0.00	+10.3 [+8.3,+12.1]
Qwen3.8-flash	0.33 / 0.28 / 0.62 / 0.12	+12.1 [+10.1,+14.0]
GPT-4.1-mini	0.58 / 0.00 / 0.52 / 0.02	+4.5 [+3.1,+5.9]
DeepSeek V4.1 Flash, provider-native reasoning	1.15 / 0.13 / 1.12 / 0.15	+6.3 [+4.7,+7.9]

Table A8: **Unrecoverable-output sensitivity.** An unrecoverable output repeats the preceding action. The interactions remain positive after excluding every agent with any such step.

Model	Reduced representation only	Reduced representation + motivation	Additive interaction (points)	Per-agent logit interaction	Full-representation interaction (points)
DeepSeek V4.1 Flash	9.0 [8.1,9.9]	16.1 [15.2,17.0]	+6.1 [+4.9,+7.3]	+0.25 [+0.03,+0.47]	+10.2
Qwen3.8-flash	6.5 [5.7,7.3]	18.4 [17.0,19.8]	+11.1 [+9.3,+12.8]	+1.02 [+0.70,+1.32]	+12.6
GPT-4.1-mini	8.3 [7.8,8.9]	9.8 [9.2,10.4]	+1.6 [+0.6,+2.5]	+0.20 [+0.03,+0.36]	+4.2

Table A9: **Reduced-coverage-representation ablation.** The ablation lists visited and unvisited positions but omits the longest-unvisited-interval and best-observation sentences. Its representation-by-motivation interaction remains positive on additive and per-agent logit scales in all three models.

Condition	DeepSeek V4.1 Flash	Qwen3.8-flash	GPT-4.1-mini
Single-peak frame only	3.1 / 81.4	10.3 / 84.6	2.5 / 80.9
Segmented frame only	6.5 / 85.0	8.8 / 85.6	4.5 / 83.3
Representation + single-peak frame	18.4 / 88.7	18.5 / 89.4	19.7 / 88.8
Representation + topology-free motivation	23.2 / 86.6	22.2 / 89.8	21.9 / 87.5
Representation + segmented frame	33.8 / 92.0	27.4 / 90.8	38.6 / 91.3
Best-found difference versus representation + motivation, controlling each agent's uncertainty percentile (pp)			
Single-peak frame	+2.2 [+0.7,+3.7]	+0.5 [−1.2,+2.2]	+1.7 [+0.1,+3.2]
Segmented frame	+5.2 [+3.3,+7.0]	−0.2 [−1.7,+1.3]	+0.6 [−2.5,+3.9]

Table A10: **Semantic-frame controls.** Each upper cell reports uncertainty percentile / best-found value (% of ceiling), with $n = 50$. The single-peak control retains the pricing cover story and removes the multi-segment topology sentences. Exploration-adjusted outcome gains are model-dependent, with the positive directional estimate concentrated in DeepSeek V4.1 Flash.

Provider-native reasoning	Measure	Neither	Representation	Motivation	Both	Interaction (points / logit)
Off	Uncertainty percentile	1.6 [1.4,1.8]	12.0 [10.9,13.2]	2.5 [2.2,3.0]	23.2 [21.9,24.6]	+10.2 [+8.8,+11.6] / +0.38 [+0.15,+0.60]
	Best-found	78.1	85.4	80.5	86.6	
On	Uncertainty percentile	0.9 [0.8,1.1]	3.8 [3.1,4.5]	1.4 [1.2,1.6]	10.8 [9.6,12.0]	+6.6 [+5.4,+7.9] / +0.90 [+0.57,+1.23]
	Best-found	79.4	83.0	81.5	88.7	

Table A11: **Provider-native reasoning robustness.** DeepSeek V4.1 Flash uses identical prompts and environment seeds with provider-native reasoning off and on ($n = 50$ per cell). Provider-native reasoning intensifies baseline locality. The interaction is smaller in percentile points but larger on the per-agent logit scale. With native reasoning off, the best-found means imply an additive interaction of -1.1 points (computed before rounding), distinct from the positive uncertainty-percentile interaction.

Label scheme	Spatial-distance slope	Spatial slope, controlling observed status	Label-number-distance slope
Original labels 1–30	−0.459 [−0.548,−0.376]	−0.405 [−0.525,−0.298]	—
Randomly permuted 1–30	−0.207 [−0.229,−0.185]	−0.135 [−0.161,−0.109]	−0.071 [−0.090,−0.050]
Random two-digit codes	−0.245 [−0.288,−0.207]	−0.148 [−0.181,−0.116]	−0.016 [−0.023,−0.010]

Model	Condition	Far-action mass, bare head (%)	Far-action mass, after rationale (%)
DeepSeek V4.1 Flash	Original instructions	2.9 [0.3,7.7]	3.8 [0.0,10.2]
	Representation + topology-free motivation	6.3 [2.8,10.5]	19.8 [10.3,31.2]
	Representation + semantic frame	3.7 [1.5,6.4]	51.5 [35.1,69.5]
GPT-4.1-mini	Original instructions	14.6 [5.7,25.4]	30.9 [17.2,45.9]
	Representation + topology-free motivation	36.2 [22.5,50.7]	56.4 [41.4,71.4]
	Representation + semantic frame	31.8 [18.4,46.0]	75.2 [61.3,88.9]

Table A12: **Token-level mechanism probes.** The first panel controls whether each candidate has already been observed. The second reports probability mass on actions at least five positions away at matched mid-environment states (DeepSeek V4.1 Flash $n = 31$; GPT-4.1-mini $n = 36$). Each candidate label is one action token scored by its log-probability; intervals are state-bootstrap 95% CIs.

Cell	Uncertainty percentile	Best-found (%)	Per-step reward (%)	Mean step distance	Re-click rate
Neither	4.8 [4.5, 5.2]	90.2 [89.0, 91.4]	65.7	2.01	0.160
Coverage representation	8.6 [7.9, 9.4]	93.0 [91.9, 94.0]	70.4	2.36	0.252
Topology-free motivation	6.1 [5.8, 6.4]	91.9 [90.7, 93.0]	65.7	2.26	0.126
Both	21.2 [20.3, 22.2]	95.7 [94.9, 96.4]	67.1	3.42	0.134
Reduced representation	7.6 [7.2, 8.1]	93.1 [91.9, 94.3]	67.0	2.59	0.160
Reduced representation + topology-free motivation	14.3 [13.7, 14.9]	94.6 [93.8, 95.4]	65.6	3.29	0.094
Humans	14.9 [12.4, 17.5]	95.6 [93.7, 97.2]	73.6	1.65	0.237

Table A13: **Two-dimensional replication and representation ablation.** Wu et al. experiment 2D, cumulative-reward × smooth-kernel condition: DeepSeek V4.1 Flash ($n = 50$ per cell) and matched humans ($n = 22$). With the full representation, the seed-paired uncertainty-percentile interaction is +11.3 [+10.2,+12.4] points and +0.82 [+0.68,+0.97] on the per-agent logit scale. The reduced representation removes the largest-block and best-observation sentences while retaining touched regions and value ranges, untouched-tile count, and untouched rows and columns; its seed-paired interaction is +5.4 [+4.6,+6.2] points and +0.46 [+0.34,+0.58] on the logit scale. Its best-found interaction is -0.2 [-2.3,+2.0] points. Step distance is Euclidean distance in grid tiles.

Model	Same-info	Same-info + M	Full R	Full $R + M$	Same-info × M	Full $R \times M$	Full − same-info, with M
DeepSeek V4.1 Flash	2.3	7.7	12.0	23.2	+4.4 [+3.4,+5.4]	+10.2 [+8.8,+11.6]	+15.6 [+14.2,+17.0]
Qwen3.8-flash	2.6	7.3	8.8	22.2	+3.8 [+2.4,+5.2]	+12.6 [+11.1,+14.1]	+15.0 [+13.5,+16.5]
GPT-4.1-mini	3.6	4.0	17.8	21.9	+0.6 [−0.0,+1.1]	+4.2 [+3.1,+5.3]	+17.9 [+16.9,+18.9]

Table A14: **Same-information control.** The control organizes observed readings like the coverage representation and is similar in length while leaving untouched positions unnamed. Entries are uncertainty percentiles; intervals are seed-paired bootstrap 95% CIs.

Model	Unc. pct., neither / R / M / both	Peak, neither / R / M / both	Representation effect	Motivation effect	Interaction
DeepSeek V4.1 Flash	1.1 / 8.6 / 1.7 / 18.3	77.1 / 84.1 / 81.3 / 84.6	+7.5 [+6.8,+8.2]	+0.6 [+0.4,+0.9]	+9.1 [+7.6,+10.4]
Qwen3.8-flash	1.9 / 6.6 / 2.5 / 19.3	79.6 / 84.6 / 83.4 / 87.1	+4.7 [+3.9,+5.5]	+0.6 [+0.0,+1.3]	+12.1 [+10.8,+13.4]
GPT-4.1-mini	2.4 / 13.5 / 2.1 / 17.7	84.9 / 86.0 / 84.6 / 86.8	+11.1 [+10.4,+11.8]	−0.2 [−0.5,+0.1]	+4.4 [+3.4,+5.5]
Humans ($n = 17$)	20.8 [15.7,26.9]	88.0			

Table A15: **Rough-landscape replication.** Wu et al.'s rough condition uses an RBF length scale of one; prompts, treatments, and evaluation otherwise match the primary design. Effects and interactions use seed-paired bootstrap 95% CIs.

Coverage source	Alone: unc. pct.	With motivation: unc. pct.	With motivation: peak	Interaction with motivation
Raw history	1.6	2.5	80.5	—
Model, free-form notes	2.7	3.1	82.6	−0.5 [−1.2,+0.1]
Model, four-field template	4.0	6.4	84.1	+1.4 [+0.5,+2.3]
Environment, coverage representation	12.0	23.2	86.6	+10.2 [+8.8,+11.6]

Table A16: **Coverage-source comparison.** DeepSeek V4.1 Flash, $n = 50$ per cell. The four-field template follows the requested format on 98% of steps; the untouched set is exact on 42%, the longest gap on 56%, and both on 34%. Jointly correct and remaining steps yield similar uncertainty percentiles. Table A17 separates source from rendering and attribution.

Coverage content	Attribution	Four-line list	Sentences
Model-written	own-notes label	6.4 [5.7,7.2]	16.9 [15.8,18.0]
	no label	6.8 [6.0,7.6]	17.8 [16.8,18.9]
Ground truth	own-notes label	6.3 [5.2,7.3]	19.3 [18.2,20.3]
	no label	6.3 [5.5,7.1]	23.2 [21.9,24.6]

Seed-paired effects in uncertainty-percentile points: sentences minus list, +12.9 [+12.2,+13.5]; ground truth minus model-written content, +1.8 [+1.1,+2.4]; no label minus own-notes label, +1.3 [+0.7,+1.9]. Within the sentence rendering, the content and label effects are +3.9 [+3.0,+4.8] and +2.5 [+1.5,+3.4]; within the list rendering, they are −0.3 [−1.1,+0.5] and +0.2 [−0.7,+1.0].

Table A17: **Availability versus use of reconstructed coverage information.** DeepSeek V4.1 Flash, one-dimensional search, topology-free motivation in every cell, $n = 50$ per cell with shared seeds. The factorial crosses coverage content, attribution, and rendering. The sentence rendering changes several linguistic properties at once, so the comparison identifies rendering-dependent use in choice rather than the effect of a single syntactic or typographic feature.

Measure	Original instructions	Representation + motivation	Difference
Uncertainty percentile	0.8 [0.8,0.9]	6.3 [5.2,7.4]	+5.4 [+4.4,+6.5]
Best-found (%)	79.3 [77.8,80.9]	85.2 [83.8,86.5]	+5.9 [+4.6,+7.1]
Per-step reward (%)	70.9 [69.5,72.3]	63.3 [62.1,64.5]	−7.5 [−8.6,−6.4]
Far-jump rate	0.001 [0.001,0.002]	0.155 [0.142,0.169]	+0.154 [+0.141,+0.168]
Re-click rate	0.506 [0.486,0.526]	0.066 [0.058,0.074]	−0.440 [−0.461,−0.418]

Table A18: **GPT-5.6 terra provider-native-reasoning diagnostic.** GPT-5.6 terra with provider-native reasoning, $n = 50$ per cell. The baseline and combined conditions provide a direct contrast.

Task (measure)	Model	Motivation text	Cells: neither / R / M / $R{+}M$	Motivation alone / interaction
Structured (unc. pct.)	DeepSeek V4.1 Flash	original	1.6 / 12.0 / 2.5 / 23.2	+1.0 / +10.2 [+8.8,+11.6]
		reduced	1.6 / 12.0 / 2.2 / 20.8	+0.7 / +8.1 [+6.5,+9.7]
	Qwen3.8-flash	original	2.8 / 8.8 / 3.7 / 22.2	+0.9 / +12.6 [+11.1,+14.1]
		reduced	2.8 / 8.8 / 2.9 / 17.3	+0.1 / +8.5 [+7.0,+10.0]
	GPT-4.1-mini	original	3.6 / 17.8 / 3.4 / 21.9	-0.1 / +4.2 [+3.1,+5.3]
		reduced	3.6 / 17.8 / 2.6 / 20.2	-1.0 / +3.3 [+2.3,+4.3]
Bandit (switch rate)	DeepSeek V4.1 Flash	original	.105 / .110 / .481 / .430	+.375 / $-.055$ [$-.079$,$-.032$]
		reduced	.105 / .110 / .310 / .290	+.205 / $-.025$ [$-.052$,+.002]
	Qwen3.8-flash	original	.107 / .121 / .709 / .679	+.602 / $-.044$ [$-.065$,$-.023$]
		reduced	.107 / .121 / .571 / .524	+.464 / $-.061$ [$-.086$,$-.035$]
	GPT-4.1-mini	original	.092 / .100 / .295 / .258	+.203 / $-.045$ [$-.069$,$-.020$]
		reduced	.092 / .100 / .215 / .194	+.123 / $-.028$ [$-.052$,$-.004$]

Table A19: **Reduced-motivation ablation.** The reduced text removes "and you go and check fairly often" while retaining high aspiration and the intrinsic value of learning. Structured-search interactions remain positive, and motivation continues to raise bandit switching. Intervals are seed-paired bootstrap 95% CIs; $n = 50$ per cell.

Model	Component	Cells: neither / R / M / $R{+}M$ (0–100 scale)	Interaction (points)
Humans ($n = 19$)	new / conditional / top quartile / product	91.4 / 22.8 / 8.0 / 20.5	—
DeepSeek V4.1 Flash	New position	56.9 / 80.1 / 71.8 / 84.9	-10.1 [-14.2,-6.0]
	Conditional percentile	1.6 / 12.0 / 2.5 / 23.2	+10.2 [+8.8,+11.6]
	Top-quartile share, unconditional	0.1 / 4.2 / 0.1 / 10.0	+5.7 [+4.8,+6.6]
	New probability × percentile	0.9 / 9.7 / 1.9 / 19.7	+9.0 [+7.9,+10.2]
Qwen3.8-flash	New position	58.9 / 75.6 / 79.0 / 91.9	-3.7 [-6.9,-0.9]
	Conditional percentile	2.8 / 8.8 / 3.7 / 22.2	+12.6 [+11.1,+14.1]
	Top-quartile share, unconditional	0.3 / 2.0 / 0.6 / 8.1	+5.9 [+4.8,+6.9]
	New probability × percentile	1.7 / 6.8 / 2.9 / 20.5	+12.5 [+11.2,+13.9]
GPT-4.1-mini	New position	97.1 / 97.7 / 96.7 / 98.5	+1.1 [+0.1,+2.2]
	Conditional percentile	3.6 / 17.8 / 3.4 / 21.9	+4.2 [+3.1,+5.3]
	Top-quartile share, unconditional	0.9 / 7.5 / 0.6 / 9.9	+2.7 [+1.8,+3.5]
	New probability × percentile	3.5 / 17.4 / 3.3 / 21.6	+4.4 [+3.3,+5.5]

Table A20: **Decomposition of the uncertainty-percentile measure.** Conditional-percentile rows reproduce the canonical primary analysis; unconditional rows use the same trajectories across all choices. Positive interactions on both unconditional measures extend the pattern beyond the conditional measure. Intervals are seed-paired bootstrap 95% CIs.

Question from raw history	DeepSeek V4.1 Flash	Qwen3.8-flash	GPT-4.1-mini
Unrevealed positions, F1	0.86	0.89	0.92
Unrevealed positions, exact set	0.31	0.59	0.60
Longest unrevealed run, exact	0.44	0.64	0.62
Longest run, $\geq 80\%$ overlap	0.62	0.72	0.80
Position of highest value seen	0.69	0.90	0.65

Table A21: **Coverage reconstruction from raw history.** Each model answers explicit questions at 40 mid-environment decision points sampled from its own baseline trajectories.

Condition, DeepSeek V4.1 Flash	Conditional uncertainty percentile	Best-found (%)
Neither	1.6 [1.4,1.8]	78.1
Topology-free motivation	2.5 [2.2,2.9]	80.5
Self-written notes	2.7 [2.4,3.0]	81.9
Self-written notes + motivation	3.1 [2.7,3.6]	82.6
Coverage representation	12.0 [10.9,13.2]	85.4
Coverage representation + motivation	23.2 [21.9,24.6]	86.6

Table A22: **Original self-written-state comparison.** In the original prompt formats, the notes-by-motivation interaction is -0.5 [-1.2,+0.1] points, compared with +10.2 points for the coverage representation. Among 12,000 notes, 83% mention unsampled positions, 67% give numeric ranges, 87% describe a local trend or peak, 50% propose a next step, and none identifies the longest gap. Table A17 shows that rendering accounts for much of this prompt-format gap.

	Destination uncertainty rank	Destination mean rank	Most-uncertain arm (%)
Humans ($n = 857$)	0.514	0.691	14.6
DeepSeek V4.1 Flash	.30 / .32 / .29 / .28	.66 / .65 / .76 / .76	1.1 / 1.4 / 7.3 / 6.7
Qwen3.8-flash	.30 / .34 / .38 / .36	.67 / .70 / .69 / .75	1.4 / 1.8 / 14.2 / 12.7
GPT-4.1-mini	.32 / .32 / .19 / .14	.63 / .65 / .78 / .79	0.9 / 1.0 / 2.3 / 1.4

Table A23: **Destinations of restless-bandit switches.** Agent cells are ordered as neither, representation, motivation, and both. Ranks among the three alternative arms run from 0 (lowest) to 1 (highest); 0.5 is random. Motivation primarily redirects switches toward high-mean arms.

Setting	Interaction, seed-paired bootstrap 95% CI
1D, DeepSeek V4.1 Flash	+10.2 [+8.8,+11.6]
1D, Qwen3.8-flash	+12.6 [+11.1,+14.1]
1D, GPT-4.1-mini	+4.2 [+3.1,+5.3]
1D, DeepSeek V4.1 Flash, provider-native reasoning	+6.6 [+5.4,+7.9]
1D, DeepSeek V4.1 Flash, reduced representation	+6.1 [+4.9,+7.3]
2D, DeepSeek V4.1 Flash, full representation	+11.3 [+10.2,+12.4]
2D, DeepSeek V4.1 Flash, reduced representation	+5.4 [+4.6,+6.2]
Restless bandit, DeepSeek V4.1 Flash	-0.055 [-0.079,-0.032]
Restless bandit, Qwen3.8-flash	-0.044 [-0.065,-0.023]
Restless bandit, GPT-4.1-mini	-0.045 [-0.069,-0.020]

Table A24: **Seed-paired factorial interactions.** Structured-search entries are uncertainty-percentile points; bandit entries are switch-rate units. The four cells share environment seeds 1000–1049; each bootstrap draw resamples seeds and carries all four observations for a seed together.

Paper label	Requested model	Model field returned by API	Evaluation dates (2026)
DeepSeek V4.1 Flash	`deepseek-chat`; later `deepseek-flash`	`deepseek-flash`; one fingerprint across reported runs	Sep 15–19; bandit baseline rerun Sep 19
Qwen3.8-flash	`qwen3.8-flash`	`qwen3.8-flash`	Sep 15–19
GPT-4.1-mini	`gpt-4.1-mini`	`gpt-4.1-mini-2025-04-14`	Sep 15–19

Table A25: **Model versions and evaluation dates.** All calls use temperature 1.0, provider-default top-p, and no API sampling seed. Maximum output is 400 tokens for choices, 8 for key-only fallbacks, and 300 for reconstruction questions and self-written notes. Provider-native reasoning is disabled in the primary design; the DeepSeek V4.1 Flash robustness enables it with an 8,000-token budget.

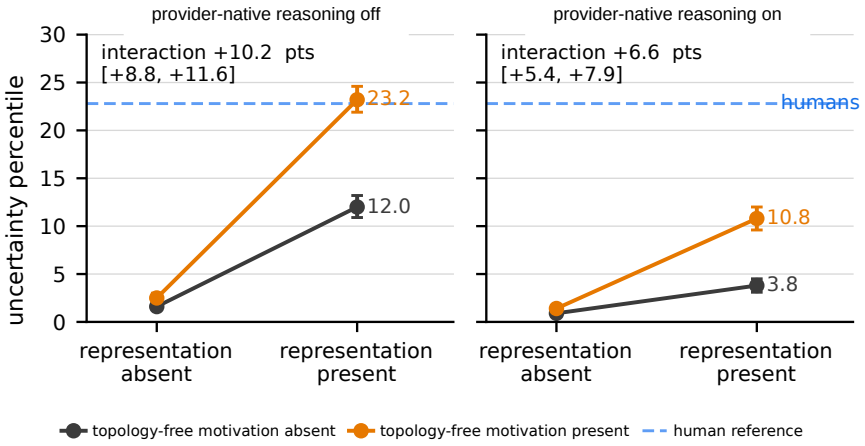


Figure A1: **A coverage representation remains effective under provider-native reasoning.** DeepSeek V4.1 Flash with identical prompts and environment seeds, provider-native reasoning off (left) and on (right), $n = 50$ agents per cell. With provider-native reasoning on, the baseline becomes more local and the joint condition reaches 10.8, with a seed-paired interaction of +6.6 [+5.4,+7.9] points.

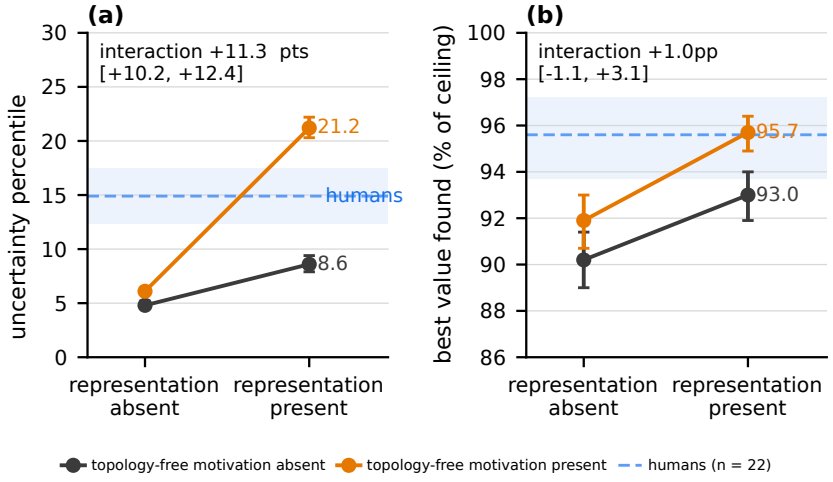


Figure A2: **The representation-by-motivation interaction replicates on a two-dimensional grid.** DeepSeek V4.1 Flash in the matched cumulative-reward, smooth-kernel condition of Wu et al. (2018), with $n = 50$ agents per factorial cell and $n = 22$ humans. (a) With the full representation, uncertainty percentile is 4.8/8.6/6.1/21.2 for neither/representation/motivation/both, giving a seed-paired interaction of +11.3 [+10.2,+12.4] points; the joint cell exceeds the human mean of 14.9. (b) Best-found value is 90.2/93.0/91.9/95.7%; the joint cell is similar to the human 95.6% (difference +0.1 points, 90% CI $[-1.5, +1.7]$). Table A13 reports the reduced-representation ablation.

D REPRESENTATION SEPARATES SEARCH EXTENT FROM SITUATION GATING

To distinguish situation-responsive exploration from a fixed exploration schedule, we fit a logistic model separately to each participant or agent:

$$\text{logit}\,\Pr(F_{iet} = 1) = \beta_{0i} + \beta_{\text{sit},i} S_{iet} + \beta_{\text{pos},i} P_{iet} + \beta_{\text{hor},i} H_{ie}. \tag{3}$$

Here F_{iet} indicates a far jump by individual i in environment e at choice t; S_{iet} is the best reward observed so far divided by the environment ceiling; P_{iet} is position within the current choice horizon; and H_{ie} indicates a ten-choice environment. A negative $\beta_{\text{sit},i}$ captures problemistic search—greater willingness to leave the current region when performance is poor—whereas $\beta_{\text{pos},i}$ captures change tied to the passage of choices.

Equation 3 separates how far search extends from when an agent leaves its current region. Representation alone moves β_{sit} in the human direction for DeepSeek V4.1 Flash (-1.92 [-3.40,-0.40]), Qwen3.8-flash (-2.09 [-3.29,-0.87]), and GPT-4.1-mini (-1.34 [-1.94,-0.73]). Motivation-only shifts are -0.13 [-1.81,+1.53], +0.84 [-0.41,+2.10], and -0.13 [-0.84,+0.56], respectively; all three interaction intervals include zero. The interaction therefore expands search reach, while the coverage representation lets departure depend on the current situation. Figure A3 shows the proof-of-concept semantic-frame comparison; Table A5 reports the clean factorial.

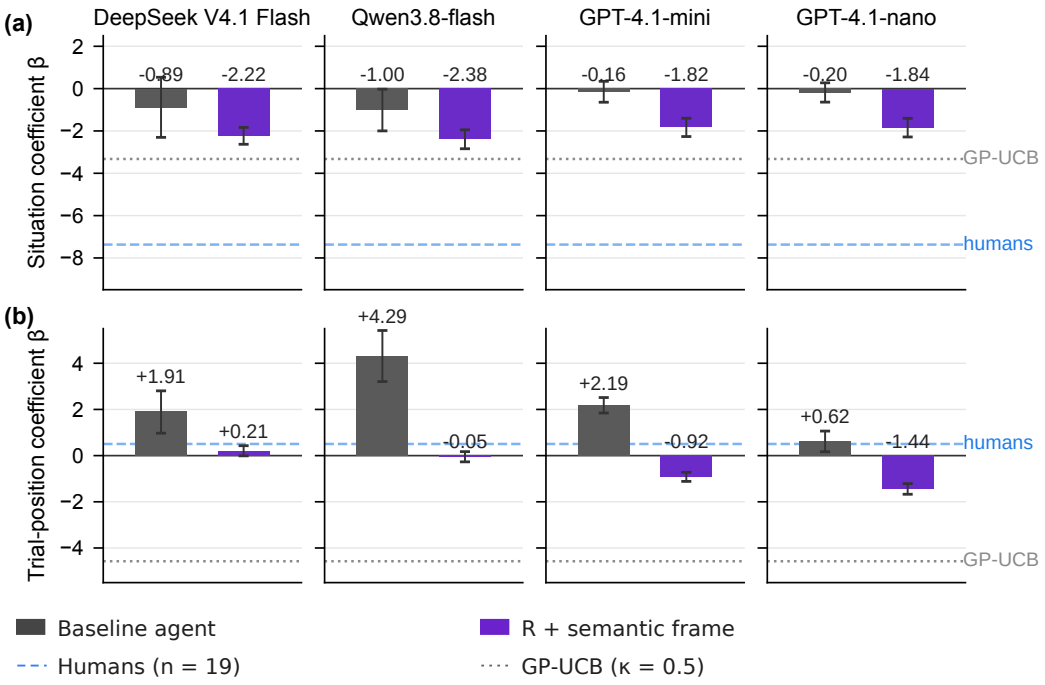


Figure A3: **Situation responsiveness and choice-position effects in agent search.** Per-agent logistic coefficients relate a distant move to the current situation (best value found so far as a fraction of the ceiling; β_{sit}) and to position within the choice horizon (β_{pos}). Purple bars show coverage representation plus the semantic-frame condition (motivation + topology), a proof-of-concept condition outside the clean factorial; dashed and dotted lines show human and GP-UCB references. Table A5 reports the clean-factorial separation between spatial reach and situation gating.

E Reference policies and token-probe model

We evaluate Gaussian-process upper-confidence-bound policies on the same structured environments as computational references. Given posterior mean $\mu_t(j)$ and standard deviation $\sigma_t(j)$, the reference selects

$$a_t^{\text{UCB}} = \arg\max_{j \in \mathcal{A}} \{\mu_t(j) + \kappa \sigma_t(j)\}, \tag{4}$$

where $\mathcal{A}$ is the set of 30 positions, with κ varied from greedy exploitation to strong uncertainty seeking. These policies place observed behavior within a standard exploration family and serve only as computational references.

Token-probe model. Token-level probes examine how locality enters the action distribution. At matched mid-environment states, we obtain action-token log-probabilities under the original labels, random permutations of labels 1–30, and random two-digit codes. For each state, candidate-action log-probability is related to both distance in the task and distance between the numeric labels:

$$\log p_t(a) = \alpha_t + \gamma_{\mathrm{sp}} d_{\mathrm{sp}}(a, a_t^{\mathrm{cur}}) + \gamma_{\mathrm{lab}} d_{\mathrm{lab}}(L(a), L(a_t^{\mathrm{cur}})) + \gamma_{\mathrm{obs}} O_t(a) + \varepsilon_{ta}. \tag{5}$$

Here a_t^{cur} is the preceding action, $L(a)$ its displayed numeric label, and α_t a state-specific intercept. The distances d_{sp} and d_{lab} are absolute differences in position and label value; $O_t(a)$ indicates whether candidate action a has already been observed. Spatial and label distance coincide under the original labels and become separately identifiable after permutation. We also sum probability over actions at least five positions away, first at a bare action head and then at the action token following the generated rationale elicited by the choice prompt. These probes compare the original instructions, the clean representation-plus-motivation intervention, and the semantic frame at matched decision points from DeepSeek V4.1 Flash and GPT-4.1-mini trajectories.