RELEVANT IS NOT ALWAYS USEFUL: TASK AND INFORMATION DEPENDENCE OF GUIDANCE FOR LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Natural-language instructions can improve LLM agents without parameter updates, but their value varies across tasks and information conditions. Existing approaches emphasize retrieving relevant guidance, leaving unclear whether it fits the task's execution demands and remains useful given what the agent already knows. We propose an evaluation framework that measures changes in task success against control instructions matched in length and format. Our central insight is that guidance must both fit the task and address an unresolved need; otherwise, it can prescribe unsuitable or unnecessary actions. We test these two factors through task-family comparisons, information and search-order interventions, and selective-delivery experiments. In ALFWorld, the same guidance changes from a 22.33-percentage-point benefit with irrelevant information to a 12.67-point loss with known target location, a 35.00-point reversal. This pattern recurs in two additional model configurations. Search-order interventions support this explanation, while independent selection gains remain uncertain. These findings motivate evaluating guidance jointly against task demands and the agent's available information, rather than task relevance or instruction adherence alone. Code and data are available anonymously at `https://anonymous.4open.science/r/review-code-anon`.

1 INTRODUCTION

As LLM agents increasingly tackle interactive, multi-step tasks, adapting them without parameter updates has become an important practical goal. A common training-free approach supplies external procedural guidance—natural-language instructions for how an agent should approach and execute a task (Shinn et al., 2023; Zhao et al., 2024; Wang et al., 2025). Such guidance is editable, reusable, and applicable to models accessible only through external interfaces. Its utility can nevertheless vary across execution contexts, making delivery decisions central to reliable agent adaptation.

Existing methods extract useful experience, organize it into reusable procedures, and retrieve or adapt those procedures for the current task (Yao et al., 2023; Shinn et al., 2023; Zhao et al., 2024; Wang et al., 2025; Lin et al., 2026; Nie et al., 2026; Pan et al., 2026). This paradigm (Paul, 2026; Li et al., 2026; Cai et al., 2026) improves guidance quality and task-level matching. Delivery also depends on incremental utility in the agent's current state: once the targeted deficit is resolved, a procedure may offer little additional benefit while continuing to incur execution costs. Estimating this state-dependent value is therefore necessary for deciding whether selected guidance should be applied. Our central insight is that guidance utility is jointly determined by **task–procedure compatibility** and **residual need**. Compatibility captures whether the procedure addresses a genuine task demand; residual need captures whether the agent still requires the information or support it supplies. Information obtained elsewhere can reduce the procedure's contribution while its prescribed actions continue to consume interaction steps and attention. This imbalance creates a **policy–state mismatch** that can reverse the procedure's net effect.

We develop a contextual-utility framework that models procedural guidance as an intervention on a frozen agent. Given a task, an information state, and a candidate procedure, guidance utility is the change in task completion relative to a surface-matched control. Task comparisons characterize compatibility, controlled information changes isolate residual need, and component interventions

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.

probe persistent execution costs. The resulting state-conditional incremental effect serves as the prediction target for guidance selection.

We instantiate the framework with search-execute guidance in ALFWorld (Shridhar et al., 2021) and task-specific procedural guidance on MiniWoB++ (Liu et al., 2018) through BrowserGym (Le Sellier de Chezelles et al., 2025). Holding task and guidance fixed shifts the main effect from $+22.33$ percentage points (pp) with irrelevant information to -12.67pp with target-location information, a -35.00pp interaction. The reversal appears at 24B and 27B, while 7B shows a different pattern. Search-order interventions support policy–state mismatch; behavior-targeted rewriting and prospective activation yield mixed or uncertain completion gains. Our contributions are:

- **When relevant guidance helps or hurts.** We identify compatibility and residual need as distinct determinants of procedural-guidance utility, supported by task heterogeneity, controlled reversals, and a model-dependent boundary.
- **A state-conditional account of guidance utility.** We formulate contextual utility through diminishing deficit-resolution benefit and persistent execution cost, with matched controls and component interventions making the account testable.
- **From mechanism tests to selective deployment.** We evaluate the mechanism and its deployment implications through information controls, procedure interventions, and prospective selection tests, exposing the limits of behavioral repair and practical activation rules.

2 Related Work

External scaffolding and harness optimization. LLM agents can improve without parameter updates by modifying their external context. ReAct, Reflexion, and Self-Refine use interleaved reasoning or verbal feedback (Yao et al., 2023; Shinn et al., 2023; Madaan et al., 2023), while prompt optimization, pipeline compilation, and harness evolution optimize prompts, programs, or runtime scaffolds (Shin et al., 2020; Zhou et al., 2023b; Khattab et al., 2024; Zhang et al., 2026b; Lin et al., 2026; Nie et al., 2026; Xu et al., 2026). Recent evaluations also distinguish the ability to modify a harness from demonstrated downstream benefit (Lin et al., 2026; Wang et al., 2026; Cho, 2026). These systems primarily ask how to construct or improve external support. Our question is complementary: after the content is fixed, what determines whether supplying it is beneficial rather than neutral or harmful?

Experience and selective procedural guidance. Voyager, ExpeL, and Agent Workflow Memory preserve skills or workflows for later retrieval (Wang et al., 2024; Zhao et al., 2024; Wang et al., 2025). ADaPT conditions decomposition on executor capability, grounded planning couples relevance with feasibility, and KnowNo uses uncertainty to request help (Prasad et al., 2024; Ahn et al., 2022; Ren et al., 2023). Recent systems additionally adapt retrieved skills or trigger retrieval on demand, moving selection toward state-aware use (Li et al., 2026; Cai et al., 2026; Zhang et al., 2026a). These studies motivate our distinction between a task's possible procedural demand and the agent's residual deficit in the current instance. Retrieval relevance alone is insufficient because a task-compatible procedure can become redundant once instance-specific evidence resolves the uncertainty that motivated it.

Conditional effects and guidance evaluation. Prompt order, formatting, and context position can alter model behavior (Lu et al., 2022; Sclar et al., 2024; Liu et al., 2024); instruction-following benchmarks assess explicit constraint satisfaction (Zhou et al., 2023a; Jiang et al., 2024; Qi et al., 2025). Evaluation variance and harness sensitivity further motivate controlled comparisons (Madaan et al., 2024; Dror et al., 2018; Wang et al., 2026; Cho, 2026). Direct studies of instruction–task interference likewise show that additional context is not uniformly beneficial (Qi et al., 2026; Gloaguen et al., 2026). This distinction matters for selective deployment: a model may follow a procedure more faithfully without improving the outcome, and high guided-arm success can coexist with negative incremental utility when the matched control performs better. We therefore evaluate context-dependent intervention effects rather than adherence or guided-arm success alone.

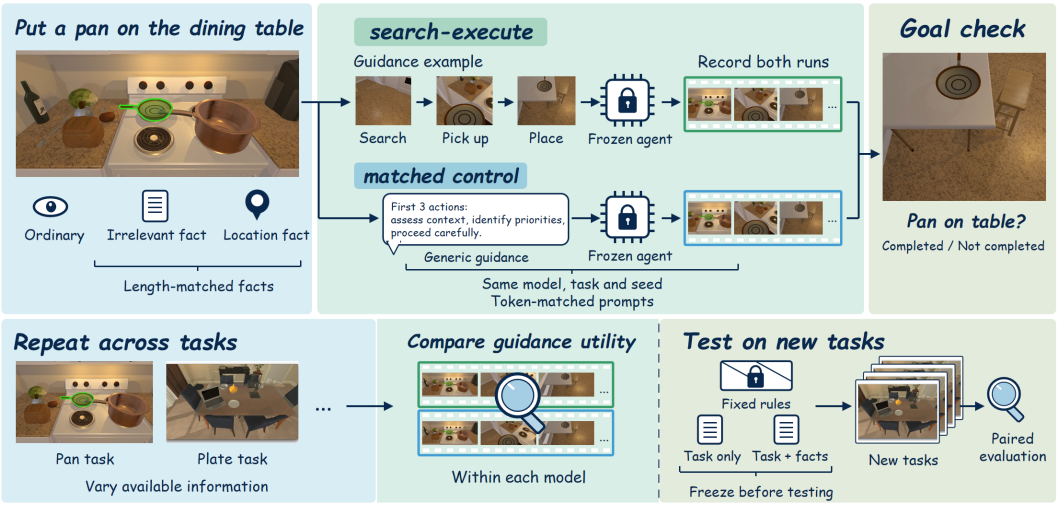

Figure 1: Overview of our Contextual-Utility Framework for state-conditional procedural guidance.

3 METHOD

We develop a framework that models procedural guidance as a context-dependent intervention on a frozen agent. Guidance utility depends on two alignments: the procedure must address the task's bottleneck, and the targeted deficit must remain unresolved in the current information state. The framework then decomposes the procedure's effect into uncertainty-reduction benefit and execution cost, explaining why useful guidance can lose value after relevant information becomes available. Section 3.1 introduces the common notation and conditional-utility estimand. Section 3.2 formalizes task–procedure compatibility and residual need. Section 3.3 derives the policy–state mismatch mechanism and its observable predictions. Section 3.4 converts these principles into a prospective activation objective. Figure 1 summarizes the matched comparisons and deployment audit.

3.1 PROBLEM DEFINITION

In this subsection, we define the notation and conditional-utility estimand used throughout the framework. For frozen model m and instance i, let ϕ_i denote the task description, k_i its task family, z_i the information available before execution, and $g \in \{0, 1\}$ the prompt condition, where $g = 1$ supplies the fixed search-execute procedure and $g = 0$ supplies the token-matched control. Let $Y_{im}(g, z_i) \in \{0, 1\}$ be benchmark-recorded success. The conditional incremental guidance utility is

$$\tau_m(k, z) = \mathbb{E}_i[Y_{im}(1, z) - Y_{im}(0, z) \mid k_i = k], \tag{1}$$

where the expectation averages over matched instances and run randomness. The quantity $\tau_m(k, z)$ measures the effect of supplying guidance to model m within task family k and information state z. In ALFWorld, $z \in \{z_{\text{ord}}, z_{\text{irr}}, z_{\text{loc}}\}$ denotes ordinary observation, a length-matched irrelevant fact, or verified target-location information.

3.2 DUAL ALIGNMENT: TASK COMPATIBILITY AND RESIDUAL NEED

In this subsection, we explain how task–procedure compatibility and residual need jointly determine $\tau_m(k, z)$. Task–procedure compatibility asks whether the ordered actions encouraged by guidance address a genuine bottleneck in task family k. A search routine fits a task that requires locating an object, whereas a task governed by a different operation or action order creates procedural misalignment. Residual need asks whether model m still lacks the information or capability supplied by the procedure under state z. Once z reveals the target location, the search routine loses its information-acquisition role, although the prompted policy can continue to execute it. The two conditions therefore shape utility through different routes. Guidance can resolve an active deficit when both conditions hold; a compatibility failure redirects behavior away from the task's bottleneck, while a residual-need failure leaves the execution cost of a procedure whose informational work has already been completed. Their combination produces the strongest mismatch between the prompted procedure and the current problem state.

We isolate the residual-need component with the specificity contrast

$$\Psi_m(k) = \tau_m(k, z_{\text{loc}}) - \tau_m(k, z_{\text{irr}}). \tag{2}$$

The contrast holds model m, task family k, guidance content, and prompt length fixed while changing whether z resolves the uncertainty targeted by search. A negative $\Psi_m(k)$ identifies lower marginal guidance utility under verified target-location information than under equally long verified but irrelevant information. Task-family comparisons characterize compatibility differences, and $\Psi_m(k)$ isolates how the remaining need for search changes guidance utility.

3.3 Information Saturation and Policy–State Mismatch

In this subsection, we explain how information saturation changes the net value of guidance and produces observable policy–state mismatch. At interaction step t, prompt condition g affects the frozen policy through its action distribution, while each action and observation updates the history:

$$a_t \sim \pi_m(\cdot \mid \phi_i, I_t, g), \qquad I_{t+1} = I_t \oplus (a_t, o_{t+1}). \tag{3}$$

This channel connects the prompt-level intervention g to the trajectories that determine $Y_{im}(g, z_i)$. When z_i leaves the target location unknown, ordered search can acquire missing evidence, reduce ambiguity, and place prerequisite actions in a productive sequence. For an informational routine s, we represent this trade-off through its conceptual net value

$$\mathcal{N}(s \mid I_t) = \underbrace{\mathbb{E}_{O_s \mid I_t}\Big[\max_a \mathbb{E}[U(a) \mid I_t, O_s]\Big] - \max_a \mathbb{E}[U(a) \mid I_t]}_{\text{uncertainty-reduction benefit}} - \underbrace{C_{\text{exec}}(s \mid I_t)}_{\text{execution cost}}, \tag{4}$$

where O_s denotes the observation produced by routine s, $U(a)$ the downstream utility of action a, and $C_{\text{exec}}(s \mid I_t)$ the cost of executing the routine from history I_t.

Moving from z_{irr} to z_{loc} reduces the uncertainty that search can resolve and therefore shrinks the benefit term in $\mathcal{N}(s \mid I_t)$. The frozen policy can still follow the prompted sequence, so its action, attention, and opportunity costs persist. The resulting policy–state mismatch arises because the procedure reflects an earlier information state and continues to induce obsolete actions. If a successful route requires L actions within horizon H and obsolete guidance adds $D \geq 0$ actions, the resulting budget-induced failure mass is

$$\Pr(L \leq H) - \Pr(L + D \leq H) = \Pr(L \leq H < L + D) \geq 0. \tag{5}$$

The extra actions can also change subsequent observations and move the agent into a less favorable state. Information saturation can therefore lower $\tau_m(k, z)$ through both budget pressure and trajectory disruption.

Testable implications. The mechanism links each experimental intervention to a specific observable consequence. Changing z from z_{irr} to z_{loc} tests whether guidance utility falls after the targeted uncertainty is resolved, which appears as a negative $\Psi_m(k)$. Removing the fixed search component then reveals whether search contributes useful structure under unresolved uncertainty and persistent overhead after saturation, which can be examined through completion, trajectory length, timeouts, and pre-target navigation. A behavior-targeted rewrite provides a complementary test: it can improve the intermediate action pattern, while terminal success remains governed by any downstream bottleneck left in the trajectory. Section 4.3 evaluates these linked predictions through information controls, search-order interventions, and rewrite experiments.

3.4 State-Conditional Guidance Selection

In this subsection, we convert conditional guidance utility into a rule for deciding whether guidance should be delivered before execution. The activation target is the sign of the guidance-minus-control effect for task description ϕ_i and information state z_i. An idealized information-aware rule enables guidance when its expected incremental utility is positive:

$$h^*_m(\phi_i, z_i) = \mathbf{1}\{\mathbb{E}[Y_{im}(1, z_i) - Y_{im}(0, z_i) \mid \phi_i, z_i] > 0\}. \tag{6}$$

The oracle rule h^*_m defines the decision target for practical selectors. A task-only rule $f(\phi_i)$ approximates compatibility from the task description and applies the same decision across information states. Conditional wording supplies z_i to the frozen model and relies on its prompted policy to suppress obsolete actions during execution. An information-aware rule $h(\phi_i, z_i)$ uses both inputs before execution, allowing the delivery decision to respond to compatibility and residual need.

Prospective evaluation. We measure an activation rule by the outcome it induces across guided and control runs. For $h_i = h(\phi_i, z_i)$ and $f_i = f(\phi_i)$, the policy value and state-aware advantage are

$$\begin{aligned} V_m(h) &= \mathbb{E}_i[h_i Y_{im}(1, z_i) + (1 - h_i) Y_{im}(0, z_i)] \,, \\ V_m(h) - V_m(f) &= \mathbb{E}_i[(h_i - f_i)\{Y_{im}(1, z_i) - Y_{im}(0, z_i)\}] \,. \end{aligned} \tag{7}$$

The second line shows that the policy gap is determined by instances where h_i and f_i choose different treatments, weighted by the realized guidance effect on those instances. We freeze both rules before observing evaluated outcomes and then compare task-only, information-aware, always-on, and always-off policies prospectively in Section 4.4.

4 Experiments

We evaluate fixed procedural guidance in ALFWorld (Shridhar et al., 2021) and MiniWoB++ (Liu et al., 2018) through BrowserGym (Le Sellier de Chezelles et al., 2025). We address the following research questions:

- **RQ1: Conditional utility.** Does utility vary with task–procedure compatibility and residual information need?
- **RQ2: Mechanistic evidence.** Do interventions on information type and search policy support the proposed mechanism?
- **RQ3: Prospective selection.** Can these conditions support reliable prospective delivery?

4.1 Experimental Setup

The ALFWorld studies compare unchanged search-execute guidance with a surface-matched control under Qwen3-32B-AWQ and separate Mistral and Gemma protocols. The MiniWoB++ studies compare task-specific or oracle-relevant guidance with matched control; browser oracle guidance supplies task-class procedural content, not an ALFWorld target-location fact. Comparisons match task draws or seeds, completion is benchmark-recorded success, and effects are intervention-minus-control risk differences in percentage points (pp). Reported 95% confidence intervals (CIs) use task-description-cluster bootstrap where available. Models, discovery runs, and confirmatory blocks remain separate; Appendix C.1 maps studies to endpoints, with protocol-level controls in Appendices B and C.2.

4.2 Conditional Utility Across Tasks and Information States

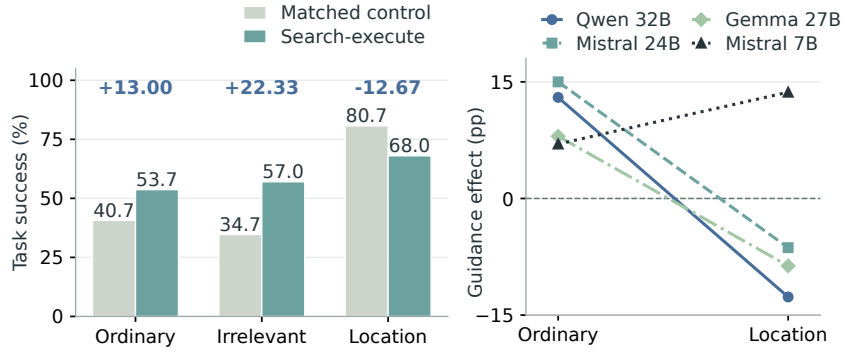

Figure 2: **Information-state reversal and model boundary.** (a) Qwen new-seed success and guidance effects. (b) Model-separated ordinary-to-location effects; connected points do not imply a model-size trend. Counts and interaction CIs appear in Table 1.

The new-seed ordinary-observation effect is $+13.00$pp, while family estimates vary in sign (Appendix Figure 10) and the held-out examine check remains inconclusive. More decisively, holding guidance fixed while changing available information reveals a crossover: ordinary observations and an equal-length irrelevant fact preserve a benefit, whereas known target location reverses it (Figure 2(a)). The

-35.00pp location-minus-irrelevant interaction (Table 1) shows that utility declines when supplied information overlaps with the procedure's function.

Mistral-24B and Gemma-27B reproduce the reversal, whereas Mistral-7B remains positive in both states (Figure 2(b); Table 1). The repeated crossover establishes an information-state boundary for three larger configurations, while the 7B result rules out a universal state-only delivery rule; Gemma also changes quantization and backend.

Table 1: **Conditional utility across models.** Paired successes and state interactions complement Figure 2.

Model	State	C	G	Δ
Qwen 32B	O	122	161	+13.00
	I	104	171	+22.33
	L	242	204	−12.67
Mistral 24B	O	38	83	+15.00
	L	164	145	−6.33
Gemma 27B	O	39	63	+8.00
	L	131	105	−8.67
Mistral 7B	O	15	36	+7.00
	L	44	85	+13.67

Interaction	Δ	95% CI
Qwen 32B: L − I	−35.00	[−50.00, −20.00]
Qwen 32B: I − O	+9.33	[+0.99, +17.93]
Mistral 24B: O − L	+21.33	[+9.23, +33.54]
Gemma 27B: O − L	+16.67	[+7.05, +27.44]
Mistral 7B: O − L	−6.67	[−21.12, +8.24]

O: ordinary; I: irrelevant fact; L: target-location fact. C/G: control/guidance successes out of 300 per state. Δ: G − C (pp); interactions compare these effects, with signs as labeled.

Models and protocols are separate; Gemma also changes quantization and backend. Qwen facts are tokenizer-matched. Only reported interaction CIs are shown; per-state CIs are not inferred. Mistral-7B counts were reconstructed from reported rates and $n = 300$.

4.3 Mechanistic Tests: Information Type and Search Policy

The proposed mechanism is that guidance helps by resolving a residual action-selection problem and loses value once its targeted uncertainty has been removed. We test this account by changing the supplied information, the prescribed search policy, and the relation between behavioral compliance and completion.

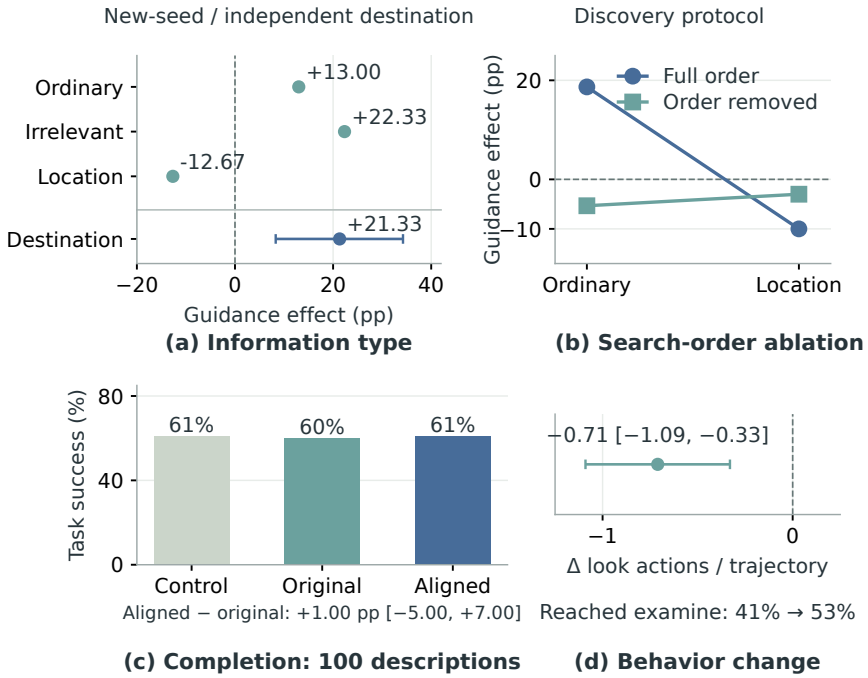

Figure 3: **Information and policy interventions.** (a) Target-location versus independent destination information. (b) Fixed-order removal. (c–d) Aligned rewriting changes behavior without a resolved completion gain. Table 2 reports the corresponding contrasts.

An independent destination fact preserves the benefit, while target-location information removes the uncertainty that the procedure is designed to resolve and reverses the effect (Figure 3(a); Table 2). Removing fixed search order changes the ordinary-minus-location interaction by -31.00pp—a 31.00pp attenuation—and independently reversing the order makes guidance adverse in both states, directly linking the crossover to the induced search policy. On new examine descriptions, aligned wording reduces look actions and increases reaching the examine stage without a resolved completion gain (Figure 3(c–d)). Together, these interventions connect effect variation to policy–state alignment and show that compliance alone does not imply utility.

Table 2: **Mechanistic interventions.** Qwen studies test information type, prescribed search order, and behavioral compliance (Figure 3).

Intervention	Contrast / outcome	Effect	95% CI
Destination fact	Guidance − control	$+21.33$	$[+8.30, +34.25]$
Remove order	Change in O − L interaction	-31.00	$[-46.58, -15.04]$
Reverse order	Guidance − control (O / L)	$-11.67/-9.33$	–
Aligned rewrite	Completion: aligned − original	$+1.00$	$[-5.00, +7.00]$
Aligned rewrite	Look actions: aligned − original	-0.71	$[-1.09, -0.33]$

Effects are pp except look actions per trajectory; O/L denote ordinary/location states. Destination, discovery order-removal, independent reversed-order, and rewrite studies are separate. Rewrite outcomes share 100 descriptions. A zero-crossing CI does not establish equivalence; – means unreported.

4.4 Prospective Selection and Its Limits

The preceding results identify conditions associated with guidance value. We evaluate whether they support delivery rules that progress from task-only prediction to conditional wording and explicit information-aware activation.

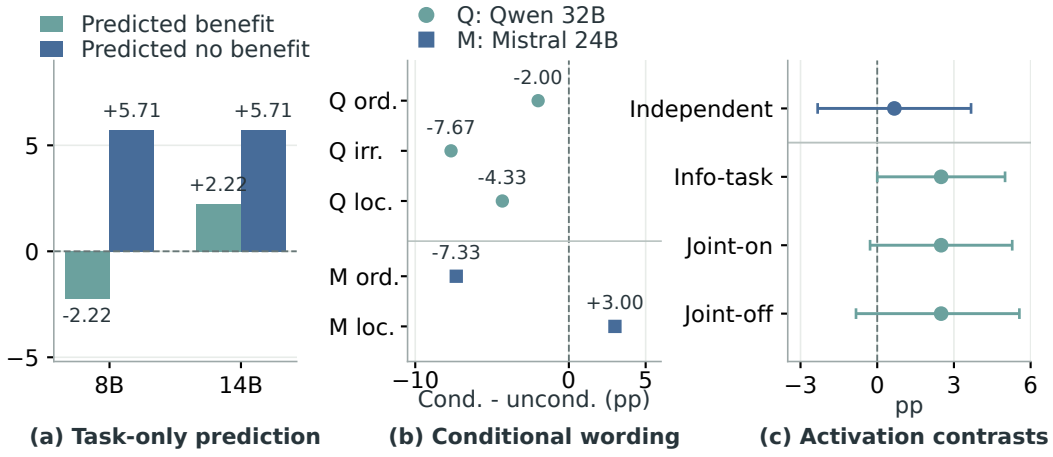

Figure 4: **Selection rules and their limits.** (a) Task-only prediction (Table 3). (b) Conditional wording. (c) Information-aware activation, with independent replication beyond the divider. Available 95% CIs are shown.

A frozen task-only rule partitions 16 unseen BrowserGym tasks, but neither model recovers the predicted benefit ranking (Table 3), and the prediction-file hash discrepancy qualifies its prospective provenance. Conditioning search order on unknown location also fails as a general repair: Qwen shows no beneficial interaction, while Mistral combines an ordinary-state cost with a smaller location-state gain (Figure 4(b)). Explicit activation from observed information state improves over task-only selection by $+2.50$pp in the primary study, but only $+0.67$pp with a zero-crossing CI independently (Figure 4(c)). Information state is therefore a better selector input than task label alone, although a reliable net deployment benefit remains unestablished.

Table 3: **Task-only prediction audit.** Neither model recovers the predicted ordering; counts supplement Figure 4(a).

Ministral-3	Predicted group	n	Control	Oracle	Effect (pp)
8B	Benefit	90	29	27	↓ 2.22
	No benefit	70	47	51	↑ 5.71
8B: benefit − no-benefit effect contrast					−7.94
14B	Benefit	90	29	31	↑ 2.22
	No benefit	70	45	49	↑ 5.71
14B: benefit − no-benefit effect contrast					−3.49

Control/oracle columns are success counts; effects are oracle − control. Each model has nine benefit and seven no-benefit tasks, ten paired seeds per task. Group-contrast CIs are unreported. The prediction-file hash differs from the pre-run validator hash; prospective provenance remains qualified.

4.5 Scope and Robustness

We test the scope of these conclusions through independent task confirmation, action-budget changes, a browser extension, and secondary format controls. The held-out examine estimate attenuates toward zero (Figure 5(a)), so unseen descriptions do not confirm a stable family-level failure; the separate new-seed breakdown appears in Appendix Figure 10. Increasing the allowance from 20 to 40 steps also does not establish a repair because both original-guidance intervals cross zero (Figure 5(b)).

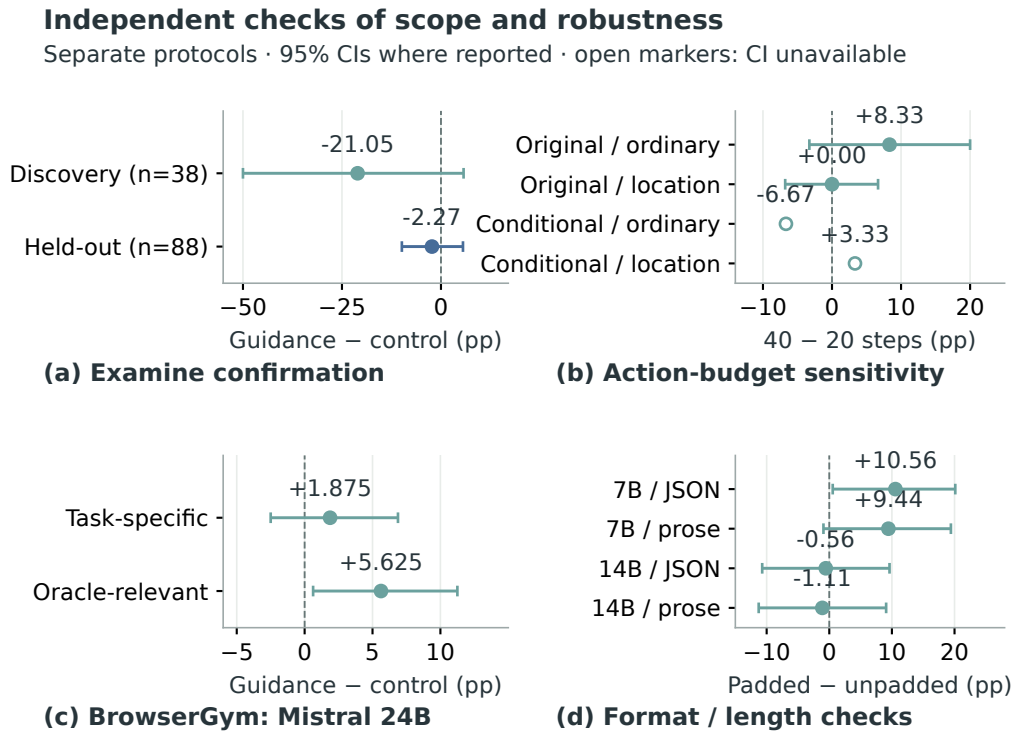

Figure 5: **Scope and robustness.** (a) Discovery and held-out examine. (b) Budget changes (40 − 20 steps); open markers indicate unreported intervals. (c) Separate 24B BrowserGym transfer. (d) Secondary padding controls confounded by length.

The 24B browser extension supports oracle-relevant guidance beyond ALFWorld, while its task-class advice differs from a target-location fact and shares controls across variants (Figure 5(c)). Validation guidance is strongly adverse for 14B under ordinary status and less so under trusted status, while 8B stays near zero (Table 4); these interactions concern a different information–procedure pair. Finally, padding effects vary across models and formats and confound rendering with token

length (Figure 5(d)), whereas the primary reversal uses tokenizer-matched facts. Protocol details and additional results appear in Appendices A.3, C, and C.2.

Table 4: **BrowserGym validation-status boundary.** Paired task-success counts and guidance effects.

Model	Information state	Control	Guidance	Effect (pp)
Ministral 8B	Ordinary status	111/160	108/160	-1.88
Ministral 8B	Trusted status	110/160	110/160	$+0.00$
Ministral 8B	State-averaged	221/320	218/320	-0.94
Trusted − ordinary interaction (CI unavailable)				$+1.88$
Ministral 14B	Ordinary status	111/160	62/160	-30.62
Ministral 14B	Trusted status	116/160	112/160	-2.50
Ministral 14B	State-averaged	227/320	174/320	-16.56
Trusted − ordinary interaction (CI unavailable)				$+28.12$

Success counts are paired within each state; each model contributes 640 trajectories. State-averaged rows reuse the two displayed states and are not independent replications. Effects compare validation guidance with control, not browser oracle guidance or ALFWorld location facts. Interactions subtract the ordinary-state effect from the trusted-state effect. No unreported CI is inferred from marginal counts.

Across these scope checks, the evidence narrows both where the mechanism transfers and what a delivery rule must estimate. Oracle-relevant browser guidance extends task–procedure compatibility beyond ALFWorld, while the validation-status contrast tests a second information–procedure pair: the same checking routine is most costly under ordinary status, while trusted validation evidence largely attenuates this cost. The budget intervention addresses a different alternative explanation. If the reversal mainly reflected a short interaction horizon, doubling the action allowance should recover a positive guidance effect; instead, the estimated changes remain compatible with no repair. The padding controls also vary across models and formats, whereas the primary comparison uses tokenizer-matched facts, so surface length does not organize the central effect.

The remaining boundaries localize the uncertainty in prospective selection. The 7B result, held-out examine attenuation, and zero-crossing activation replication show that a state cue cannot be converted directly into a universal on/off rule. A useful selector must jointly estimate whether the current observation has already resolved the uncertainty targeted by the procedure, how strongly the procedure will redirect the model's action policy, and whether the resulting search cost is justified by residual uncertainty. These requirements explain why task labels alone fail to recover the expected ordering and why information-aware activation improves the primary study without yielding a stable replicated gain. Taken together, the scope analyses support conditional utility and the policy–state mechanism while marking calibration as the remaining deployment problem. They therefore turn the failure cases into a concrete design constraint: expose guidance only when its targeted uncertainty remains active and its induced policy is expected to reduce, rather than repeat, search.

5 Conclusion

We develop a contextual-utility framework for frozen LLM agents that separates task–procedure compatibility from residual information need through matched interventions and prospective selection tests. In ALFWorld, fixed guidance changes from $+22.33$pp with irrelevant information to -12.67pp once the target location is known, yielding a -35.00pp interaction. Information-type and search-order interventions support a policy–state mismatch account: guidance helps while it resolves residual uncertainty and becomes costly when it induces redundant search after that uncertainty is removed. This reversal recurs in three larger model configurations and transfers to an oracle-relevant browser setting, while the 7B boundary, held-out attenuation, and independently replicated activation gain of only $+0.67$pp (95% CI $[-2.33, +3.67]$) show that neither task labels nor information state alone provide a universal delivery rule. The practical implication is to evaluate guidance as a state-dependent intervention: prospective systems should estimate its incremental utility from the procedure, the current observation, and the action policy it will induce, and should withhold it when those signals indicate that its informational role has already been satisfied.

REPRODUCIBILITY STATEMENT

This work spans several runtime environments and open-weight agents from 4B to 32B. The legacy `Qwen3-4B` experiments ran on an NVIDIA A100-PCIE-40GB under vLLM v0.26.0. The scaffold-content axis (`Qwen2.5-7B`, `Llama-3.1-8B`, and `Qwen2.5-14B` at 8192 context) ran on one RTX 4090 under vLLM 0.6.4.post1, Python 3.12.3, and `ALFWorld` 0.4.2; the older `Qwen2.5-32B` point used a 4096-token ceiling and is excluded from pooled inference. The `Qwen3-32B` `ALFWorld` studies used the official dense AWQ 4-bit checkpoint, 8192 context, and vLLM 0.10.2. The confirmatory additions comprise a new-seed 1,800-episode 2×3 specificity test, a 300-episode examine rewrite validation, a 720-episode activation study, and a 600-episode activation replication. A separate Mistral Small 3.2 24B ALFWorld replication contributes 1,200 episodes and is not pooled with Qwen. Four further blocks add 4,200 non-overlapping episodes: a 1,200-episode Gemma 3 27B replication, a 600-episode reversed-search-order control, a 600-episode destination-fact control, and 1,800 new unconditional/conditional M3 arms. The conditional analysis spans 3,600 episodes because its placebo and original-M3 arms reuse the existing 1,800-episode specificity shard. The `BrowserGym`/`MiniWoB++` studies ran BF16 `Ministral-3-8B/14B` and `Mistral Small 3.2 24B` on separate 16-task, ten-seed, three-arm protocols; these benchmark-specific estimates are not pooled with `ALFWorld`. The independent `BrowserGym` prediction record has a disclosed hash-integrity mismatch and is treated as prospectively timestamped rather than cryptographically verified. A separate `Qwen3-14B-AWQ` real-subprocess-I/O check evaluates harness fidelity under clean and accumulated scratch state; it is not pooled with task-success contrasts. Experiment-to-endpoint lineage and supporting protocol details appear in Appendices C.1 and C.2. The previously consolidated artifact retains 7,420 primary episodes, including per-episode outcomes and trajectories for the earlier Qwen3 and `BrowserGym` follow-ups, manifests, prompt hashes, preregistration records, logs, interaction outputs, and analysis scripts. The 8,820 non-overlapping confirmatory episodes listed above are reported in the September 25, 2026 evidence bundle and its September 26 supplement and are not silently folded into the earlier archive count. It does not contain `ALFWorld` axis per-episode model-output text or the raw legacy judgment records; the original format table is therefore not fully recomputable. These gaps and cross-environment differences are limitations.

Ethics statement and broader impact. This work involves no human subjects, no private data, and no deployment of autonomous agents outside controlled `ALFWorld` and `BrowserGym`/`MiniWoB++` simulations. Delivery policies can change agent behavior unevenly across tasks and models, so aggregate gains should not be treated as evidence of uniform reliability. The experiments do not evaluate deployment safety or safety-critical compliance.

AI use statement. AI systems were used in three distinct roles, itemized per ICLR guidance. **(i) Adherence judging:** `GPT-5.6-Sol` supported baseline validation and `DeepSeek v4-flash` produced legacy format-comparison scores. **(ii) Directive authoring:** a language-model assistant helped draft harness directives, which the authors reviewed and fixed before use. **(iii) Manuscript preparation:** a language-model assistant helped draft and edit prose. The authors verified the experimental design, code, analyses, and claims.

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Yuxuan Cai, Wei Li, Jie Zhou, Qin Chen, Xin Li, Bo Zhang, and Liang He. Ask only when needed: Proactive retrieval from memory and skills for experience-driven lifelong agents. *arXiv preprint arXiv:2604.20572*, 2026.

Yong-eun Cho. It's not the capability: Harness sensitivity is non-monotone across LLM agent tiers. *arXiv preprint arXiv:2605.26731*, 2026.

Rotem Dror, Gili Baumer, Segev Bogorad, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 1383–1392, 2018.

Thibaud Gloaguen, Niels Mündler, Mark Müller, Veselin Raychev, and Martin Vechev. Evaluating AGENTS.md: Are repository-level context files helpful for coding agents? *arXiv preprint arXiv:2602.11988*, 2026. URL https://arxiv.org/abs/2602.11988v2. Version 2, June 23, 2026.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. Compiling declarative language model calls into state-of-the-art pipelines. In *International Conference on Learning Representations*, 2024.

Thibault Le Sellier de Chezelles, Maxime Gasse, Alexandre Lacoste, Massimo Caccia, Alexandre Drouin, Léo Boisvert, Megh Thakkar, Tom Marty, Rim Assouel, Sahar Omidi Shayegan, Lawrence Keunho Jang, Xing Han Lù, Ori Yoran, Dehan Kong, Frank F. Xu, Siva Reddy, Graham Neubig, Quentin Cappart, Russ Salakhutdinov, and Nicolas Chapados. The BrowserGym ecosystem for web agent research. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=5298fKGmv3.

Qinfeng Li, Dalin He, Yuntai Bao, Ying Yang, Ruoxi Chen, Xinyan Yu, Lizhou Liang, Ge Su, Wenqi Zhang, and Xuhong Zhang. SkillAligner: Treating retrieved skills as adaptable drafts at execution time. *arXiv preprint arXiv:2608.06880*, 2026.

Minhua Lin, Juncheng Wu, Zijun Wang, Zhan Shi, Yisi Sang, Bing He, Zewen Liu, Tianxin Wei, Zongyu Wu, Zhiwei Zhang, Dakuo Wang, Xiang Zhang, Benoit Dumoulin, Cihang Xie, Yuyin Zhou, Suhang Wang, and Hanqing Lu. Harness updating is not harness benefit: Disentangling evolution capabilities in self-evolving LLM agents. *arXiv preprint arXiv:2605.30621*, 2026.

Evan Zheran Liu, Kelvin Guu, Panupong Pasupat, Tianlin Shi, and Percy Liang. Reinforcement learning on web interfaces using workflow-guided exploration. In *International Conference on Learning Representations*, 2018. arXiv:1802.08802.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 8086–8098, 2022.

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 (NeurIPS)*, 2023. arXiv:2303.17651.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks. In *NeurIPS 2024 Workshops: RegML*, 2024. arXiv:2406.10229.

Jun Nie, Yonggang Zhang, Jun Song, Qianshu Cai, Dahai Yu, Yike Guo, Xinmei Tian, and Bo Han. TTHE: Test-time harness evolution. *arXiv preprint arXiv:2607.08124*, 2026.

Shuai Pan, Yixiang Liu, Jiaye Gao, Te Gao, Weiwen Liu, Jianghao Lin, Zhihui Fu, Jun Wang, Weinan Zhang, and Yong Yu. SkillMAS: Skill co-evolution with LLM-based multi-agent system. *arXiv preprint arXiv:2605.09341*, 2026.

Debjyoti Paul. Context assembly as the controlled variable: A control-theoretic view of harness policies for frozen LLM agents. *arXiv preprint arXiv:2607.25408*, 2026.

Archiki Prasad, Alexander Koller, Mareike Hartmann, Peter Clark, Ashish Sabharwal, Mohit Bansal, and Tushar Khot. ADaPT: As-needed decomposition and planning with language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4226–4252, 2024. doi: 10.18653/v1/2024.findings-naacl.264.

Yunjia Qi, Hao Peng, Xiaozhi Wang, Amy Xin, Youfeng Liu, Bin Xu, Lei Hou, and Juanzi Li. AGENTIF: Benchmarking instruction following of large language models in agentic scenarios. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. Spotlight; arXiv:2505.16944.

Yunjia Qi, Hao Peng, Xintong Shi, Amy Xin, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. On the paradoxical interference between instruction-following and task solving. *arXiv preprint arXiv:2601.22047*, 2026.

Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 661–682, 2023.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *Proceedings of the 12th International Conference on Learning Representations*, 2024. arXiv:2310.11324.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4222–4235, 2020.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Coté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv:2305.16291.

Yike Wang, Huaisheng Zhu, Zhengyu Hu, Yige Yuan, Zhengyu Chen, Shakti Senthil, Hannaneh Hajishirzi, Yulia Tsvetkov, Pradeep Dasigi, and Teng Xiao. Rethinking the evaluation of harness evolution for agents. *arXiv preprint arXiv:2607.12227*, 2026.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2409.07429.

Tianshi Xu, Huifeng Wen, and Meng Li. Adapting the interface, not the model: Runtime harness adaptation for deterministic LLM agents. *arXiv preprint arXiv:2605.22166*, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the 11th International Conference on Learning Representations*, 2023. arXiv:2210.03629.

Haozhen Zhang, Quanyu Long, Jianzhu Bao, Tao Feng, and Weizhi Zhang. MemSkill: Learning and evolving memory skills for self-evolving agents. *arXiv preprint arXiv:2602.02474*, 2026a.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Reddy Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Han-Chen Li, Urmish Thakker, James Zou, and Kunle Olukotun. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations (ICLR)*, 2026b. arXiv:2510.04618.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*, 2024. arXiv:2308.10144.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. IFEval: Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023a.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *Proceedings of the 11th International Conference on Learning Representations*, 2023b.

A Content and Task Evidence

A.1 Content-Isolation Controls

The semantic-matched ALFWorld control compares task-relevant Search–execute directives with a tokenizer- and surface-matched Matched control. On Qwen2.5-7B, Search–execute succeeds on 49/300 episodes (16.33%) and the Matched control on 25/300 (8.33%), a difference of $+8.00$ percentage points (pp; Fisher $p = 0.004054$). The arms have three directives, 42 tokens, 29 words, seven imperatives, and identical placement. The result isolates task relevance from directive count, placement, and measured length. The bare-versus-generic contrast is $+1.00$pp (not significant), consistent with generic filler being inert.

An independent fresh-seed extension gives a pooled Search–execute-versus-generic difference of $+7.80$pp (paired McNemar $p \approx 4 \times 10^{-8}$; Fisher $p = 0.000207$). In the fresh four-arm read, real, scrambled, generic, and bare content succeed on 49/300, 40/300, 27/300, and 28/300 episodes, respectively; the primary real-versus-generic contrast is $+7.33$pp. The bare arm is not used for the isolation claim because it is not length matched.

The same content-isolation design on AWQ Llama-3.1-8B yields 38/300 real, 32/300 generic, 28/300 scrambled, and 26/300 bare successes. The primary real-versus-generic estimate is $+2.00$pp (Fisher $p = 0.5251$; 95% CI $[-5.30, +9.25]$pp). This is a non-detection at the tested sample size, not evidence that task relevance is absent on Llama.

On BrowserGym/MiniWoB++, the tokenizer-matched task-specific versus Matched control contrast is not detected in four separately analysed Ministral instances: 8B pilot 0.00pp, 8B follow-up -1.25pp, 14B pilot $+2.00$pp, and 14B follow-up -2.50pp. These instances are not pooled. A prospectively locked oracle-information follow-up gives oracle-minus-Matched control $+2.50$pp at 8B (CI $[-6.25, +11.25]$pp) and $+12.50$pp at 14B (CI $[+2.50, +22.50]$pp; Holm-adjusted $p = .0618$). The pattern motivates an information-utility hypothesis but does not establish a universal cross-family effect.

A.2 Task-Clustered Sensitivity Analysis

The label C2 denotes the paper's second content axis: a multi-seed, task-clustered sensitivity analysis. It is not a model name, benchmark, or additional treatment arm. For the multi-seed Qwen2.5-7B content axis, task-clustered estimates are the primary sensitivity analysis. The Search–execute-versus-generic contrast remains positive after clustering, while task-family estimates vary substantially. The 14B replication is structure-matched to the 14B baseline and gives Search–execute-versus-bare $+18.89$pp (Fisher $p = .00048$) and Search–execute-versus-generic $+15.56$pp ($p = .0043$). These are within-family scale checks, not cross-family claims.

The component ablation is descriptive. Removing the exploration directive changes success from 18.9% to 12.2% (-6.7pp), while removing the ordering directive changes it to 24.4% ($+5.6$pp). The three-seed sign test for the exploration removal is $p = .25$. The matched three-directive control is matched in directive and word count, not full-prompt token count.

The $2 \times 2 \times 2$ content-by-format factorial has 720 episodes. Search–execute succeeds on 67/360 (18.61%) and generic content on 28/360 (7.78%), a marginal difference of $+10.83$pp (95% CI $[+3.92, +17.52]$pp; Fisher $p = 2.30 \times 10^{-5}$). All format effects and content-by-format interactions have intervals including zero. We report format effects as unresolved at this resolution, rather than as absent.

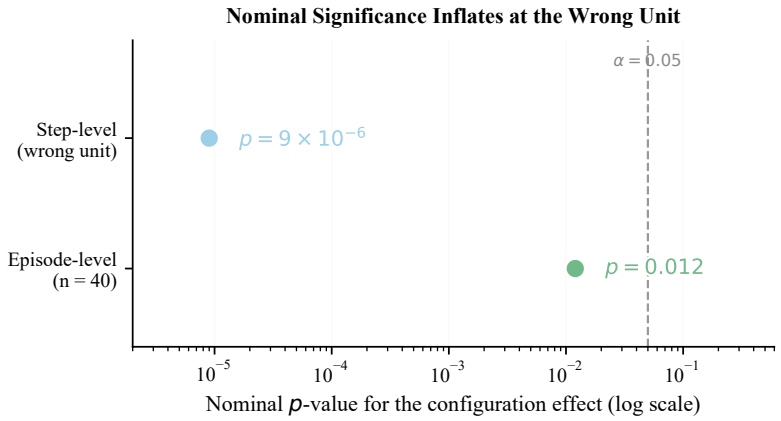

Figure 6: Episode-level versus step-level pseudo-replication sensitivity.

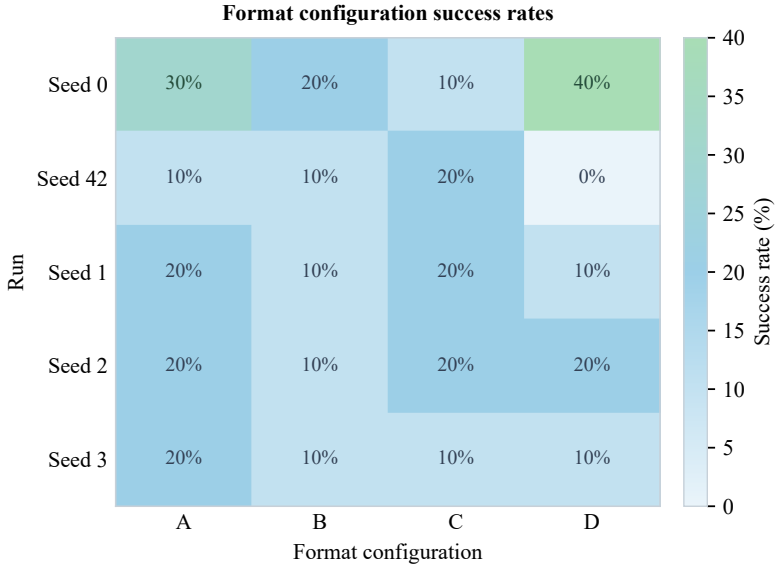

Figure 7: Format configuration success rates across runs.

A.3 Task-Level Prediction Audit

The frozen BrowserGym rule labelled nine tasks as predicted benefit and seven as predicted no benefit. The scored target was oracle-relevant guidance minus Matched control; task-specific guidance was secondary. Table 5 gives the paired oracle risk difference for the 14B and 8B strata.

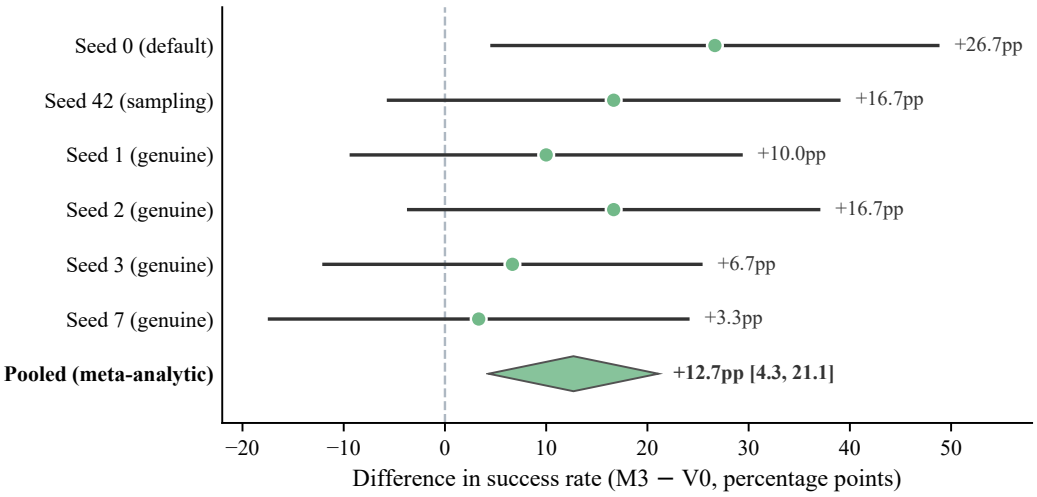

Figure 8: Forest plot of the task-clustered content effect across sensitivity runs.

Table 5: Task-level oracle-minus-Matched-control audit (percentage points).

Task	Prediction	14B RD	8B RD	Diagnostic
click-shades	benefit	0	0	floor
click-dialog-2	benefit	0	0	ceiling
enter-date	no benefit	+40	+50	date-format semantics
click-test-2	no benefit	0	0	ceiling
book-flight	benefit	+10	-10	model reversal
form-sequence-3	benefit	0	0	ceiling
enter-password	no benefit	0	0	ceiling
use-slider-2	benefit	0	0	floor
find-word	benefit	0	0	low headroom
find-greatest	benefit	+10	0	low headroom
read-table-2	no benefit	0	0	ceiling
find-midpoint	no benefit	0	0	stable baseline
email-inbox-star-reply	benefit	-20	-10	instruction mismatch
social-media-all	benefit	+20	0	model-specific
email-inbox-delete	no benefit	0	-10	model divergence
click-tab-2-hard	no benefit	0	0	floor

Across 320 paired task–model units, oracle succeeds on 158 versus 150 Matched control episodes: RD $+2.50$pp, McNemar $p = .1338$, task-cluster CI $[-2.19, +9.06]$pp. The predicted-benefit group averages 0.00pp and the predicted-no-benefit group $+5.71$pp; the preregistered interaction is -5.71pp (CI $[-20.24, +3.81]$pp), with 8/16 directional hits. The rule therefore fails as a deployment selector. The failure is structured: enter-date requires ISO date formatting, star-only email instances conflict with a fixed instruction, and six tasks are at floor or ceiling.

B Information-State and Transport Evidence

B.1 September 2026 Control Experiments

The Qwen3-32B information-state control crossed the Matched control and Search–execute with ordinary observation, an equal-length irrelevant true fact, and the true target location. Search–execute-minus-Matched control is $+13.00$pp, $+22.33$pp, and -12.67pp, respectively. The location-minus-irrelevant interaction is -35.00pp (task-description-cluster bootstrap 95% CI $[-50.00, -20.00]$pp); all injected facts are tokenizer-exactly matched. This supports decision-relevant information as a moderator rather than a generic extra-text effect.

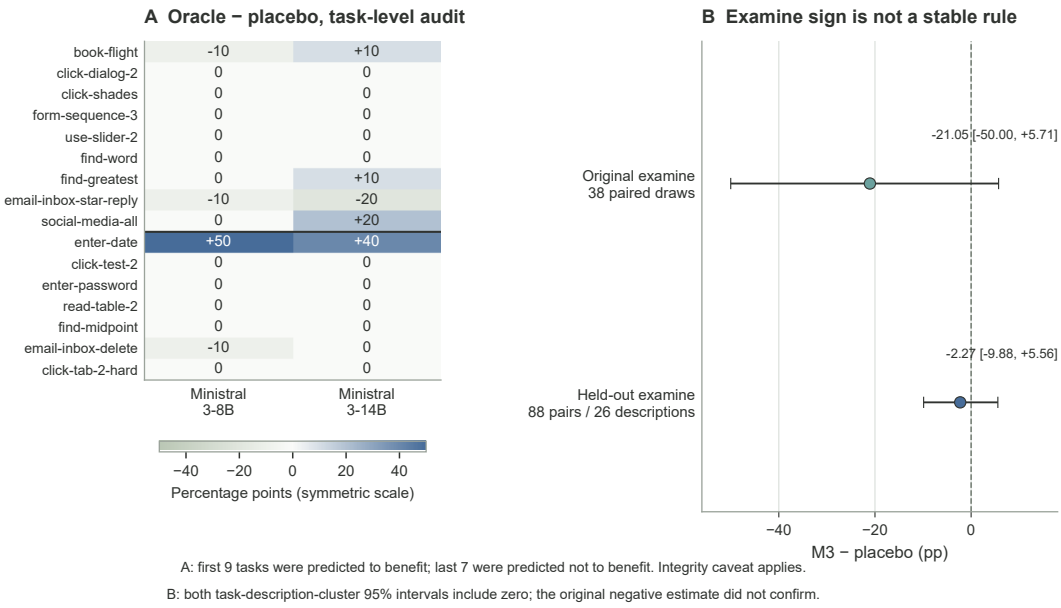

Figure 9: Task-level prediction audit and independent examine confirmation. Oracle-minus-Matched-control effects are shown for the 16-task prediction set; the held-out examine estimate remains compatible with zero.

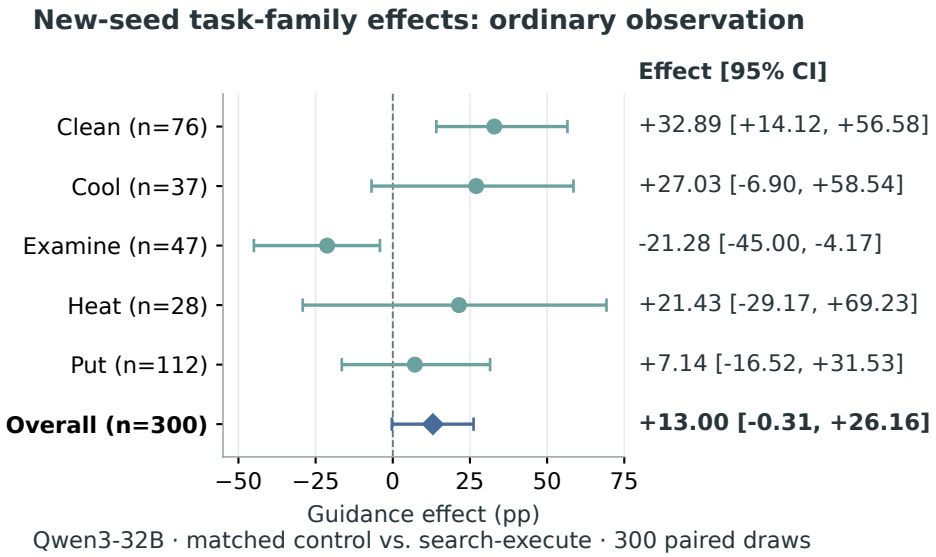

Figure 10: New-seed task-family effects under ordinary observation. Points show search-execute minus matched-control success; bars are task-description-cluster 95% CIs. The overall diamond aggregates the five displayed families.

Table 6: Information-state effects and interactions.

Condition	Matched control	Search–execute	Search–execute−Matched control
Qwen3-32B ordinary	122/300	161/300	+13.00pp
Qwen3-32B irrelevant fact	104/300	171/300	+22.33pp
Qwen3-32B target location	242/300	204/300	-12.67pp
Qwen3-32B high deficit	111/300	167/300	+18.67pp
Qwen3-32B low deficit	234/300	204/300	-10.00pp
Mistral Small 24B ordinary	38/300	83/300	+15.00pp
Mistral Small 24B location	164/300	145/300	-6.33pp
Gemma 3 27B ordinary	39/300	63/300	+8.00pp
Gemma 3 27B location	131/300	105/300	-8.67pp

Deleting only the fixed search-order instruction changes the interaction from $+28.67$pp to -2.33pp, an attenuation of -31.0pp (CI $[-46.58, -15.04]$pp). This is consistent with a directive-induced search policy becoming redundant or harmful when the target location is already known; it is a mechanism diagnostic, not a complete causal decomposition.

The original Qwen3-32B 2×2 information-deficit factorial gives high-deficit Search–execute-minus-Matched control $+18.67$pp and low-deficit -10.00pp, for an interaction of $+28.67$pp (task-cluster CI $[+11.64, +45.57]$pp). The P1 examine diagnostic finds 38 paired examine draws: Search–execute delays the first examine by 1.818 steps and adds 1.842 look actions (CI $[+0.364, +3.390]$), with no increase in manipulation actions. The diagnostic is descriptive and does not establish mediation.

B.2 Preregistered V4 Follow-Ups

The new-seed Qwen3-32B 2×3 test contains 1,800 episodes. Under ordinary observation, Search–execute-minus-Matched control is $+13.00$pp; with an irrelevant fact it is $+22.33$pp; with the true target location it is -12.67pp. The primary location-minus-irrelevant interaction is -35.00pp (CI $[-50.00, -20.00]$pp).

The Mistral Small 3.2 24B replication contains 1,200 episodes. Search–execute-minus-Matched control is $+15.00$pp under ordinary observation and -6.33pp with target location; the interaction is $+21.33$pp (CI $[+9.23, +33.54]$pp). The examine-aligned three-arm validation contains 300 episodes: Matched control 61%, original Search–execute 60%, aligned rewrite 61%; aligned-minus-original is $+1.00$pp (CI $[-5.00, +7.00]$pp). The rewrite changes examine behaviour but does not improve completion.

B.3 Gemma 3 27B Information-Deficit Replication

Gemma 3 27B provides an independent 2×2 transport check with 1,200 formal episodes. Under ordinary observation, the Matched control succeeds on 39/300 (13.00%) and Search–execute on 63/300 (21.00%), for $+8.00$pp. With the true target location, the Matched control succeeds on 131/300 (43.67%) and Search–execute on 105/300 (35.00%), for -8.67pp. The original-minus-location interaction is $+16.67$pp (task-description-cluster 95% CI $[+7.05, +27.44]$pp). Gemma changes model family, quantization, and inference backend together, so this is a transport check rather than an isolated model-only comparison.

B.4 Information-State Control Decomposition

The information controls preserve the Search–execute/Matched control contrast while changing what the agent already knows.

The location-minus-irrelevant interaction is -35.00pp (CI $[-50.00, -20.00]$pp); the high-versus-low deficit interaction is $+28.67$pp (CI $[+11.64, +45.57]$pp). Deleting the fixed search-order instruction changes the latter interaction to -2.33pp, an attenuation of -31.0pp (CI $[-46.58, -15.04]$pp). Reversing the order gives risk differences of -11.67pp under ordinary observation and -9.33pp with location supplied. A destination fact produces $+21.33$pp (CI $[+8.30, +34.25]$pp), showing that the harmful crossover is specific to target-location knowledge rather than truthful text in general. These

Table 7: Ordinary-to-location interaction across model configurations.

Model	Ordinary RD	Location RD	Interaction
Qwen3-32B-AWQ	+13.00pp	-12.67pp	+25.67pp
Mistral Small 3.2 24B	+15.00pp	-6.33pp	+21.33pp
Gemma 3 27B	+8.00pp	-8.67pp	+16.67pp
Mistral-7B BF16	+7.00pp	+13.67pp	-6.67pp

Table 8: Experimental lineage and primary endpoints.

Experiment	Model / benchmark	Episodes	Primary endpoint
Semantic-matched isolation	Qwen2.5-7B / ALFWorld	300/arm	Search–execute minus Matched control
Fresh-seed extension	Qwen2.5-7B / ALFWorld	300/arm	Independent-seed content contrast
Cross-family isolation	Llama-3.1-8B / ALFWorld	300/arm	Transport of content contrast
Information-deficit factorial	Qwen3-32B / ALFWorld	1,200	Search–execute-by-deficit interaction
Location-specificity test	Qwen3-32B / ALFWorld	1,800	Location-minus-irrelevant interaction
Mistral/Gemma replications	24B / 27B ALFWorld	1,200 each	Ordinary-minus-location interaction
Prediction audit	Ministral-3 / BrowserGym	960	Frozen task-level oracle rule
Activation tests	Qwen3-32B / ALFWorld	720+600	Information-aware policy contrast

controls identify alignment between induced policy and current information state as the operative interaction.

B.5 Cross-Model Transport Detail

Qwen3-32B, Mistral Small 24B, and Gemma 3 27B reproduce the qualitative reversal; Mistral-7B does not. The three larger configurations therefore establish a repeated information-state interaction, while the smaller-model result defines a clear model boundary.

C Strategy and Visual Evaluation

C.1 Strategy Evaluation: September 2026 Control Experiments

The frozen examine-off prospective rule reaches 43.10% success, compared with 44.83% for always-on Search–execute and 30.17% for always-off Matched control. It improves over always-off but does not improve over always-on and does not validate simple examine-based gating.

The 720-episode information-aware activation study compares information-only, joint task-and-information, task-only, always-on, and always-off policies. Information-only and joint each reach 56.94%; the three fixed baselines each reach 54.44%. Joint-minus-always-on is $+2.50$pp (CI $[-0.28, +5.28]$pp), and joint-minus-always-off is $+2.50$pp (CI $[-0.83, +5.56]$pp). An independent 600-episode check gives information-aware 54.67% versus task-only 54.00% ($+0.67$pp; CI $[-2.33, +3.67]$pp). These are directionally encouraging but do not establish a stable deployment rule.

Experimental Lineage and Endpoint Map

Table 8 maps each retained experiment to its estimand and evidence role. The sequence moves from content isolation to information intervention, cross-model transport, and prospective selection.

Repeated task draws are reported as extensions or transport checks, not as independent confirmations. Each row has one declared primary endpoint; trajectory measures and task-family decompositions are secondary diagnostics.

C.2 Detailed Task, Selection, and Robustness Results

This section retains detailed results moved out of the main-text evidence sequence. Task-family estimates are descriptive starting points; the BrowserGym prediction audit and transfer extension are different task sets and must not be combined. The main figures reproduce retained reported statistics, not new bootstrap analyses.

Table 9: Task-family effects in the original 300-pair study.

		Success % (wins)			
Family	n	**Matched control**	**Search-execute**	**RD (pp)**	**95% CI**
		Qwen3-32B-AWQ / ALFWorld			
Overall	300	36.67 (110)	57.67 (173)	↑ **21.00**	[+9.47, +32.47]
clean	71	33.80 (24)	73.24 (52)	↑ 39.44	[+19.28, +60.92]
cool	44	18.18 (8)	43.18 (19)	↑ 25.00	[−2.33, +51.35]
examine	38	81.58 (31)	60.53 (23)	↓ 21.05	[−50.00, +5.71]
heat	33	21.21 (7)	39.39 (13)	↑ 18.18	[−10.00, +55.17]
put	114	35.09 (40)	57.90 (66)	↑ 22.81	[+4.35, +41.32]

n is the number of paired draws; parentheses contain successful episodes. CIs are task-description-cluster bootstrap intervals. The gray examine row has a negative point estimate but a zero-crossing CI; it is not a validated exclusion rule.

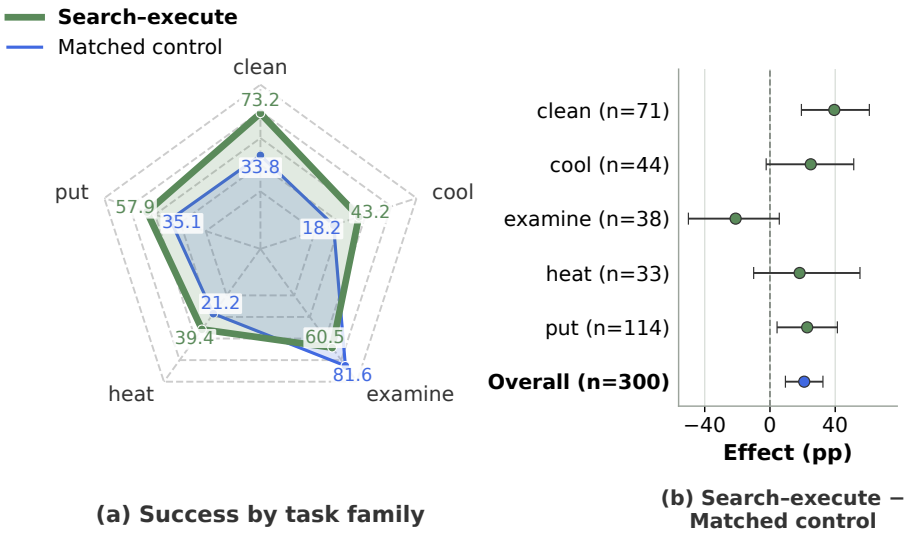

Figure 11: **Original task-family characterization.** The radar uses a common 0–100% success scale; the paired-effect panel retains reported 95% CIs. Examine is a negative point estimate with a zero-crossing CI, not a validated exclusion rule. The independent confirmation is shown in Figure 9.

Table 3 in the main text reports these estimates; the supplementary protocol details are retained here.

Table 10: Independent 24B transfer and model-specific format controls. These studies are not a common-sample algorithm ranking.

Model	Contrast	Effect (pp)	95% CI
A. Content transfer and information-interaction follow-up			
Mistral Small 3.2 24B	Task-specific	↑ 1.875	$[-2.500, +6.875]$
Mistral Small 3.2 24B	Oracle-relevant	↑ 5.625	$[+0.625, +11.250]$
Mistral-7B	Information interaction	-6.67	$[-21.12, +8.24]$
B. Secondary, token-length-confounded format evidence			
Qwen2.5-7B	JSON padding	↑ 10.56	$[+0.55, +20.12]$
Qwen2.5-7B	Prose padding	↑ 9.44	$[-0.93, +19.42]$
Qwen2.5-14B	JSON padding	↓ 0.56	$[-10.71, +9.62]$
Qwen2.5-14B	Prose padding	↓ 1.11	$[-11.27, +9.08]$
Five-run format ranking: Kendall's $W = 0.21$; 200 episodes, four task orderings.			

Protocols remain separate; zero-crossing intervals are not equivalence results.

Model-separated extension counts. In the independent 16-task extension (160 task–seed units per model and arm), Ministral-3-8B matched control, task-specific and oracle arms succeed on 89, 88 and 96 units; the corresponding 14B counts are 91, 94 and 97. Task-specific effects are -0.625pp and $+1.875$pp, and oracle effects are $+4.375$pp and $+3.750$pp. The separate 24B counts are 88, 91 and 97. These model-separated counts replace the unexplained "8B + 14B" pooled rows in the former main robustness figure; no model-specific paired CI is inferred from counts alone.

Counting and provenance limits. The retained raw audit verifies 640 BrowserGym validation trajectories per model (1,280 combined). Raw run/reuse accounting for the newer activation replication remains unavailable: its 328/600 and 324/600 outcomes do not resolve the source's "600-episode" label. The earlier 8,820-episode non-overlapping aggregate is therefore not asserted as verified. First-examine mean-step denominators are also unverified; the main figure uses per-trajectory look change and the all-episode fraction reaching examine instead. Figure sources, study identifiers, uncertainty and limitations are recorded in `figures/experiment_v2/source/evidence.json`.