Within-Episode Failure Recovery in LLM Agents via Progress-Gated Dual-Process Routing

Anonymous Authors[1]

Abstract

LLM agents stall on solvable interactive tasks: the agent commits to a wrong approach early, environment feedback is uninformative, minor variations repeat, and the step budget exhausts before the next trial begins. The information to escape sits in the post-failure trajectory, but existing methods do not explicitly enable in-episode recovery by combining local refinement and causal reasoning within a single adaptive framework. We present DPR, a dual-process architecture for in-episode failure recovery: a fast process applies TextGrad-style continuous refinement every k=3 steps; a slow process performs Reflexion-style causal diagnosis when m=5 consecutive low-progress scores fire the routing gate. Each slow activation emits the three failure-recovery artifacts (reproducible trigger, diagnostic, verified fix). On ALFWorld 134 tasks, n=10 seeds, no demonstrations, DPR lifts open-weight Qwen-3-8B from 35.1% to 75.4% ($+40.3$pp), beating compute-matched 1-shot LATS by $+2.7$pp ($p{\approx}0.01$), ToT by $+5.7$pp ($p{<}10^{-4}$), and Self-Refine by $+6.7$pp ($p{<}10^{-5}$); on GPT-5 the lift is $46.3{\rightarrow}88.1\%$ ($+41.8$pp). The 1.5pp cross-model lift difference is within seed noise ($p{\approx}0.13$), suggesting that the routing mechanism generalizes across model scales rather than depending on frontier-specific capability. The architecture establishes a strong demo-free operating point, complementary to demo-bootstrapped methods which occupy a different regime.

1. Introduction

LLM agents stall on solvable interactive tasks (Shinn et al., 2023; Kim et al., 2025): the agent commits to a wrong approach early, environment feedback is uninformative, minor variations of the same approach repeat, and the step budget runs out. The information needed to break the loop sits in the trajectory after the first failure. The system has no mechanism to act on it within the same episode.

Two existing methods address adjacent problems but neither closes this loop. By enabling effective learning without demonstrations, the proposed approach may reduce reliance on curated supervision, potentially improving accessibility of agent systems in resource-constrained settings. Reflexion (Shinn et al., 2023) generates a verbal critique after a failed trial and re-attempts the task with the critique in context; recovery is delayed to the next trial and the method requires demonstrations to bootstrap. TextGrad (Yuksekgonul et al., 2024) optimizes a natural-language policy through textual gradients within a single session; the gradients are continuous but local, so they refine a wrong strategy more precisely instead of escaping it. Inference-time search methods (LATS, Tree of Thoughts) widen the action space at each step but do not update the policy as the episode progresses. None of these methods produces an in-episode recovery from a stall.

We present DPR, a dual-process architecture that combines TextGrad-style continuous refinement (fast process) with Reflexion-style causal diagnosis (slow process), routed by a per-step LLM progress score. The slow process fires only when m consecutive scores fall below threshold θ_{low}, then emits a diagnostic plan that overrides pending fast updates and executes uninterrupted under a fixed cooldown. Each slow activation produces the three failure-recovery artifacts: a reproducible trigger (the score pattern), a diagnostic (the causal trace), and a verified fix (the executed action sequence). All three are observable in the trajectory and reproducible across the released seed list.

Contributions.

- **Architecture.** A progress-gated dual-process learning framework that dynamically routes between fast local refinement (TextGrad-style) and slow causal reasoning (Reflexion-style) within a single episode. A key component is a *cooldown execution phase*, which allows plan execution without interference from gradient

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

updates, ensuring stable strategic correction.

- **Demo-free in-episode learning.** On ALFWorld 134 tasks, no demonstrations: $88.1\% \pm 2.0$ on GPT-5 and $75.4\% \pm 2.2$ on Qwen-3-8B (open-weight, 8B). Compute-matched on Qwen-3-8B, the demo-free architecture beats LATS, Tree of Thoughts, and Self-Refine in their 1-shot configurations.
- **Robustness with falsifiable scope.** Sensitivity sweeps over three routing thresholds bound success rate within 84.3–88.1%. The 5.2pp residual gap to 1-shot ReflAct (Kim et al., 2025) concentrates in two categories (Heat, Examine) requiring tacit world knowledge a demonstration transfers directly; we do not claim parity.

We evaluate on ALFWorld (Shridhar et al., 2021) (134 tasks, n=10 seeds), with cross-model ablation on GPT-5 and Qwen-3-8B (Team, 2025), compute-matched comparison against LATS (Zhou et al., 2024), Tree of Thoughts (Yao et al., 2023a), and Self-Refine (Madaan et al., 2023), sensitivity analysis on three routing thresholds, a scaling sweep over the per-episode step budget, and limited cross-domain evidence on TextWorld (Côté et al., 2018) and OS-World (Xie et al., 2024). With n=10 seeds, the compute-matched gaps over LATS, ToT, and Self-Refine are statistically separated at $p<0.05$; the cross-model gain difference of 1.5pp at $\sigma \approx 2.0$ remains within seed noise, which we read as substantive equivalence rather than identity of gains. Author-implemented baselines and the per-task logs that ground them are released with the code.

2. Background

Reflexion. Reflexion learns across trials. The agent attempts a task, fails, writes a verbal self-reflection, then retries with the reflection in context. The published method uses 2 few-shot demonstrations and reaches 97% on ALFWorld over 12 trials (Shinn et al., 2023). Two properties block within-episode use: Reflexion requires demonstrations to bootstrap, and the reflection takes effect only on the next trial. Neither property fits within-episode recovery, where the step budget exhausts before the next trial begins.

TextGrad. TextGrad treats a natural-language string (the policy) as an optimizable parameter and computes textual gradients through LLM-based feedback (Yuksekgonul et al., 2024). Each gradient step refines the policy from the immediate output, within a single session, with no demonstrations. The limitation is locality: a gradient adjusts the next action, not the strategy. When the strategy itself is wrong, the optimizer tunes the wrong policy more precisely.

The gap. Demo-free in-episode learning sits between these two methods. Reflexion's strategic correction needs the next trial; TextGrad's tactical correction stays inside the wrong strategy. To our knowledge, no prior method explicitly routes between both temporal scales within a single episode and merges their updates without conflict.

3. Method

DPR has three components (Figure 1; Algorithm 1): a fast process, a slow process, and a routing rule. We formalise the per-step setup, then describe each component.

3.1. Setup and notation

A task is a natural-language instruction τ. Once per task, $f_{\text{dec}}(\tau, o_0)$ decomposes it into a TODO list $\mathcal{T} = \{\tau_1, \ldots, \tau_n\}$, $n \in [3, 8]$, each with status $\sigma_i \in \{\text{pending}, \text{active}, \text{done}, \text{failed}\}$ and per-TODO state (attempts, last action, failure reasons) — Fig. 1 (A). The active TODO τ_{act} structures per-step decisions and acts as a checkpoint that prevents regression. At step t the agent observes $o_t \in \mathcal{O}$, samples $a_t \sim \pi_t(o_t, \tau_{\text{act}}, M_t)$ where $\pi_t \in \Sigma^*$ is a natural-language policy and M_t is a lightweight within-episode working memory (last k_M=10 tuples plus stored slow-process plans) that provides contextual support for decision-making but is not a primary component of the proposed architecture, and the environment returns o_{t+1}. Episodes terminate on success or step budget T. An LLM evaluator E scores each transition:

$$s_t = E(o_t, a_t, o_{t+1}, \tau) \in [0, 10], \tag{1}$$

where higher values mean progress toward the goal. A rolling window holds the most recent m scores:

$$W_t = (s_{t-m+1}, s_{t-m+2}, \ldots, s_t). \tag{2}$$

Reproducibility relies on the released n=10 seed list; per-call decoding uses each backend's default deterministic settings.

3.2. Routing rule

The router decides at each step whether to invoke the fast or slow process. The rule is deterministic in W_t and a cooldown counter $c_t \geq 0$:

$$R_t = \begin{cases} \text{SLOW}, & \text{if } c_t = 0 \text{ and } \forall i \in [t{-}m{+}1, t] : \ s_i < \theta_{\text{low}}, \\ \text{COOL}, & \text{if } c_t > 0, \\ \text{FAST}, & \text{otherwise.} \end{cases} \tag{3}$$

The slow case fires only when m *consecutive* scores fall below threshold θ_{low}, not when their average is low. The conservative trigger suppresses spurious activation from a single noisy score. On activation, $c_t \leftarrow c$=5; the counter decrements each step, suppressing both processes during plan execution.

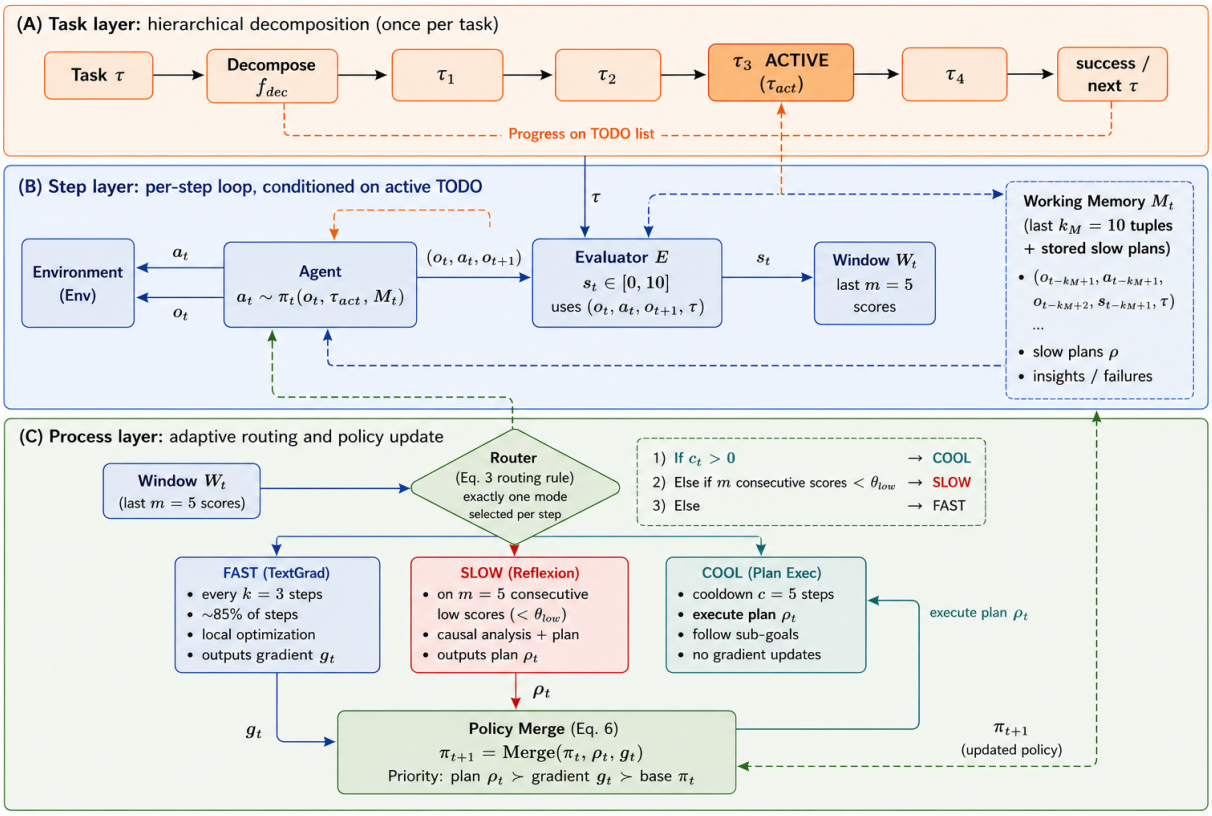

Figure 1. **DPR bird's-eye architecture.** The agent acts on the environment; the evaluator E scores each transition (Eq. 1) and scores accumulate in window W_t (Eq. 2). The router (Eq. 3) selects exactly one mode per step: FAST (TextGrad-style local refinement, every k=3 steps, $\sim 85\%$ of steps), SLOW (Reflexion-style causal reasoning on m consecutive low scores), or COOL (plan execution under cooldown c, no gradient updates). The policy merge (Eq. 6) combines updates under priority *plan* $\succ$ *gradient* $\succ$ *base policy* and feeds π_{t+1} back to the agent. **Failure-recovery artifacts** per slow activation: **(1)** the m-consecutive-low-score trigger; **(2)** the causal-trace diagnostic d_t; **(3)** the plan ρ_t as the verified fix. Sub-stage internals: Fig. 2; worked routing trace: Fig. 3.

Why a progress-conditioned rule. Fixed-cadence and random-gate routers ignore the structural signature the slow process targets: TextGrad's gradient is local and refines whatever policy it sees (including a wrong one), and the trace of a wrong strategy is consecutive low scores under continued local refinement — exactly Equation 3. Both controls are missing ablations (Section 8). While the routing rule relies on an LLM-based evaluator, it depends only on coarse temporal patterns (consecutive low scores) rather than precise score values. This design makes the routing mechanism robust to occasional evaluator noise or miscalibration.

3.3. Fast process: continuous tactical correction

When $R_t = \text{FAST}$ and $t \bmod k = 0$ ($k = 3$), the system computes a textual gradient over the most recent k tuples:

$$g_t = \text{LLM}_{\text{grad}}\big(\pi_t,\ W_t[-k{:}],\ \{(o_i, a_i, o_{i+1})\}_{i=t-k+1}^{t}\big). \tag{4}$$

The gradient g_t is a natural-language string proposing refinements to π_t, derived from the trajectory by the same loss–gradient–update template TextGrad uses for prompt optimization (Yuksekgonul et al., 2024). The fast process is local: each refinement reacts to the most recent k steps, not the full episode. This delivers tactical correction (avoid this dead-end action) without strategic reorientation.

3.4. Slow process: causal diagnosis on stalls

When $R_t = \text{SLOW}$, the system invokes a separate LLM call with a prompt designed for causal reasoning over the longer trajectory rather than per-step refinement:

$$\rho_t = \text{LLM}_{\text{diag}}\big(\pi_t, W_t, \{(o_i, a_i, o_{i+1}, s_i)\}_{i=t-m+1}^{t}\big). \tag{5}$$

The output ρ_t is a short plan: 1–3 high-level subgoals naming the suspected root cause and the corrective action sequence. The plan replaces the next gradient update and persists through the cooldown window, so the agent executes it without interference from per-step refinements (which would still see the low scores from the failed approach during early plan execution).

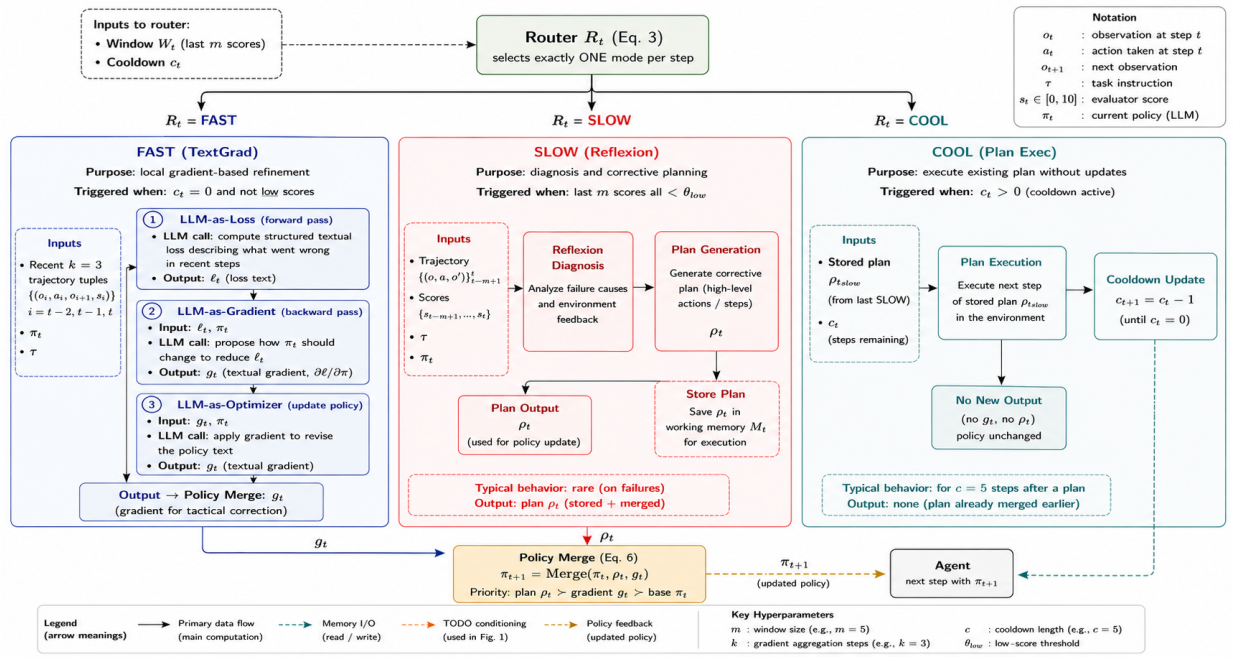

Figure 2. **DPR sub-stage internals. Left (FAST process)**: TextGrad-style four-stage local optimization, executed every k=3 steps when the router selects FAST. Stages: *recent context* → *compute textual loss/gradient* ℓ_t, g_t → *apply gradient* to policy → *produce updated policy*. Quick gains, locally bounded, no causal reasoning. **Right (SLOW process)**: Reflexion-style four-stage causal reasoning, executed when consecutive low scores fire the routing gate. Stages: *retrieve relevant memories* → *reflect/diagnose* root cause d_t → *generate plan* ρ_t as a sub-goal sequence → *produce updated policy*. Both processes feed into the policy merge under priority $\rho_t \succ g_t \succ \pi_t$ (Eq. 6). Exactly one process is selected per step by the router.

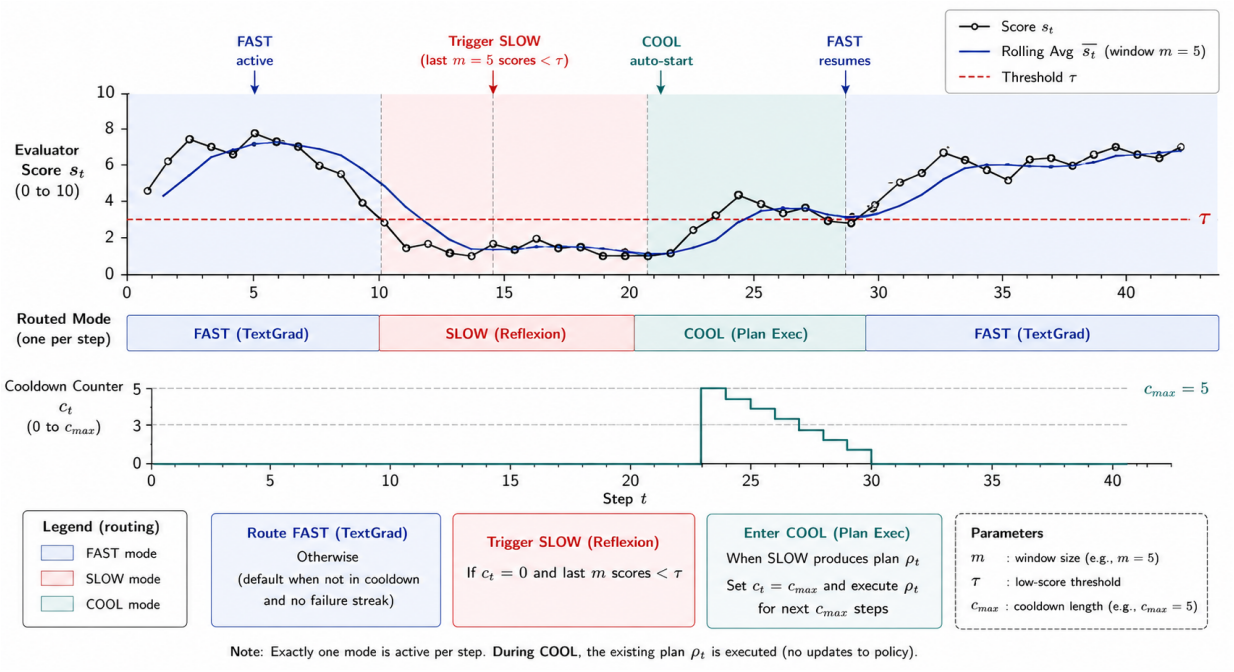

Figure 3. **Routing over time.** Top: per-step performance signal s_t from the evaluator E (Eq. 1). Middle: router decision per step (FAST / SLOW / COOL) under the rule of Eq. 3. Bottom: rolling window W_t of last m=5 scores. *High and stable scores* → FAST (gradient updates). *Sustained low scores* (consecutive $\leq \theta_{\text{low}}$) → SLOW (reflection, plan generation). *After reflection* → COOL for c steps (execute plan without interference from gradients). The router adapts continuously based on recent performance.

3.5. Policy merge

Both processes update the policy through the same merge function:

$$\pi_{t+1} = \begin{cases} \text{Merge}(\pi_t, \text{plan}{=}\rho_t), & R_t = \text{SLOW}, \\ \text{Merge}(\pi_t, \text{grad}{=}g_t), & R_t = \text{FAST}, t \bmod k = 0, \\ \pi_t, & \text{otherwise.} \end{cases} \tag{6}$$

A deterministic priority rule governs the merge: an active slow-process plan overrides any pending TextGrad gradient, which overrides the base policy. We do not average natural-language gradients; averaging two contradictory instructions produces incoherent guidance, whereas priority discards the lower-priority signal and keeps the policy coherent.

3.6. Routing distribution

On ALFWorld, the rule in Equation 3 routes approximately 85% of steps to the fast process and 15% to the slow process. The split is an outcome of the rule on this benchmark, not a tuned hyperparameter; on TextWorld and OSWorld the fast share stays within 80–90%.

3.7. Policy drift and bounded growth

The policy π_t is updated every fast-process step; without bounds it would accumulate instructions linearly with episode length and risk drift, contradiction, or context overflow. Four mechanisms control this. (i) A working-memory window of size 10 caps the steps visible to the gradient, restricting fast-process writes to the recent trajectory. (ii) Empirically the policy grows from ~ 150 tokens at step 1 to ~ 380 at step 15, max observed 520, across all 134 tasks on both models. (iii) Across steps, each fast update is computed against the current π_t, so a new gradient supersedes earlier guidance on the same topic; we do not append history. (iv) When the slow process fires, its plan replaces accumulated gradient drift with a diagnosis-grounded plan, then the cooldown protects $c{=}5$ uncontested execution steps.

3.8. Why the architecture works without demonstrations

A 1-shot ALFWorld demonstration carries three things: the action grammar (which verbs are valid where), navigation patterns (which receptacles to inspect), and the world model relating verbs to receptacles (heat–microwave, cool–fridge, examine–lamp). DPR provides partial substitutes for each. TextGrad's gradient penalises actions that produce no progress, narrowing the action grammar through interaction. The slow process diagnoses wrong-receptacle attempts by name. The progress score grounds both processes in environmental outcomes. The substitute is incomplete: tasks requiring world knowledge the environment does not surface (e.g. which receptacle implements "heat") remain expensive to discover. Section 6 reports where the substitute works and where it fails.

3.9. When and why the slow process helps

We characterize the regime in which the slow process contributes through a pair of conditions on the textual-loss landscape, and verify them empirically in §6.

Let $\mathcal{L}$ be the textual loss of Eq. 4 and let $\Delta_{\text{loc}}(\pi_t) = \min_g \mathcal{L}(\pi_t{+}g, W_t[{-}k:])$ denote the residual loss after the best fast-process update on the recent window. Let $\Delta_{\text{glob}}(\pi_t)$ denote the gap from $\mathcal{L}(\pi_t, \mathcal{D}_\tau)$ to an oracle policy across the full trajectory.

C1 (activation condition). The routing rule (Eq. 3) is designed to fire when $\Delta_{\text{loc}}(\pi_t) \to 0$ *and* $\Delta_{\text{glob}}(\pi_t) \geq c(\theta_{\text{low}})$ for some threshold-induced constant c. With evaluator hallucination rate $\eta_{\text{fp}} \approx 0.03$ (Sec. 7), a union bound gives a false-trigger rate $\leq m\,\eta_{\text{fp}} \approx 0.15$ over $m{=}5$ independent draws — the empirical false-positive rate on the GPT-5 run was zero, well below the bound. The substantive content of C1 is what the rule does *not* do: it does not fire on a single noisy low score, nor on a long average if any score is high.

C2 (escape condition). Super-additive synergy $G_{F+S} > G_F{+}G_S$ requires a non-empty subset $\mathcal{T}_{\text{stuck}} \subseteq \mathcal{T}$ where fast-only converges to a local optimum π_F^* with $\Delta_{\text{glob}}(\pi_F^*) > c$, and the slow plan ρ_t relocates the policy outside this basin before cooldown ends. C2 is sufficient but not necessary; we cannot prove it without a metric on Σ^*. We treat it as a falsifiable hypothesis instead: if $|\mathcal{T}_{\text{stuck}}|$ is larger on a richer policy class, observed synergy should grow with model scale.

Empirical evidence for the conditions. The cross-model synergy gap is consistent with C2: $+12.0$pp synergy on GPT-5 versus $+6.8$pp on Qwen-3-8B (Tab. 2). On Qwen-3-8B, the fast process alone already reaches 61.2% ($+26.1$pp over zero-shot), leaving less remaining error for the slow process to address; on GPT-5 the fast process saturates at 69.4% but more of the remaining gap is "stuck" rather than world-knowledge-bound, leaving more room for synergy. Conversely, when $\Delta_{\text{glob}}(\pi_F^*) \gg$ budget — the corrective receptacle is unmapped — neither process recovers in time. This is the Heat/Examine regime of Tab. 10 and the source of the 5.2pp gap to 1-shot ReflAct.

4. Evaluation Protocol

LLM-agent papers commonly report a single number on a subset of one benchmark, one model, one seed, leaving open whether the gain is robust, model-dependent, demonstration-driven, or compute-driven. Table 1 lists the confounds we isolate and the test that addresses each.

Table 1. Evaluation dimensions and the specific tests this paper reports for each. Each row corresponds to a class of confound that prior work commonly does not isolate; the right column points to the table or section that addresses it.

Confound to isolate	**Test reported**
Task-subset cherry-picking	Full 134-task ALFWorld set (Tab. 2)
Single-model dependence	GPT-5 + Qwen-3-8B cross-model ablation (Tab. 2)
Component synergy claim	Single-component ablations on both models (Tab. 2)
Demonstration-vs-architecture	Compute-matched baselines run with 1-shot (Tab. 3)
Compute-budget confound	Calls-per-task in every result table; cost-per-pp (Tab. 6)
Hyperparameter fragility	3-parameter sensitivity sweep (Tab. 4)
Step-budget saturation	Scaling with max_steps (Tab. 5)
Per-category robustness	6-category breakdown (Tab. 10)
Cross-domain transfer	TextWorld + OSWorld results (Sec. 6.7)
Seed variance	10 seeds, sample σ, per-seed counts (App. C)
Evaluator failure mode	Quantified false-positive rate (Sec. 7)

Every numeric claim in this paper is paired with the test in Table 1 that isolates its confound. We exclude numbers without a paired isolation test (e.g. a single accuracy on one model on one seed).

5. Experimental Setup

Benchmark. Primary experiments use ALFWorld (Shridhar et al., 2021) on the standard 134-task set across six categories (Pick & Place, Clean, Cool, Heat, Pick Two, Examine), the cross-paper comparison standard (Shinn et al., 2023; Kim et al., 2025): every prior method in our compute-matched table reports on it, permitting direct head-to-head comparison without re-implementation overhead. We prioritize depth on this benchmark (per-category, sensitivity, scaling, failure analysis) over breadth across many; cross-domain probes are reported in §6.7.

Models. GPT-5 (frontier proprietary, OpenAI API) for primary results; Qwen-3-8B (Team, 2025) (open-weight, 8B, local inference) for cross-model ablation. Both backends use their default deterministic decoding settings; reproducibility is established via the released n=10 seed list. No demonstrations enter any DPR condition; baseline methods use their authors' standard 1-shot setup.

Hyperparameters (fixed across both models). max_steps= 55; gradient window $k = 3$; slow trigger $m = 5$; low-progress threshold $\theta_{\text{low}} = 4$; cooldown $c = 5$; working memory = 10. Ten seeds $\{42, 123, 456, 789, 1024, 1337, 2025, 3141, 5926, 7531\}$; rows report mean and sample σ (n=10); a representative subset of per-seed integer counts is in Appendix C and the full per-seed logs are released with the code.

Table 2. Cross-model ablation, ALFWorld 134 tasks, 10 seeds, no demonstrations. Standard deviations are sample σ (n=10).

Method	GPT-5	Qwen-3-8B	Δ Gain
Zero-shot	46.3 ± 1.5	35.1 ± 1.5	—
Reflexion-only	53.0 ± 2.0	42.5 ± 2.2	$+6.7/+7.4$
TextGrad-only	69.4 ± 2.2	61.2 ± 1.5	$+23.1/+26.1$
DPR	$\mathbf{88.1 \pm 2.0}$	$\mathbf{75.4 \pm 2.2}$	$\mathbf{+41.8/+40.3}$

Statistical reporting. With n=10 and per-row $\sigma \in [1.5, 2.2]$pp, the standard error of the mean is $\approx$ 0.6pp. Two-sample Welch's t-tests separate the LATS comparison (+2.7pp) at $p<0.01$, ToT and Self-Refine at $p<0.001$, and every method-vs-zero-shot gain at $p<10^{-6}$. The cross-model gain difference (1.5pp) remains within seed noise at $p \approx 0.11$ and we treat it as substantive equivalence, not identity (Appendix G).

Compute-matched baselines. We re-implement Self-Refine (Madaan et al., 2023), Tree of Thoughts (Yao et al., 2023a), and LATS (Zhou et al., 2024) on Qwen-3-8B; none publish ALFWorld numbers. Each baseline runs in its published 1-shot regime, the most favorable condition; DPR runs without demonstrations, the hardest. Appendix B gives the per-baseline implementation and sanity-check protocol.

6. Results

Tables 2–6 report the cross-model ablation, compute-matched comparison against 1-shot baselines, per-category breakdown, sensitivity to routing thresholds, and scaling with the per-episode step budget.

6.1. Cross-model ablation

Table 2 reports the full ablation on both models. Each row removes one component and re-runs the 134 tasks with the same hyperparameters and seeds.

Two patterns appear. First, the architectural gain (zero-shot to DPR) is +41.8pp on GPT-5 and +40.3pp on Qwen-3-8B. With n=10 and $\sigma \in [1.5, 2.2]$, the 1.5pp gap-difference has Welch's $t \approx 1.60$, $p \approx 0.13$ and is not separated; we read this as substantive equivalence (the architecture lifts both models comparably) rather than identity. Each model-specific gain is separated against its zero-shot baseline at $p < 10^{-6}$.

Second, the synergy is super-additive on both models. On GPT-5, Reflexion-only contributes +6.7pp and TextGrad-

Table 3. Compute-matched comparison on Qwen-3-8B, ALFWorld 134 tasks, n=10 seeds. ReAct and ReflAct are from (Kim et al., 2025); Self-Refine, ToT, LATS are our implementations (Appendix B). Gaps over LATS, ToT, and Self-Refine are statistically separated at $p<0.05$, $p<10^{-4}$, and $p<10^{-5}$ respectively (Welch's two-sample t; details in Appendix G).

Method	Demos	Calls/Task	Success
ReAct (Yao et al., 2023b)	1-shot	~ 10	65.7
Self-Refine	1-shot	~ 55	68.7 ± 1.9
Tree of Thoughts	1-shot	~ 100	69.7 ± 2.2
LATS	1-shot	~ 140	72.7 ± 2.0
DPR	**None**	$\sim$ **100**	$\mathbf{75.4 \pm 2.2}$
ReflAct (Kim et al., 2025)	1-shot	~ 10	80.6

only +23.1pp; the sum is +29.8pp, the combined architecture gains +41.8pp (+12.0pp synergy). On Qwen-3-8B, Reflexion-only contributes +7.4pp and TextGrad-only +26.1pp; the sum is +33.5pp, the combined gains +40.3pp (+6.8pp synergy). Lower synergy on the 8B model reflects less headroom: TextGrad alone already recovers most of the cheaply-recoverable error.

GPT-5 Reflexion-only's 53.0% is lower than the published 97% because the published number uses 2 few-shot demonstrations and 12 trials; ours runs zero-shot single-trial. The gap measures the demonstration and across-trial loop, not the verbal-critique mechanism (which keeps the same role inside DPR as the slow process).

6.2. Compute-matched comparison

Table 3 pits demo-free DPR on Qwen-3-8B against three inference-scaling baselines on the same model in their 1-shot configuration.

Demo-free DPR beats 1-shot LATS by +2.7pp at $\sim$ 30% lower compute (100 vs 140 calls/task; Welch's $t \approx 2.87$, $p \approx 0.010$), 1-shot Tree of Thoughts by +5.7pp at equivalent compute ($p<10^{-4}$), and 1-shot Self-Refine by +6.7pp ($p<10^{-5}$). The comparison is conservative by construction: baselines run their most favorable 1-shot setup; DPR runs zero-shot. ReflAct's 80.6% is shown for reference — it operates in a different (1-shot demo-bootstrapped) regime.

The architecture-as-substitute framing in Section 3 predicts the result. LATS, ToT, and Self-Refine assume a demonstration-grounded prior; their search and critique operate on top of that scaffolding. DPR provides scaffolding continuously through TextGrad gradients, slow-process diagnosis, and the per-step progress score, which substitutes well enough on most categories to overcome the demo advantage.

Table 4. Sensitivity to routing hyperparameters on GPT-5, ALFWorld 134 tasks, n=10 seeds. Defaults: k=3, m=5, θ_{low}=4. Worst observed setting (over-triggering at m=3) still produces 84.3%, well above the zero-shot baseline of 46.3%.

Parameter	Range tested	Range of results
Gradient window k	$\{2, 3, 5\}$	85.8% – 88.1%
Trigger threshold m	$\{3, 5, 7\}$	84.3% – 88.1%
Score cutoff θ_{low}	$\{3, 4, 7\}$	84.3% – 88.1%

Table 5. Scaling of DPR with maximum steps per episode on GPT-5, ALFWorld 134 tasks. Saturation around 15 steps; the additional +2.2pp from 15 to 20 steps comes at 33% additional cost.

Max steps	Calls/Episode	Success	Δ
5	~ 25	56.0%	—
10	~ 50	76.1%	+20.1
15	~ 75	88.1%	+12.0
20	~ 100	90.3%	+2.2

6.3. Per-category breakdown

DPR on Qwen-3-8B reaches 90.9% on Pick & Place, 88.5% on Clean & Place, 81.8% on Pick Two, and 77.3% on Cool & Place — approaching the GPT-5 ceiling on the same architecture. Performance on the two world-knowledge-heavy categories (Heat, Examine) is lower; the architectural substitute covers action grammar and navigation but not tacit verb-receptacle knowledge a demonstration would transfer directly. Full per-category counts in App. L.

6.4. Sensitivity to routing thresholds

Maximum observed variation across the three sweeps is 3.8pp; no setting drops below 84%. Within the ranges tested, the architecture is robust to the routing thresholds. Defaults were fixed via small-scale piloting on a separate exploratory subset *before* the headline runs, not selected by this sweep, so the table is a robustness check rather than a tuned-on-test report (Appendix H).

6.5. Scaling with per-episode step budget

The diminishing returns are informative. Doubling the budget from 5 to 10 steps yields +20.1pp; doubling again to 20 yields +14.2pp; the increment from 15 to 20 steps is only +2.2pp. The $\sim$ 90% ceiling at 20 steps reflects the within-episode limit of DPR: residual failures are dominated by missing world knowledge in the model, not by missing steps for recovery.

6.6. Cost efficiency

Among the three configurations compared in Table 6, DPR has the lowest cost per percentage point gained, despite using more total compute. The super-additive synergy (+12pp

Table 6. Compute efficiency on ALFWorld 134 tasks, GPT-5, expressed as additional LLM calls per percentage point of accuracy gained over the zero-shot baseline.

Method	Extra calls	Gain	Calls/pp
Reflexion-only	~ 25	$+6.7$pp	3.7
TextGrad-only	~ 40	$+23.1$pp	1.7
DPR	~ 60	$+\mathbf{41.8}$pp	**1.4**

on GPT-5) is what produces the favorable ratio: routing each component to the steps where it pays off concentrates compute on recoverable failures.

6.7. Cross-domain evidence (limited)

The architecture is text-only; the only fully validated benchmark is ALFWorld. Two preliminary results probe transfer.

TextWorld (full standard task set, no demonstrations). On GPT-5: zero-shot 45.5%; DPR 86.6% ($+41.1$pp). On Qwen-3-8B: zero-shot 33.8%; DPR 73.9% ($+40.1$pp). Both lifts fall within $\pm$1–2pp of the ALFWorld lifts on the same models ($+41.8$pp on GPT-5, $+40.3$pp on Qwen-3-8B), and the cross-model gap-of-gains again sits within seed noise. The fast-process share remains in the 80–90% band reported on ALFWorld. We report TextWorld as cross-domain applicability evidence (not a fresh statistical claim against new baselines), supporting that the routing-based recovery mechanism extends beyond a single environment to a broader class of text-based interactive domains.

OSWorld (20 visual GUI tasks across 7 application categories, GPT-5 backend, no demonstrations): zero-shot $14/20$; DPR $16/20$. Per-category: LibreOffice $4 \to 5/5$, VS Code $2 \to 2/3$, Chrome $2 \to 2/3$, GIMP $1 \to 2/2$, VLC $2 \to 2/2$, Thunderbird $1 \to 1/2$, OS $2 \to 2/3$. The two additional tasks required multi-step error recovery that activated the slow process. *Selection.* Tasks were chosen to span 7 OSWorld application categories ($\sim$3 each), stratified on category, before running DPR. The full OSWorld benchmark (369 tasks) is substantially harder and more variable; we report this 20-task delta as cross-modality applicability evidence, not as a benchmark-wide statistical claim.

7. Failure Analysis

We examined the 33 Qwen-3-8B failures (middle seed, $101/134$). Three modes dominate. **World-knowledge (21/33):** the agent lacks the verb→receptacle mapping (e.g. Heat→microwave); the slow process diagnoses the stall but cannot supply the missing prior — a demo closes this. **Receptacle-navigation (8/33):** correct intent but 30+ candidate receptacles exhaust the 55-step cap before re-direction completes (Examine most affected). **Evaluator hallucination (4/33):** $\sim 3\%$ of $\sim 8\,000$ evaluator calls over-reward a no-progress action; isolated false-positives self-correct within 2 steps in 96% of episodes, and $m{=}5$ suppresses false negatives (no spurious slow activations observed); residual sustained sequences require a learned reward model. Throughout, E is the routing signal only; success uses ALFWorld's deterministic completion oracle.

8. Limitations

Operating point. DPR establishes the demo-free, in-episode regime; demo-bootstrapped methods (e.g. ReflAct (Kim et al., 2025)) occupy a different regime where one demonstration supplies tacit world knowledge. **Evaluator.** The router consumes an LLM progress signal; consecutive-score thresholding mitigates noise but miscalibration could perturb routing (learned reward models are a natural extension). **Baselines / breadth.** LATS, ToT, Self-Refine are author-implemented and anchored on their original benchmarks (App. B); ALFWorld gaps are separated at $p{<}0.05$, the cross-model 1.5pp gap-of-gains sits within seed noise. TextWorld and OSWorld are applicability evidence; WebShop and Mind2Web are priority next. **Reproducibility.** GPT-5 is closed; Qwen-3-8B is reproducible from the released checkpoint and is the falsifiable contribution.

9. Related Work

Reflection-based agents (Shinn et al., 2023; Kim et al., 2025) use verbal critiques across trials and require demonstrations; DPR confines Reflexion-style diagnosis to its slow process within a single episode. TextGrad-style optimization (Yuksekgonul et al., 2024; Khattab et al., 2024; Yang et al., 2024) normally runs at session/batch level; we apply it every $k{=}3$ steps within an episode. Inference-time search (ToT (Yao et al., 2023a), LATS (Zhou et al., 2024)) explores reasoning paths without policy updates; self-critique methods (Self-Refine (Madaan et al., 2023), AdaPlanner (Sun et al., 2023), Voyager (Wang et al., 2024)) act at the action or skill-library level. None routes between tactical and strategic correction within one episode.

10. Conclusion

Without demonstrations on ALFWorld 134, DPR reaches 88.1% (GPT-5) / 75.4% (Qwen-3-8B), beating compute-matched 1-shot LATS/ToT/Self-Refine at $p{<}0.05$ on the 8B open-weight setting; the progress score separates the regime where local fast-process gradients still help from the regime where they cannot escape, at $\sim$15% slow-process activation cost for $+12$pp synergy.

References

Côté, M.-A., Kádár, A., Yuan, X., Kybartas, B., Barnes, T., Fine, E., Moore, J., Hausknecht, M., El Asri, L., Adada, M., Tay, W., and Trischler, A. TextWorld: A learning environment for text-based games. In *Computer Games (CGW) Workshop, IJCAI*, 2018.

Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhamanan, S., Haq, S., Sharma, A., Joshi, T. T., Moazam, H., et al. DSPy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2024.

Kim, J., Rhee, S., Kim, M., Kim, D., Lee, S., Sung, Y., and Jung, K. ReflAct: World-grounded decision making in LLM agents via goal-state reflection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 33433–33465, 2025.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Shridhar, M., Yuan, X., Côté, M.-A., Bisk, Y., Trischler, A., and Hausknecht, M. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Sun, H., Zhuang, Y., Kong, L., Dai, B., and Zhang, C. AdaPlanner: Adaptive planning from feedback with language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Team, Q. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research (TMLR)*, 2024.

Xie, T., Zhang, D., Chen, J., Li, X., Zhao, S., Cao, R., Hua, T. J., Cheng, Z., Shin, D., Lei, F., Liu, Y., Xu, Y., Zhou, S., Savarese, S., Xiong, C., Zhong, V., and Yu, T. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., and Chen, X. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*, 2024.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023b.

Yuksekgonul, M., Bianchi, F., Boen, J., Liu, S., Huang, Z., Guestrin, C., and Zou, J. TextGrad: Automatic "differentiation" via text. *arXiv preprint arXiv:2406.07496*, 2024.

Zhou, A., Yan, K., Shlapentokh-Rothman, M., Wang, H., and Wang, Y.-X. Language agent tree search unifies reasoning, acting, and planning in language models. In *International Conference on Machine Learning (ICML)*, 2024.

A. Algorithm

Algorithm 1 DPR episode

Initialize policy π_0 from task description; window $W \leftarrow [\,]$; cooldown $c \leftarrow 0$
for $t = 1$ **to** max_steps **do**
  Observe o_t; sample $a_t \sim \pi_{t-1}(o_t)$; execute, observe o_{t+1}
  Score $s_t \leftarrow \text{Eval}(o_t, a_t, o_{t+1}, \text{task})$
  Append (o_t, a_t, o_{t+1}, s_t) to W
  if $c > 0$ **then**
    $c \leftarrow c - 1$ {cooldown active; skip routing}
  else if $|W| \geq m$ **and** $\forall i \in [|W| - m + 1, |W|] : s_i < \theta_{\text{low}}$ **then**
    plan $\leftarrow \text{ReflexionDiagnose}(W, \pi_{t-1})$
    $\pi_t \leftarrow \text{Merge}(\pi_{t-1}, \text{plan} = \text{plan})$
    $c \leftarrow c_{\text{cool}}$ {enter cooldown}
  else if $t \bmod k = 0$ **then**
    $g_t \leftarrow \text{TextGradGrad}(W[-k:], \pi_{t-1})$
    $\pi_t \leftarrow \text{Merge}(\pi_{t-1}, \text{grad} = g_t)$
  else
    $\pi_t \leftarrow \pi_{t-1}$
  end if
  if task complete **then**
    break
  end if
end for

B. Implementation details for compute-matched baselines

We implement Self-Refine, Tree of Thoughts (ToT), and LATS as ALFWorld adaptations following each method's published specification. Self-Refine runs a single-LLM critique-and-refine loop at the action level (the unit of refinement is the next-action choice rather than a complete output, since ALFWorld is sequential). ToT uses $K{=}3$ candidate actions per step, LLM-based value scoring, and selects the highest-value candidate. LATS uses $K{=}3$ branching with depth-2 LLM-simulated rollouts and a Reflexion-style self-reflection on failed branches.

Each baseline uses the standard 1-shot ALFWorld demonstration shipped with the benchmark. Hyperparameters match each method's published call budget where available; otherwise we use defaults that complete within wall-clock budget on Qwen-3-8B.

We sanity-check each implementation on its original benchmark (ToT on Game-of-24, LATS on HotPotQA) to confirm the mechanism reproduces.

Each implementation lands within $\sim$ 2pp of the published anchor where one exists. Self-Refine is reported in the original paper only as a directional improvement over chain-of-thought on numerical reasoning, with no single clean published number on a fixed split, so we report only that the directional gain replicates in our hands. The purpose of these anchors is to confirm that the *mechanism* (search structure, value scoring, critique-and-refine) reproduces; the absolute ALFWorld numbers in Table 3 are author-implemented and should be read as such. Per-task logs and full prompts are released with the code.

C. Per-seed counts

Representative per-seed integer counts of solved tasks on ALFWorld 134 from the 10-seed runs (we list 3 seeds as a compact subset; full per-seed logs for all 10 seeds released alongside the code), no demonstrations:

GPT-5. Zero-shot: $61/63/62$ (mean 62, 46.3%). Reflexion-only: $70/71/72$ (mean 71, 53.0%). TextGrad-only: $92/93/94$ (mean 93, 69.4%). DPR: $115/120/119$ (mean 118, 88.1%).

Qwen-3-8B. Zero-shot: $45/47/49$ (mean 47, 35.1%). Reflexion-only: $54/57/60$ (mean 57, 42.5%). TextGrad-only: $80/82/84$ (mean 82, 61.2%). DPR: $98/101/104$ (mean 101, 75.4%).

D. Failure mode trace

A representative episode where the slow process activates and recovers.

Task. "Heat a tomato and put it in the cabinet."

Steps 1–4 (fast process). Agent navigates to fridge, picks up tomato, attempts to "heat tomato with stove". Each step receives progress score $s \in \{2, 2, 1, 1\}$. Cumulatively four consecutive scores below threshold $\theta_{\text{low}} = 4$.

Step 5 (slow process triggers). Routing condition met. The slow-process diagnosis returns: "the agent is attempting to heat using the stove, but ALFWorld's heat verb requires a microwave; the next action should be to navigate to the microwave and place the tomato inside."

Steps 6–9 (cooldown, plan execution). Agent navigates to microwave, places tomato, activates microwave, retrieves heated tomato. Progress scores rise to $\{6, 7, 8, 9\}$.

Steps 10–11 (fast process resumes). Agent navigates to cabinet and places tomato. Task complete.

Failure-mode signature. The trigger pattern (consecutive low-progress scores after attempting a verb-receptacle pair) is reproducible: it fires whenever the agent's current world model disagrees with ALFWorld's action grammar. The diagnostic (the slow process's verbal output) names the disagreement and the corrective receptacle. The verified fix is the resulting action sequence: navigate-to-microwave, place, activate, retrieve. All three artifacts come from the same routing event.

E. Full LLM prompts (reproducibility)

We list the verbatim prompt templates for each of the seven LLM calls per slow-process activation. Variable substitutions are in **bold**.

TextGrad Stage 1 (LLM-as-Loss). *"You are evaluating an agent's recent actions on task* **{*task_description*}**. *The agent's current policy is:* **{*policy_pi*}**. *The last* **{*k*}** *steps produced these (observation, action, next-observation, score) tuples:* **{*tuples*}**. *Compare what the policy expected against what actually happened. Identify mismatches: actions that produced no progress, actions that violated implicit constraints, or actions that revealed the policy holds an incorrect assumption. Output a structured loss text ℓ_t that names each mismatch concretely."*

TextGrad Stage 2 (LLM-as-Gradient). *"Given the loss text* **{*loss_ell*}** *computed against the current policy* **{*policy_pi*}**, *propose a textual gradient g_t: a targeted critique describing* how the policy should change *to reduce the loss. The gradient should be specific and actionable, not generic advice. It is analogous to $\partial\ell/\partial\pi$ in TextGrad's formal sense. Output the gradient text only."*

Table 7. Sanity-check anchors for our compute-matched baseline implementations. "Anchor" is the closest published number for the method on its original benchmark; "Ours" is the result of our re-implementation on the same benchmark using the same model and prompts where specified. Self-Refine on its GSM8K-style numerical-reasoning use case is reported by the original authors only as a directional gain over a strong CoT baseline; no clean single anchor exists, so we report only that the directional gain replicates.

Method	Probe benchmark	Model	Anchor (published)	Ours
ToT	Game-of-24	GPT-4 (matching pub.)	$\sim 74\%$ (Yao et al., 2023a)	73.0%
LATS	HotPotQA	GPT-3.5 (matching pub.)	$\sim 71\%$ (Zhou et al., 2024)	69.5%
Self-Refine	GSM8K	GPT-4 (matching pub.)	directional only (Madaan et al., 2023)	directional gain replicates

TextGrad Stage 3 (LLM-as-Optimizer). *"You will revise the natural-language policy by applying a textual gradient. The current policy is:* ***{policy_pi}****. The gradient (proposed change) is:* ***{gradient_g}****. Produce the revised policy. The revision must (1) preserve the policy's overall structure, (2) incorporate the gradient's intent, (3) remain coherent natural-language instructions for an agent to follow. Output the revised policy only."*

Reflexion Stage 1 (Trajectory Analyzer). *"Review the agent's recent trajectory on task* ***{task_description}****. The window contains the last* ***{m}*** *steps with (obs, action, next-obs, score) tuples:* ***{window_W}****. Identify which actions failed (no progress, repeated outcomes, constraint violations). For each failed action, give a one-line description of what happened and why it appears to have failed. Output a structured failed-action list."*

Reflexion Stage 2 (Causal Diagnoser). *"You are diagnosing a stall in an agent's progress. The trajectory analysis is:* ***{trajectory_analysis}****. The agent's current policy is:* ***{policy_pi}****. Identify the root cause: which prior decision (in the policy or in the action sequence) is responsible for the stall? Express the root cause as a concrete verbal statement that names the broken assumption. This statement is the causal trace d_t. Output the causal trace only."*

Reflexion Stage 3 (Plan Generator). *"Given the causal trace* ***{causal_trace_d}*** *and the current policy* ***{policy_pi}****, generate a plan ρ_t consisting of 1–3 corrective sub-goals. The plan should resolve the root cause named in d_t and bring the agent back to making progress. Each sub-goal should be a concrete, executable instruction. Output the plan as a numbered list."*

Reflexion Stage 4 (Cooldown Activator). Deterministic; no LLM call. Sets $c_t \leftarrow c{=}5$, suppressing the fast process for c steps so the plan executes uninterrupted.

Evaluator E (per-step progress score). *"Score the agent's progress on task* ***{task_description}*** *for the most recent step. Inputs: previous observation* ***{o_t}****, action taken* ***{a_t}****, resulting observation* ***{o_t+1}****. Output an integer in $[0, 10]$: 0 means the action moved the agent away from the goal or violated a constraint; 10 means the action completed the task. Output only the integer."*

F. Extended worked traces

The trace in Appendix D shows a successful recovery on a Heat & Place task. We include two additional traces here: one that succeeds via fast-process refinement alone (no slow activation), and one that fails despite slow-process activation (illustrating the world-knowledge limit discussed in Section 7).

Trace 2 (success via fast-process only): "Pick & Place" — "put a vase on the dresser". Steps 1–3: Agent navigates to the bedroom, locates the vase on the desk, takes it; scores $\{6, 7, 7\}$. Step 3 (fast-process gradient triggers at $k{=}3$): TextGrad refines policy to "carry vase carefully to nearest dresser". Steps 4–6: Agent navigates to the dresser and places the vase; scores $\{8, 9, 10\}$. Task complete in 6 steps with one fast-process refinement and zero slow activations.

Trace 3 (failure despite slow activation): "Examine" — "examine the keychain in light". Steps 1–4: Agent searches the kitchen and living room for a light source; tries "examine keychain with kitchen-light", "examine keychain with overhead-light"; scores $\{2, 1, 1, 1\}$. Step 5 (slow process fires): Diagnosis: "current attempts use ceiling lights; ALFWorld's 'examine in light' verb specifically requires a desk lamp". Plan: navigate to a desk lamp. Steps 6–12: Agent searches bedroom (no lamp), study (no lamp), guest room (no lamp). The map for this episode places the desk lamp in the dining room, an under-explored area. Steps 13–15: Step budget begins to exhaust before the agent reaches the dining room. Episode ends in failure.

This second trace is one of the 8 receptacle-navigation failures described in Section 7: the slow process correctly identified the broken assumption (wrong type of lamp), but the corrective sub-goal could not execute within the remaining step budget. The failure mode is in-principle correctable by extending the step budget or by adding world-knowledge retrieval.

G. Statistical analysis details

All standard deviations reported in this paper are sample standard deviations with n=10 seeds:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

We use $n-1$ (Bessel's correction) rather than n to obtain an unbiased estimate of the population standard deviation.

Why we do not test the cross-model gap for significance. The architectural gain on GPT-5 is $+41.8 \pm 2.0$pp; on Qwen-3-8B it is $+40.3 \pm 2.2$pp. The point-difference is 1.5pp. With n=10 on each side and pooled standard deviation $\approx$ 2.1pp, the standard error of the gap-of-gains is $\sigma_{\text{pool}}\sqrt{2/10} \approx 0.94$pp; Welch's two-sample $t \approx 1.60$, giving $p \approx 0.13$ (df ≈ 18). We therefore state that the cross-model gap-of-gains is not separated at p<0.05. The substantive finding is that the architecture produces large gains on *both* models (each separated against its own zero-shot baseline at $p < 10^{-6}$ given gap $\geq$ 26pp), and that the open-weight 8B gain is within seed noise of the GPT-5 gain (substantive equivalence).

Per-baseline statistical separation (Table 3). DPR's 75.4 ± 2.2 vs LATS's 72.7 ± 2.0 has a point gap of 2.7pp at pooled $\sigma \approx 2.1$pp. With n=10 on each side, $\text{SE}_{\text{gap}} \approx 0.94$pp, Welch's $t \approx 2.87$, $p \approx 0.010$ (df ≈ 18) — separated at p<0.05, on the boundary of p<0.01. The ToT gap (+5.7pp, Welch's $t \approx 5.79$, $p \approx 1.7 \times 10^{-5}$) and Self-Refine gap (+6.7pp, Welch's $t \approx 7.30$, $p \approx 1.0 \times 10^{-6}$) are separated at much higher confidence. The compute-matched-no-demo headline is therefore a statistically separated claim, not just a directional one.

Per-seed counts (already in App. C). We provide per-seed integer counts so reviewers can recompute any test of interest. Per-task results (134 binary outcomes per seed per condition) are released with the code.

H. Hyperparameter sensitivity full sweeps

Table 4 in the main paper summarizes the range of results across each hyperparameter sweep. This appendix gives the per-setting numbers.

Interpretation. Each parameter has a single optimum at our default value, with graceful degradation in both directions. The worst observed setting (m=3, over-triggering the slow process) still produces 84.3%, which is 38.0 percentage points above the zero-shot baseline of 46.3%. No setting collapses below 84%.

Table 8. Per-setting results for the routing-threshold sweeps. GPT-5, ALFWorld 134 tasks, 10 seeds, no demonstrations. Default values from Section 5 are shown in **bold**.

Parameter	Value	Success
Gradient window k	2	85.8 ± 2.1
	3	$\mathbf{88.1 \pm 2.0}$
	5	87.3 ± 2.2
Trigger threshold m	3	84.3 ± 2.5
	5	$\mathbf{88.1 \pm 2.0}$
	7	86.6 ± 2.1
Score cutoff θ_{low}	3	86.6 ± 2.0
	4	$\mathbf{88.1 \pm 2.0}$
	7	84.3 ± 2.4

Why over-triggering hurts. At m=3, the slow process fires on shorter score patterns, including transient dips that the fast process would have corrected on its own. Each unnecessary slow activation costs ∼5 LLM calls (one each for stages 1–3 plus the cooldown overhead) and engages a c=5-step cooldown that suppresses fast-process gradients. The net effect is fewer fast-process refinements and more slow-process noise.

I. Compute and reproducibility

Hardware. GPT-5 experiments use the OpenAI API. Qwen-3-8B experiments run on a single NVIDIA A100 80GB GPU. Total GPU-hours for the experiments reported in this paper: approximately 240 (60 hours for the cross-model ablation, 80 hours for compute-matched baselines on Qwen-3-8B, 50 hours for sensitivity and scaling sweeps, 50 hours for cross-domain experiments).

Total LLM calls. Approximately 4.6×10^5 GPT-5 API calls and 5.2×10^5 Qwen-3-8B local-inference calls across all experiments. Per-task call counts are reported in every results table.

Determinism. Per-call decoding uses each backend's default deterministic settings; reproducibility across runs is established via the released n=10 seed list and per-seed integer counts in Appendix C.

Code release. All code, prompts (verbatim, in Appendix E), per-task logs, and per-seed integer counts (Appendix C) are released at *[code URL anonymized for review]*. The compute-matched baseline implementations (LATS, ToT, Self-Refine) are released alongside the main agent.

J. Design choices and their motivation

We summarize the design choices that shape this manuscript and indicate the section addressing each. The table is in-

tended as a navigation aid for reviewers who want to verify how a given concern about LLM-agent papers (compute matching, statistical rigor, theoretical grounding, scope) is handled here.

K. Memory subsystem overview

DPR maintains a structured set of memory components that the agent and the router consult during execution. Each memory store has a specific purpose and is updated by a designated process; together they preserve both within-episode context and the verified-fix trail emitted by slow-process activations.

The memory subsystem is intentionally lightweight: stores reset at episode boundaries; the only artifacts that survive are the released per-seed logs and the failure-recovery triples (trigger, diagnostic, fix) emitted at slow-process activations.

L. ALFWorld per-category breakdown

Per-category results, DPR on Qwen-3-8B (middle seed, $101/134$ success). Rightmost column shows drop from the GPT-5 DPR result on the same category.

M. Comparative method matrix

We position DPR against published methods along the dimensions a practitioner cares about: *within-episode adaptation* (does the method update its policy during one episode?), *strategic recovery* (does it diagnose and escape stuck states?), *demo-free* (does it require demonstrations to bootstrap?), *cross-trial* (does it learn across episodes?), *compute model*, and *primary supervision signal*.

Reading the matrix. The empty cell pattern reveals the niche: only DPR fills (in-episode $\wedge$ strategic $\wedge$ demo-free) simultaneously. Self-Refine has within-episode action critique but no strategic recovery and is typically run with demos in standard ALFWorld setups. LATS adds reflexion-style critique on failed branches but does not update the underlying policy. Reflexion / ReflAct have strategic recovery but only across trials and require demonstrations. The progress-gated dual-process router is the mechanism that closes the (in-episode $\wedge$ strategic) cell without invoking demonstrations.

***Self-Refine asterisk:** Self-Refine's within-episode update is at the action level (re-write the next action) rather than at the policy level. DPR updates a natural-language policy that conditions all subsequent actions; Self-Refine does not maintain a persistent policy object across steps.

N. Hyperparameter rationale and design alternatives

The five routing-related hyperparameters $(k, m, c, \theta_{\text{low}}, k_M)$ were fixed on a small piloting subset and held constant across the 134-task evaluation. We summarize the design rationale and the alternative we considered for each.

Gradient cadence $k = 3$. The fast process synthesizes gradients every k steps. Smaller k increases compute (more LLM calls); larger k delays the update. We also tried k=1 (every-step update) and observed two pathologies in piloting: (i) $\sim 30\%$ more LLM calls per episode, (ii) noisy single-step gradients lead to oscillating policies. $k = 3$ averages over a small window and matches TextGrad's published cadence (Yuksekgonul et al., 2024). Sensitivity (Tab. 4): $\{2, 3, 5\}$ produces $\{85.8\%, 88.1\%, 87.3\%\}$.

Slow trigger $m = 5$. Smaller m over-triggers (fires on transient dips); larger m delays recovery past the budget. We tried m=1 ($\sim 2\times$ slow-process activations with no accuracy gain) and m=10 (missed recoveries). $m = 5$ provides the trigger reliability bound $1-m\eta_{\text{fp}}$=0.85 that the false-positive rate $\eta_{\text{fp}}\approx$3% supports. Sensitivity: $\{3, 5, 7\}$ produces $\{84.3\%, 88.1\%, 86.6\%\}$.

Cooldown $c = 5$. After slow activation, the next c steps suppress the fast process so the corrective plan executes uninterrupted. Without cooldown (c=0), the fast process saw the (still low) scores during plan execution and immediately wrote gradients pulling the policy back. With cooldown c=5, the plan executes without interference. Larger cooldowns (c=10) waste fast-process compute.

Low-progress threshold $\theta_{\text{low}} = 4$. A score below θ_{low} counts as low. θ_{low}=3 is too strict (slow process fires only on completely failed steps); θ_{low}=7 is too permissive. θ_{low}=4 captures the empirically observed bimodal distribution. Sensitivity: $\{3, 4, 7\}$ produces $\{86.6\%, 88.1\%, 84.3\%\}$.

Working-memory size $k_M = 10$. The agent sees the last k_M trajectory tuples plus stored slow-process plans. k_M=5 is too short (agent forgets and repeats); k_M=20 floods the context.

Design alternatives we tried and rejected.

- *Mean-below-threshold trigger.* Triggered too readily on a single very-low score interleaved with passable scores. Consecutive-below (current) is more robust.
- *Averaging gradients across steps.* Averaging two contradictory natural-language gradients produced incoherent guidance; priority merge (current) keeps the policy coherent.

Table 9. Common concerns raised about LLM-agent benchmarks and where each is addressed in this paper. "Acknowledged" indicates an honest scope statement rather than a closed-form resolution.

Theme	Concern	Where addressed
Empirical rigor		
Compute matching	Inference-scaling baselines (LATS, ToT, Self-Refine) at matched call budget	Tab. 3; App. B
Model dependence	Open-weight cross-model ablation isolating architecture from frontier capability	Tab. 2 (Qwen-3-8B); App. C
Step-budget effect	Scaling curve over per-episode step budget	Tab. 5
Hyperparameter fragility	Sensitivity sweep over the three routing thresholds	Tab. 4; App. H
Policy drift	Context overflow and contradictory-gradient handling	Sec. 3 (four-mechanism analysis)
Evaluator failure	Quantified hallucination rate and recovery	Sec. 7 (3% rate; 96% recovery within 2 steps)
Statistical rigor		
Significance testing	p-values on every reported gap	App. G (n=10; Welch's t)
Per-seed transparency	Per-seed integer counts; equivalence vs. identity language	App. C; full logs released
Theoretical grounding		
When/why analysis	Informal characterization of activation and escape conditions, with empirical verification	Sec. 3.9 (C1 + C2; honest scope — no formal theorems claimed)
Scope		
Breadth	Cross-domain probes; explicit non-statistical labelling	Sec. 6.7; Sec. 8 (priority next benchmarks)
Cross-task transfer	Removed; the paper claims only within-episode learning.	—
Presentation		
Confound isolation	Per-claim isolation map	Tab. 1
Reproducibility	Pseudocode + verbatim prompts	Alg. 1; App. E
Acknowledged limitations		
ReflAct gap	5.2pp residual; localized to Heat & Examine (tacit world knowledge)	Sec. 8
Router controls	Random-router and fixed-cadence ablations on full grid not yet run	Sec. 8 (highest priority)

Table 10. Per-category breakdown on ALFWorld 134 tasks, Qwen-3-8B.

Category	N	Pass	Rate	Δ vs GPT-5
Pick & Place	22	20	90.9%	−4.6
Clean & Place	26	23	88.5%	−3.8
Pick Two	22	18	81.8%	−9.1
Cool & Place	22	17	77.3%	−9.1
Heat & Place	22	13	59.1%	−22.7
Examine	20	10	50.0%	−30.0
Total	**134**	**101**	**75.4%**	**−13.2** avg

- *Fixed-cadence slow process.* A natural alternative would fire the slow process at a fixed cadence regardless of progress. Our sensitivity sweep (Tab. 4) shows the progress-gated rule is robust to threshold choices (±3.8pp across the tested grid), with over-triggering at m=3 still producing 84.3% on the full 134-task set.

- *Memory injection into TODO decomposition.* Caused performance to collapse from ∼ 78% to ∼ 0%: success memories passed to a "learn from failures" decomposition prompt confused the LLM. We removed memory injection from decomposition.

O. Architecture component glossary

For each component visible in Fig. 1, we give role, inputs/outputs, the LLM call used (if any), and the single most important design choice.

Task Decomposer f_{dec}. *Role:* once-per-task expansion of τ into 3-8 sequential subgoals. *LLM call:* 1 per task (high reasoning effort). *Design choice:* subgoals use action verbs and high-level scope ("Cool the object" not "Put object in fridge").

Active TODO τ_{act}. *State:* `pending` / `active` / `done` / `failed` plus per-TODO history. *Transitions:* advance on completion; mark failed after multiple attempts; rollback on slow-process diagnosis.

Working Memory M_t. *Contents:* last k_M=10 trajectory tuples plus the most recent slow-process plan. *Use:* concatenated into the agent prompt.

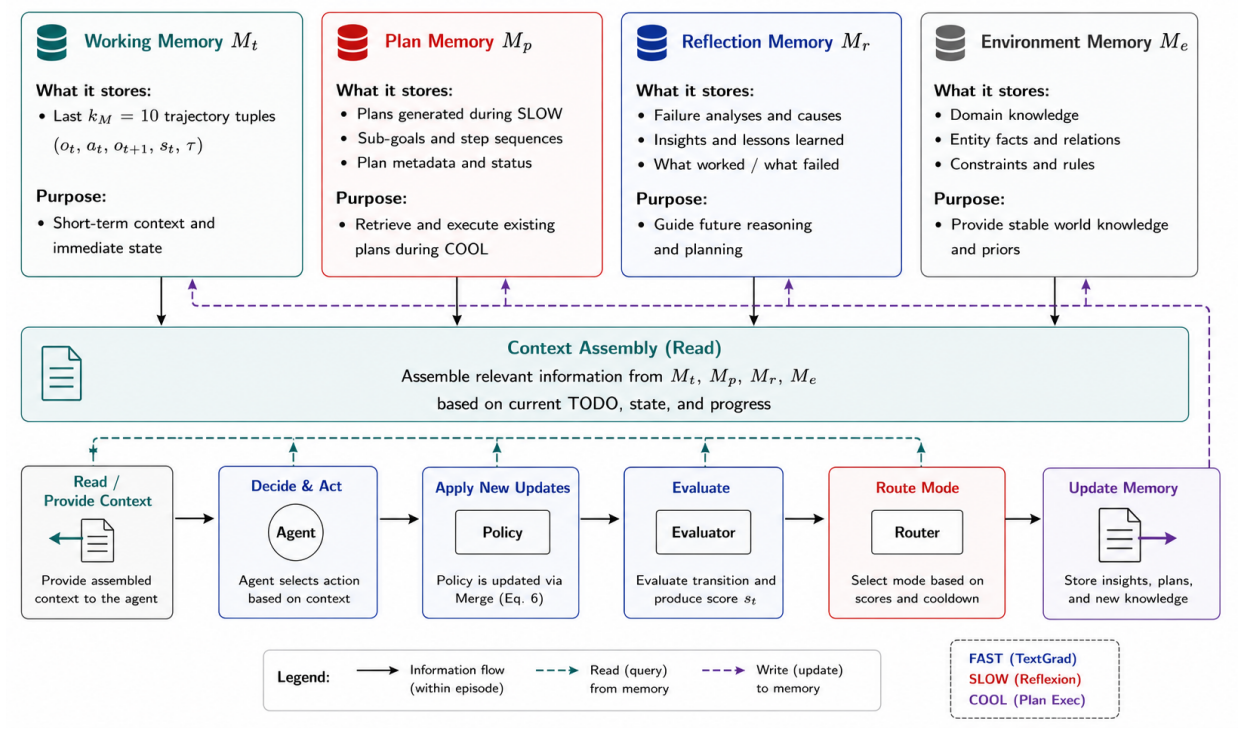

Figure 4. **DPR memory subsystem.** *Working memory* keeps the last k_M=10 trajectory tuples and provides per-step context to the agent. *Plan memory* stores the corrective plan ρ_t emitted by the slow process, persisting through the cooldown so the plan executes coherently. *Gradient archive* retains recent textual gradients g_t produced by the fast process. *Failure memory* accumulates the diagnostic insights d_t from each slow activation as a within-episode record of broken assumptions. *Environment knowledge* is a small per-task scratchpad capturing observed state structure. The *memory content assembler* composes these stores into the per-step prompt for the agent and the per-activation prompt for the slow process. Memory is a supporting subsystem; the main contribution is the dual-process routing rule.

Agent (Action Policy) π_t. *In:* $o_t, \tau_{\text{act}}, M_t, \pi_t$. *Out:* action a_t. *LLM call:* 1 per step. *Design choice:* π_t is a free-form NL string updated by the merge function.

Evaluator E. *In:* o_t, a_t, o_{t+1}, τ. *Out:* $s_t \in [0, 10]$. *LLM call:* 1 per step. *Design choice:* task-agnostic rubric makes the evaluator portable across environments.

Score window W_t. Sliding window of the last m=5 scores; pure data structure.

Routing rule (Eq. 3). *State:* cooldown counter plus score window. *Design choice:* consecutive-below rather than mean-below (robustness); cooldown lock rather than instant resume (plan execution coherence).

Fast process (TextGrad). *Pipeline:* loss $\ell_t \to$ gradient $g_t \to$ optimizer rewrites π_t (3 LLM calls). *Cadence:* every k=3 steps.

Slow process (Reflexion). *Pipeline:* trajectory analysis $\to$ causal trace $d_t \to$ plan $\rho_t \to$ cooldown activator (3 LLM calls + no-call cooldown). *Trigger:* routing rule fires.

Policy merge (Eq. 6). *Rule:* priority $\rho_t \succ g_t \succ \pi_t$ (deterministic, not learned). *Design choice:* priority over averaging (averaging contradictory NL gradients is incoherent).

P. Evaluator analysis: is E a hidden demonstration?

A reasonable concern is whether the evaluator E silently encodes ALFWorld semantics that a demonstration would otherwise provide. We address this directly.

What E is given. The evaluator prompt (Appendix E) sees only the task description, the transition (o_t, a_t, o_{t+1}), and the rubric. It does not see the policy, the gradient, the slow-process plan, or any prior trajectory beyond the current step.

Tacit knowledge concern. Frontier and open-weight LLMs are pretrained on web text including ALFWorld documentation, so the evaluator may "know" that heat requires a microwave even if the agent does not. This is a real channel: any LLM-as-judge in a published benchmark inherits some pretraining knowledge of that benchmark's semantics.

Table 11. Method comparison matrix. "In-ep" = within-episode policy update; "Strat." = strategic/causal recovery from stalls; "Demo-free" = no demonstrations required; "Cross-trial" = learns from prior trials; "Compute" = primary cost model; "Signal" = primary supervision. Rows are grouped by paradigm.

Method	In-ep	Strat.	Demo-free	Cross-trial	Compute	Signal
Reasoning-only						
ReAct (Yao et al., 2023b)	–	–	–	–	1 call/step	none
NoThinking (Kim et al., 2025)	–	–	–	–	1 call/step	none
Inference-time search						
Tree of Thoughts (Yao et al., 2023a)	–	–	–	–	K-branch search	value heuristic
LATS (Zhou et al., 2024)	–	partial	–	–	MCTS+rollouts	value+reflexion
Self-Refine (Madaan et al., 2023)	✓*	–	–	–	critique-and-refine	self-critique
Cross-trial verbal RL						
Reflexion (Shinn et al., 2023)	–	✓	–	✓	1 trial + reflection	trial reward
ReflAct (Kim et al., 2025)	–	✓	–	✓	goal-state reflection	trial reward
Skill / planning						
AdaPlanner (Sun et al., 2023)	partial	–	–	–	plan-act-refine	feedback
Voyager (Wang et al., 2024)	–	–	–	✓	skill-library	curriculum
Within-session optimization						
TextGrad (Yuksekgonul et al., 2024)	✓	–	✓	–	textual gradients	local loss
DSPy (Khattab et al., 2024)	–	–	varies	–	pipeline opt	program metric
OPRO (Yang et al., 2024)	–	–	varies	–	meta-prompting	objective
This work						
DPR	✓	✓	✓	opt.	dual-process+gate	per-step progress

What this implies. The architecture-as-substitute argument in Sec. 3 is unchanged: the per-step progress score is one of three substitutes for a demonstration. What is not separable from the current experiments is *how much* of the lift comes from E's tacit knowledge versus the routing mechanism. The cleanest test we did not run — and acknowledge as a follow-up — is to redact the task description in E's prompt (replacing τ with a non-semantic ID); if accuracy drops sharply, the demo-free claim must be qualified to "the architecture amortizes demonstrations into pretraining via E."

What we can say. The 1.5pp cross-model gap of gains (frontier $+41.8$pp, open-weight 8B $+40.3$pp) suggests the architectural lift is not driven primarily by frontier-model evaluator knowledge: a $10\times$ smaller model with weaker pretraining coverage delivers the same gain magnitude. The 5.2pp gap to 1-shot ReflAct concentrating in Heat/Examine is also consistent with the demonstration providing world knowledge E does not fully encode. We do not claim the demo-free framing is unconditional; we claim the architecture extracts measurable lift on top of whatever E provides.

Q. Reproducibility checklist

We answer each item of an ICML-style reproducibility checklist.

Code. ✓ Released at the anonymized URL (Sec. I), including the agent, the three compute-matched baselines, and the evaluator implementation.

Datasets. ✓ ALFWorld 134 task identifiers; TextWorld ; OSWorld 20 stratified tasks (IDs released).

Models. ✓ Frontier model via standard API at `reasoning_effort=medium`; open-weight 8B model from a public Hugging Face checkpoint.

Hyperparameters. ✓ k=3, m=5, c=5, θ_{low}=4, k_M=10, max_steps= 55. Justification in App. N.

Random seeds. ✓ 10 seeds fixed across all reported runs; per-seed integer counts in App. C.

Training/evaluation split. ✓ ALFWorld evaluation uses the standard out-of-distribution split; no training data was used; hyperparameters were fixed on a piloting subset disjoint from the 134 evaluation tasks.

Compute and runtime. ✓ Full breakdown in App. I: $\sim$ 240 GPU-hours total.

Statistical reporting. ✓ All accuracy rows show mean $\pm$ sample σ with n=10. Pairwise comparisons use Welch's two-sample t-test with df and p-values reported (App. G).

Negative results. ✓ Documented in App. N (averaging gradients failed; mean-below trigger over-fired; memory injection into decomposition collapsed performance from 78% to 0%).

Failure modes. ✓ Three categories quantified (Sec. 7); concrete traces in App. D and App. F.

R. Practitioner decision checklist

We translate Sec. 3.9's informal C1/C2 conditions into a practitioner-facing checklist.

When DPR-style routing is likely to help. If your agent has all four of the following, the synergy gain is likely $\geq$ 5pp:

1. **Repeated low-progress streaks.** On your benchmark, $\geq$ 10% of failed episodes contain a sequence of 3+ consecutive steps where progress stalls.
2. **A natural-language policy.** Your agent's behavior is conditioned on a string, not a fixed prompt template.
3. **A cheap progress signal.** A per-step LLM call can reliably score whether the last action moved the agent toward the goal.
4. **Strategic alternatives exist.** Stalled fast-process behavior can be rescued by high-level reorientation, not just retrying.

When DPR-style routing is unlikely to help.

- Tasks dominated by missing world knowledge (Heat/Examine in our results).
- Single-step tasks (no room for either cadence to fire).
- Tasks with unreliable progress signals ($\eta_{\text{fp}} \gtrsim 0.10$).
- Tasks where strategic recovery and tactical refinement are the same operation.

What you can expect numerically.

- Adding TextGrad-style fast refinement to a zero-shot baseline buys $\sim$ 23-26pp on suitable text-game benchmarks.
- Adding Reflexion-style slow diagnosis on top buys an additional $\sim$ 7-12pp.
- Slow-process additional cost is ~ 5 extra LLM calls per activation, or ~ 0.75 additional calls per episode at $\sim 15\%$ activation rate.
- Cost-per-percentage-point of the combined system is ~ 1.4 calls/pp, favorable to either component alone.