FROM KNOWLEDGE TO DECISIONS: REPAIRING EVIDENCE SEEKING AFTER FULL-EVIDENCE ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Full-evidence adaptation can transfer non-uniformly across evidence-conditioned competence and interactive evidence acquisition. On HotpotQA, it raises Oracle exact match from 58.6% to 68.0% but lowers Interactive exact match from 40.6% to 32.0%. Restoring inherited Base acquisition while retaining the adapted reader raises Interactive EM to 46.1%, showing that useful reader competence survives independently of the degraded adapted acquisition behavior. We call this a *competence–policy transfer gap*: adaptation improves evidence-conditioned competence, but that competence is not reliably translated into the decisions required to acquire useful evidence.

Functional decoupling does not make inherited acquisition optimal. K2D-ALIGN (**K**nowledge-to-**D**ecision Alignment) addresses the residual decision problem: when should retained reader competence justify a beneficial deviation from inherited evidence seeking? Matched continuation identifies Base-relative opportunities for learning root proposals; deployment-time admission authorizes replacement or executes exact Base fallback. Admission compares candidate and Base terminal shadow continuations before execution, trading additional model-side inference for conservative replacement.

On HotpotQA, K2D-ALIGN improves Interactive EM from 46.1% to 52.3%; under matched initialization, Base-relative supervision contributes a further 3.9 points over the zero-update proposal. On 872 held-out cases, the frozen system gains 2.3 EM points over Base+FE and exceeds an approximately compute-matched Base Best-of-2 control by 3.1 points. Across four evidence-seeking benchmarks, it improves macro-average EM/F1 from 30.7/41.5 to 32.6/43.4. These results connect functional diagnosis to localized acquisition repair while preserving the reader competence that already transferred.

1 INTRODUCTION

Tool-using agents rely on distinct functions: using available evidence, acquiring new evidence, deciding when to stop, and deciding whether a proposed action should replace inherited behavior. Adaptation need not improve these functions uniformly. Supplied-evidence accuracy can therefore increase even when the interactive behavior required to obtain that evidence degrades.

We observe this divergence in interactive retrieval. On RAG-Gym/HotpotQA (Xiong et al., 2026; Yang et al., 2018), full-evidence SFT raises Oracle EM from 58.6 to 68.0 but lowers Interactive EM from 40.6 to 32.0. Restoring the original BASE acquisition policy while retaining the adapted reader recovers Interactive EM to 46.1 (Figure 2a). The adapted reader therefore remains useful, but its improved evidence-conditioned competence is no longer reliably translated into the acquisition decisions needed to expose that competence at deployment. We call this divergence a *competence–policy transfer gap*.

Functional substitution first asks *what transferred*: does competence acquired under full-evidence adaptation survive independently of degraded acquisition behavior? The 32.0-to-46.1 substitution establishes this separation in HotpotQA. Functional decoupling is the diagnosis, not the learning contribution.

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.

The next question is *where beneficial deviation remains*. Inherited acquisition is a reference, not an optimum. We compare alternative first retrieval actions from the same question-level state under a shared adapted reader and frozen downstream policy. Matched terminal improvement identifies localized Base-relative opportunity; a continuous collection advantage weights the retained training pairs.

Finally, *when should deployment realize that deviation*? We introduce K2D-ALIGN (**K**nowledge-to-**D**ecision Alignment)[1] to align evidence-seeking decisions with retained reader competence. It learns root proposals from Base-relative opportunities, then authorizes replacement using two terminal shadow continuations and reader comparison. Otherwise it executes exact Base fallback. The adapted reader and inherited downstream policy stay fixed.

Component substitution, matched target comparisons, and frozen held-out and four-benchmark evaluations follow these three questions. HotpotQA is the primary harmful-divergence regime; FHIR shows that the relative direction of functional transfer is setting-dependent. Transfer diagnostics also distinguish available proposal opportunity from the gain realized by admission.

Contributions.

1. **Functional diagnosis.** We demonstrate non-uniform transfer across evidence-conditioned competence and acquisition behavior, and isolate retained reader competence through component substitution.
2. **Base-relative opportunity.** We formulate matched continuation to identify beneficial localized deviations under a shared downstream policy.
3. **Knowledge-to-decision alignment.** We introduce K2D-ALIGN, which learns root proposals and conservatively authorizes replacement while preserving the adapted reader, inherited continuation, and exact Base fallback. Frozen evaluations measure the resulting system's held-out and transfer behavior.

2 RELATED WORK

Evidence-conditioned adaptation and functional transfer. Retrieval-augmented generation combines parametric generation with external evidence (Lewis et al., 2020), while RAFT adapts models to answer from relevant documents among distractors (Zhang et al., 2024). PaST separates knowledge updates from RL-acquired skills in continual adaptation (Tang et al., 2026). Our focus is complementary: we ask whether adaptation transfers uniformly across agent functions. Through component substitution, we separate evidence use from evidence acquisition and test whether improved evidence-conditioned competence remains expressed through the policy that must acquire evidence at deployment.

Interactive retrieval and policy optimization. ReAct and IRCoT interleave reasoning with retrieval (Yao et al., 2023; Trivedi et al., 2023); Self-RAG learns when to retrieve and reflect (Asai et al., 2024), and Adaptive-RAG selects retrieval strategies by question complexity (Jeong et al., 2024). RAG-Gym provides process supervision for information-seeking agents (Xiong et al., 2026). More recent methods optimize interactive retrieval directly: Search-R1 and ReSearch train reasoning–search trajectories with RL (Jin et al., 2025; Chen et al., 2025); RAG-RL combines RL with curriculum learning for multi-hop QA (Huang et al., 2026); and Interact-RAG trains an expanded retrieval action space with SFT and RL (Hui et al., 2026). In contrast, K2D-ALIGN repairs a diagnosed post-adaptation mismatch by intervening only at the first retrieval decision while preserving the adapted reader and inherited downstream acquisition policy.

Counterfactual and localized credit assignment. Counterfactual methods localize credit through matched alternatives: C3 compares context-matched messages under fixed continuation (Chen et al., 2026), while CRAFT uses sibling rollouts for token-level credit assignment (Meng & Chen, 2026). K2D-ALIGN applies this principle to a Base-relative first retrieval action: the inherited action is the

[1]We use "knowledge" operationally to refer to the evidence-conditioned competence retained by the adapted reader, rather than to claim an increase in the model's parametric factual knowledge.

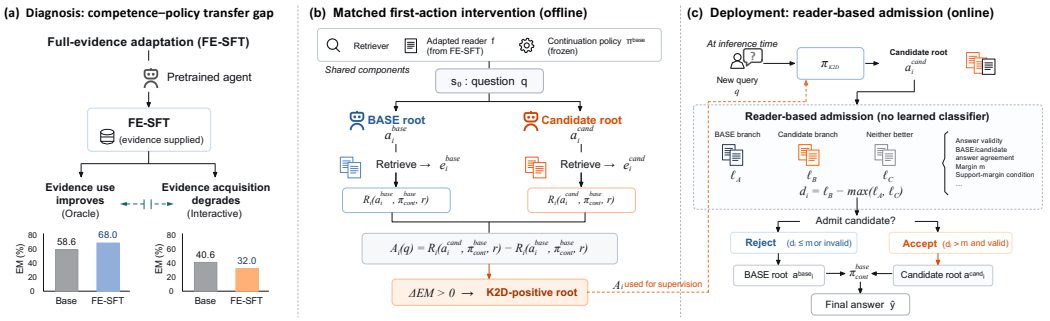


Figure 1: **Overview of K2D-ALIGN.** Offline, matched continuation identifies Base-relative positive roots for supervision. At inference, reader-based admission selects between the learned proposal and exact BASE fallback under the same frozen continuation.

anchor, downstream acquisition is fixed, and only beneficial deviations supervise localized repair. Our full-evidence diagnosis does not imply the same gap under alternative adaptation regimes such as RAFT. Likewise, unlike methods that directly train or control retrieval policies, K2D-ALIGN targets post-adaptation diagnosis and localized replacement. Our empirical controls are frozen, interface-compatible scaffolds and post-adaptation baselines rather than reproductions of trained interactive-search policies or an exhaustive RL-agent leaderboard.

Appendix B compares these approaches by adaptation target and credit unit.

3 K2D-ALIGN

K2D-ALIGN localizes post-adaptation repair to the first retrieval decision while preserving the adapted reader and inherited downstream acquisition policy. Its two stages are shown in Figure 1. Offline, matched continuation compares candidate roots against the inherited root under a fixed downstream policy and provides Base-relative supervision. At inference, reader-based admission selects between the learned proposal and exact BASE fallback.

3.1 SEPARATING EVIDENCE USE FROM ACQUISITION

We separate the adapted agent into a reader that answers from available evidence and an acquisition policy that searches for evidence and decides when to stop. Although both share a backbone in our implementation, their behavior can be controlled independently. We retain the full-evidence-adapted reader and restore the frozen BASE policy for acquisition after the first retrieval decision.

The reader is adapted on complete evidence with

$$\mathcal{L}_{\text{read}} = -\log p_{\theta_r}(y \mid q, \mathcal{E}_{\text{full}}), \tag{1}$$

using a low-rank adapter over the shared backbone (Hu et al., 2022). Candidate roots are therefore compared under the same reader, retriever, and downstream BASE policy, isolating the effect of root replacement.

3.2 MATCHED FIRST-ACTION INTERVENTION

We intervene at the first retrieval action, where candidate and BASE actions originate from the same question-level state. This provides a controlled boundary for measuring the effect of one acquisition decision under a fixed downstream policy.

For case i, let a_i^{BASE} denote the first retrieval action proposed by BASE and a a candidate action. We evaluate both with non-mutating shadow rollouts using the same retriever, adapted reader, and frozen BASE continuation policy. The actions may retrieve different evidence and induce different subsequent states, but all decisions after the root follow the same continuation policy.

We define the Base-relative root advantage as

$$A_i(a) = R_i(a, \pi^{\text{cont}}_{\text{BASE}}, r_{\text{FE}}) - R_i(a_i^{\text{BASE}}, \pi^{\text{cont}}_{\text{BASE}}, r_{\text{FE}}), \tag{2}$$

where R_i is the terminal utility of the resulting rollout. A positive $A_i(a)$ denotes higher collection utility under matched continuation; terminal EM and validity determine pair inclusion separately.

This advantage measures intervention opportunity independently of proposal and admission: weak positive opportunity indicates a proposal bottleneck, whereas unrealized positive opportunity indicates an admission bottleneck.

Evaluable outcomes are used only offline for target construction and are never exposed to deployment-time admission.

3.3 BASE-RELATIVE ROOT SUPERVISION

Matched continuation provides supervision for beneficial deviations from inherited behavior. For each case, we retain candidate roots that improve over the corresponding BASE action:

$$\mathcal{D}_{\text{K2D}} = \left\{ (i, a_i^+, a_i^{\text{BASE}}) : \Delta\text{EM}_i(a_i^+) > 0,\ V_i(a_i^+) = 0,\ a_i^+ \neq a_i^{\text{BASE}} \right\}. \tag{3}$$

Here $\Delta\text{EM}_i(a) = \text{EM}_i(a) - \text{EM}_i(a_i^{\text{BASE}})$, and $V_i(a)$ counts invalid-action, budget-overflow, and post-budget-action indicators, so $V_i(a) = 0$ means valid execution. We call the retained candidates *K2D-positive* roots. Unlike supervision based on absolute trajectory success, this criterion labels an action positive only when its substitution improves the matched downstream outcome.

We fit the proposal policy using root-only weighted DPO: $A_i(a_i^+)$ determines each retained pair's continuous weight, not its inclusion. Root-action SFT is an optimizer control; implementation details are given in Appendix D. For controlled attribution, we also construct Direct and Relation targets without outcome-based selection, and an Absolute-outcome target that selects terminally correct roots without requiring improvement over BASE. RQ3 compares these targets from the same proposal initialization under an otherwise fixed deployment interface.

3.4 PROPOSAL AND ADMISSION

Proposal generates an alternative root action; admission determines whether it replaces the inherited action. At inference, K2D-ALIGN evaluates the candidate and BASE roots using matched shadow rollouts. The admission rule receives the question, retrieved evidence summaries, and reader answers from both branches, but no gold answers, gold titles, offline rewards, or external verifier.

The reader compares the two branches using a three-way prompt: A (BASE is better supported), B (the candidate is better supported), or C (neither is reliably better). Let ℓ_A, ℓ_B, and ℓ_C denote the corresponding reader-adapter log probabilities. We define

$$d_i = \ell_B - \max(\ell_A, \ell_C). \tag{4}$$

For branch $b \in \{B, +\}$, let $h^b_{i,L}$ be the reader's log probability of pointwise support label L: supported (A), contradicted (B), or insufficient evidence (C). Define

$$\rho_i^b = h^b_{i,A} - \max(h^b_{i,B}, h^b_{i,C}). \tag{5}$$

These pointwise labels are distinct from the pairwise response classes above; pairwise C is a critic response, not a third rollout. Invalid or abstaining branch answers receive $\rho_i^b = -10$ in the implementation. Let W_i^+ indicate a valid, non-abstaining candidate answer. The fixed-margin admission rule is

$$\begin{aligned} \Gamma_i &= \left[\text{norm}(\hat{y}_i^+) = \text{norm}(\hat{y}_i^B)\right] \vee [d_i > m] \vee [\rho_i^B < 0 \leq \rho_i^+], \\ a_i^{\text{exec}} &= \begin{cases} a_i^+, & V_i(a_i^+) = 0 \ \wedge \ W_i^+ \ \wedge \ \Gamma_i, \\ a_i^{\text{BASE}}, & \text{otherwise.} \end{cases} \end{aligned} \tag{6}$$

The selected margin is $m = 1.5$. The second case executes exact Base fallback.

All subsequent search and stop decisions follow the frozen BASE policy:

$$\pi_{\text{K2D-ALIGN}}(a_t \mid s_t) = \begin{cases} \pi_{\text{selected}}(a_0 \mid s_0), & t = 0, \\ \pi_{\text{BASE}}(a_t \mid s_t), & t > 0. \end{cases} \tag{7}$$

Thus the intervention changes only the root action; the adapted reader and downstream acquisition policy remain fixed. Forced candidate execution measures available opportunity, whereas routed execution measures the gain realized by admission.

3.5 Interaction and computation

Admission evaluates two non-mutating terminal shadow continuations, rooted at the candidate and Base actions, followed by reader comparison and live execution. We distinguish committed-trajectory retrievals, total retrieval-index invocations including speculative shadows, and model-side calls. Non-mutating shadows still perform retrieval work; this matters for external, metered, or rate-limited retrievers. Historical executed-retrieval logs alone do not determine total index workload. Appendix D.3 specifies the accounting.

4 Experiments

4.1 Experimental setup

Our primary study uses Qwen3-4B on RAG-Gym/HotpotQA with a frozen full-evidence reader, BM25 retrieval ($k = 32$), the native search, stop and answer interface, deterministic decoding, and a five-search budget. Construction and evaluation are strictly separated. The reader is trained on 8,192 cases, while root candidates are collected from 256 disjoint training cases (1,024 sampled roots; 24 retained Base-relative pairs). Configuration selection uses only 128 fixed development cases. The resulting reader, proposal, admission rule, and decoding configuration are then frozen before four no-retuning replications, the 872-case held-out HotpotQA evaluation, and all transfer evaluations. We evaluate this system on 872 disjoint HotpotQA cases and under a common 1,625-case protocol spanning HotpotQA, 2WikiMultiHopQA, MuSiQue, and Bamboogle. We report answer EM/F1, supporting-title recall and precision, executed retrievals, and model-side calls. Unless stated otherwise, comparisons are paired on the same cases. Complete data splits, training settings, baseline protocols, statistical procedures, and interface contracts are provided in Appendices C, D, and G. Invalid actions are scored rather than used as an evaluation gate.

4.2 RQ1: Do adaptation gains transfer uniformly across agent functions?

Evidence use improves while acquisition degrades. Full-evidence adaptation raises Oracle EM from 58.6 to 68.0 but lowers Interactive EM from 40.6 to 32.0 on the same 128 HotpotQA cases (Figure 2a). Replacing FE-SFT acquisition with the inherited Base policy while retaining the adapted reader recovers Interactive EM to 46.1. The reader-side gain therefore survives substitution, localizing the observed degradation to evidence acquisition.

Acquisition traces change accordingly. FE-SFT retrieves complete support less often and stops earlier than BASE (Table 1); these statistics are descriptive rather than causal. RQ3 isolates the supervision-specific increment under matched initialization and downstream components.

4.3 RQ2: Where can acquisition repair be localized?

The root provides a controlled repair boundary. The first retrieval action is the earliest point at which candidate and Base actions share the same state and can be compared under a common frozen continuation. In the 16-case location pilot, a root-trained adapter is transferred zero-shot to later states; root-only intervention is the only tested location with a positive point estimate, although the root–step-1 contrast remains uncertain (95% CI $[-12.5, +25.0]$). We therefore use the root for the present implementation. This pilot does not rule out separately trained intermediate-state interventions.

Repair changes the executed root queries. Table 1c shows that selected roots are longer and contain substantially more question–query relation and entity coverage than the inherited roots, while using fewer executed retrievals. These surface differences characterize the resulting intervention rather than establish its causal mechanism.

(a) Functional factorization, 16 cases

Acquisition	Reader	Changed	EM	Recall	Queries
BASE	BASE	–	31.25	68.75	1.50
BASE	FE-SFT	0/16	**50.00**	68.75	1.50
FE-SFT	FE-SFT	13/16	31.25	50.00	1.25
Process-SFT	FE-SFT	16/16	31.25	56.25	2.00

(b) Acquisition traces, 128 cases

Trace (%)	BASE	FE-SFT
First-query gold hit	91.4	85.2
Full support retrieved	50.8	39.1
One-query stop	46.9	78.9
Second-hop completion	42.4	21.3

(c) Root-query surface audit, 128 cases

	Interactive outcome			Executed query surface			
System	EM	Recall	Queries	1-query (%)	Query length	Relation coverage	Entity coverage
BASE	40.6	72.3	1.8	46.9	5.4	27.9	68.6
Base+FE	46.1	72.3	1.8	46.9	5.4	27.9	68.6
K2D-ALIGN	**52.3**	**76.6**	1.5	60.9	11.9	76.3	89.8
Paired change	+6.25	+4.3	−0.3	+14.0	+6.5	+48.4	+21.2

Table 1: **Localizing and repairing the acquisition-side mismatch.** (a) Reader substitution isolates acquisition changes. (b) FE-SFT retrieves complete support less often and stops earlier; second-hop completion conditions on partial first-query hits ($n = 92$ BASE, 75 FE-SFT). (c) K2D-ALIGN changes root-query behavior while improving Interactive EM and recall over Base+FE. Statistics in (b–c) are descriptive.

	Answer quality		Acquisition			EM gain (pp)	
Root target	EM	F1	Recall	Prec.	Queries	vs Base+FE	vs zero-update
Zero-update initializer	48.44	62.41	76.17	6.15	1.53	+2.34	0.00
Direct	48.44	62.93	76.17	6.14	1.53	+2.34	0.00
Relation	48.44	62.93	75.78	6.12	1.51	+2.34	0.00
Absolute-outcome	48.44	62.41	76.17	6.15	1.53	+2.34	0.00
K2D-positive	**52.34**	64.41	76.56	6.18	1.47	+6.25	**+3.91**

Table 2: **Matched-initialization root-target attribution.** All rows share the inherited initialization, admission rule, reader, retriever, continuation, and evaluator; trained variants use the same 24 states and 48-update budget. Only K2D-positive supervision improves EM beyond the zero-update initializer ($+3.91$; five wins, zero losses; paired 95% bootstrap CI $[+0.78, +7.81]$, $P_{\text{boot}}(\Delta\text{EM} > 0) = 0.994$).

4.4 RQ3: CAN LOCALIZED REPAIR IMPROVE ON INHERITED ACQUISITION?

Localized repair improves interactive performance. Table 3 compares K2D-ALIGN with the matched Base+FE reference and retrieval baselines under the same adapted reader. K2D-ALIGN improves Interactive EM from 46.1 to 52.3 and F1 from 58.6 to 64.4, while supporting-title recall increases from 72.3% to 76.6%. Direct and Oracle performance remain unchanged, and all acquisition decisions after the root continue to follow the frozen BASE policy. The recovery therefore comes from the localized acquisition intervention rather than replacing the transferred reader or downstream policy.

Base-relative supervision contributes beyond initialization. Under the matched conditions in Table 2, Direct, Relation, and Absolute-outcome targets remain at the 48.44-EM zero-update level. Only Base-relative K2D-positive supervision reaches 52.34 EM. Its $+3.91$-point increment is development-specific; the full $+6.25$ system gain also includes the inherited proposal/admission configuration.

Development selection. The selected development system gains $+6.25$ EM (paired 95% CI $[+2.34, +10.94]$), with eight repairs and no regressions against Base+FE. Because the admission margin $m = 1.5$ was selected on these cases, this complete-system contrast is descriptive rather than

	Reader		Interactive		Evidence seeking		Efficiency		Gain vs Base+FE	
Method	Direct	Oracle	EM	F1	Gold-title recall	Precision	Live queries	Direct calls	ΔEM	ΔRecall
BASE	17.2	58.6	40.6	52.4	72.3	5.4	1.76	5.52	–	–
FE-SFT	18.0	68.0	32.0	42.6	64.1	5.9	1.22	4.44	−14.1	−8.2
Base+FE	18.0	68.0	**46.1**	58.6	72.3	5.4	1.76	5.52	0.0	0.0
RAG-once	18.0	68.0	39.1	48.7	69.1	6.3	1.00	2.00	−7.0	−3.1
Prompt-RW	18.0	68.0	46.9	57.4	74.2	6.0	1.58	5.16	+0.8	+2.0
ReAct-style	18.0	68.0	44.5	56.0	71.5	5.6	1.80	5.59	−1.6	−0.8
RootOnly	18.0	68.0	46.9	58.0	75.8	5.7	1.54	5.08	+0.8	+3.5
K2D-ALIGN	18.0	68.0	**52.3**	**64.4**	**76.6**	6.2	1.47	N/A	**+6.25**	+4.3

Table 3: 128-case HotpotQA development results (seed 13; five-search budget). Base+FE is the matched-reader reference; Prompt-RW is prompted rewriting and RootOnly preserves the answer. Live queries and direct calls exclude shadow computation. The 52.34-EM K2D run lacks a complete call trace; Figure 2d uses a separate 50.78-EM audit. Paired ReAct-style CIs are in Appendix G.

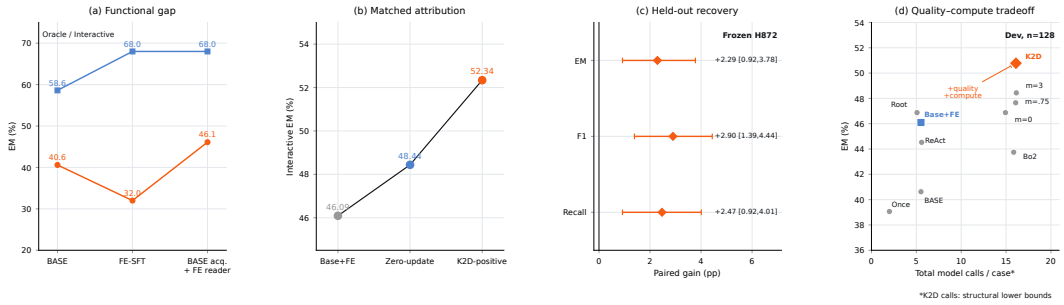


Figure 2: Development and frozen evaluation. (a) Functional gap. (b) Development EM decomposition: +2.34 from inherited configuration and +3.91 from supervision. (c) H872 paired gains (95% CIs). (d) EM versus model calls in a separate 50.78-EM trace-complete audit; K2D cost is lower-bounded. The 52.34-EM run in (b) lacks a complete cost trace.

held-out evidence. RQ4 evaluates the frozen configuration outside development. Component and admission ablations are reported in Appendix D.1.

4.5 RQ4: DOES LOCALIZED REPAIR SURVIVE HELD-OUT AND COMPUTE-MATCHED CONTROLS?

The gain persists on held-out HotpotQA. With the configuration frozen after development, K2D-ALIGN improves Interactive EM over Base+FE from 40.83 to 43.12 on 872 disjoint cases (+2.29 points; 95% CI $[+0.92, +3.78]$), with 31 repairs and 11 regressions (Table 4). F1 increases from 53.13 to 56.03 and supporting-title recall from 74.77% to 77.24%, while executed retrievals decrease by 0.196 per case. Relative to directly deployed FE-SFT (32.34 EM), the frozen system gains +10.78 EM (95% CI $[+7.80, +13.76]$), with 142 repairs and 48 regressions. Replication statistics are reported in Appendix A.

Frozen controls separate training from generic lookahead. On the same H872 cases, zero-update admission reaches 41.75 EM, a +0.92-point gain over Base+FE, while trained K2D-ALIGN reaches +2.29. Their +1.37-point aggregate difference has no separate paired-CI claim. Base Best-of-2 samples two frozen Base roots with the same critic and admission rule. It uses 15.8 model calls per case versus 16.1 for K2D-ALIGN, but attains only 40.03 EM; K2D-ALIGN exceeds this approximately compute-matched control by +3.1 EM (paired 95% CI $[+1.4, +4.8]$). Thus neither the inherited proposal/admission system without updates nor the tested extra-branching control accounts for the full frozen gain (Appendix A.1).

	Answer and evidence			Cost	EM vs Base+FE		Paired outcomes
Method	EM	F1	Recall	LM calls	ΔEM	95% CI	Rep./Reg.
Base+FE	40.83	53.13	74.77	6.1	–	–	–
Zero-update + admission	41.75	54.26	–	–	+0.92	[0.31,2.15]	19/11
Base Best-of-2	40.03	52.51	74.37	15.8	-0.80	[-2.5,+0.9]	–
K2D-ALIGN	**43.12**	**56.03**	**77.24**	16.1	**+2.29**	[0.92,3.78]	**31/11**

Table 4: **Frozen H872 evaluation and alternative-explanation controls.** All rows use the same 872 cases and adapted reader. EM differences and paired 95% bootstrap CIs use Base+FE as reference; repairs/regressions count wrong-to-correct/correct-to-wrong changes against it. Zero-update retains the proposal/admission interface without root updates; Base Best-of-2 tests approximately matched inference lookahead. LM calls are per case; unrecorded quantities are left blank, not imputed.

Compute qualifies the recovery. The comparable H872 call counts do not make K2D-ALIGN an efficiency method. A separate trace-complete 128-case development audit records at least 16.10 model calls and 4.73 index invocations per K2D case versus 5.52 and 1.76 for Base+FE; mean wall time rises from 12.00 to 15.67 seconds. Its measured point is not Pareto-dominated among measured development methods on EM/calls, F1/calls, EM/time, or recall/index invocations (Figure 2d; Appendix 4). This audit does not cover H872 or the later 52.34-EM replay.

Independent root-repair confirmation. We use the 24 retained pairs in Table 2 specifically for controlled supervision attribution, and separately test reproducibility with an independently collected, implementation-audited set of 192 balanced preferences (96 candidate-winning and 96 Base-winning). Across two sealed-v3 partitions (256 paired cases), frozen repair improves EM by 3.91 and F1 by 3.52 points while reducing executed retrievals by 0.332 per case ($P_{\text{boot}}(\Delta\text{EM} > 0) = 0.9555$; $P_{\text{boot}}(\Delta\text{F1} > 0) = 0.9607$). This independently collected confirmation tests reproducibility of root repair, not the target-specific attribution made by the 24-pair comparison.

The frozen system transfers across four retrieval benchmarks. Under the shared 1,625-case protocol, K2D-ALIGN reaches 32.6/43.4 macro-average EM/F1 versus 30.7/41.5 for Base+FE (Table 5). It leads on three of four benchmarks in EM and all four in F1; MuSiQue EM falls by 0.6 points. FE-SFT obtains 26.1/35.2, and unconditional Scaled Root-DPO 26.2/33.8. The system gain requires more LM calls (17.71 versus 6.08) with similar executed retrievals (1.968 versus 2.038). Protocol details are in Appendix C.

4.6 RQ5: WHERE DOES FUNCTIONAL TRANSFER BREAK DOWN?

Native-corpus check and functional boundaries. The four-benchmark MuSiQue arm uses the HotpotQA index to test frozen cross-corpus behavior. With BM25 rebuilt from MuSiQue contexts on a separate 128-case slice, K2D-ALIGN reaches 26.56/40.91 EM/F1 versus 22.66/37.16 for ReAct-style and 25.00/39.16 for IRCoT-style, with zero validity violations (Appendix G, Table 24). These scaffold differences are descriptive; their paired EM intervals include zero. Other diagnostics show why transfer need not be uniform: FHIR factorization raises Interactive accuracy by 23.44 points but Oracle accuracy by only 5.21 without running K2D; MuSiQue's FE-SFT-relative root-repair contrast is zero at four hops after positive two- and three-hop contrasts; and Hermes forced candidates gain 15.63 EM points (95% CI $[+3.13, +28.13]$) while admission uses only 7/32 and realizes none of the gain (Figure 3; Appendix F). These are mechanism and boundary diagnostics, not pooled efficacy estimates.

5 SCOPE AND LIMITATIONS

The strongest mechanism and supervision-attribution evidence comes from the controlled Qwen3-4B/HotpotQA study. The frozen four-benchmark evaluation broadens system-level evidence, while FHIR, Qwen2.5, Qwen3-8B, MuSiQue depth, and Hermes primarily characterize functional boundaries. In particular, the evaluated 8B shard shows only a small directional gain ($+0.78$ EM). We therefore do not claim model-, dataset-, or framework-independent gains.

Method	HotpotQA EM	HotpotQA F1	2Wiki EM	2Wiki F1	MuSiQue EM	MuSiQue F1	Bamboogle EM	Bamboogle F1	Macro Avg. EM	Macro Avg. F1	Cost LM	Cost Ret.
BASE	32.6	46.5	17.6	31.2	11.6	18.3	27.2	40.3	22.3	34.1	6.08	2.038
FE-SFT	35.4	47.2	32.8	40.4	9.0	15.9	27.2	37.2	26.1	35.2	5.04	1.527
BASE acq. + FE reader	41.4	54.5	40.2	48.6	**13.2**	20.4	28.0	42.6	30.7	41.5	6.08	2.038
Scaled Root-DPO	36.8	46.6	35.0	42.4	11.4	17.5	21.6	28.9	26.2	33.8	6.08	2.040
K2D-ALIGN	**43.0**	**55.8**	**41.8**	**50.3**	12.6	**20.8**	**32.8**	**46.8**	**32.6**	**43.4**	17.71	1.968

Table 5: **Frozen four-benchmark evaluation.** Unweighted macro averages over a shared 1,625-case protocol. Executed retrievals exclude speculative shadows. MuSiQue uses the Wikipedia-HotpotQA index and therefore measures cross-corpus, not native-corpus, transfer.

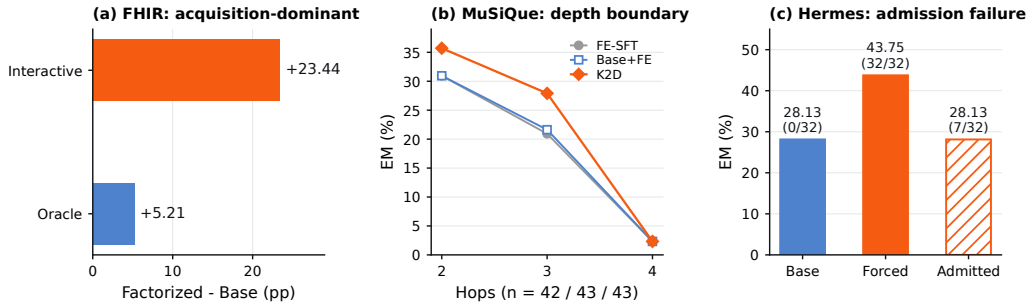


Figure 3: **Complementary functional boundaries, not pooled efficacy results.** (a) FHIR factorization on 64 cases (three runs; no K2D). (b) MuSiQue EM by hop depth on a separate 128-case slice. (c) Hermes forced candidates versus admission on 32 cases; parentheses show candidate use. Exact values and protocols are in Appendix F.

K2D-ALIGN trades speculative model and index work for acquisition recovery. Complete compute accounting is available for the matched development audit; the later selector replay lacks a corresponding full cost trace. We therefore make no inference-efficiency claim.

Finally, we intervene only at the first retrieval decision and use offline evaluable outcomes to identify beneficial roots. The 24-state attribution is development-specific and does not establish the same supervision advantage for other state sets, models, or later decisions. Extending repair beyond the root and improving admission transfer remain open problems.

6 CONCLUSION

Full-evidence adaptation can improve evidence use while degrading interactive acquisition. Component substitution localizes this competence–policy gap; K2D-ALIGN then preserves the adapted reader, learns Base-relative first-action proposals under matched continuation, and admits candidates with exact Base fallback. Matched development controls isolate an incremental root-supervision gain beyond inherited configuration, while frozen HotpotQA evaluation confirms the complete system's improvement. The four-benchmark result is broadly positive but non-uniform, and FHIR and Hermes show that reader–acquisition mismatch, proposal opportunity, and admission realization need not coincide.

These results support factorized repair, not universal transfer or inference efficiency. Later-action repair and more reliable admission without losing exact Base fallback remain open. Post-adaptation agents should be judged by *what transferred, where repair is available, and whether deployment realizes it.*

ETHICS STATEMENT

The FHIR diagnostic uses patient-disjoint cases from MIMIC-IV-FHIR and reports aggregate results; no patient records are reproduced here. The evaluated agent performs read-only retrieval and answer generation. These constraints limit the scope of the analysis but do not establish clinical suitability.

AI USE STATEMENT

Generative AI tools assisted with language editing, code organization, and early visual prototyping. All scientific decisions, experiments, analyses, numerical claims, and final figures were produced or verified by the authors, who take responsibility for the manuscript.

REPRODUCIBILITY STATEMENT

We provide the data splits, model and adapter configurations, retrieval interfaces, prompts, training hyperparameters, evaluation protocols, and statistical procedures needed to reproduce the reported experiments in Appendices C–G. The experimental pipeline fixes the reader before root collection, constructs Base-relative supervision only from disjoint training cases, selects the proposal and admission configuration on the designated development set, and freezes all resulting choices before held-out and transfer evaluation. Appendix C records case-set and interface identifiers together with the frozen evaluation configurations, while Appendix D specifies training, decoding, admission, metric, bootstrap, and computation-accounting procedures. Reported evaluations preserve per-case traces and use paired comparisons on fixed case IDs where applicable. An anonymized repository containing the implementation, configuration files, evaluation manifests, and reproduction scripts is available at https://anonymous.4open.science/r/k2d-align-EFC7.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. ReSearch: Learning to reason with search for LLMs via reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2858. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/7b3a0d3a0864b66f229c0a9c84f9e950-Abstract-Conference.html.

Yanjun Chen, Yirong Sun, Hanlin Wang, Jinghan Wang, Xinming Zhang, Xiaoyu Shen, Wenjie Li, and Wei Zhang. Exact is easier: Credit assignment for cooperative LLM agents. *arXiv preprint arXiv:2603.06859*, 2026. doi: 10.48550/arXiv.2603.06859. URL https://arxiv.org/abs/2603.06859.

Zhiyuan He, Siwei Zhang, Zhiwen Zhou, Yuqing Yang, Yu Kang, Yuge Zhang, Luna K. Qiu, Tin Yan Tsui, Jiahang Xu, and Chong Luo. Agent lightning v1.0: Towards harnessed agentic RL. *arXiv preprint arXiv:2608.17528*, 2026. doi: 10.48550/arXiv.2608.17528. URL https://arxiv.org/abs/2608.17528.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jerry Huang, Siddarth Madala, Risham Sidhu, Cheng Niu, Hao Peng, Julia Hockenmaier, and Tong Zhang. Tackling distractor documents in multi-hop QA with reinforcement and curriculum learning. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 5548–5561. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-eacl.294. URL https://aclanthology.org/2026.findings-eacl.294/.

Yulong Hui, Chao Chen, Zhihang Fu, Yihao Liu, Jieping Ye, and Huanchen Zhang. Interact-RAG: Reason and interact with the corpus, beyond black-box retrieval. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b61f288da3c106f65d57b0d45b470b6b-Abstract-Conference.html.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.389. URL https://aclanthology.org/2024.naacl-long.389/.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O. Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Rwhi91ideu.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Zibin Meng and Kani Chen. CRAFT: Counterfactual credit assignment from free sibling rollouts for self-distilled agentic reinforcement learning. *arXiv preprint arXiv:2606.29476*, 2026. doi: 10.48550/arXiv.2606.29476. URL https://arxiv.org/abs/2606.29476.

Pingzhi Tang, Yiding Wang, and Muhan Zhang. Knowledge is not enough: Injecting RL skills for continual adaptation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11969–11997. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.550. URL https://aclanthology.org/2026.acl-long.550/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.557. URL https://aclanthology.org/2023.acl-long.557/.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Yin Fang, Haolin Liu, Yifan Yang, Fangyuan Chen, Zhixing Song, Dengyu Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. Supervising the search process produces reliable and generalizable information-seeking agents. *arXiv preprint arXiv:2502.13957*, 2026. URL https://arxiv.org/abs/2502.13957.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. RAFT: Adapting language model to domain specific RAG. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=rzQGHXNReU.

Zhi Zheng, Rongsheng Chen, Yunpeng Ba, Zhenkun Wang, Yee Whye Teh, and Wee Sun Lee. Agentic ESOpt: Fine-tuning long-horizon LLM agents with minimal GPU requirements. *arXiv preprint arXiv:2608.17310*, 2026. doi: 10.48550/arXiv.2608.17310. URL https://arxiv.org/abs/2608.17310.

A Replication and Held-Out Statistics

Eval.	Reference	r	n	EM	ΔEM	Rec.	ΔRec.	Purpose
Dev., budget 5	Base+FE	1	128	52.34	+6.25	76.56	+4.30	selection
Dev., budget 4	Base+FE	2	128	47.27 ± 0.55	$+12.50 \pm 3.31$	74.81 ± 0.28	$+12.70 \pm 0.83$	replication
Dev., budget 8	Base+FE	2	128	48.83 ± 0.55	$+13.29 \pm 5.52$	75.39 ± 0.00	$+13.09 \pm 1.93$	replication
Held-out	**Base+FE**	**1**	**872**	**43.12**	+2.29	**77.24**	+2.47	**intervention**
Held-out	FE-SFT (direct)	1	872	43.12	+10.78	77.24	+15.37	end-to-end repair

Table 6: Replication and held-out results for the selected configuration. r is the number of independent runs, "Cases" is the paired set size, and Δ uses the reference in column two. The two held-out rows report the incremental K2D–Base+FE effect and the separate K2D–direct-FE repair effect. Values are mean $\pm$ sample standard deviation for $r \geq 2$; held-out cases remain excluded from selection.

A.1 Frozen held-out attribution and compute controls

The H872 attribution protocol compares Base acquisition + FE reader, the zero-update inherited proposal with frozen admission, and K2D-Align on the same 872 IDs. All settings are inherited from development. Repairs and regressions are paired against Base+FE; EM differences use 50,000 paired bootstrap draws and percentile 95% intervals. The zero-update comparison isolates the effect of learned root supervision within the fixed proposal/admission system.

Method	EM	F1	ΔEM vs. Base+FE	95% CI	Repairs	Regressions
Base acq. + FE reader	40.83	53.13	–	–	–	–
Zero-update + admission	41.75	54.26	+0.92	[0.31, 2.15]	19	11
K2D-Align	43.12	56.03	+2.29	[0.92,3.78]	31	11

Table 7: Frozen H872 attribution under a shared reader, continuation, and admission configuration. Differences and repair counts use Base+FE.

The high-compute Base Bo2 control reuses the development selector-only mechanism: two independent frozen-Base root samples (temperature 0.7, top-p 0.95, top-k 20), each completed with Base continuation and the same FE reader, followed by the same A/B/C critic and fixed $m = 1.5$ admission. No K2D proposal adapter is loaded. H872 IDs, BM25 interface ($k = 32$), five-search budget, reader, prompts, and downstream decoding are identical to the frozen H872 protocol. It is an approximately inference-compute-matched design; realized model calls, rather than the two-branch structure alone, determine the degree of matching. Table 8 separates quality from executed retrievals, total index invocations, and model calls. Each paired comparison uses the same 50,000 resampled case indices and percentile 95% intervals.

(a) Quality and inference resources

Method	EM	F1	Recall	Exec. retrievals	Total index calls	Model calls
Base acq. + FE reader	40.83	53.13	74.77	2.04	2.04	6.1
Base Bo2 + reader selection	40.03	52.51	74.37	1.96	3.92	15.8
K2D-Align	43.12	56.03	77.24	1.84	4.90	16.1

(b) Paired Base Bo2 differences

Reference	ΔEM	95% CI	ΔF1	95% CI	ΔRecall	95% CI
Base acq. + FE reader	-0.8	[-2.5, +0.9]	-0.6	[-2.0, +0.8]	-0.5	[-2.0, +1.1]
K2D-Align	-3.1	[-4.8, -1.4]	-3.5	[-5.0, -2.0]	-2.9	[-4.5, -1.3]

Table 8: H872 high-compute Base Bo2 protocol. Resource counts are per case; total index invocations include speculative shadows. Panel (b) is Bo2 minus the named reference. Recorded execution counts alone do not supply total retrieval workload.

Four-benchmark compute-control protocol. The same frozen Base Bo2 mechanism can be evaluated on the 1,625-case protocol in Table 5, using each benchmark's prescribed IDs and shared retrieval interface, without a proposal adapter or benchmark-specific tuning. EM/F1 and available support recall, executed retrievals, total index invocations, model calls, and paired differences against Base+FE and K2D separate generic lookahead compute from K2D-specific proposal learning. This protocol is independent of the reported four-benchmark results.

B DESIGN-SPACE COMPARISON

Line of work	Primary adaptation target	Credit / control unit	Relationship to K2D-ALIGN
RAFT (Zhang et al., 2024)	Evidence-conditioned reader	Training documents with distractors	Reader adaptation; no acquisition-transfer estimate.
IRCoT / Self-RAG (Trivedi et al., 2023; Asai et al., 2024)	Retrieve–reason behavior	Generated retrieval trajectory	Retrieval control; K2D freezes continuation.
Adaptive-RAG (Jeong et al., 2024)	Retrieval strategy selection	Query complexity / strategy	Strategy choice; K2D compares to a Base root.
Search-R1 / ReSearch / RAG-RL (Jin et al., 2025; Chen et al., 2025; Huang et al., 2026)	Full interactive policy	End-to-end rollout reward	Whole-policy RL; K2D tests root-relative credit at one boundary.
Agent Lightning (He et al., 2026)	Harness-connected policy	Trainer-observed agent traces	RL infrastructure; K2D defines a Base-matched intervention.
Agentic ESOpt (Zheng et al., 2026)	Full model trajectory	Perturbation-level return	Trajectory optimizer; K2D fixes one intervention.
Counterfactual credit (Chen et al., 2026; Meng & Chen, 2026)	Messages or tokens	Matched alternatives	Shared logic; K2D targets evidence acquisition.
K2D-ALIGN	Reader + root action	Base root vs. candidate root	Base is anchor, continuation, and fallback.

Table 9: Design-space comparison by adaptation target and credit unit. Numerical results in Table 3 use the matched native interface.

C REPRODUCIBILITY DETAILS

The development case file has SHA-256 prefix `54422f...acf57d`; all selection settings share interface hash `03cec...0659031c`. The frozen 872 K2D and direct-FE traces use this same native-interface hash. The later same-reader Base+FE control is checked field-by-field against this shared task contract (case IDs/hash, corpus, BM25 k, budget, native action schema, decoding, and token limits). Its evaluator manifest additionally serializes disabled root-proposal fields, so its full manifest hash is reported rather than treated as an equality requirement. The MuSiQue bridge case file has SHA-256 prefix `0a87c6...5dfe167ad` and uses the same native interface hash. We preserve per-case traces, paired summaries, and interface-validity records for every reported result. All numerical tables are derived from these records. The 2Wiki question file has SHA-256 prefix `13c458...a7ef` and uses the same interactive interface hash. The HotpotQA canonical-baseline bootstrap and admission-cost summary is stored at `runs/iclr/k2d_align_hotpot_baseline_bootstrap_cost_20260814.json`. The prompted-query-rewrite baseline is stored at `runs/iclr/k2d_align_prompted_rewrite_baseline_hotpot128_seed13_b5`.

Frozen community evaluation. The community matrix follows `configs/community_evaluation_protocol_v1.json`: 500 HotpotQA, 500 2WikiMultiHopQA, 500 MuSiQue, and 125 Bamboogle cases, for 1,625 cases per method. All rows use the same native interactive interface, BM25 over Wikipedia-HotpotQA ($k = 32$), five-search budget, and seed 13. Thus the 500-case MuSiQue arm tests cross-corpus/out-of-index retrieval transfer. Per-case records and summaries are stored by method and benchmark under `runs/community_aligned_k2d_v1/final_eval/`; the protocol manifest records the common interface contract. EM and F1 macro averages are unweighted means over the four benchmark values. LM calls are the per-case recorded model-call mean; executed retrievals are the corresponding per-case retrieval-query mean. Invalid actions are retained in the per-case records and scored by the common evaluator; they are not used as an exclusion or evaluation gate. The frozen rows use the BASE, FE-SFT, BASE acquisition + FE reader, Scaled Root-DPO, and K2D-ALIGN checkpoints/configurations specified in `configs/community_final_eval_queue_v1.json`; that configuration also records the Root-DPO and FE-SFT adapter sources.

Evaluation-set construction and transfer boundaries. The primary HotpotQA development matrix uses 128 fixed case IDs under seed 13 and budget five. The 872-case HotpotQA slice is disjoint from the development matrix; each fixed evaluation arm is executed once without changing the selected configuration. The Bamboogle, 2WikiMultiHopQA, and MuSiQue evaluations reuse the selected HotpotQA-trained reader, root adapter, admission margin, native action schema, and decoding rule without parameter updates or margin retuning. The 2Wiki comparison changes the question distribution while keeping the HotpotQA retrieval corpus and BM25 interface fixed. The separate 128-case MuSiQue comparison changes both questions and corpus: we use 128 fixed cases from a public source-compatible MuSiQue-Answerable development mirror and build a deterministic paragraph-pool BM25 index only from MuSiQue contexts. Same-ID ReAct-style and IRCoT-style runs use that same reader, retriever, native action schema, and budget. The hop-stratified MuSiQue profile uses a separate deterministic sample with 42/43/43 cases from the 2/3/4-hop strata and evaluates all methods on the same IDs. The fixed-query BGE substitution evaluates the 225 policy queries emitted by the frozen 128-case HotpotQA rollouts against a retrieved-candidate pool of 8,572 documents; it is a ranking-backend sensitivity check conditional on those queries, not a full-corpus retriever-transfer evaluation.

Hermes external diagnostic. The Hermes cross-framework diagnostic uses the official Hermes v0.20.5 checkout at revision `64a6f42` with Qwen3.5-35B-A3B-FP8, Hermes-native MCP search, BM25 top-32 retrieval, and a four-search cap. The deterministic HotpotQA split allocates 64 training cases, 16 calibration cases, and 32 fresh held-out evaluation cases after excluding earlier Hermes protocol IDs. The candidate branch executes a relation-preserving first retrieval root; the Base branch executes the frozen Hermes root. Both branches then hand control to a fresh frozen Base Hermes continuation for all later search, stop, and answer actions. The admission rule observes no gold answer, supporting title, EM/F1, or held-out outcome at deployment. All reported held-out branches have valid root divergence, one finish, no invalid actions, no overflow, and no post-budget actions.

D Implementation and Statistical Details

This appendix gives formal definitions and implementation details for the description. For case i, let $a_i^B \equiv a_i^{\text{BASE}}$ denote the decoded Base root and let $\mathcal{C}_i$ be the deduplicated candidate roots. Every candidate and the Base root is evaluated in a non-mutating shadow branch using the native retriever/summarizer, the same frozen Base continuation, and the same reader. The selected root is then executed on the live interface. We write

$$U_i(a) = R_i\big(a, \pi_{\text{BASE}}^{\text{cont}}, r_{\text{FE}}\big), \qquad A_i(a) = U_i(a) - U_i(a_i^B), \tag{8}$$

where R_i is the collection utility from Eq. 9. Thus, A_i is a matched, Base-relative utility difference used to weight retained training pairs. The K2D-positive selection criterion itself is the matched terminal-answer improvement $\Delta\text{EM}_i(a) > 0$, defined below. During collection, the behavior adapter samples four root actions per case on 256 training cases (1,024 root rollouts total, temperature 0.7). Each sampled root is paired with the deterministic Base root anchor for the same case; both branches then use the same terminal shadow procedure with the frozen Base continuation policy and frozen full-evidence reader. The collection reward is

$$\begin{aligned} R_i(a, \pi_{\text{BASE}}^{\text{cont}}, r_{\text{FE}}) = {} & 1.00R_{\text{ans}} + 0.75R_{\text{bridge}} + 0.75R_{\text{cont}} + 0.25R_{\text{rel}} + 0.75R_{\text{stab}} \\ & - 1.00C_{\text{invalid}} - 0.02N_{\text{query}} - 0.05N_{\text{repeat}}, \end{aligned} \tag{9}$$

Operational reward definitions. Let $L_i^r(E)$ be the mean token log probability assigned by reader r to the gold-answer JSON target `{"predicted_answer":answer}` after the question and evidence-prefix prompt. Readers $r \in \{\text{FE}, \text{BASE}\}$ score the same target. Write $E_0 = \varnothing$, E_1 for the root-retrieval prefix, E_T for terminal evidence, and $E_\star$ for full evidence. With $c(x) = \text{clip}(x, -1, 1)$,

the collection code defines

$$\begin{aligned}
\phi_i(E) &= c\left(\frac{L_i^{\mathrm{FE}}(E) - L_i^{\mathrm{FE}}(E_0)}{\max(|L_i^{\mathrm{FE}}(E_\star) - L_i^{\mathrm{FE}}(E_0)|, 0.05)}\right), \\
D_i(E) &= [L_i^{\mathrm{FE}}(E) - L_i^{\mathrm{FE}}(E_0)] - [L_i^{\mathrm{BASE}}(E) - L_i^{\mathrm{BASE}}(E_0)], \\
R_{\mathrm{bridge}} &= c\left(\frac{D_i(E_T)}{\max(|D_i(E_\star)|, 0.05)}\right), \qquad R_{\mathrm{ans}} = \mathbb{1}[\mathrm{norm}(\hat{y}_i) = \mathrm{norm}(y_i)], \\
R_{\mathrm{cont}} &= \phi_i(E_T) - \phi_i(E_{\min(1,T)}), \qquad R_{\mathrm{stab}} = -\max\left(0, \max_{0 \le t < T} \phi_i(E_t) - \phi_i(E_T)\right), \\
R_{\mathrm{rel}} &= 0.7\frac{|\mathcal{L}(q_i) \cap \mathcal{T}(a)|}{\max(1, |\mathcal{L}(q_i)|)} + 0.3\frac{|\mathcal{Q}(q_i) \cap \mathcal{T}(a)|}{\max(1, |\mathcal{Q}(q_i)|)}.
\end{aligned} \tag{10}$$

For $T = 0$, the earlier-prefix maximum is $\phi_i(E_0)$. Here $\mathcal{T}(a)$ contains case-folded query tokens and $\mathcal{Q}(q_i)$ contains question tokens outside the implementation's fixed stopword list. $\mathcal{L}(q_i)$ is the first eight distinct lowercase, non-stopword, non-numeric question tokens of length at least three, excluding capitalized entity-like tokens; if empty, it uses the first eight non-stopword question tokens. The exact tokenizer and stopword list are in `src/k2d_gap/rag_gym_policy.py`, function `relation_preservation`. Non-search roots receive $R_{\mathrm{rel}} = 0$. This lexical term scores the root query, not retrieved-document support. C_{invalid} counts invalid native actions; N_{query} counts branch search calls and N_{repeat} counts repeated queries. Normalization uses the empty and full evidence scores only offline. These definitions follow `scripts/run_rag_gym_root_grpo.py` and `src/k2d_gap/rag_gym_grpo.py`.

Reward coefficients. The coefficients in Eq. 9 were specified once as heuristic design constants and were not tuned against development or held-out performance. They encode a coarse preference ordering for offline candidate construction rather than an optimized or calibrated task reward.

Retained-pair weights clip positive Base-relative advantage at 5.0 and use a minimum normalization scale of 0.1, as in Eq. 13. No gold title, answer, or utility is placed in an inference prompt.

The reward fields are computed from the per-case trace and expose the components of the scalar. R_{ans} is the final exact-match indicator; R_{cont} is final answer utility minus root-only answer utility; R_{bridge} is the Base-anchored difference-in-differences change in evidence utility; R_{rel} is the question–root-query relation-preservation score; and $R_{\mathrm{stab}} = -\max(0, \max_{t<T} \phi_i(E_t) - \phi_i(E_T))$ penalizes a regression from the best earlier evidence prefix within that branch. Table 10 summarizes these fields. The cost terms are C_{invalid} (invalid native actions), N_{query} (environment search calls), and N_{repeat} (repeated-query calls). All fields are logged before aggregation, so candidate ranking and subsequent analysis use the same definitions.

Reward field	Operational definition
R_{ans}	Final exact-match indicator.
R_{bridge}	Base-anchored difference-in-differences evidence-utility change.
R_{cont}	Final answer utility minus root-only answer utility.
R_{rel}	Lexical question–root-query relation preservation (Eq. 10).
R_{stab}	Prefix-potential regression penalty in Eq. 10.
C_{invalid}	Count of invalid native actions; N_{query} and N_{repeat} count search and repeated-query calls.

Table 10: Per-case collection-reward fields computed from native interaction traces.

The validity indicator is

$$V_i(a) = \mathbb{1}[\text{invalid action}] + \mathbb{1}[\text{budget overflow}] + \mathbb{1}[\text{post-budget action}], \tag{11}$$

and the training mask is

$$z_i(a) = \mathbb{1}\left[\Delta\mathrm{EM}_i(a) > 0 \ \wedge \ V_i(a) = 0 \ \wedge \ a \neq a_i^B\right]. \tag{12}$$

This is the K2D-positive filtering rule used in the main text: a candidate is retained only when replacing the Base root increases terminal exact match under matched continuation, while remaining valid and distinct from the Base action: the training set is $\mathcal{P} = \{(i, a) : z_i(a) = 1\}$ after duplicate

removal, and contains 24 pairs. For each retained pair, our weighted-DPO implementation uses the weight

$$w_i(a) = \frac{\mathrm{clip}(A_i(a), 0, 5)}{\max(0.1,\ |\mathcal{P}|^{-1} \sum_{(j,b)\in\mathcal{P}} \mathrm{clip}(A_j(b), 0, 5))}. \tag{13}$$

The denominator is computed once from the training collection and is not recomputed on held-out data.

Equivalently, for each training case the root candidate set is

$$\begin{aligned} \mathcal{C}_i &= \{a_i^B\} \cup \{a_{i,1}, a_{i,2}, a_{i,3}, a_{i,4}\}, \\ \mathcal{P} &= \{(i, a_{i,k}) : \Delta\mathrm{EM}_i(a_{i,k}) > 0,\ V_i(a_{i,k}) = 0,\ a_{i,k} \neq a_i^B\}, \end{aligned} \tag{14}$$

after duplicate-action removal. Thus $|\mathcal{P}| = 24$ is the final K2D-positive training set selected by matched terminal-EM improvement; $A_i(a)$ does not determine pair inclusion.

For a selected pair, let π_0 be the frozen initial behavior adapter used as the DPO reference. We use the weights $w_i(a)$ from Eq. 13; pair inclusion is determined by $\Delta\mathrm{EM}_i(a) > 0$, while $A_i(a)$ controls the relative weight assigned to retained pairs. The implemented proposal-fitting objective is the root-only weighted DPO loss

$$\begin{aligned} \mathcal{L}_{\mathrm{K2D}} = -\frac{1}{|\mathcal{P}|} \sum_{p=1}^{|\mathcal{P}|} w_{i_p}(a_p^+) \log \sigma\Big(\beta\big[&\log \pi_{\theta_p}(a_p^+|s_{0,i_p}) - \log \pi_{\theta_p}(a_p^-|s_{0,i_p}) \\ &- \log \pi_0(a_p^+|s_{0,i_p}) + \log \pi_0(a_p^-|s_{0,i_p})\big]\Big), \end{aligned} \tag{15}$$

with $\beta = 0.1$, learning rate 2×10^{-6}, eight epochs, gradient accumulation four, and maximum gradient norm 1.0. Pair p belongs to case i_p. Here a_p^+ is the selected candidate root and a_p^- is the matched Base root; π_0 log-probabilities are cached before updating θ_p. The loss is applied only to root-action tokens, and this is the complete loss used for the selected root-policy update.

For the implementation-audited sealed-v3 confirmation, we retain the same matched terminal-EM preference criterion but expand the collection to include both preference directions. Candidate-winning cases satisfy $\Delta\mathrm{EM}_i(a) > 0$; Base-winning cases satisfy $\Delta\mathrm{EM}_i(a) < 0$, with the ordered pair reversed for training. The resulting balanced set contains 192 retained pairs (96 per direction). Weights use the magnitude of the corresponding Base-relative utility difference.

At inference, the reader critic compares the terminal outputs of the two non-mutating shadow branches. It receives the question, each branch's evidence summaries, and each reader answer, and emits one of three labels: A (Base better supported), B (candidate better supported), or C (neither reliably better). With label log probabilities (ℓ_A, ℓ_B, ℓ_C), its margin is the d_i defined in Eq. 4. Admission follows Eq. 6 with $m = 1.5$ and the pointwise margins in Eq. 5; subsequent actions follow the Base policy at the resulting state, as in Eq. 7.

For completeness, the paired comparison statistic is computed from per-case metric values. For $j \in \{\mathrm{EM}, \mathrm{Recall}\}$,

$$\widehat{\Delta}_j = \frac{1}{n} \sum_{i=1}^{n} \left(M_{i,j}^{\mathrm{K2D}} - M_{i,j}^{\mathrm{control}}\right). \tag{16}$$

For bootstrap draw b, indices $I_b = (i_1, \ldots, i_n)$ are sampled with replacement from $\{1, \ldots, n\}$ and $\widehat{\Delta}_j^{(b)}$ is evaluated with the same paired differences. The bootstrap directional probability for a comparison is

$$P_{\mathrm{boot},j} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\left[\widehat{\Delta}_j^{(b)} > 0\right], \tag{17}$$

and the selection criterion requires positive paired answer and recall differences, no incumbent regression, and $\sum_i V_i = 0$. The development comparison uses $B = 50{,}000$ paired draws; the canonical baseline and MuSiQue comparisons use $B = 200{,}000$. Confidence intervals use the corresponding percentile bootstrap over the same paired differences.

The secondary metrics are computed directly from the same trace. If G_i is the set of gold supporting titles and H_i is the set of titles returned before the stop action, then

$$\text{Recall}_i = \frac{|H_i \cap G_i|}{|G_i|}, \qquad \text{Precision}_i = \frac{|H_i \cap G_i|}{\max(1, |H_i|)}, \qquad Q_i = \sum_t \mathbb{1}[a_t = \texttt{search}]. \quad (18)$$

Exact match is $\mathbb{1}[\text{norm}(\hat{y}_i) = \text{norm}(y_i)]$ and token F1 uses the harmonic mean of token-level precision and recall after the same normalization. The model-call count is $M_i = M_i^{\text{policy}} + M_i^{\text{reader}} + M_i^{\text{critic}}$, where shadow calls made only for admission are included, and wall time is $T_i = t_i^{\text{end}} - t_i^{\text{start}}$ on the same process boundary. Table entries report the arithmetic means $n^{-1}\sum_i Q_i$, $n^{-1}\sum_i M_i$, and $n^{-1}\sum_i T_i$; no cost metric is normalized by a method-specific number of successful answers.

The fixed execution order is therefore: (1) train the reader adapter on the 8,192 training cases with complete evidence; (2) collect four root samples per case on 256 disjoint training cases at temperature 0.7; (3) evaluate each root and its Base anchor through matched non-mutating shadow branches using the native retriever/summarizer and identical frozen Base continuation; (4) compute U_i, A_i, V_i, and z_i from the trace; (5) train the root-only DPO adapter on the resulting 24 pairs; (6) check decoded actions on 16 cases; (7) evaluate the 128-case development matrix; and (8) fix the resulting configuration before replication, held-out, transfer, or baseline evaluation. This order prevents later evaluations from changing training or admission choices.

Component	Setting
Backbone and roles	Qwen3-4B; one shared backbone with separate reader and acquisition LoRA adapters; reader trains on 8,192 full-evidence examples and acquisition trains only on root tokens.
Reader adapter	LoRA rank 16, alpha 32, dropout 0.05, target modules $q/k/v/o$, learning rate 10^{-4}, one epoch, batch size 4, gradient accumulation 8, maximum sequence length 1,024.
Candidate collection	256 cases, four root candidates per case, temperature 0.7, deterministic Base anchor, 1,024 total root rollouts, duplicate removal, 24 retained positive pairs.
Collection reward	Weights $(1.00, 0.75, 0.75, 0.25, 0.75, -1.00, -0.02, -0.05)$ in the order of Eq. 9; advantage cap 5.0 and minimum normalization scale 0.1.
Policy update	Root-only DPO, $\beta = 0.1$, learning rate 2×10^{-6}, 8 epochs, gradient accumulation 4, maximum gradient norm 1.0, cached Base reference log-probabilities.
Admission and interface	Three-way pairwise reader-utility critic with margin grid $\{0, 0.75, 1.5, 3.0\}$; selected value $m = 1.5$; two non-mutating shadow rollouts plus one critic call per root decision; BM25 $k = 32$; deterministic decoding; search budget 5; later actions use Base fallback.
Selection and statistics	16-case decoded-action check, then 128-case paired comparison; positive paired answer and recall differences; zero invalid/overflow/post-budget actions; $B = 50{,}000$ bootstrap draws for the selected development comparison.
Transfer and baselines	No parameter changes on held-out HotpotQA, Bamboogle, 2Wiki, MuSiQue, ReAct-style, or IRCoT-style; canonical baseline reports use the fixed 5-search budget and $B = 200{,}000$ paired draws.

Table 11: Training, selection, and evaluation settings fixed before held-out and transfer evaluation.

The admission rule, reward weights, adapter learning rates, candidate count, prompts, and budgets are fixed before held-out and transfer evaluation.

Development selection versus confirmation. We separate hyperparameter selection from held-out evaluation. Table 12 records the order. Directional probabilities on the development matrix quantify within-comparison uncertainty and do not constitute independent confirmation after selecting the margin. The 872-case evaluation is the first held-out result with all method settings fixed. Four additional seed–budget runs reuse the development case IDs and measure training replication without retuning. Bamboogle, 2Wiki, and MuSiQue are transfer evaluations and do not feed back into any selection decision.

D.1 Root controls and component ablations

Direct SFT imitates the first valid sampled non-BASE root; Relation SFT selects the root with greatest question–query relation coverage. Neither uses reward, reader outcome, gold answer/title, retrieved documents, or candidate–BASE advantage. Fresh-initialization versions have 252 available valid states, compared with 24 canonical positive states in the inherited versions. All receive 48 updates. Initialization and available states therefore confound cross-block contrasts in Table 13; Table 2 provides matched attribution.

Absolute-outcome uses the same 24 canonical states, candidate pool, and validity filter as K2D-positive, selecting a terminally correct candidate by absolute downstream utility instead of $U(a) -$

Stage	Data and decision	Inferential role
Reader and candidate construction	8,192 training cases; 256 disjoint candidate-collection cases; 1,024 root rollouts; 24 retained pairs	Training inputs; no evaluation case enters construction
Development selection	Seed 13, budget 5, 128 fixed cases; margin grid $\{0, 0.75, 1.5, 3.0\}$ and one two-step control	Configuration-selection evidence; $m = 1.5$ is the first choice satisfying the paired criterion
Training replication	Seeds 17 and 29 at budgets 4 and 8 on the same 128 case IDs	Four replications with no admission retuning
Held-out confirmation	872 new HotpotQA cases; one selected configuration	Primary held-out confirmation of the HotpotQA claim
Transfer evaluations	Bamboogle, fixed-corpus 2Wiki, and independent-corpus MuSiQue	Question/corpus transfer evidence; no feedback into selection

Table 12: Evaluation ledger separating development selection, replication, held-out confirmation, and transfer.

Root-control setting	EM	F1	Recall	Prec.	Queries
Base+FE (no root proposal adapter)	46.09	58.61	72.27	5.40	1.76
Direct root SFT (fresh init.)	46.88	58.95	71.88	5.44	1.70
Relation root SFT (fresh init.)	46.88	58.95	72.27	5.44	1.75
Absolute-outcome root SFT (inherited init.)	48.44	62.41	76.17	6.15	1.53
K2D-positive weighted root DPO (inherited init.)	52.34	64.41	76.56	6.18	1.47

Table 13: **Fresh- versus inherited-initialization root controls.** Direct and Relation are outcome-free controls trained from a fresh LoRA with 252 available valid states; Absolute-outcome and K2D-positive start from the inherited behavior adapter and use the same 24 canonical states. All trained rows receive 48 updates. Because initialization and state availability differ across the two blocks, the 46.88-versus-48.44 contrast does not identify a target-construction effect. The matched-initialization comparison in Table 2 is the basis for target attribution.

$U(a_B)$. The component matrix separately retains the admission grid, optimizer alternatives, shared adapter, Base Best-of-2, and two-step control. These controls describe the design space without replacing the matched root-target comparison.

D.2 Protocol details for compacted result controls

The main Results section reports only the role of several diagnostics and controls. We keep the operational details here so that the shorter presentation does not remove reproducibility information. Unless otherwise stated, the controls below use the same 128 HotpotQA development case IDs, Qwen3-4B backbone, frozen full-evidence reader, BM25 $k = 32$ retriever, native `search`/`stop`/`answer` interface, deterministic decoding, and five-search budget as Table 3.

The 16-case factorization probe uses a fixed subset of development IDs and no additional training. Reader-substitution rows retain the same acquisition trace and replace only the terminal reader; acquisition-substitution rows decode a different acquisition adapter and then use the same full-evidence reader. A trace is counted as changed when the decoded native action sequence differs from the corresponding BASE sequence. The first-query surface audit uses the first live `search` string from each trace. Query length, one-query rate, and relation/entity coverage are computed from the question text and decoded root query only; no gold titles, gold answers, or retrieved documents are used in this audit.

For admission evaluation, every candidate decision first constructs the two non-mutating terminal branches in Eq. 8. The pairwise grid uses margins $m \in \{0, 0.75, 1.5, 3.0\}$ with the deterministic rule in Eq. 6. The no-critic answer-preserving admission rule admits a valid candidate only when the candidate and BASE shadow branches produce the same normalized reader answer; otherwise it falls back to the BASE root. The selected $m = 1.5$ setting is the first grid value satisfying positive paired

(a) Fixed admission grid

Metric	No critic	Pairwise 0.75	Pairwise 1.5	Pairwise 3.0	Two-step
Paired Δ recall (pp)	$+2.73$	$+2.73$	**+3.91**	$+2.73$	$+2.34$
Development-only bootstrap direction	0.873	0.882	0.974	0.923	0.797
Trace	valid	valid	valid	valid	2 invalid

(b) Component ablation under the same interface

Component variant	EM	F1	Recall	Queries	Recorded calls	Sec.	Inv. (%)
AP admission ($m = 0$)	46.88	58.71	75.00	1.40	4.80	14.18	0
Support margin	46.88	59.49	75.00	1.40	–	–	0
BASE Bo2	43.75	56.05	71.88	1.73	15.85	–	0
High margin ($m = 3.0$)	48.44	60.29	75.00	1.61	6.22	14.99	0
Uniform-weight DPO	48.44	62.93	76.17	1.57	6.14	13.96	0
Root-action SFT	48.44	62.41	76.17	1.61	6.22	14.15	0
Shared adapter	48.44	59.46	75.39	1.61	6.20	14.91	0
Two-step direct	41.41	53.03	74.61	1.44	4.86	6.46	1.56
K2D-Align	**52.34**	**64.41**	**76.56**	1.47	–	–	0

Table 14: Admission and component ablations (128 HotpotQA cases). Panel **(a)** records the fixed admission grid; pairwise columns use $m = 0.75, 1.5, 3.0$. Panel **(b)** varies proposal, critic, objective, admission, and continuation with interface and cases fixed. The K2D row uses the same fixed admission rule as Table 2. Recorded calls exclude shadows except for Base Bo2, whose 15.85 includes branch construction. The 52.34-EM K2D replay has no matching complete cost trace, and Bo2 wall time is unavailable. Invalid values in panel (b) are percentages (Appendix D.3).

answer and recall differences with zero invalid, overflow, and post-budget actions. No admission margin is retuned in the replication, held-out, 2Wiki, Bamboogle, or MuSiQue evaluations.

For the fixed-query BGE substitution, BM25 recall on the frozen emitted queries is 72.27 and BGE recall is 73.44, a paired increase of 1.17 points with six gain cases and three loss cases. BGE precision is 6.05. Because the policy queries and candidate-pool documents are fixed before ranking, the analysis only checks whether the observed generated queries collapse under a different local ranker.

Analysis	Fixed protocol	Stored artifact
Outcome-free root SFT	Direct SFT takes the first valid sampled non-BASE root; relation SFT maximizes question–query relation coverage. Both draw only executable candidate roots from the frozen collection and never read reward, reader outcome, gold answer/title, retrieved evidence, or candidate–BASE advantage. Each uses 252 targets, a fresh adapter, 48 updates, the same reader, Base continuation, evaluator, and case IDs as the root-SFT ablation.	`runs/iclr/k2d_align_non_outcome_root_sft_v1/`
Absolute-outcome root SFT	Reuses the 24 frozen K2D-positive training state IDs, the same four-way candidate pool and executable-root validity filter, and selects one candidate per state with terminal reader EM $= 1$ and maximum absolute downstream utility. It never uses $U(a) - U(a_B)$ for target construction. It uses 24 targets, 48 updates, the same reader, BM25 retriever, Base continuation, evaluator, and inference admission.	`runs/iclr/k2d_absolute_outcome_root_sft_v1/`
Same-initialization target controls	Direct, relation, Absolute-outcome, and K2D-positive target rules use the same 24 frozen positive-state IDs and the same behavior-adapter initialization. A zero-update initializer is evaluated alongside the 48-update trained variants; reader, BM25 retriever, frozen Base continuation, pairwise admission rule ($m = 1.5$), and 128-case evaluator are fixed.	`runs/iclr/k2d_matched_init_target_controls_v1/`
Support-margin control	Uses the stored candidate/BASE terminal shadows, removes the pairwise critic, and admits a candidate by the deterministic support rule with minimum support improvement 0.05 and support tolerance 0.1. It selects 84 candidate roots and 44 BASE roots.	`runs/iclr/k2d_align_ablation_v1/support_heuristic_shadow_matched.json`
Support-crossing admission	Admits a valid candidate when the BASE support margin is negative and the candidate support margin is non-negative, with differing normalized answers.	`runs/iclr/k2d_matched_init_target_controls_v1/`
Base Best-of-2	Two independently sampled BASE root proposals are evaluated with BASE continuation and the same A/B/C reader critic used by K2D. No policy adapter, reward, answer, title, or collection utility enters the proposal source. This is a matched two-draw control, not a larger Best-of-N curve.	`runs/iclr/k2d_align_base_bestof2_pairwise_v1/`
Pair-count ablation	The 24 retained positive pairs are frozen in their recorded order. Root-only DPO adapters are trained on the first 4, 8, and 12 pairs with the same initialization, DPO hyperparameters, admission rule, and 128-case evaluator; the 24-pair row is the canonical run.	`runs/iclr/k2d_pair_scaling_v1/pair_scaling_summary.json`
Budget-2 matched run	The selected configuration is re-evaluated on the same fixed cases with the live search cap set to two. The reader, root adapter, admission margin, retriever, prompts, and statistics are unchanged; this is a budget stress test, not a new training run.	Same trace schema and paired evaluator as Table 3
Fixed-query BGE substitution	The generated policy queries are frozen and evaluated against a local BGE FlatIP ranking backend over the 8,572 documents already retrieved by the 128-case rollouts. It covers 225 emitted queries and is therefore a bounded candidate-pool substitution, not a full-corpus retriever-transfer evaluation.	`artifacts/iclr/k2d_bge_fixed_query_substitution_18218_v1/`

Table 15: Protocols for controls summarized in the main Results section. The table lists fixed components, varied components, and the corresponding artifact.

D.3 COST ACCOUNTING

We report executed retrievals, total retrieval-index invocations, and model-side computation as three resources. Total index invocations count live and shadow requests, with cache hits identified separately when logged; they are not inferred from live counts. Live interaction counts native search calls on the branch whose root is executed and is the quantity constrained by the retrieval budget. The full K2D protocol additionally constructs two non-mutating terminal shadows, scores branch support, applies one pairwise critic, and executes the selected root. These events are not free merely because a shadow retrieval is cached before live execution.

The standard evaluator retained a direct-execution counter (6.141 calls and 4,136.6 input tokens per K2D case), but this counter omits part of the shadow-side construction. We therefore reconstruct a conservative full-branch lower bound from the saved trace: root proposals, both shadow continuations and summaries, their terminal answers, support/pairwise critic calls, and final live execution. This gives at least 16.10 calls per K2D case, 2.00 terminal shadows, and 2.97 reader-critic calls. A separately compute-logged K2D trace subset indicates that this structural reconstruction can miss bookkeeping calls (mean 0.30 calls), so it is deliberately not used to claim an exact or lower total cost. Complete shadow input/output token totals were not retained for all historical traces; the table marks these as N/A rather than creating a tokenization proxy. The complete audit corresponds to 50.78 EM at $m = 1.5$; the later 52.34-EM development replay has no matching full cost trace.

Table 16: Trace-complete development quality–compute audit on 128 fixed cases. The K2D row is the measured $m = 1.5$ audit operating point (50.78 EM), not the later 52.34-EM replay. Calls include the two-shadow construction and are a structural lower bound for K2D. "N/A" means unlogged, not zero.

(a) Quality and branch structure

Method	EM	F1	Recall	Live q.	Shadows	Critic
Base policy + FE reader	46.1	58.6	72.3	1.76	0.0	0.00
ReAct	44.5	56.0	71.5	1.80	0.0	0.00
Retrieve-once	39.1	48.7	69.1	1.00	0.0	0.00
Root-DPO, no selector	46.9	58.0	75.8	1.54	0.0	0.00
Selector-only Base Best-of-2	43.8	56.0	71.9	1.73	2.0	2.96
K2D-ALIGN ($m = 1.5$ audit)	50.8	62.9	76.2	1.57	2.0	2.97

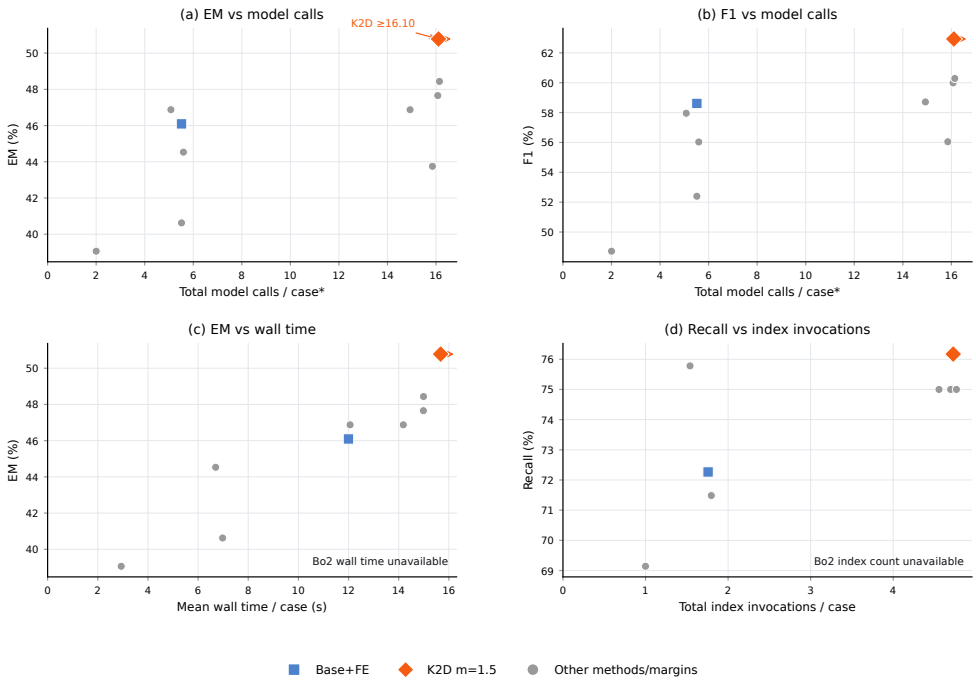
(b) Cost and selector risk

Method	Calls	Index	Mean s	p50/p90 s	Act./harm	Invalid
Base policy + FE reader	5.52	1.76	12.00	10.6/19.9	N/A	0.0
ReAct	5.59	1.80	6.70	5.6/11.2	N/A	0.0
Retrieve-once	2.00	1.00	2.93	2.5/4.4	N/A	0.0
Root-DPO, no selector	5.08	1.54	12.06	9.5/21.0	N/A	0.0
Selector-only Base Best-of-2	15.85	N/A	N/A	N/A	16.4/2.3	0.0
K2D-ALIGN ($m = 1.5$ audit)	≥ 16.10	4.73	15.67	13.8/24.5	68.8/0.0	0.0

Accounting note. "Live q." counts executed retrievals; "Index" includes live and speculative shadow invocations. K2D calls are reconstructed from saved proposals, two shadow continuations, summaries, terminal answers, critics, and live execution. Complete shadow token totals were not logged. "Act./harm" is available only for matched selector comparisons: percentage of changed roots / percentage with lower paired terminal EM.

E STRICT SAVED-STATE LOCATION PILOT

The location pilot tests zero-shot transfer of a root-specialized adapter to downstream states. We examine where the already-trained root adapter can be inserted without changing the prefix state. We reuse the fixed 16-case factorization IDs, saved BASE traces, native action parser, retriever, budget, and deterministic decoding. At each available saved BASE state, we evaluate the next BASE action and require it to match the action recorded in the trace before decoding the candidate action. This check passes on all 40 available states (16 at step 0, 16 at step 1, 6 at step 2, and 2 at step 3). For each single-step condition, the candidate acts once at the designated saved state and frozen BASE continues thereafter. *First uncertain* selects the available state with the smallest frozen-BASE next-token top-1/top-2 log-probability margin, with earliest-step tie breaking; it uses no threshold or outcome-based choice.



Admission-margin tradeoff on the same 128 cases

Margin m	EM	F1	Model calls†	Total index calls
0	46.88	58.71	$\geq$ 14.93	4.55
0.75	47.66	59.99	$\geq$ 16.07	4.70
1.5 (selected)	**50.78**	**62.93**	$\geq$ 16.10	4.73
3	48.44	60.29	$\geq$ 16.14	4.77

Figure 4: Trace-complete 128-case development quality–compute audit. Panels show (a) EM and (b) F1 versus total model calls, (c) EM versus mean wall time, and (d) recall versus total index invocations. The selected K2D operating point is not Pareto-dominated among measured development methods on these axes. †K2D call counts are trace-reconstructed structural lower bounds; unavailable Bo2 wall time and index counts are omitted. This audit is separate from frozen H872 quality confirmation and the later 52.34-EM selector replay.

Table 17 reports the pilot. Single-step comparisons have an identical pre-intervention state, but their terminal outcomes are paired downstream outcomes after different actions, not same-state causal estimates. The multi-step and full-continuation rows are stricter still: only the root state is shared, because later histories already diverge. The root–step-1 EM difference is $+6.25$ points (95% CI $[-12.5, +25.0]$), while step 2 has only six comparable states. The root–later-state contrast remains inconclusive. The pilot supports root localization for the present implementation but does not rule out separately trained intermediate-state interventions.

Location	Strict shared state	n_{eff}	ΔEM (pp)	95% CI (pp)	Invalid (%)
t_0 only	yes	16	+6.25	[-12.5, +25.0]	0.00
t_1 only	yes	16	0.00	[0.0, 0.0]	0.00
t_2 only	yes	6	0.00	[-50.0, +50.0]	0.00
First uncertain	yes	16	0.00	[0.0, 0.0]	0.00
t_0+t_1	t_0 only	16	-6.25	[-31.25, +18.75]	6.25
Full candidate continuation	t_0 only	16	-6.25	[-31.25, +18.75]	6.25

Table 17: Saved-state intervention-location pilot on 16 fixed cases. “Invalid” is the final live-trajectory rate; the pilot audits intervention boundaries rather than separately trained later-step effects.

(a) Depth boundary on MuSiQue

Hop	n	FE-SFT	Base+FE	IRCoT-style	K2D	Δ vs. FE
2	42	30.95	30.95	35.71	**35.71**	+4.76
3	43	20.93	21.63	16.28	**27.91**	+6.98
4	43	2.33	2.33	0.00	2.33	0.00

(b) Opportunity versus realization on Hermes

Route	EM	F1	Candidate use
Base branch	28.13	40.67	0/32
Forced candidate	**43.75**	**53.38**	32/32
Admission-routed	28.13	40.54	7/32

Table 18: Detailed transfer-boundary values for Figure 3. MuSiQue is a hop-stratified 128-case diagnostic; Hermes uses 32 cases. The settings and denominators are not pooled.

F Additional Transfer Analyses

The following analyses examine transfer and boundary conditions beyond the primary HotpotQA comparison. Each evaluation retains its task-specific interface and information conditions.

F.1 FHIR as a direct-evidence acquisition diagnostic

FHIR-AgentBench provides a complementary external diagnostic regime. In this MIMIC-IV-FHIR slice, the relevant structured resources often contain the answer directly. We therefore ask a more specific question: can functional factorization activate clean evidence acquisition when the evidence-to-answer mapping is comparatively direct? We use a patient-disjoint 64-case evaluation set and three independent replications at two adjacent tool budgets with the same case IDs, interface, and decoding. The factorized arm uses a tool-first acquisition adapter initialized from the matched full-evidence reader; Base uses no adapter. This diagnostic does not include root DPO and shadow admission and is not a full K2D-ALIGN algorithm evaluation. Answer correctness is scored with the official-compatible binary evaluation prompt and the `deepseek/deepseek-chat` judge. Results from the alternative upstream $o4$-mini evaluator are not combined with these measurements.

The judge receives only the prescribed question and answer record; gold resource identifiers and retrieval traces are excluded.

Table 19 reports the results. The factorized arm improves Interactive accuracy by $+21.88$–$+26.56$ points in both budgets and all three replications, and all 64 cases in every condition issue a valid query. Retrieval F1 is 0.4485 in every setting, with no invalid, overflow, repeated, or post-budget event. In contrast, the corresponding full-evidence-reader Interactive control queried zero times in the earlier matched run. Thus the medical result provides an acquisition-to-answer gain under this evaluator. Oracle gains are only $+3.13$–$+6.25$ points, so G_{FHIR} is negative in all settings. FHIR therefore gives a complementary boundary condition: the acquisition bottleneck can dominate even when the Oracle-side gain is small. It supports functional separation between reader use and acquisition, but not the positive-gap diagnostic observed on HotpotQA as a universal prerequisite for useful factorization.

(a) Interactive answer

Bud.	r	Base-I	Fact.-I	ΔI
4	3	21.88 ± 0.00	45.31 ± 2.71	$+23.44 \pm 2.71$
8	3	21.88 ± 0.00	45.31 ± 2.71	$+23.44 \pm 2.71$

(b) Oracle answer and transfer gap

Bud.	r	Base-O	Fact.-O	ΔO	G_{FHIR}
4	3	85.94 ± 0.00	91.15 ± 1.80	$+5.21 \pm 1.80$	-18.23 ± 2.39
8	3	85.94 ± 0.00	91.15 ± 1.80	$+5.21 \pm 1.80$	-18.23 ± 2.39

(c) Retrieval and validity

Bud.	r	Query cases	Ret. F1	Violations
4	3	64/64	0.4485 ± 0.0000	0
8	3	64/64	0.4485 ± 0.0000	0

Table 19: FHIR direct-evidence diagnostic on a fixed 64-case MIMIC-IV-FHIR slice. Entries are mean $\pm$ sample standard deviation over three runs per budget; ΔI and ΔO are Factorized–Base paired changes and $G_{\mathrm{FHIR}} = \Delta O - \Delta I$. Negative G_{FHIR} means the Interactive gain exceeds the Oracle gain.

Separate monolithic FE-SFT diagnostic. The earlier 64-case, budget-four record reports Base/FE-SFT Interactive accuracy of 29.6875/17.1875% and Oracle accuracy of 65.625/79.6875%, with zero FE-SFT queries. That run used an 8,192-token interface and the `deepseek-v4-pro` judge; 121 of 151 Base calls were rejected by bounded context delivery. It is a separate interface-feasibility diagnostic, not a matched numerical row for the factorized protocol above. The record is `runs/remote/fhir_story_controls_seed13/story_controls_summary.json`.

F.2 Additional backbone calibration

We also evaluated a separately trained Qwen2.5-7B reader on the fixed 2Wiki and independent MuSiQue case files. The calibration uses the same retrieval budget and a prompted JSON action interface. Qwen2.5 is an earlier-generation Qwen2 checkpoint, so this experiment is reported as a capacity/interface calibration alongside the same-generation Qwen3 results. The Base-versus-full-evidence comparison did not support policy transfer: the 2Wiki gap was non-positive, while the MuSiQue difference was small and each arm contained one invalid action. We therefore treat this experiment as backbone calibration rather than efficacy evidence.

Corpus	n	Base EM	Base+FE EM	ΔEM	Recall	Invalid (count/n)
2Wiki	128	41.41	39.84	-1.56	69.73	0/128
MuSiQue	128	11.72	13.28	$+1.56$	62.89	1/128

Table 20: Qwen2.5-7B backbone calibration on 2Wiki and MuSiQue. The reader was trained independently, with no Qwen3 adapter or root candidate reused. These rows are supplementary calibration evidence; primary efficacy claims use Qwen3-4B.

F.3 Same-generation Qwen3-8B capacity stress test

To probe whether the direction survives a larger, same-generation backbone, we evaluated Qwen3-8B with the same full-evidence reader adapter, native action interface, BM25 corpus, retrieval budget, and FE-relative candidate construction. This experiment increases the disjoint candidate-collection pool from 256 to 1,024 training states and yields 37 strict answer-improving cases under the same validity and matched-continuation filters. Evaluation uses the first 128 cases of an independent development shard that is disjoint from every training shard and from the original Qwen3-4B development matrix; no admission threshold or test-case result was retuned.

Table 21 shows a positive directional change over the matched Qwen3-8B FE-policy control: EM, F1, and supporting-title recall rise by 0.78, 1.24, and 1.56 points, respectively, while mean environment-query cost decreases by 0.039. However, each arm has one invalid or post-budget event, and the paired probabilities are 0.602, 0.775, and 0.867 for EM, F1, and recall. We therefore report this as a directional capacity stress test, not as confirmatory efficacy evidence; the primary claims rely on Qwen3-4B.

Backbone / policy	n	EM	F1	Recall	Prec.	Env. queries	Invalid
Qwen3-8B FE-policy control	128	52.34	57.41	69.92	5.37	1.836	1
K2D-Align (1,024 states)	128	**53.13**	**58.65**	**71.48**	**5.54**	**1.797**	1
Paired change	–	+0.78	+1.24	+1.56	+0.17	-0.039	–

Table 21: Same-generation Qwen3-8B capacity stress test on an independent 128-case shard. The control and K2D run share model, reader, corpus, interface, and budget; only the FE-relative acquisition adapter differs. This directional stress test is reported separately from the primary Qwen3-4B estimate.

G Interface-Compatible Baseline Protocols

Table 22 specifies four paper-local, matched-interface instantiations of retrieval controls. Retrieve-once, Prompted query rewrite, and ReAct-style form the HotpotQA baseline rows; IRCoT-style is evaluated on the MuSiQue same-ID comparison, where its explicit missing-relation scaffold is defined. All baseline methods use the fixed case matrix for their respective corpus, the same Qwen3-4B model and full-evidence reader, the corpus specified for each evaluation (Wikipedia_HotpotQA

for HotpotQA and Bamboogle; MuSiQue contexts for the 128-case MuSiQue comparison), BM25 with $k = 32$, native `search`/`stop`/`answer` tools, temperature zero, and no acquisition adapter. A retrieval budget of five counts search calls only; the final answer turn is an additional turn, giving a maximum of seven agent turns. The observation supplied to the policy is the frozen Base summarizer's passage summary, and the answer is decoded by the same frozen full-evidence reader used by K2D-ALIGN.

Baseline	Action rule	Query budget	Policy prompt / termination protocol
Retrieve-once (single-shot RAG)	Deterministically call `search` on the question at the first policy turn, then call `stop`.	1 search (at most 2 agent turns including stop)	No learned acquisition prompt. The original question is passed verbatim to BM25; the frozen reader answers from the resulting observation.
Prompted query rewrite	On the first policy turn, ask BASE to emit one self-contained `search` query preserving named entities, relations, and constraints; later turns use the native BASE continuation.	5 searches (at most 7 agent turns including answer)	The first turn exposes only the original question and allows only `search`; subsequent turns receive the same summarized search history as the native BASE policy.
ReAct-style (matched native)	At each policy turn, emit exactly one native `search` or `stop`; stop when the evidence is sufficient or the five-search budget is exhausted.	5 searches (at most 7 agent turns including answer)	The user message appends the original question, remaining search calls, and frozen summarized search history; answer text is disallowed before `stop`.
IRCoT-style (matched native)	At each policy turn, identify the next missing relation and emit one atomic native `search` or `stop`; stop when the chain is complete or the five-search budget is exhausted.	5 searches (at most 7 agent turns including answer)	The scaffold asks for an internal iterative retrieval chain and one atomic sub-question per turn; answer text is disallowed before `stop`.

Table 22: Matched native-interface protocols for Retrieve-once, Prompted query rewrite, ReAct-style, and IRCoT-style. All share the reader, retriever, corpus, information boundary, and budget; prompts are native-tool adaptations of the cited methods.

Exact baseline prompts. Prompted query rewrite uses the first-turn instruction: "Rewrite the original question into one self-contained search query that preserves all named entities, relations, and constraints. On this policy turn call search exactly once with only that query. Do not answer, stop, explain, or emit plain text." ReAct-style uses: "Think step by step about the current situation. Call search for one precise next query without repeating history, or call stop when the evidence is sufficient. On this policy turn call exactly one available native tool. Never emit an answer or plain text." IRCoT-style uses: "Use iterative retrieval chain-of-thought: reason about the original multi-hop question, identify the next missing relation or sub-question, and retrieve evidence for that sub-question. On each turn call search for exactly one atomic sub-question without repeating history, or call stop when the reasoning chain is complete. Keep the reasoning internal. On this policy turn call exactly one available native tool. Never emit an answer or plain text."

Search-O1-style and Research-style prompts. Both scaffolds prepend "You are an information-seeking policy." Search-O1-style then uses: "Reason step by step over the information-seeking history. Call search for one precise missing fact without repeating a query, or call stop when the evidence is sufficient." Research-style uses: "Reason toward an answer, identify the first claim not grounded in the information-seeking history, and call search for that atomic missing fact. Call stop only when no answer-critical claim remains unverified." Both append: "On this policy turn call

Method	n	EM	F1	ΔEM	ΔF1
Base	109	26.61	39.85	–	–
Base policy + FE reader	109	28.44	41.77	–	–
K2D-ALIGN	109	33.94	48.16	+5.50	+6.39

Table 23: Bamboogle transfer with the HotpotQA-trained system and the same Wikipedia_HotpotQA BM25/native interface. K2D is compared with the matched Base-policy + full-evidence-reader control; all traces are valid. This sealed 109-case slice excludes the 16-case action check from the official 125 questions; Table 5 uses all 125 under its separate common protocol.

exactly one available native tool. Never emit an answer or plain text." The user message contains the original question, remaining search calls, and query/summary history. The common five-search cap and terminal FE reader apply. ReAct-style and IRCoT-style likewise suppress explicit reasoning output; all four are interface-compatible prompting scaffolds over a frozen backbone.

Method	n	EM	F1	Recall	Env. queries	Invalid
Base	128	17.19	30.77	76.17	1.81	0
Base + FE reader	128	19.53	34.78	76.17	1.81	0
ReAct-style (same IDs)	128	22.66	37.16	76.56	2.34	0
IRCoT-style (same IDs)	128	25.00	39.16	79.69	2.46	0
K2D-ALIGN	128	26.56	40.91	78.52	1.84	0

Table 24: MuSiQue transfer on 128 fixed cases with a BM25 corpus built only from MuSiQue contexts. The HotpotQA corpus and index are not reused; K2D, ReAct-style, and IRCoT-style share the native interface and case IDs, with zero validity violations.

MuSiQue scaffold paired bootstrap. Using the same 128 case IDs and 200,000 paired bootstrap draws, K2D–ReAct-style has $+3.91$ EM points (95% CI $[-1.56, +10.16]$, $P_{\text{boot}}(\Delta\text{EM} > 0) = 0.880$) and $+3.76$ F1 points (95% CI $[-1.50, +9.10]$, $P_{\text{boot}}(\Delta\text{F1} > 0) = 0.920$), while reducing environment queries by 0.492 per case (95% CI $[-0.711, -0.289]$). These paired results support the positive K2D transfer direction on the evaluated MuSiQue slice. Against full-budget IRCoT-style, K2D–IRCoT-style is $+1.56$ EM points (95% CI $[-3.91, +7.03]$, $P_{\text{boot}}(\Delta\text{EM} > 0) = 0.655$), $+1.75$ F1 points (95% CI $[-3.93, +7.46]$, $P_{\text{boot}}(\Delta\text{F1} > 0) = 0.728$), and -0.62 environment queries (95% CI $[-0.84, -0.41]$). Under the cost-matched two-search cap, IRCoT-style obtains 22.66 EM, 35.63 F1, 74.61 title recall, and 1.56 realized queries with zero interface violations. K2D therefore gains $+3.91$ EM (95% CI $[-0.78, +9.38]$, $P_{\text{boot}}(\Delta\text{EM} > 0) = 0.956$), $+5.28$ F1 (95% CI $[+0.03, +10.67]$, $P_{\text{boot}}(\Delta\text{F1} > 0) = 0.976$), and $+3.91$ recall points (95% CI $[+0.39, +7.81]$, $P_{\text{boot}}(\Delta\text{Rec} > 0) = 0.986$), while using 1.84 realized queries.

Acquisition scaffold	n	EM	F1	Recall	Env. queries	Invalid
Native policy	128	46.09	58.61	72.27	1.76	0
Retrieve-once (single-shot RAG)	128	39.06	48.72	69.14	1.00	0
ReAct-style (matched native)	128	44.53	56.03	71.48	1.80	0
Search-O1-style	128	44.53	55.82	74.22	2.27	0
Research-style	128	42.97	55.81	76.17	2.43	0

Table 25: Matched acquisition-scaffold comparison on 128 HotpotQA cases. All methods share the full-evidence reader, BM25 corpus, native interface, and retrieval budget; all traces are valid.

HotpotQA baseline comparison and computational cost. On the same 128 HotpotQA case IDs, paired bootstrap gives K2D–ReAct-style $+7.81$ EM points (95% CI $[+1.56, +14.84]$, $P_{\text{boot}}(\Delta > 0) = 0.987$), $+8.38$ F1 points (CI $[+2.24, +14.84]$, $P_{\text{boot}}(\Delta > 0) = 0.997$), and $+5.08$ gold-title-recall points (CI $[+1.17, +8.98]$, $P_{\text{boot}}(\Delta > 0) = 0.993$). Against Retrieve-once, the corresponding EM/F1 gains are $+13.28/+15.70$ points (95% CIs $[+6.25, +20.31]$ and $[+8.91, +22.83]$, $P_{\text{boot}}(\Delta > 0) = 0.9998/1.0$). These paired baseline comparisons provide the interface-compatible acquisition reference for the fixed protocol. The cost accounting separates live interaction from admission computation. The trace-complete $m = 1.5$ executor records 1.57 live environment queries, 6.14 direct calls, and 15.67 seconds per K2D case; the direct counter omits the full construction of both shadow branches. Saved-branch accounting instead reconstructs *at least* 16.10 model calls per case, including root proposals, both terminal continuations, branch summaries and answers, reader-critic calls, and live execution (Appendix D.3). ReAct-style records 1.80 live environment

queries, 5.59 direct calls, no admission branches, and 6.70 seconds. Thus the interaction result is a lower live search count, not a claim of lower total computation.