THE UNIT OF MEMORY SUPERVISION IS THE SET

Anonymous authors
Paper under double-blind review

ABSTRACT

Memory selectors for language agents score each candidate item on its own, by relevance, salience, or recency. But a memory helps only through the set it forms, and a set's value is rarely the sum of its items, so which items to keep is a question about the set. We show that item-level supervision cannot in general identify the optimal set. Two value functions can have identical add-one and delete-one effects for every fact yet have disjoint optimal budgeted sets, so no rule restricted to those add-one and delete-one signals can avoid set regret. On a sealed split of 250 MuSiQue questions with exact subset enumeration, a development-selected rank-sum of these singleton signals leaves 0.18 answer-F1 regret against the optimal set, while a set-derived item score leaves 0.006. Changing only the training label from salience to gold supporting sets, at matched architecture, data, and budget, raises exact match from 0.107 to 0.345. On 500 fresh MuSiQue questions with retrieved candidates, supporting-set supervision beats both add-one item supervision and dense retrieval. The supervision must observe value at the set level, but deployment need not: set-level value can be distilled into per-item scores, preserving item-wise selection at test time. When downstream value is interactive, the unit of supervision is the set.

1 INTRODUCTION

A language agent cannot keep everything it has seen. Under a fixed budget it selects a few items from a larger pool and passes them downstream, to a reader that answers a question, a planner that acts, or a tool that is called (Packer et al., 2023; Park et al., 2023; Zhong et al., 2024; Shinn et al., 2023; Lewis et al., 2020; Yao et al., 2023; Schick et al., 2023; Ichter et al., 2023; Huang et al., 2023). The standard recipe scores each candidate on its own, by retrieval similarity (Robertson & Zaragoza, 2009; Xiao et al., 2024), salience, recency, or a learned relevance head (Xu et al., 2024), and keeps the top few. Retaining more context carries its own cost, since longer inputs increase processing time and can lower accuracy (Liu et al., 2024). Item-wise scoring works when items help independently, but can miss combinations needed for a decision. Three passages can each look relevant, yet their individual relevance scores need not reveal which particular pair supplies the evidence needed for the answer. In a multi-hop question, for example, the answer may require one passage that names a bridge entity and another that resolves it, while a third, equally on-topic passage adds nothing once that pair is present.

This failure is built into the supervision, not the model. A memory method can cheaply measure how downstream value changes when one item is added to an empty set or removed from the full set: the add-one and leave-one-out effects used by relevance and attribution scores. We call these the *singleton signals*, and they pin down only the boundary of a set's value. Two value functions can agree on every singleton signal at every item yet have disjoint optimal budgeted sets (Figure 1a). A selector trained on such signals must therefore choose the wrong set on some inputs, leaving a *set regret* that no rule restricted to these singleton signals can drive to zero.

The obstruction lies in where supervision observes value, not in how the selector represents it. Once set-level value has been observed, it can be compressed into a per-item score. For the example in Figure 1a, how often an item appears in an optimal budgeted set already ranks the right items first (Figure 1c). Deployment can stay a plain top-B ranking, keeping the B highest-scoring items. What item-level supervision cannot do is recover that score from singleton signals alone. Because the missing information is in the signal and not the model, a set-aware architecture fed only singleton signals inherits the same limit.

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.

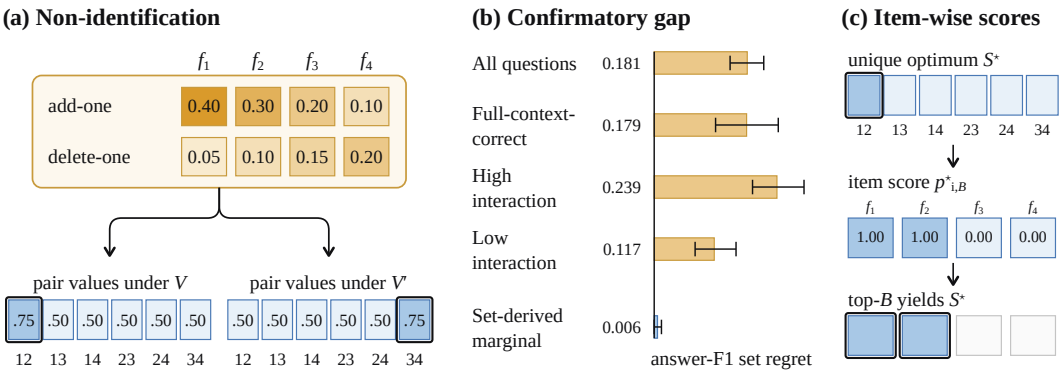

Figure 1: **Identical item-wise signals, different optimal sets.** (amber: singleton-derived; blue: set-derived). **(a)** Two value functions V, V' have identical per-fact add-one and delete-one effects, with pairwise-distinct (add-one, delete-one) pairs, yet disjoint unique budget-2 optima: $\{f_1, f_2\}$ under V and $\{f_3, f_4\}$ under V'. Thus even exact item-wise evidence need not identify the optimal set (Example 1, Theorem 1). **(b)** All 250 sealed MuSiQue questions, pooled $B = 2, 3$: rank-sum fusion of exact singleton signals loses 0.181 answer-F1 against the optimal set; optimal-set inclusion frequencies lose 0.006 (no learning). High/low: maximum pairwise gain beyond separate gains, split at the development-set median. Full-context-correct: full-pool F1 ≥ 0.5. Error bars: question-clustered 95% CIs (Table 1). **(c)** For the unique optimum $S^\star = \{f_1, f_2\}$ in the left witness, the optimal-family marginal is $p^\star_{i,B} = (1, 1, 0, 0)$, whose top-$B$ ranking recovers $S^\star$. Set-derived information can therefore be represented by item-wise scores without requiring set-valued deployment (Section 5).

We make this precise and measure it. A set of items is scored by the value of the decision it supports at the true state, which makes relevance decision-theoretic rather than semantic (Section 2). Throughout, *set regret* is the decision value a selector loses against the *optimal set*, the value-maximizing subset within budget found by exact search.

Singleton signals cannot identify the set (theory). For n candidates and each admissible budget B, we construct value functions with identical singleton signals at every item but a combinatorial family of distinct optimal sets. Any rule reading only these signals returns the optimal set with probability at most $1/\binom{n}{B}$ under a uniform prior, so a strictly positive set regret is unavoidable, and it persists even when every item's signature is distinct (Section 3). Finding a smallest decision-sufficient set is additionally NP-hard, with Set-Cover inapproximability (Feige, 1998; Dinur & Steurer, 2014).

The gap is large on held-out language (confirmatory). On a sealed held-out split of 250 MuSiQue questions, exact subset enumeration with a frozen deterministic reader gives 0.18 answer-F1 regret against the optimal set $[0.15, 0.21]$ for exact singleton-based selection, versus 0.006 for a score read off the optimal sets. The shortfall concentrates where value is interactive, 0.24 regret versus 0.12 under a threshold fixed on development data, both excluding zero at budgets $B \in \{2, 3\}$ (Figure 1b). A separate learning experiment changes only the training label from salience to gold supporting sets; matched architecture, data, and budget yield exact match of 0.107 versus 0.345 (Section 6).

The supervision effect persists under controlled decorrelation. On synthetic decisions that deliberately decorrelate value from surface salience, a value-supervised selector reaches near-zero out-of-distribution regret (0.023), while a matched salience-supervised selector stays two orders of magnitude worse (8.3). Neither stronger prompting nor a frontier model substitutes for the supervision signal (Section 7).

Set-derived supervision can be represented by an item-wise score at deployment. The per-item score read off the optimal sets recovers the optimal set to within 0.006 regret on MuSiQue, illustrating how set-level information supports item-wise selection. The identifying information lives at the set, and once observed it can be carried by an item-wise selector. Set targets are constructed offline; deployment requires no subset search. What item-level *supervision* cannot do is recover that score from singleton signals alone. Set-conditioned deployment may help when learned item scores fail to capture joint utility. Learned set supervision also outperforms dense retrieval on 500 fresh questions under matched retrieval training.

2 PROBLEM AND DEFINITIONS

We study memory selection as choosing a subset of candidate facts under a budget, scored by the value of the decision it supports. Let $F(h) = \{f_1, \ldots, f_m\}$ be the candidate facts for an instance h, and let $V_h(S)$ be the *decision value* of keeping subset $S \subseteq F(h)$: the value of the downstream decision produced from S, scored at the true state (larger is better), in the task's own units (answer F1 on MuSiQue; the solver's utility on the synthetic and planning suites, so V_h and the regret below are in those raw units). With a budget B on the kept set, we use two optimal families,

$$O_{\leq B}(h) = \arg\max_{|S|\leq B} V_h(S), \qquad O_B(h) = \arg\max_{|S|=B} V_h(S), \tag{1}$$

and the corresponding *set regret* $G_B(S) = \max_{|S'|=B} V_h(S') - V_h(S)$. We report the fixed-cardinality family O_B in MuSiQue because the decoder returns exactly B facts; the non-identification theory applies to either feasible family. Two per-fact quantities summarize what a selector sees one fact at a time: the *add-one* signal $a_i = V_h(\{f_i\}) - V_h(\varnothing)$ and the *delete-one* signal $d_i = V_h(F) - V_h(F \setminus \{f_i\})$. These are the add-one and leave-one-out effects used by relevance and attribution methods, which we call the *singleton signals*: marginal effects anchored at the empty and full sets. Against them we compare the *optimal-family marginal*, a set-derived item score (the uniform inclusion frequency over the optimal sets),

$$p^{\star}_{i,B} = \frac{1}{|O_B(h)|} \sum_{S \in O_B(h)} \mathbf{1}[f_i \in S], \tag{2}$$

a per-fact score that, unlike the singleton signals, is derived from the set-level target. A selector is evaluated by the set regret it incurs, not by the relevance of its facts. The non-identification results (Section 3) treat V_h as an arbitrary set function; our experiments instantiate it with solver- or reader-defined downstream value.

The value $V_h(S)$ is instantiated per suite by the downstream task: token-F1 of a fixed reader's answer on the multi-hop question-answering tasks, and a solver's realized value on the controlled synthetic decisions. The synthetic instantiation treats a dropped fact as unobserved rather than false. A retained set S induces a minimax-regret decision over the resulting uncertainty set, and $V_h(S)$ is that decision's realized value at the true state. Forgetting thus widens an uncertainty set rather than editing the world; this decision-theoretic construction is given in Appendix C, and a value-preservation bound over the language abstraction in Appendix B. The non-identification results below establish the obstruction even within the restricted class of monotone set functions, which only strengthens them. The empirical reader-defined value functions need not be monotone, since added context can lower answer quality; no empirical result assumes monotonicity.

3 ANCHORED SINGLETON SIGNALS CANNOT IDENTIFY THE SET

A memory method can cheaply obtain one number per fact: how the decision changes when that fact is added or removed. We ask whether these singleton signals, however perfectly measured, suffice to recover the budgeted set a decision needs. They do not, and the obstruction is informational rather than a limitation of any particular model.

Definition 1 (Pointwise selector)**.** *A selector is* pointwise *if each fact's inclusion is a threshold on a score that depends on the value function only through that fact's singleton signals, equivalently the first-order data $V_h(\varnothing)$, $V_h(F)$, $V_h(\{f_i\})$, $V_h(F \setminus \{f_i\})$.*

Definition 2 (Anchored observation)**.** *The* anchored observation *of a value function collects exactly these first-order signals,*

$$\mathcal{A}(V) = \Big(V(\varnothing),\, V(F),\, \big\{V(\{f_i\}),\, V(F \setminus \{f_i\})\big\}_{f_i \in F}\Big),$$

the empty- and full-set values together with each fact's add-one and leave-one-out effect. This is a deliberately strong signal. It grants the selector exact *counterfactual effects that item-wise memory methods typically only approximate, through relevance, attribution, or marginal-contribution proxies. A selector is $\mathcal{A}$-*measurable *if it depends on V only through $\mathcal{A}(V)$; every pointwise selector is $\mathcal{A}$-measurable, so a limit proved for all $\mathcal{A}$-measurable selectors holds for pointwise scoring in particular, and granting the selector this much information only strengthens it.*

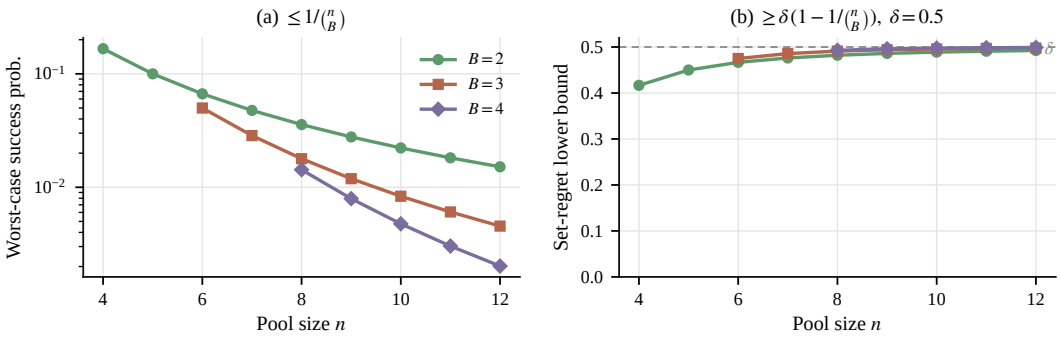

Figure 2: **The ambiguity is combinatorial** (Theorem 1, exact bounds; budget B in green/clay/plum). **(a)** Any selector that reads only the anchored observation $\mathcal{A}(V)$ is optimal with worst-case probability at most $1/\binom{n}{B}$, which collapses as the candidate pool n grows (log scale; $1/495$ at n=12, B=4). **(b)** The guaranteed set-regret lower bound $\delta\,(1-1/\binom{n}{B})$ rises to the full gap δ. These curves plot the theorem's guarantees directly; no empirical estimates are involved.

Example 1 (Identical signals, distinct signatures, disjoint optima)**.** Take $F = \{f_1, f_2, f_3, f_4\}$ under budget $B = 2$. Both worlds share $V(\varnothing) = 0$, singleton values $(0.4, 0.3, 0.2, 0.1)$, leave-one-out values $V(F \setminus \{f_i\}) = (0.95, 0.90, 0.85, 0.80)$, and $V(F) = 1$; they differ only on pairs. World I sets $V(\{f_1, f_2\}) = 0.75$; world II sets $V(\{f_3, f_4\}) = 0.75$; every other pair is 0.50 in both. The anchored signals are therefore identical across the two worlds and, unlike the usual complementarity construction, *pairwise distinct* across facts: $(a_i, d_i) = (0.40, 0.05), (0.30, 0.10), (0.20, 0.15), (0.10, 0.20)$. Both value functions are monotone. Yet the unique budget-2 optimum is $\{f_1, f_2\}$ in world I and the disjoint $\{f_3, f_4\}$ in world II. A pointwise selector reads the same input in both worlds, returns one set, and is optimal in at most one; no tie among facts is involved.

Theorem 1 (Combinatorial non-identification)**.** *For $n = |F|$ and $2 \leq B \leq n-2$ there is an anchored observation a whose fiber $\{V : \mathcal{A}(V) = a\}$ contains $\binom{n}{B}$ monotone value functions with pairwise-distinct unique budget-B optima. Consequently any deterministic $\mathcal{A}$-measurable budget-B selector is optimal for at most one member of this family; under a uniform prior over it, any randomized such selector returns the optimal set with probability at most $1/\binom{n}{B}$. The family can be chosen so that the expected set regret satisfies $\mathbb{E}\, G_B \geq \delta\big(1 - 1/\binom{n}{B}\big)$ for any $0 < \delta < 1$.*

Proof sketch. For each B-subset T let $V_T(S) = |S| + \delta\,\mathbf{1}[S = T]$, $0 < \delta < 1$. Each V_T is monotone; because T is neither $\varnothing$, a singleton, an $(n-1)$-set, nor F, all V_T share the same $\mathcal{A}(V)$; and T is the unique budget-B optimum of V_T. An $\mathcal{A}$-measurable rule sees identical input for every T, so it commits to a single distribution q over B-sets. Averaging over a uniform T, its success probability is $\binom{n}{B}^{-1} \sum_T q(T) = \binom{n}{B}^{-1}$, and each remaining outcome costs δ, so $\mathbb{E}\, G_B = \delta\big(1 - \binom{n}{B}^{-1}\big)$. □

In other words, the anchored observations constrain only boundary evaluations of the set function: the empty and full sets, the singletons, and the leave-one-out sets. The budget-relevant interior, meaning which subsets of intermediate size a decision actually needs, can vary freely. Which *combinations* of facts are jointly sufficient is exactly this interior, and the anchored signals do not carry it. Decision sufficiency is therefore irreducibly a property of subsets. The witness of Example 1 shows the obstruction is not an artifact of tied singletons. It persists when every (a_i, d_i) is distinct, and it is margin-robust rather than knife-edge, surviving every sufficiently small perturbation of the values the anchored signals do not observe (Corollary 1). The full construction is in Appendix Z.6.

The set objective is also computationally hard. Even given the full value function, finding the smallest set preserving full-information decision value is intractable.

Proposition 1 (Hardness)**.** *Finding a minimum decision-sufficient set is NP-hard, and with N requirements to cover it is NP-hard to approximate within $(1-\varepsilon)\ln N$ for every $\varepsilon > 0$ unless $\mathrm{P} = \mathrm{NP}$. On the Set-Cover subclass used in the reduction a value-graded greedy attains the matching $\ln N + 1$ factor (Appendix Z.5), but on general instances forward greedy and anonymous threshold selection on singleton (first-order) signals have unbounded approximation ratios (Appendix Z.7).*

Table 1: Sealed replication on 250 held-out MuSiQue questions with exact subset enumeration and a deterministic reader. Exact singleton signals leave large answer-F1 regret; optimal-family marginals nearly close it. "Sel. %" is the fraction of questions whose budget-B subsets differ in reader F1. At B=1, add-one values identify the optimum (set-derived regret 0.000); nonzero singleton regret comes from the development-selected fusion, not non-identification. The theorem addresses $2 \leq B \leq n-2$.

Budget	Questions	Singleton regret [95% CI]	Set-derived marginal [95% CI]	Sel. %
B=1	250	0.068 [0.044, 0.095]	0.000 [0.000, 0.000]	68.4
B=2	250	0.190 [0.149, 0.231]	0.004 [0.000, 0.012]	80.4
B=3	250	0.172 [0.133, 0.212]	0.008 [0.000, 0.020]	80.4
Pooled $B\in\{2,3\}$	250×2	0.181 [0.148, 0.213]	0.006 [0.000, 0.014]	–

Interpretation. Non-identification says the cheap signal is the wrong unit; hardness says the exact set cannot be found by tractable global optimization. Together they point to a practical design: supervise offline against an optimal set, then deploy selection without test-time subset search.

When surface carries no value, scale does not substitute. The gap does not close with a larger selector. When the decision-critical fact is statistically independent of surface features, every salience- or prompt-based selector that keeps k of n facts incurs expected regret $\Delta\,(1-\frac{1}{K})(1-\frac{k}{n})$ on a K-way decision with error cost Δ, independent of selector capacity (Appendix Z.8). A value-supervised selector, trained on the expected regret of forgetting each fact, identifies the pivotal fact and attains zero. The controlled suites of Section 7 show a two-orders-of-magnitude gap that the prompting and model-scale baselines we test do not close.

Remark (a positive boundary). The obstruction is not universal. Under structural regimes in which the budget-B optimum is determined by the anchored observations, item-wise supervision is sufficient; our theorem concerns the absence of such structure. Complementary interactions, where facts that change value little in isolation are jointly decisive, provide one concrete regime in which the budget-relevant interior is not identified by boundary observations. Multi-hop questions can exhibit this bridging structure; Section 4 measures the induced gap on held-out language.

4 The Gap Is Large on Held-Out Language

To quantify the regret of singleton-based selection on real language, we evaluate a split of 250 MuSiQue multi-hop questions held sealed until protocol and analysis freeze, disjoint from all method-development splits. For each question we enumerate every subset at the evaluated budgets within a six-paragraph pool containing all gold supports plus BM25 distractors. A frozen deterministic reader (Qwen3-8B, greedy decoding) defines $V(S)$ as answer token-F1. Against the optimal set we compare (i) the development-selected rank-sum (sum of ranks) of the *exact* add-one and delete-one signals and (ii) the set-derived score $p^\star$. Boundary evaluations supply the empty, full, singleton, and delete-one values.

Singleton supervision leaves a large regret while $p^\star$ nearly removes it (Table 1). Pooled over budgets $B \in \{2,3\}$, the singleton-fusion selection leaves 0.181 answer-F1 regret against the oracle, with a question-clustered 95% interval of $[0.148, 0.213]$, and $p^\star$ leaves 0.006. The shortfall is where the theory locates it. We score each question by its strongest pairwise interaction, $\mathrm{I}(V) = \max_{i<j}\big[V(\{i,j\}) - V(\{i\}) - V(\{j\}) + V(\varnothing)\big]$, the largest second-order (non-additive) coupling between two facts, and split at a threshold τ fixed to the development-set median. On the interaction-heavy half the singleton regret is 0.239 [0.191, 0.291] against 0.117 [0.079, 0.159] on the rest. On the full-context-correct subset (full-pool F1 ≥ 0.5), the regret remains 0.179 [0.119, 0.241]. Neither the gap nor its concentration depends on the two development-fixed choices: it persists across every pre-specified fusion of the add-one and delete-one signals, including the per-question strongest, which still leaves 0.079 regret, and grows monotonically as the interaction threshold is swept (Appendix W).

The binary interaction split understates the structure of this gap. In Appendix W we reconstruct reader-defined set value from progressively higher-order subset observations. Recovery improves

with order across three readers (Qwen3-8B, Tulu-3-8B, gemma-2-9b) and all MuSiQue hop strata, where first-order observations often admit an optimal set but rarely guarantee one. We use interaction order only diagnostically; it is not reasoning depth, and hop count does not determine the reconstruction order required once non-unique optima are accounted for.

5 SET-DERIVED SUPERVISION

The obstruction is a property of the supervision signal, so the remedy is to supervise from sets. For each instance we identify the optimal family $O_B(h) = \arg\max_{|S|=B} V_h(S)$ by exact search over subsets, and read off a per-fact target: the set-derived item score $p^\star_{i,B}$ of Eq. 2, the frequency of fact i across optimal sets.

The supervision unit need not equal the deployment unit. The target $p^\star$ is itself a per-item score, so the set-level signal can be represented as an item-wise selector that ranks facts and keeps the top B at test time, without subset search. When the budget-B optimum is unique, this ranking recovers it exactly; with non-unique optima the uniform marginal is a heuristic, which on MuSiQue leaves only 0.006 regret (Table 1). Whether set-level supervision can be learned into an item-wise selector on unseen instances is a separate empirical question. Section 6 tests this with independently annotated supporting-set supervision; Appendix W reports a matched test using reader-derived $p^\star$. Deployed selectors never invoke the exact oracle defining O_B; enumeration is confined to offline supervision and diagnostic evaluation. Appendix C gives the search procedure, its minimum-cardinality variant, tie-breaking, and a greedy fallback for large pools.

Why the set is the right unit, empirically. On Synthetic training instances, decision sufficiency is non-additive: the union of individually relevant facts (those with $a_i > 0$) is insufficient in 58% of instances, and the minimum decision-sufficient set is non-unique in 38% (Table 22; Appendix F). We therefore score kept sets by the regret they induce rather than by exact match to a tie-broken oracle. Holding solver and budget fixed and varying only how facts are chosen, only the set-derived criterion is consistently sufficient. Independent per-fact value leaves mean regret 3.4 to 3.8 and fails on one instance in four, a salience-like prior leaves about 5.9, and importance prompting about 7.0 (Appendix J). The theorem shows singleton-derived scores are not sufficient once budget-relevant value is carried by interactions; a per-fact score read off the set-level target can be.

6 LEARNING DECISION-SUFFICIENT MEMORY ON MULTI-HOP QA

The principle is to keep what changes the decision rather than what looks relevant. We test this on MuSiQue multi-hop question answering (Trivedi et al., 2022), where the ground truth is annotated independently of our solver. Each of 600 questions is paired with 20 candidate paragraphs, a few of which are *supporting* (the facts a correct multi-hop derivation must use) and the rest distractors. A fixed reader (Qwen3-8B) produces the answer from its input paragraphs, which constitute memory; a decision-sufficient memory keeps the supporting set. We compare against feeding all paragraphs (full context), salience retrieval (BM25 (Robertson & Zaragoza, 2009) and dense embeddings at several depths), recency (the most recent k), and padding the supporting set with retrieved paragraphs (Appendix R).

Gold decision-sufficient memory is both accurate and compact: the supporting set answers 42.2% exactly from 344 tokens, whereas full context reaches only 16.8% from 2447 tokens. The accuracy loss with more context is consistent with long-context degradation (Liu et al., 2024). On BabiLong, a 9-token decision-sufficient memory holds 0.945 across 1k–8k contexts while full context drops to $\sim$0.5 (Appendix U). On replayed ScienceWorld (Wang et al., 2022) trajectories, decision-sufficient oracle memory keeps 1.8 of 8.3 past observations and preserves the gold action, which recency and salience at that budget preserve on only 70% and 60% of steps (Appendix V). All learned selection baselines use the same per-instance budget and the same fixed greedy-decoded reader. Decision-sufficient selection, supervised by the dataset-provided gold supporting set, answers 0.345 exact match (EM), nearly double dense retrieval and triple BM25 at matched budget, while an identically-trained salience-label selector reaches only 0.107, isolating the label as the operative lever (Figure 3, Table 2). The paired gaps over the retrieval baselines are $+0.18$ and $+0.24$ EM, intervals excluding zero, and the ordering holds across three readers and replicates on HotpotQA (Appendices R, T).

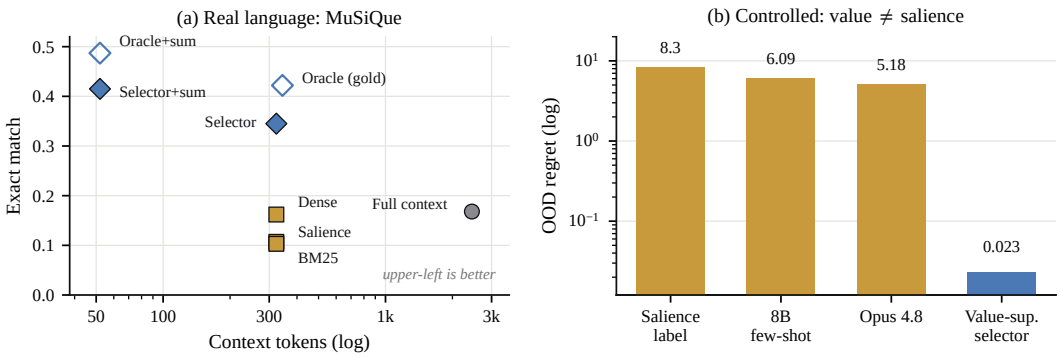

Figure 3: **The gap is large on real language and reproduces under controlled decorrelation** (amber: item-level/salience; blue: set-level/decision-sufficient). **(a)** On MuSiQue the decision-sufficient family sits upper-left, more answers from fewer tokens: the learned selector reaches 0.345 at 323 tokens and its query-focused summary 0.415 at ∼50. Dense retrieval (0.162), salience-label selection (0.107), and BM25 (0.103) trail at the same budget and full context reaches only 0.168 from 2447 tokens (open markers are oracle references; Table 2). **(b)** On controlled synthetic decisions that decorrelate value from surface salience, out-of-distribution regret for a salience-label selector (8.3), few-shot prompting (6.09), and a frontier model (5.18) stays two orders of magnitude above a value-supervised selector (0.023; Appendix K).

Table 2: **Matched per-instance budget on MuSiQue** (n=600, fixed Qwen3-8B reader). At matched budget, decision-sufficient selection nearly doubles dense retrieval and triples BM25; identical training (data, architecture, budget) with salience labels yields 0.107. MS-MARCO neural rerankers do not close the gap even with three paragraphs. **Bold**/underline: best/second. High-div: dense retrieval misses gold support; gold-recall: fraction of gold support retained.

Selection (matched budget)	EM	EM (high-div, 81%)	gold-recall
Supporting-set label (selector)	**0.345**	**0.320**	**0.828**
Salience label (same data, arch, budget)	0.107	0.093	0.422
Dense retrieval	<u>0.162</u>	0.085	0.570
BM25	0.103	0.087	0.394
Cross-encoder reranker (MS MARCO)	0.140	0.099	0.563
monoT5 reranker (MS MARCO)	0.157	<u>0.118</u>	<u>0.596</u>
Oracle (gold support)	0.422	0.409	1.000

On the 81% of questions where dense retrieval misses gold support the lead widens to 3.8× (0.320 vs 0.085), most on the deepest four-hop chains where salience misses the bridging facts.

A learnable decision-sufficient selector. Here decision sufficiency is the dataset-provided gold supporting set a correct multi-hop derivation must use, rather than a solver-verified minimum, so the supervision is independent of our solver. We train a selector (Qwen3-4B with quantized low-rank adapters, QLoRA) on these labels to predict the kept set from the question and candidate paragraphs, with no solver at test time. It reaches 34.5% EM at under one-third of full-context tokens (82% of the gold-support reference), keeping 2.5 paragraphs (323 tokens) at precision 0.86/recall 0.83 on held-out supporting facts (Appendix R). Gold supporting-set supervision thus trains a solver-free selector on real language (a label-free self-counterfactual variant is evaluated in Appendix W).

Selection is upstream of compression. Summarizing the learned selector's set raises exact match from 0.345 to 0.415 at ∼50 tokens, above summarizing full context, while token-level pruning collapses to 5.5% (summarized gold support: 0.487). What to keep precedes how to compress, and established compressors at their own budgets stay below the learned selection (Table 3).

The same supervision–scale pattern appears on real data. On MuSiQue, a frontier model (Claude Opus 4.8) only matches the 4B selector (0.350 vs 0.345 EM), contrasting with the roughly

Table 3: **Selection is upstream of compression** (same Qwen3-8B reader). Established compressors stay below summarized selection on both datasets. Summarized selection beats summarizing full context on MuSiQue ($+0.058$ $[.018, .098]$) and ties it on the shorter HotpotQA contexts (-0.018 $[-.047, .010]$). Oracle rows use the gold supporting set; the token column counts reader context only (query-focused summarizer protocol in Appendix R). RECOMP's abstractive model is trained on HotpotQA (out-of-domain on MuSiQue). †: LLMLingua-2 on HotpotQA over n=300, else n=600. Best deployable per dataset **bold**. Slashes separate MuSiQue / HotpotQA settings.

	MuSiQue		HotpotQA	
Method	EM	Tokens	EM	Tokens
LLMLingua-2 (target 200)	0.055	239	0.240†	224
LongLLMLingua (Jiang et al., 2024) (Llama-2-7B, tgt 256 / 228)	0.132	271	0.337	270
RECOMP abstractive (out- / in-domain)	0.200	65	0.453	58
Summarize full context (60-word)	0.357	48	**0.560**	40
Selector (ours; HotpotQA zero-shot)	0.345	323	0.498	228
Selector + query-focused summary (ours)	**0.415**	52	0.542	47
Gold supporting set (oracle ref.)	0.422	344	0.537	223
+ query-focused summary (oracle ref.)	0.487	52	0.588	47

two-orders-of-magnitude synthetic gap (Table 10). Capability can substitute for value supervision when surface cues carry the needed signal, but in the surface-decorrelated suites no prompting or scale baseline we test substitutes for it. The same selector transfers zero-shot to HotpotQA (0.498 EM at 228 tokens, on par with full context at $\sim 6\times$ fewer tokens; Table 3, Appendix T).

The solver–reader boundary. The theory's label-defining solver is assumed to satisfy monotone sufficiency; the consuming reader can be non-monotone and lose accuracy with added context. Appendix Z.9 bounds this: if forgotten facts perturb reader action values within ε at decision margin γ, its regret is at most 2ε and exactly zero below $\gamma/2$. Because readers degrade with distractors, supplying only the decision-sufficient set can push this distortion below the margin.

7 Controlled Mechanism

Suites and setup. Our three controlled suites are Synthetic, MiniGrid (Chevalier-Boisvert et al., 2023), and PDDLGym (Silver & Chitnis, 2020), with decision value deliberately decorrelated from surface salience. Instances have 20 shuffled, anonymously indexed candidate facts with distractors. Counterfactual regret is exact, and evaluation uses out-of-distribution (OOD) and held-out splits. Inputs omit the target action, preventing prompt copying. Qwen3-4B/8B selectors (Yang et al., 2025) are trained with QLoRA (Dettmers et al., 2023) with matched interfaces and only supervision varied (Appendices E, G, S, Y).

The failure reproduces under controlled decorrelation. In the Synthetic OOD matched value-versus-salience control, a value-supervised selector reaches near-zero regret (0.023) while the matched salience-label selector stays two orders of magnitude worse (8.3), with the gap due to fact choice, not count. The ordering holds across seeds and backbones; stronger prompting and larger or frontier models do not substitute for the supervision signal (Appendices S, K). Appendices I, N, X, L, M report full per-suite results, held-out failure modes, objective-shift stress testing, a vision-language diagnostic, and cost robustness.

8 Set Supervision Beyond Enumerable Pools

Although the set-derived target nearly closes oracle regret (Section 4), learned singleton- and set-derived selectors show no detectable difference on held-out enumerable pools (Δ=0.001, $[-0.031, 0.034]$; Appendix W). We transfer them without retraining to larger, imperfect global dense-retrieval pools of 50 and 100 candidates from $\sim$21k deduplicated passages. Fixing retriever, pool, reader, and budget isolates supervision construction. Both trail dense top-B, but set-derived

Table 4: **The supervision-unit distinction extends beyond enumerable pools** (Qwen3-8B reader, B=3, 430 held-out questions; selector rows average three seeds). Only the supervision unit differs; set-derived F1 leads every seed. Both trail dense top-B: a supervision contrast, not end-to-end retrieval gain. "R_n": all gold retrieved; question-clustered 95% CIs. The gold-support row keeps every retrieved gold paragraph and is not budget-matched. Pool and per-seed details: Appendix W.

	n=50		n=100	
Policy (keep B=3)	EM	F1	EM	F1
Dense top-B (retriever)	0.119	0.223	0.119	0.223
Item-derived selector	0.081	0.153	0.067	0.132
Set-derived selector	0.094	0.169	0.104	0.173
Oracle (gold support)	0.305	0.435	0.333	0.469
Δ F1 (set $-$ item), all	$+0.016$ $[-0.004, 0.036]$		$+0.040$ $[0.021, 0.063]$	
Δ F1 (set $-$ item), R_n	$+0.018$ $[-0.017, 0.054]$		$+0.061$ $[0.030, 0.093]$	

F1 leads all six seed comparisons, significantly at n=100 overall and on the retrievable subset (Table 4; Appendix W). This transfer moves from curated six-paragraph pools to global retrieval; we next train on candidates drawn from the same global retrieval pipeline.

With matched retrieval training, set-derived supervision outperforms the dense top-B policy. We evaluate both selectors on 500 fresh, outcome-sealed MuSiQue questions, disjoint from all earlier splits and uninspected until this measurement. Using the same global top-100 training pools, architecture, retrieved-candidate construction, training budget, and reader, we train an add-one item selector and a supporting-set selector (implementation, candidate cap, and training-only quantization audit: Appendix W). Sealed answer-F1 is 0.276 for set supervision, exceeding item supervision (0.204) and dense top-B (0.235): $\Delta_{\text{set-item}} = +0.072$ $[0.050, 0.095]$ and $\Delta_{\text{set-dense}} = +0.041$ $[0.013, 0.070]$, both excluding zero (question-clustered, three seeds); the item-supervised selector does not exceed the retriever (-0.031 $[-0.063, +0.002]$). Item labels are reader-derived add-one values, whereas set labels are dataset-provided supporting sets, so label source and unit vary together. The frozen transfer above holds the label source fixed and varies only the unit.

9 Related Work

Memory and retrieval commonly use independent similarity, salience, or recency scores, or item-wise context pruning (Lewis et al., 2020; Robertson & Zaragoza, 2009; Xiao et al., 2024; Packer et al., 2023; Park et al., 2023; Zhong et al., 2024; Shinn et al., 2023; Maharana et al., 2024; Wu et al., 2025; Jiang et al., 2023; Pan et al., 2024; Li et al., 2023; Mu et al., 2023; Ge et al., 2024; Xu et al., 2024; Zhang et al., 2023). Set-aware scoring and set-conditioned utility model joint sufficiency or compatibility (Lee et al., 2025; Jiang et al., 2026; Song & Lee, 2026), complementing our identification question (Appendix W). Decision-aware memory, data valuation, and interaction-aware attribution study downstream value or complementarity (Srivastava, 2026; Wang et al., 2024; Heo et al., 2026). We establish memory-specific non-identification: exact anchored add-one and delete-one evaluations can agree across monotone value functions with distinct optimal budgeted sets. We measure the resulting regret on sealed language data.

10 Conclusion

Anchored item-wise signals need not identify the memory set a decision needs, and on sealed MuSiQue the resulting regret is large and interaction-localized (Theorem 1, Section 4). A set-derived item score nearly closes this gap. When trained on a realistic retrieval distribution, set-supervised selection outperforms add-one item supervision and the dense top-B policy on a fresh sealed set (Section 8). The theory identifies a limit of the observed supervision signal; the learned experiments test whether set-derived targets generalize to new questions. Together, they motivate obtaining supervision from jointly retained facts while keeping deployment inexpensive, since an item-wise policy can carry set-level information without online subset search. Deployment may remain item-wise, but when downstream value is interactive, the unit of supervision is the set.

AI USE STATEMENT

Generative AI assistants were used for limited assistance with coding, language editing, and internal consistency checks. The authors take responsibility for all content of this work, including its claims, proofs, and experimental results.

ETHICS STATEMENT

This work studies memory selection for language agents using publicly available benchmarks (MuSiQue, HotpotQA, BabiLong) and synthetic decision suites; it involves no human subjects, no personally identifiable data, and no new data release. The method chooses which existing items a downstream reader sees rather than generating content, so we foresee no ethical risks beyond those already present in the underlying models and datasets.

REPRODUCIBILITY STATEMENT

All trained selectors are open-weight (Qwen3-1.7B/4B/8B) with the QLoRA settings given in the experimental setup and the appendix, and the counterfactual-forgetting oracle (Algorithm 1) and all benchmark generators are deterministic given fixed seeds. The synthetic decision suites, the MiniGrid- and PDDLGym-derived suites, and the real multi-hop question-answering setups (MuSiQue and HotpotQA, Section 6), and long-context BabiLong (Appendix U) are described in the main text and appendix; evaluation uses solver- and annotation-grounded regret and exact-match accuracy. We fix random seeds and report variance across seeds and backbone sizes (Section K). The appendix lists training hyperparameters, prompt templates, the exact oracle, dataset statistics, and the code and release plan (Appendix Q); the frontier model (Claude Opus 4.8) appears only as a prompting baseline.

REFERENCES

David Abel, David Ellis Hershkowitz, and Michael L. Littman. Near optimal behavior via approximate state abstraction. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

David Andre and Stuart J. Russell. State abstraction for programmable reinforcement learning agents. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, 2002.

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. BabyAI: A platform to study the sample efficiency of grounded language learning. In *International Conference on Learning Representations*, 2019.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid and miniworld: Modular and customizable reinforcement learning environments for goal-oriented tasks. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. doi: 10.52202/075280-3209.

Vašek Chvátal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):233–235, 1979. doi: 10.1287/moor.4.3.233.

Thomas L. Dean and Robert Givan. Model minimization in markov decision processes. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, 1997.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, 2023. doi: 10.52202/075280-0441.

Irit Dinur and David Steurer. Analytical approach to parallel repetition. In *Proceedings of the Symposium on Theory of Computing (STOC)*, pp. 624–633, 2014. doi: 10.1145/2591796.2591884.

Uriel Feige. A threshold of ln n for approximating set cover. *Journal of the ACM*, 45(4):634–652, 1998. doi: 10.1145/285055.285059.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 162–169, 2004.

Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. In *International Conference on Learning Representations*, 2024.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003. doi: 10.1016/S0004-3702(02)00376-4.

Jaeseung Heo, Kyeongheung Yun, Youngbin Choi, Sehyun Hwang, Jungseul Ok, and Dongwoo Kim. Interaction-aware influence functions for group attribution. *arXiv preprint arXiv:2605.15675*, 2026.

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 *International Conference on Learning Representations*, 2022.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 1769–1782, 2023.

Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 287–318, 2023.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.18653/v1/2023.emnlp-main.825.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677, 2024. doi: 10.18653/v1/2024.acl-long.91. URL https://aclanthology.org/2024.acl-long.91/.

Yi Jiang, Sendong Zhao, Jianbo Li, Bairui Hu, Yanrui Du, Haochun Wang, and Bing Qin. OptiSet: Unified optimizing set selection and ranking for retrieval-augmented generation. *arXiv preprint arXiv:2601.05027*, 2026.

Richard M. Karp. Reducibility among combinatorial problems. In Raymond E. Miller and James W. Thatcher (eds.), *Complexity of Computer Computations*, pp. 85–103. Plenum Press, 1972. doi: 10.1007/978-1-4684-2001-2_9.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. doi: 10.52202/079017-3381.

Dahyun Lee, Yongrae Jo, Haeju Park, and Moontae Lee. Shifting from ranking to set selection for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. doi: 10.18653/v1/2025.acl-long.861.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020.

Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for mdps. *Proceedings of the International Symposium on Artificial Intelligence and Mathematics*, 2006.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.18653/v1/2023.emnlp-main.391.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2024. doi: 10.18653/v1/2024.acl-long.747.

Jesse Mu, Xiang Lisa Li, and Noah Goodman. Learning to compress prompts with gist tokens. In *Advances in Neural Information Processing Systems*, 2023. doi: 10.52202/075280-0848.

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

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. doi: 10.18653/v1/2024.findings-acl.57.

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

Balaraman Ravindran and Andrew G. Barto. An algebraic approach to abstraction in reinforcement learning. In *Proceedings of the Twelfth Yale Workshop on Adaptive and Learning Systems*, pp. 109–114, 2003.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, 2023. doi: 10.52202/075280-2997.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023. doi: 10.52202/075280-0377.

Tom Silver and Rohan Chitnis. PDDLGym: Gym environments from PDDL problems. In *International Conference on Automated Planning and Scheduling Workshop on Bridging the Gap Between AI Planning and Reinforcement Learning*, 2020.

Mooho Song and Jay-Yoon Lee. Retrieving a set, not independent passages: Set-level compatibility learning for efficient set exploration. *arXiv preprint arXiv:2607.05712*, 2026.

Saksham Sahai Srivastava. Causal intervention-based memory selection for long-horizon LLM agents. *arXiv preprint arXiv:2605.17641*, 2026.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475.

Jiachen T. Wang, Tianji Yang, James Zou, Yongchan Kwon, and Ruoxi Jia. Rethinking data Shapley for data selection tasks: Misleads and merits. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 52033–52063, 2024.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Cote, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. doi: 10.18653/v1/2022.emnlp-main.775.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. doi: 10.52202/068431-1800.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations (ICLR)*, 2025.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general chinese embeddings. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649, 2024. doi: 10.1145/3626772.3657878.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with compression and selective augmentation. In *International Conference on Learning Representations*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. 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 (EMNLP)*, 2018. doi: 10.18653/v1/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 *International Conference on Learning Representations*, 2023.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, 2023. doi: 10.52202/075280-1506.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2024. doi: 10.1609/aaai.v38i17.29946.

A Appendix Overview

This appendix expands the proof sketch, experimental settings, full ablation tables, the visual-language diagnostic, prompt and interface details, qualitative examples, and reproducibility details.

B Value-Preservation Bound

Keeping only the facts that preserve a decision's value connects to the classical view of value-preserving state abstraction, MDP homomorphisms, and bisimulation (Dean & Givan, 1997; Givan et al., 2003; Ferns et al., 2004; Li et al., 2006; Andre & Russell, 2002; Ravindran & Barto, 2003); the bound below states this preservation over a generated language state.

Let $\phi(h)$ map histories to language abstractions. Write R for reward, P for the transition law, D_{TV} for total variation distance, and $\gamma \in [0, 1)$ for the discount factor. Assume that histories sharing an abstraction have bounded one-step reward error and bounded next-abstraction transition error:

$$|R(h, a) - R(h', a)| \le \epsilon_R$$

and

$$D_{\mathrm{TV}}(P(\phi(H_{t+1}) \mid h, a), P(\phi(H_{t+1}) \mid h', a)) \le \epsilon_P.$$

Let the maximum absolute reward be $R_{\max}$. Standard simulation-lemma arguments for approximate state abstraction and bisimulation (Ferns et al., 2004; Li et al., 2006; Abel et al., 2016) imply that the value error accumulated from reward mismatch is bounded by $\epsilon_R/(1-\gamma)$, while transition mismatch changes the distribution over future abstract states by at most ϵ_P per step and propagates value error bounded by $O(\gamma R_{\max}\epsilon_P/(1-\gamma)^2)$. Therefore the policy optimized over abstract language states satisfies

$$V^*(h) - V^{\pi_\phi}(h) \le O\left(\frac{\epsilon_R}{1-\gamma} + \frac{\gamma R_{\max}\epsilon_P}{(1-\gamma)^2}\right).$$

The role of counterfactual forgetting is to construct instance-level supervision that approximates this value-preservation criterion without requiring hand-written equivalence classes.

C Exact Forgetting Oracle

Algorithm 1 gives the exact minimum-cardinality forgetting oracle for offline supervision. Here $V^*_{M_F}(s)$ and $Q^*_{M_F}(s, a)$ are the full model's optimal state and action values. It enumerates fact subsets in nondecreasing cardinality and returns the first whose robust (minimax-regret) action is ϵ-value-preserving over the induced uncertainty set, so the returned label is exactly minimal. When $|F(h)|$ is too large for exhaustive search, we replace the inner enumeration with greedy deletion: starting from all facts, repeatedly drop the fact whose removal least increases robust regret while the result stays within ϵ. The greedy variant is linear per pass but approximate; it is not used for the synthetic exact-oracle results (the ScienceWorld labels of Appendix V instead use a greedy backward-elimination oracle, described there as approximately minimal). In practice the cardinality-increasing search terminates as soon as a minimal set is found (3–5 facts in our suites), so it is cheap despite the worst-case exponential bound. For very large fact sets, submodular, integer-programming, or evolutionary search offer approximate acceleration that trades exact minimality for speed.

Algorithm 1 Exact Minimum-Cardinality Forgetting Oracle

Input: instance h with latent state $s^\star$, candidate facts $F(h)$, full model M_F, tolerance ϵ
for $k = 0$ **to** $|F(h)|$ **do**
  for all $z \subseteq F(h)$ with $|z| = k$, in fixed feature order **do**
    $\Omega(z) \leftarrow \{s : s \text{ agrees with every fact in } z\}$
    $a^{\mathrm{rob}} \leftarrow \arg\min_a \max_{s \in \Omega(z)} [V^*_{M_F}(s) - Q^*_{M_F}(s, a)]$
    $b \leftarrow \max_{s \in \Omega(z)} [V^*_{M_F}(s) - Q^*_{M_F}(s, a^{\mathrm{rob}})]$
    if $b \le \epsilon$ **then**
      return minimum-cardinality oracle z and action a^{rob}
    end if
  end for
end for

D SUFFICIENCY FOR AN OBJECTIVE FAMILY

A trained action policy can match a decision-sufficient kept set on regret when the objective is fixed. The two objects separate once the downstream objective is not known at compression time. Consider an instance with full state $s^\star$, candidate facts F, and a family of objectives $\mathcal{O}$ (e.g., reward reweightings or risk levels), each $o \in \mathcal{O}$ inducing optimal value V_o^* and action values Q_o in the full model. For $z \subseteq F$, let $\Omega(z)$ be the states consistent with z and

$$a_o^{\mathrm{rob}}(z) \in \arg\min_a \max_{s \in \Omega(z)} \big(V_o^*(s) - Q_o(s,a)\big).$$

Call z *ϵ-family-sufficient* if $\max_{o \in \mathcal{O}} \big(V_o^*(s^\star) - Q_o(s^\star, a_o^{\mathrm{rob}}(z))\big) \le \epsilon$. We consider the *post-compression* setting: z is produced before o is revealed, and the decision is then made from z alone, as in memory compression.

Proposition. If no single action is optimal for all objectives (in particular, when two objectives $o, o' \in \mathcal{O}$ have different *unique* optimal actions at $s^\star$, $a_o^*(s^\star) \neq a_{o'}^*(s^\star)$), then any policy that commits to a single action before o is revealed incurs strictly positive regret for some $o \in \mathcal{O}$, whereas a family-sufficient kept set (the $\epsilon = 0$ case) incurs zero regret for every $o \in \mathcal{O}$. Hence a minimum-cardinality family-sufficient set is a sufficient statistic for the decision family $\{(M, o) : o \in \mathcal{O}\}$; an action label is sufficient only for a single objective.

Proof. A committed action a has worst-case regret $\max_o \big(V_o^*(s^\star) - Q_o(s^\star, a)\big)$. Since the optima disagree, at least one of o, o' has its optimal action, assumed unique here, different from a, so that objective has strictly positive regret at a. A family-sufficient z instead defers to $a_o^{\mathrm{rob}}(z)$ after o is revealed, which is optimal for each o, so its regret is zero for every o. The argument extends to $\epsilon > 0$ under the stronger hypothesis that no single action is ϵ-optimal for all of $\mathcal{O}$. □

We instantiate $\mathcal{O}$ as a risk-sensitivity family and verify the predicted separation empirically: a family-sufficient selector keeps a slightly larger state but preserves the decision across the family, while same-objective action cloning and a single-objective selector both incur substantial regret once the objective shifts.

E DATASET CONSTRUCTION

Each instance consists of candidate facts, a high-level action set, hidden decision variables, a solver-backed full model, and a solver-generated oracle kept set. Candidate facts include both decision-relevant and decision-irrelevant facts. Irrelevant facts are not arbitrary noise; many are plausible natural-language observations such as colors, unused objects, domain comments, type labels, or route descriptions that are semantically meaningful but do not change the optimal decision.

Synthetic. The controlled suite contains 30,000 train, 4,000 dev, and 16,000 total test instances across IID (same distribution as training), neutral (independent distractors), OOD (reversed distractor correlations), and held-out task families. Train families include lava shortcuts, locked doors, and fragile bridges. The held-out family is probe-or-act.

MiniGrid. The grounded suite builds on MiniGrid/BabyAI-style environments (Chevalier-Boisvert et al., 2019; 2023) and contains 12,000 train, 1,500 dev, and 6,000 test instances. It includes key-door, lava-gap, and route-risk structures. The held-out family is unlock.

PDDLGym. The planning suite uses PDDLGym-style planning domains (Silver & Chitnis, 2020) with stochastic/reward wrappers and contains 12,000 train, 1,500 dev, and 6,000 test instances. The held-out family is miconic.

F DECISION-SUFFICIENCY STRUCTURE

Table 5 gives the per-family decomposition of the structure summarized in Table 22, on the Synthetic train split. All three train families show the same pattern, so the aggregate effect is not driven by

a single family; the held-out family (probe-or-act) behaves similarly (complementarity 0.60, non-uniqueness 0.44). Among non-unique instances the mean number of distinct minimum decision-sufficient sets is 2.4–2.7.

Table 5: Per-family decision-sufficiency structure (Synthetic train split). Complementarity: individually relevant facts jointly fail sufficiency. Non-unique: multiple minimum sufficient sets. Oracle size: mean minimum cardinality.

Family	Oracle size	Complementarity	Non-unique
lava_shortcut	3.68	0.603	0.415
locked_door	3.85	0.565	0.345
fragile_bridge	3.77	0.586	0.380

Complementarity persists when the tolerance is relaxed. At tolerance ϵ, a fact counts as individually relevant when its leave-one-out sensitivity exceeds ϵ, and the union of these facts is sufficient when its regret is at most ϵ; both tests change with ϵ. As ϵ grows the minimum subset shrinks, but the gap to marginal relevance widens: complementarity is 0.585 at $\epsilon = 0$, 0.598 at $\epsilon = 0.5$, 0.694 at $\epsilon = 1$, and 0.729 at $\epsilon = 2$.

Worked examples. *Complementarity.* In a locked-door instance the individually relevant facts are `door_locked`, `key_reachable`, `timeout_tight`, and `bridge_fragile`; a minimum sufficient kept set additionally requires `shortcut_blocked`, which is decision-irrelevant in isolation but required jointly. *Non-uniqueness.* A fragile-bridge instance admits multiple minimum decision-sufficient sets of size four; two of them, {`shortcut_has_lava`, `door_locked`, `door_route_long`, `bridge_fragile`} and {`shortcut_has_lava`, `door_locked`, `timeout_tight`, `bridge_fragile`}, both yield zero regret.

G Training Details

We train Qwen3-4B and Qwen3-8B selectors (Yang et al., 2025) with QLoRA and LoRA adapters (Dettmers et al., 2023; Hu et al., 2022). The configuration is:

- 4-bit NF4 quantization with double quantization.
- LoRA rank $r = 16$, alpha $= 32$, dropout $= 0.05$.
- Target modules: q, k, v, o projections and gate/up/down MLP projections.
- Learning rate $2 \cdot 10^{-4}$, cosine scheduler, warmup ratio 0.03.
- One epoch.
- Batch size 2 with gradient accumulation 8 for the 4B selector; gradient checkpointing enabled for the 8B selector.

The configuration was selected by development performance. We report sensitivity to rank, data size, epochs, model size, and the per-fact memory cost λ. The synthetic main selectors (Table 7) use one epoch. The matched value-vs-salience comparison (Appendix S) and the MuSiQue selector (Appendix R) use separate two-epoch runs.

H Qualitative Kept Sets

Table 6 gives representative OOD cases. The direct-action baseline sees the same candidate facts but makes a high-regret choice; the selector keeps the value-determining facts and leaves action selection to the downstream solver.

Table 6: Representative OOD kept sets. Direct and Selector columns give realized decision regret; the selected facts preserve decision value.

Bench.	Oracle facts	selector-selected facts	Direct	Selector
Synthetic	fast route avoids lava; shortcut surface is dry; direct passage is passable; bridge may fail	same four facts	110.0	0.0
MiniGrid	enough time remains; safe detour is short; direct route crosses lava	same three facts	94.0	0.0
PDDLGym	no risky transition on direct plan; short-plan preconditions hold; relevant action is reliable	same three facts	116.0	0.0

I Full Main Results

Table 7 is the main OOD summary across the three synthetic-family suites (all baselines and mean facts kept); Table 8 expands it with per-split (OOD and held-out) numbers.

Table 7: OOD regret and mean facts kept across the three suites. Top block: reference bounds (oracle, all observations); middle: selection under a matched interface and solver; bottom: Direct-SFT (supervised fine-tuning), a same-data trained action model that ties the selector on a fixed objective but yields no inspectable state (Section X); it was trained and evaluated on Synthetic only. Dashes: no facts are kept, or not run. **Bold**/underline: best/second-best.

	Synthetic		MiniGrid		PDDLGym	
Method	Reg.	Facts	Reg.	Facts	Reg.	Facts
Oracle	0.000	3.78	0.000	4.24	0.000	3.66
All obs.	0.000	20.0	0.000	20.0	0.000	20.0
Random	8.274	3.78	8.922	4.24	3.577	3.66
Zero-shot	9.688	2.97	8.714	3.43	3.626	2.28
Direct (prompt)	41.12	–	23.04	–	21.00	–
Reasoning	18.59	–	16.37	–	18.36	–
Value-sup. 4B	<u>0.032</u>	3.94	<u>0.284</u>	4.26	<u>0.127</u>	3.58
Value-sup. 8B	**0.023**	3.93	**0.210**	4.25	**0.066**	3.61
Direct-SFT	0.000	–	–	–	–	–

Table 8 includes OOD and held-out results. Confidence intervals for learned models are across seeds when three runs are available. Single-run LLM baselines report example-level bootstrap intervals in the generated artifact files.

Table 8: Expanded split-level results. Regret: mean decision loss; Net: mean regret plus λ times facts kept ($\lambda = 0.05$); Zero: zero-regret fraction; Runs: training runs. Dashes: Net is undefined for direct action, which keeps no facts.

Benchmark	Split	Method	Runs	Regret	Net	Zero
Synthetic	OOD	Direct action, Qwen3-4B	1	41.118	–	0.134
Synthetic	OOD	Zero-shot select IDs	1	9.688	9.837	0.394
Synthetic	OOD	Selector, Qwen3-4B	3	<u>0.032</u>	<u>0.229</u>	<u>0.987</u>
Synthetic	Held-out	Selector, Qwen3-4B	3	**0.025**	**0.216**	**0.993**
MiniGrid	OOD	Direct action, Qwen3-4B	1	23.040	–	0.166
MiniGrid	OOD	Zero-shot select IDs	1	8.714	8.886	0.262
MiniGrid	OOD	Selector, Qwen3-4B	3	**0.284**	**0.497**	**0.952**
MiniGrid	Held-out	Selector, Qwen3-4B	3	<u>0.533</u>	<u>0.752</u>	<u>0.871</u>
PDDLGym	OOD	Direct action, Qwen3-4B	1	20.996	–	0.246
PDDLGym	OOD	Zero-shot select IDs	1	3.626	3.740	0.448
PDDLGym	OOD	Selector, Qwen3-4B	3	**0.127**	**0.306**	**0.965**
PDDLGym	Held-out	Selector, Qwen3-4B	3	<u>0.758</u>	<u>0.954</u>	<u>0.677</u>

Table 10: On Synthetic OOD, prompting and model scale stay two orders of magnitude above the trained value-supervised selector, under a common candidate pool and selection interface.

OOD selection	Regret	Facts
8B few-shot	6.09	4.98
8B self-cons.	6.23	4.57
14B few-shot	4.89	5.80
Opus 4.8	5.18	3.84
Value-sup. 4B	0.032	3.94
Value-sup. 8B	**0.023**	3.93

J Selection Criterion at Fixed Budget

Holding the downstream solver and the selection budget fixed (every rule keeps exactly the oracle cardinality, except *All relevant*, which keeps all facts) and varying only how facts are chosen isolates the selection criterion from the mere act of keeping few facts (Table 9). Ranking by independent per-fact value leaves mean regret 3.42/3.80 (IID/OOD) and is zero-regret on 75% of instances; a frequency-prior salience score leaves 5.85/5.89 (42%); an LLM prompted to keep the most important facts does no better (6.97/7.49, 39%); random selection is worse still (8.39/8.85, 36%). Only the combinatorial value oracle reaches zero regret at the oracle budget; keeping all facts also suffices but forgoes compression. Salience, whether a global prior or an LLM's own importance judgment, tracks what looks relevant, not what changes downstream value.

Table 9: Selection criterion with solver and budget fixed. Mean regret (in-distribution / out-of-distribution) and zero-regret rate (in-distribution); only the combinatorial value oracle is consistently sufficient. Every rule keeps the oracle cardinality except *All relevant (no compr.)*, which keeps all facts.

Selection criterion	IID	OOD	Zero-reg.
Value oracle (comb.)	**0.00**	**0.00**	**1.00**
Anchored marginal (top-k)	3.42	3.80	0.75
Freq. prior (salience)	5.85	5.89	0.42
LLM salience (prompt)	6.97	7.49	0.39
Random	8.39	8.85	0.36
All relevant (no compr.)	0.00	0.00	1.00

K Fair Prompting and Model-Scale Ablations

Tables 11 and 12 provide the full fair-prompting results, using the same candidate facts and action interfaces as the primary baselines but adding three-shot demonstrations, self-consistency over five sampled selections (Wang et al., 2023; Wei et al., 2022), and a larger (14B) open model. They test whether the gap is due to weak prompting rather than value-supervised selection. Table 10 summarizes the Synthetic OOD split, including a frontier prompting baseline (Claude Opus 4.8): prompting and scale stay two orders of magnitude above the trained selector.

Robustness and transfer. Across two seeds and two backbones (Qwen3-4B/8B), decision-sufficient selection stays near zero regret on IID and OOD, degrades gracefully as distractors grow from 20 to 200, and is stable to LoRA rank, epoch count, and the cost weight λ; held-out domains that introduce new relevance structure are the hardest transfer case. A separate analysis shows that the kept set, unlike a distilled action, remains sufficient when the objective is revealed only after compression (Appendix M, X).

Table 11: Fair prompting and model-scale ablations on Synthetic. Larger open models and stronger prompting improve over naive direct action, but remain far above the selector regret. Dashes: Net is undefined for direct action, which keeps no facts.

Benchmark	Model	Split	Mode	Regret	Net	Zero
Synthetic	Qwen3-8B	OOD	few-shot direct action	21.128	–	0.413
Synthetic	Qwen3-8B	OOD	few-shot select IDs	6.090	6.339	0.327
Synthetic	Qwen3-8B	OOD	self-consistency select IDs	6.228	6.457	0.333
Synthetic	Qwen3-14B	OOD	few-shot direct action	17.101	–	0.310
Synthetic	Qwen3-14B	OOD	few-shot select IDs	**4.887**	**5.177**	**0.473**
Synthetic	Qwen3-14B	OOD	self-consistency select IDs	9.443	9.564	0.353
Synthetic	Qwen3-8B	Held-out	few-shot direct action	15.556	–	0.353
Synthetic	Qwen3-8B	Held-out	few-shot select IDs	**4.704**	**4.954**	0.383
Synthetic	Qwen3-8B	Held-out	self-consistency select IDs	5.017	5.250	0.357
Synthetic	Qwen3-14B	Held-out	few-shot direct action	14.241	–	0.333
Synthetic	Qwen3-14B	Held-out	few-shot select IDs	4.981	5.243	**0.510**
Synthetic	Qwen3-14B	Held-out	self-consistency select IDs	9.090	9.179	0.373

Table 12: Fair prompting ablations on external benchmark-derived suites using Qwen3-8B. Prompting produces valid JSON but remains high-regret relative to the trained selector. Dashes: undefined for direct action, which keeps no facts.

Benchmark	Split	Mode	Regret	Net	Zero	Facts
MiniGrid	OOD	few-shot direct action	14.053	–	0.297	–
MiniGrid	OOD	few-shot select IDs	**6.093**	**6.327**	**0.340**	4.66
MiniGrid	OOD	self-consistency select IDs	6.413	6.627	0.330	4.27
MiniGrid	Held-out	few-shot direct action	14.337	–	0.267	–
MiniGrid	Held-out	few-shot select IDs	**5.853**	**6.093**	**0.353**	4.79
MiniGrid	Held-out	self-consistency select IDs	6.103	6.329	0.340	4.51
PDDLGym	OOD	few-shot direct action	22.747	–	0.313	–
PDDLGym	OOD	few-shot select IDs	**3.367**	**3.584**	**0.480**	4.36
PDDLGym	OOD	self-consistency select IDs	3.443	3.650	0.467	4.14
PDDLGym	Held-out	few-shot direct action	29.713	–	0.310	–
PDDLGym	Held-out	few-shot select IDs	**2.110**	**2.336**	**0.647**	4.51
PDDLGym	Held-out	self-consistency select IDs	2.180	2.390	0.643	4.20

L VISUAL-LANGUAGE DIAGNOSTIC

Table 13 reports the MiniGrid visual-language diagnostic. The same open vision-language model (VLM) is evaluated with image+text and text-only inputs, under zero-shot and few-shot prompting. Text-only prompting yields lower regret than image+text prompting under both zero-shot and few-shot conditions on these splits.

Table 13: Qwen3-VL-4B MiniGrid diagnostic. Image+text prompting is not sufficient to recover decision-sufficient kept sets; text-only prompting is stronger in this interface.

Split	Variant	Input	Regret	Net	Zero	Facts
OOD	image+text zero-shot	image+text	5.920	6.227	0.390	6.14
OOD	text-only zero-shot	text	4.860	5.270	0.460	8.21
OOD	image+text few-shot	image+text	5.680	5.975	0.390	5.89
OOD	text-only few-shot	text	3.970	4.277	0.440	6.14
Held-out	image+text zero-shot	image+text	8.260	8.518	0.340	5.17
Held-out	text-only zero-shot	text	5.420	5.818	0.470	7.96
Held-out	image+text few-shot	image+text	7.840	8.063	0.270	4.46
Held-out	text-only few-shot	text	5.330	5.601	0.380	5.42

M COST SENSITIVITY

We evaluate net regret with $\lambda \in \{0, 0.01, 0.05, 0.1\}$ (Table 23). The selector's net regret grows smoothly because its learned kept-set size is close to the oracle minimal size. All-observation base-

lines preserve value but pay the memory cost on every one of their facts, so their net regret grows fastest as λ increases.

N Held-Out Failure Modes

Held-out failures are concentrated in instances where the selector misses a newly important relevance pattern. In PDDLGym heldout, frequent missed oracle features include inspection-resolved preconditions and resource reachability, while extra selected features often come from familiar shortcut/safety patterns. This supports the interpretation that the model transfers over known decision factors but struggles when a held-out domain changes which kinds of facts are value-determining.

O Qualitative Examples

MiniGrid OOD. In one instance, the optimal action is a detour. The oracle kept set contains: enough time remains, the safe detour is short, and the direct route crosses lava. The selector selects exactly these facts, while direct action prediction chooses the direct route and incurs high regret. The example illustrates why reading the full observation is not equivalent to identifying value-determining facts.

PDDLGym OOD. In a blocks-style planning instance, the oracle kept set includes that the direct plan avoids risky transitions, the short plan preconditions are satisfied, and the relevant action is reliable. The selector selects these facts and preserves the short-plan action. A direct LLM action baseline selects a resource plan, incurring high regret.

P Prompt Templates

The selector prompt contains: task family, candidate high-level actions, candidate observed facts with IDs, and an instruction to return valid JSON with selected fact IDs and a rationale. The model is not asked to choose the action directly. Direct-action baselines receive the same action list and facts but must return an action JSON. Reasoning-action baselines receive an additional instruction to reason briefly before returning an action. Zero-shot select-ID baselines receive the same fact-selection interface but no solver-generated fine-tuning. Few-shot baselines prepend three training demonstrations with gold selected fact IDs or gold actions, and self-consistency samples five fact-selection outputs before majority aggregation.

Fact-selection prompt skeleton.

> You are evaluating a sequential decision problem. Use only the observed facts and listed high-level actions. Select the smallest set of observation IDs needed to preserve the best action. Do not select decorative, naming, ordering, color, or spuriously correlated facts unless they affect action value. Return only valid JSON with keys `selected_fact_ids` and `rationale`.

Direct-action prompt skeleton.

> You are evaluating a sequential decision problem. Use only the observed facts and listed high-level actions. Choose the single best high-level action. Return only valid JSON with key `action`.

Q Reproducibility

We provide anonymized code as supplementary material: the benchmark generators, the counterfactual-forgetting oracle, the training and evaluation scripts for every baseline, and the result-collection scripts that produce each reported table and figure; a de-anonymized public release follows on publication. All learned components are open-weight (Qwen3-1.7B/4B/8B) fine-tuned with QLoRA, and the benchmark generators and oracle are deterministic given fixed seeds, so every synthetic number is reproducible end to end; the frontier proprietary model (Claude Opus 4.8) is used

only as a prompting baseline. Experiments run under Python 3.11 with PyTorch, Transformers, PEFT, and bitsandbytes.

R MATCHED-BUDGET VALUE VS. SALIENCE ON MUSIQUE

The MuSiQue selector is trained on the human supporting-fact annotations, which mark the decision-sufficient set: the paragraphs a correct multi-hop derivation must use. It is trained exclusively on the official MuSiQue training split and evaluated on 600 questions from the validation split; the two sets are disjoint by question id. At each MuSiQue question the trained selector keeps k paragraphs; we give dense retrieval and BM25 the same per-instance budget k and feed every selection to the same greedy-decoded Qwen3-8B reader, isolating selection quality from budget and reader. Table 2 reports exact match over the 600 questions. Value-based selection roughly doubles dense retrieval and triples BM25 at identical budget, and on the 81% of questions where dense retrieval drops a gold-supporting paragraph the gap widens to $3.8\times$. Paired bootstrap intervals (10k resamples) place the value$-$dense gap at $+0.183$ $[+0.147, +0.220]$ and value$-$BM25 at $+0.242$ $[+0.203, +0.280]$, both excluding zero. The advantage over dense retrieval is largest on the deepest four-hop chains (Table 14), where salience most often fails to recover the bridging facts. MS-MARCO-trained neural rerankers (a MiniLM cross-encoder and monoT5) rank paragraphs by learned relevance; at an equal-or-larger budget they reach 0.14–0.17 EM, comparable to dense retrieval and below value-based selection (Table 2). These rerankers add generic relevance supervision and neural ranking capacity yet do not close the gap. A same-data control isolates the cause directly: a selector trained on identical data, architecture, per-instance budget, and schedule (Qwen3-4B, 2 epochs) but on BM25-salience labels instead of decision-sufficient labels reaches only 0.107 EM (Table 2), a $3.2\times$ drop from flipping the label alone. This mirrors the synthetic label-vs-training confound (Appendix S): the decision-sufficient label, not supervised training or model scale, drives the result.

Query-focused summarization (Table 3). The summarize-then-answer rows use the same Qwen3-8B model as a separate summarizer. Given the question and the passages, it greedily produces a summary of at most 60 words with the prompt "Summarize the passages in at most 60 words, keeping only information needed to answer the question. Question: {q} ... Passages: {ctx}"; the summarizer receives the question but never the gold answer, and emits a summary rather than an answer. A separate greedy reader pass then answers the original question from the summary. Learned-summary rows summarize the trained selector's output ($0.345 \rightarrow 0.415$ EM); oracle-reference rows summarize the dataset-provided gold supporting set ($0.422 \rightarrow 0.487$). Token counts are reader-context tokens; the summarizer pass is additional compute, so these rows compare context compression rather than inference cost.

The full matched-budget head-to-head is Table 2 in the main text; the hop-depth breakdown follows.

Table 14: Exact match by hop depth at matched budget. The advantage over dense retrieval is largest on the deepest four-hop questions.

Hop depth	n	Value-based	Dense	BM25
2-hop	340	**0.368**	0.176	0.097
3-hop	157	**0.293**	0.153	0.115
4-hop	103	**0.350**	0.126	0.107

The advantage is also reader-agnostic. Feeding the same matched-budget selections to three different reader models preserves the value-over-salience ordering at every reader (Table 15), confirming that the gain comes from *what is selected* rather than from a particular reader.

Table 15: Reader-agnostic generality on MuSiQue (matched budget, n=600). The identical selections are fed to three reader models; value-based selection beats both salience retrievers at every reader.

Reader	Value-based	Dense	BM25	Oracle
Qwen3-8B	**0.345**	0.162	0.103	0.422
Llama-3.1-Tulu-8B	**0.438**	0.232	0.153	0.510
Gemma-2-9B	**0.450**	0.245	0.153	0.565

Frontier comparison protocol. The frontier baseline quoted in Section 6 uses the same reader and interface as the trained selector. Claude Opus 4.8 is prompted to select supporting paragraphs from the same candidate pool; it keeps 2.33 paragraphs (302 tokens) on average, and its selections are scored by the same greedy-decoded Qwen3-8B reader. It reaches 0.350 EM, essentially tied with the 4B selector's 0.345 at 323 tokens under this matched reader and interface. Capability substitutes for value supervision on MuSiQue precisely because the decision value is already carried by the paragraph text.

Sealed confirmatory replication. Table 1 reports a frozen-protocol replication on 250 held-out MuSiQue questions, with exact enumeration at budgets $B \in \{2, 3\}$ over a pool of $K = 6$ candidate paragraphs (all gold supporting paragraphs, filled to six with BM25-ranked distractors) and a deterministic reader (4-bit Qwen3-8B, greedy decoding), run after the protocol was fixed. A selection given exact singleton signals, using the development-selected rank-sum fusion of the add-one and delete-one effects, leaves a large answer-F1 regret against the optimal set at every budget, while a scalar read off the optimal-set family stays near zero; the pooled $B \in \{2, 3\}$ row is the 0.18 versus 0.006 separation quoted in the abstract and Section 1. Stratifying by a pairwise-interaction threshold fixed on development data, the singleton-supervision regret is 0.239 $[0.191, 0.291]$ on the high-interaction half and 0.117 $[0.079, 0.159]$ on the low-interaction half (pooled $B \in \{2, 3\}$), the localization predicted by the theory: the gap concentrates where fact values interact.

The cache evaluates all subsets of sizes $0, 1, 2, 3$, the full pool, and each delete-one subset (49 subsets per question). The rank-sum score is $\mathrm{rank}(a_i) + \mathrm{rank}(d_i)$: each rank orders (signal, i) increasingly, and the final top-B selection breaks score ties by increasing index. For $p^\star$, the frozen evaluation breaks equal-marginal ties by decreasing index. These fixed rules give the per-budget results and question-clustered intervals in Table 1.

S VALUE VS. SALIENCE SUPERVISION, BOTH TRAINED

To isolate the effect of the supervision signal on the synthetic suites, we train a second selector under conditions identical to the value-oracle selector of this section (Qwen3-4B, 30k examples, 2 epochs, LoRA rank 16, same optimizer and budget) but on *salience* labels (frequency-prior top-k) instead of value-oracle labels. Both selectors use 2 epochs; the synthetic main selectors (Table 7) use the separate 1-epoch configuration in Appendix G. The two label sets select disjoint facts, overlapping on 0% of training instances. Table 16 evaluates both selectors identically. The value-supervised selector reaches near-zero regret, while the salience-supervised one, despite identical training access, stays two orders of magnitude worse: the lever is the supervision signal (value vs. salience), not training.

Table 16: Value- vs. salience-supervised selector: identical 4B/30k/2ep/LoRA-16, differing only in the label (value oracle vs. frequency prior). Both are trained, so the gap of over two orders of magnitude isolates the supervision signal, not training access. This is a 2-epoch run; its value-oracle row is the run's own 2-epoch evaluation, distinct from the 1-epoch synthetic main table (Table 7).

Supervision (both trained)	IID regret	OOD regret
Value oracle (selector)	**0.012**	**0.023**
Frequency-prior salience	<u>8.00</u>	<u>8.33</u>

T SECOND REAL BENCHMARK: HOTPOTQA

To test that the multi-hop QA finding is not specific to MuSiQue, we repeat the frontier on HotpotQA (Yang et al., 2018) (600 questions, the same Qwen3-8B reader and protocol). Table 17 shows an identical pattern: the decision-sufficient supporting set gives the highest accuracy at the smallest context; full context is less accurate at roughly six times the tokens; salience retrieval (BM25, dense) and recency do not reach it; and padding the decision-sufficient set with retrieved paragraphs again degrades accuracy. The effect of value-based selection on real multi-hop QA therefore replicates across two independent datasets.

Table 17: HotpotQA multi-hop QA (600 questions, Qwen3-8B reader). Decision-sufficient memory is the most accurate and the most compact, replicating the MuSiQue result; full context is less accurate despite far more tokens, and salience retrieval and recency do not reach the decision-sufficient point. The MuSiQue-trained selector transferred zero-shot (no HotpotQA labels) reaches 0.498 EM at 228 tokens, between full context and the gold-support reference. Acc. is containment accuracy (the normalized gold answer appears in the prediction); EM is normalized exact match.

Memory	EM	Acc.	Tokens
Full context	0.483	0.642	1396
Decision-suff. (gold support)	**0.537**	**0.727**	**223**
Selector (zero-shot transfer)	0.498	0.682	228
BM25 (k=8)	0.450	0.583	1122
Dense (k=8)	0.477	0.630	1072
Recency (k=8)	0.418	0.557	1118
Padding (k=5)	0.510	0.652	682

U Long-Context Memory: BabiLong

We further test decision-sufficient memory on BabiLong (Kuratov et al., 2024), where bAbI reasoning facts are hidden in long distractor contexts of increasing length. Figure 4 shows that an oracle decision-sufficient memory (the gold supporting fact) answers correctly from 9 tokens at every context length (1k to 8k, accuracy 0.945), while feeding the full long context (the standard setting) stays near 0.50 and degrades as the context grows from 1k (0.520) to 8k (0.475), despite hundreds of times more tokens. Decision relevance, not length, determines what the reader needs.

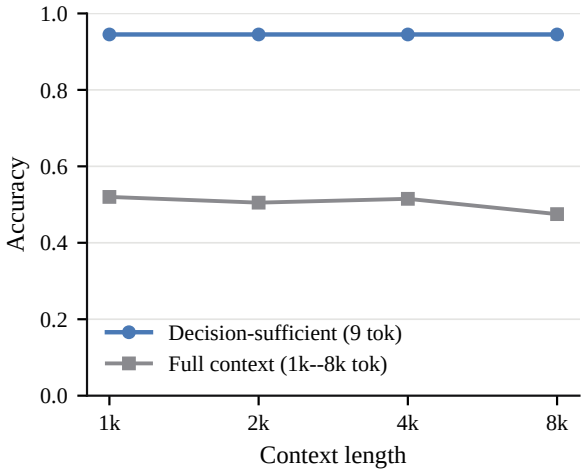

Figure 4: The long-context wall on BabiLong. As the haystack grows from 1k to 8k tokens, full-context accuracy stays near 0.50 and declines, while a 9-token decision-sufficient memory holds 0.945 at every length. What the reader needs is fixed by decision relevance, not context length.

V Real Long-Horizon Agent: ScienceWorld

To test decision-sufficient memory on a real long-horizon agent rather than question answering, we replay gold trajectories on ScienceWorld (Wang et al., 2022) and, at each step the model can already solve with full memory, ask how much past observation can be forgotten while still predicting the gold next action. Restricting to these solvable steps controls out agent competence and isolates the memory question. Across 849 such steps, a decision-sufficient memory (an approximately minimal subset of past observations that alone preserves the predicted (gold) action, found by greedy backward elimination) retains 1.8 of 8.3 past observations (189 vs 342 tokens) and preserves the next action by construction. At the *same* retained budget, recency (a sliding window) preserves it on only 70% of steps and BM25 salience on 60%. Even on real multi-step agent trajectories, value-based retention forgets far more than recency or salience while keeping the decision.

W Extended Analyses: Larger Pools, Robustness, and Reconstruction Order

Gold-free self-supervised selection. Without any solver or gold labels, one can keep a paragraph iff removing it changes the reader's own answer (a self-counterfactual signal). On MuSiQue this reaches EM 0.220 while keeping 6.8 paragraphs, above full context (0.168) but far below the gold supporting-set reference (0.422 at 2.4 paragraphs), consistent with the non-additivity of Table 22. Even so, this gold-free decision-value signal exceeds salience supervision (the same-data BM25-label selector reaches only 0.107 at a tighter budget, Table 2), suggesting that decision-value signals can help even without gold labels, at a budget these two selectors do not match. A matched test qualifies this. A LoRA-tuned Qwen3-1.7B item scorer trained from reader-derived $p^\star$ labels, compared against one trained from singleton supervision at matched architecture, data, budget, and three seeds, shows no aggregate advantage. Pooled over $B \in \{2, 3\}$ on 180 held-out questions the set-regret difference is 0.001 with a question-clustered 95% interval of $[-0.031, +0.034]$, within its interval on both the interaction-heavy (-0.003) and low-interaction ($+0.006$) strata. Identification does not imply learnability from surface form: the set-derived target is representable and recovers the oracle when read off directly (the 0.006 marginal of Section 4), but *learning* to predict reader-derived $p^\star$ from surface features did not, in this matched setting, generalize better than singleton supervision. The learned real-language gain of Section 6 comes from dataset-provided gold supporting-set labels, a distinct and stronger supervision source. These are a good but not by-construction-optimal proxy for the reader-derived $p^\star$: all 250 sealed questions contain their complete gold support in the candidate pool. The comparison uses the 204 questions (82%) with at most three gold-supporting paragraphs, matching the cache's exhaustive coverage of subsets up to size three. Among these, the gold set belongs to the reader-defined optimal family at the same cardinality in 71% of questions (median reader-regret 0, mean 0.13).

Globally retrieved candidate pools. This part gives the retrieval construction and per-seed detail behind Table 4. We deduplicate all MuSiQue paragraphs to $21{,}023$ canonical passages by normalized (title, text) and, for each of 430 held-out questions disjoint from the selectors' training questions, retrieve the top 50 and 100 by a dense encoder (BGE-base). A gold-supporting paragraph counts as retrieved only if its canonical passage appears in the deduplicated top-n. The retriever is imperfect: support recall is 0.751 at $n{=}50$ and 0.813 at $n{=}100$, and all required evidence is present for 48% and 59% of questions respectively (the retrievable subset R_n). The two frozen selectors from the enumerable protocol, singleton-derived and set-derived (reader-derived $p^\star$), score every retrieved paragraph and keep the top $B{=}3$; the reader (Qwen3-8B) and the B-paragraph budget are identical across selectors and the dense top-B policy, while the gold-support reference keeps every retrieved gold paragraph and is not budget-matched. The set-over-item F1 gap is positive in all six seed-level comparisons ($n{=}50$: $+0.031, +0.005, +0.013$; $n{=}100$: $+0.043, +0.041, +0.037$) and its question-clustered interval excludes zero at $n{=}100$, both overall and on R_n (Table 4). Both learned selectors remain below the dense top-B baseline throughout, so the result is transfer of the supervision-unit distinction under a large curated-to-global candidate shift, not an end-to-end retrieval improvement.

Training on the retrieval distribution. This part details the matched-training result of Section 8. We reserve a fresh split of MuSiQue questions disjoint from all earlier splits (seed frozen before any statistic was viewed): 800 for training and 500 held outcome-sealed for evaluation; both get global top-100 pools built as above. Two selectors are trained under one recipe (Qwen3-1.7B, QLoRA, three seeds, identical schedule), differing only in the per-candidate target. Each selector puts a one-logit scoring head on the model and reads one prompt per candidate (question, budget, candidate passage); training applies a pairwise logistic loss to candidates of the same question whose targets differ by more than 10% of that question's target range (LoRA rank 16, three epochs, learning rate 2×10^{-4}), and at test time every candidate is scored independently and the top B are kept. The item arm uses the add-one singleton value $V(\{i\})$ (the single-passage reader F1 of candidate i), and the set arm uses gold supporting-set membership. Questions with no gold retrieved are dropped from training. *Training candidate construction.* Each training question uses eight candidates, all retrieved gold plus the highest-ranked dense negatives, selected by the same deterministic rule for both arms and fixed before any evaluation outcome was inspected; evaluation uses the full top-100 pools. *Add-one label precision.* The add-one labels were computed with a 4-bit copy of the fixed

reader; a training-only audit on 40 questions found close agreement with the bf16 reader's candidate ordering (median Spearman 0.897, mean top-3 overlap 0.692, argmax agreement 70%), and the evaluation reader is bf16. A separate training-only audit relabels 336 questions outside this fresh split with the bf16 reader (median Spearman 0.89 over the 273 questions with nonconstant values in both label vectors; mean top-3 overlap 0.68 over all 336). On the sealed 500 (budget B=3), answer-F1 is 0.276 (set), 0.235 (dense top-B), 0.204 (item), 0.488 (gold-support reference); per seed the set arm scores 0.304/0.241/0.282 and the item arm 0.180/0.240/0.192. The two contrasts are $\Delta_{\text{set-item}} = +0.072$ $[0.050, 0.095]$ and $\Delta_{\text{set-dense}} = +0.041$ $[0.013, 0.070]$ (both exclude zero); on the retrievable subset (n=292) they are $+0.079$ $[0.047, 0.111]$ and $+0.034$ $[-0.009, 0.075]$. For these intervals, each question's F1 is first averaged over the three seeds and questions are then resampled (5,000 draws), so the intervals cover question sampling for these trained selectors, not retraining; the per-seed values above show the variation across seeds. The set-derived learned selector's advantage over the add-one target does not resolve the scalability of a reader-derived $p^\star$ oracle (gold support is a distinct, dataset-provided supervision source; the 71% overlap above stands), but under a fixed dense-retrieval pipeline it outperforms both the matched item arm and the retriever's own top-B policy.

Robustness of the sealed gap to fusion and threshold. The sealed measurement of Section 4 fixes, on development data, both how the add-one and delete-one signals are fused into a single ranking and the pairwise-interaction threshold that splits questions. We check that neither choice drives the result, on the same 250 held-out questions, pooled over $B \in \{2, 3\}$ with question-clustered intervals.

The identification gap persists across the evaluated fusion rules. Across the pre-specified alternatives, taking only the add-one signal, only the delete-one signal, their rank-sum, or convex combinations at $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$, the singleton selection leaves between 0.154 and 0.228 answer-F1 regret against the optimal set, and every interval excludes zero (Table 18). Even choosing, per question, the strongest of these fusions after the fact leaves 0.079 $[0.059, 0.101]$: none of the evaluated fusion rules closes the gap, even when chosen per question.

The concentration in interaction-heavy questions persists across the threshold sweep. Sweeping the interaction threshold across quantiles $q \in \{0.3, \ldots, 0.7\}$, the singleton regret on the interaction-heavy stratum rises monotonically from 0.225 to 0.299 while the low-interaction stratum stays between 0.115 and 0.131; the gap grows from $+0.104$ to $+0.168$ and every stratum interval excludes zero. The shortfall tracks interaction structure across the whole sweep.

Table 18: **The sealed singleton regret is robust to how the anchored signals are fused** (250 held-out questions, pooled $B \in \{2, 3\}$, question-clustered 95% intervals). Every fusion leaves a large regret against the optimal set; the last row takes the per-question best fusion after the fact.

Singleton fusion	Regret vs. optimal set [95% CI]
Add-one only	0.228 [0.193, 0.266]
Delete-one only	0.193 [0.159, 0.231]
Rank-sum	0.181 [0.148, 0.213]
Convex (λ=0.25)	0.155 [0.124, 0.187]
Convex (λ=0.5)	0.154 [0.123, 0.186]
Convex (λ=0.75)	0.159 [0.128, 0.191]
Per-question strongest fusion	0.079 [0.059, 0.101]

Graded recovery with interaction order. The binary interaction split of Section 4 treats a question as interaction-heavy or not; the underlying structure is graded. We probe it with *canonical order-r reconstruction* of reader-defined set value. A set function of interaction degree at most r is exactly reconstructed from its values on subsets of size at most r (Möbius truncation); we use $\widehat{V}_r(S) = \sum_{T \subseteq S,\, |T| \le r} \sum_{U \subseteq T} (-1)^{|T|-|U|} V(U)$. Taking its budget-$B$ argmax family $\mathcal{A}_r$, we report $P_{\text{guaranteed}}(r)$, the fraction of questions for which $\mathcal{A}_r$ is contained in the optimal-set family O_B, i.e. order-r observations recover an optimal set under any tie-break. This is a statement about canonical order-r reconstruction under reader-defined value, not an information-theoretic necessity, and the endpoint r=B is trivially 1.

Recovery improves progressively with interaction order, and the pattern holds across all three readers (Qwen3-8B, Tulu-3-8B, gemma-2-9b) and every MuSiQue hop stratum (Figure 5, Table 19). At first order the reconstruction leaves a wide gap: an optimal set is often *possible* in the order-1 argmax family ($P_{\text{possible}} = 0.84$ to 0.93) yet not *guaranteed* ($P_{\text{guaranteed}} = 0.22$ to 0.41), a reconstruction-level identification gap of $+0.52$ to $+0.62$ at $B{=}5$; higher orders close progressively more of the gap in aggregate. The reader-macro curves increase monotonically, while individual reader×hop strata can exhibit local non-monotonicity; the graded pattern is a property of the decision problem's value geometry rather than of a single reader.

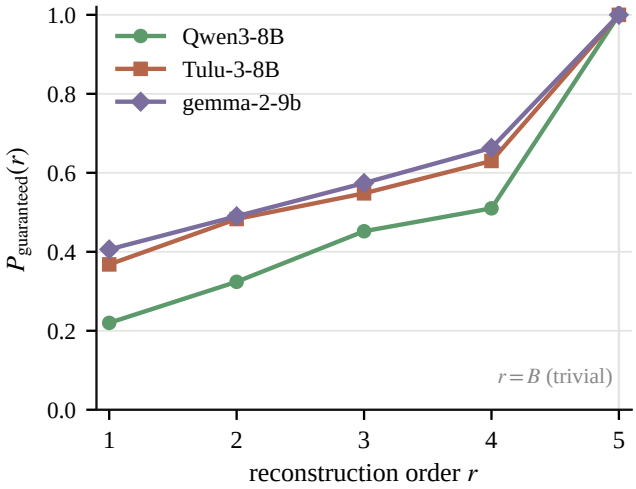

Figure 5: **Optimal-set recovery improves progressively with interaction order across readers** (MuSiQue, $K{=}10$, budget $B{=}5$, three readers over 451 questions). $P_{\text{guaranteed}}(r)$ is the fraction of questions for which the order-r reconstruction argmax family lies inside the optimal-set family; $r{=}B$ is the trivial endpoint. The plotted values measure recovery by the specified reconstruction rule.

Table 19: $P_{\text{guaranteed}}(r)$ per hop stratum and reader (budget $B{=}5$; $r{=}5{=}B$ is trivially 1.0 and omitted). Recovery increases with r overall, with occasional local non-monotonicity within a reader×hop cell (e.g. Tulu three-hop at $r{=}3$). Hop count does not order the required reconstruction level: Qwen3-8B recovers more on four-hop than on three-hop questions at every order, and Tulu and gemma recover more on three-hop than on two-hop questions, so interaction order is not reasoning depth.

Hop	Reader	$r{=}1$	$r{=}2$	$r{=}3$	$r{=}4$
2-hop	Qwen3-8B	0.175	0.275	0.455	0.480
	Tulu-3-8B	0.230	0.375	0.520	0.565
	gemma-2-9b	0.330	0.435	0.530	0.645
3-hop	Qwen3-8B	0.246	0.329	0.431	0.497
	Tulu-3-8B	0.527	0.599	0.581	0.701
	gemma-2-9b	0.461	0.539	0.605	0.683
4-hop	Qwen3-8B	0.274	0.429	0.488	0.607
	Tulu-3-8B	0.381	0.512	0.548	0.643
	gemma-2-9b	0.476	0.524	0.619	0.667

Positioning against set-aware selection methods. Item-wise retrievers, set-aware selectors, and this work occupy different axes (Table 20). Set-aware selection methods change how candidates are scored jointly at deployment: SETR (Lee et al., 2025) identifies a query's information requirements and selects a set that meets them, and OptiSet (Jiang et al., 2026) builds set-level preference labels from the generator's set-conditioned utility. We instead ask what information the supervision signal must carry to identify the decision-sufficient set, a target that can be read off into item scores. These are complementary: a set-aware architecture and set-derived supervision can be combined. The matched swap of Section 6 holds architecture, data, and budget fixed while changing the supervision target from salience to supporting-set membership; because target source and unit change together,

we read it as a practical supervision comparison rather than a pure isolation of the supervision unit, which the reader-derived matched test of Appendix W addresses more directly. The non-identification result characterizes the set-level information needed when anchored item signals leave the optimum undetermined.

Table 20: **Supervision and deployment in memory selection.** The table separates the supervision signal, the deployment scoring rule, and the question addressed by each approach.

Approach	Supervision signal	Deployment	Question posed
BM25 / dense	item relevance	item-wise	Which items are similar?
SETR	requirement-set labels	set-aware	Which set meets the needs?
OptiSet	set-conditioned utility	set-aware	Which set maximizes utility?
Ours	decision-value set target	item- or set-wise	What supervision suffices?

X Objective Shift and Action Cloning

The kept set is a sufficient statistic for a *family* of objectives, whereas a distilled action is sufficient for only one. If objectives have distinct unique optimal actions, any policy that commits to an action before the objective is revealed incurs positive regret for some objective, while a family-sufficient kept set incurs zero; a minimum family-sufficient state is thus a sufficient statistic for the family (proof and bounds in Appendix D, Theorem 2). Under a shift in hazard-penalty multiplier from β=1 to β=3 revealed only after compression, the family-sufficient selector stays near-oracle while a single-objective selector and a trained action policy break (Table 21).

Table 21: Objective shift on the synthetic `v2-hard` suite: only the family-sufficient kept set survives an objective (β=3) revealed after compression.

Method	IID β1	IID β3	OOD β1	OOD β3
Family-sufficient	<u>0.02</u>	**0.03**	0.03	**0.04**
Single-objective	0.04	<u>0.97</u>	<u>0.02</u>	<u>0.84</u>
Action cloning	**0.01**	3.00	**0.00**	2.09

Objective family (Table 21). The family is a risk-sensitivity parameter β (written ρ in our code) scaling hazard penalties on the synthetic `v2-hard` suite. Family-sufficient labels are the union, over a grid $\beta \in \{0.6, 1.0, 1.8, 3.0\}$, of the per-objective minimum-sufficient kept sets; single-objective labels use β=1 alone. All selectors are Qwen3-4B with QLoRA (r=16, α=32, 1 epoch); the action-cloning baseline is a direct action head trained on the same base model and data size. The kept set is formed before β is revealed and evaluated at the training objective β=1 and the shifted β=3 (n=800 per cell, greedy decoding). Mean regret (IID β1 / IID β3 / OOD β1 / OOD β3): family-sufficient $0.019/0.029/0.032/0.036$, single-objective $0.043/0.968/0.021/0.843$, action-cloning $0.013/3.001/0.000/2.089$. The family-sufficient set keeps one extra fact on average (4.89 vs 3.86).

Action cloning at a fixed objective. On the fixed training objective the direct action head is competitive (`direct_action` suite: IID regret 0.010, OOD 0.000, held-out family 0.006; n=1000 each), matching the low regret of decision-sufficient selection in that setting. Uninformed action policies on the same scale (keeping only distractors, or acting from the training-distribution action prior) incur regret $\approx$ 8–13. The action head keeps no fact set, so it carries no inspectable state and, as above, breaks once the objective shifts.

Y Additional Tables and Statistics

This section collects dataset statistics, held-out shift metrics, and cost-sensitivity numbers referenced from the main text.

The combinatorial structure of decision sufficiency is stable across splits (Table 22).

Table 22: Decision sufficiency is combinatorial (Synthetic). *Complementarity*: fraction of instances where the union of individually relevant facts is not sufficient. *Non-unique*: fraction admitting more than one minimum decision-sufficient set. *Oracle size*: mean cardinality of the minimum sufficient kept set. Per-family breakdown in Table 5.

Split	Oracle size	Complementarity	Non-unique
Train	3.77	0.585	0.380
IID	3.81	0.573	0.375
OOD	3.78	0.571	0.377

Table 23 reports net regret as the memory cost λ grows.

Table 23: Net-regret sensitivity for the selector Qwen3-4B on OOD splits. Net regret rises smoothly with the memory cost.

Bench.	$\lambda = 0$	.01	.05	.10
Synthetic	0.032	0.072	0.229	0.426
MiniGrid	0.284	0.327	0.497	0.710
PDDLGym	0.127	0.163	0.306	0.486

The selector is robust across seeds and backbones (Table 24).

Table 24: Robustness (Synthetic): regret $\pm$ std over 2 seeds, 2 backbones (zero-regret ≥ 0.97).

Split	4B	8B
IID	$0.057_{\pm.013}$	$0.053_{\pm.010}$
Neutral	$0.025_{\pm.006}$	$0.030_{\pm.002}$
OOD	$0.038_{\pm.017}$	$0.020_{\pm.013}$
Held-out	$0.018_{\pm.021}$	$0.029_{\pm.000}$

The selector trained at $N{=}20$ candidate facts degrades gracefully as N grows to 200 (Table 25).

Table 25: Scaling candidates $20{\rightarrow}200$ (Synthetic); the selector trained at $N{=}20$ degrades gracefully.

Candidate facts N	20	50	100	200
IID regret	0.064	0.044	0.082	0.237
OOD regret	0.024	0.135	0.084	0.236
Zero-regret	0.982	0.980	0.970	0.946
Retained facts	3.83	3.89	4.27	4.65

The selector ordering is stable under one-factor sweeps of LoRA rank, training-data size, and epoch count (Table 26).

Table 26: Hyperparameter sensitivity of the Qwen3-4B selector (mean regret, IID / OOD, $n{=}500$ each). One-factor-at-a-time sweeps around the base configuration (30k examples, LoRA rank 16, 1 epoch, seed 0). Every setting stays far below both zero-shot prompting (9.69 OOD) and the salience-supervised selector (8.33 OOD, Table 16); regret rises only in the undertrained 10k cell.

Factor	Setting	IID	OOD
LoRA rank (30k, 1ep)	$r{=}8$	0.055	0.078
	$r{=}16$	0.042	0.021
	$r{=}32$	0.048	0.013
Data size (r16, 1ep)	10k (undertrained)	0.319	0.710
	20k	0.123	0.083
	30k	0.042	0.021
Epochs (30k, r16)	1 epoch	0.042	0.021
	2 epochs	0.012	0.023

Table 27 lists the dataset sizes for each suite.

Table 27: Dataset sizes. Test combines IID, neutral, OOD, and held-out splits.

Benchmark	Train	Dev	Test	Held-out
Synthetic	30000	4000	16000	probe_or_act
MiniGrid	12000	1500	6000	unlock
PDDLGym	12000	1500	6000	miconic

Held-out PDDLGym introduces genuinely new relevance structures rather than only new compositions of known factors (Table 28).

Table 28: Held-out (HO) shift relative to training. ActTV: total variation of optimal-action frequencies; FeatJS: Jensen–Shannon divergence of oracle-feature frequencies; Unseen: fraction of unseen oracle-feature sets. MG: MiniGrid; PDDL: PDDLGym.

Bench.	Split	Reg.	ActTV	FeatJS	Unseen
Synthetic	OOD	0.032	0.018	0.000	0.000
Synthetic	HO	0.025	0.075	0.003	0.023
MG	OOD	0.284	0.024	0.000	0.001
MG	HO	0.533	0.082	0.018	0.256
PDDL	OOD	0.127	0.033	0.000	0.000
PDDL	HO	0.758	0.297	0.093	0.629

Figure 6 places each suite on a diagnostic map of held-out transfer.

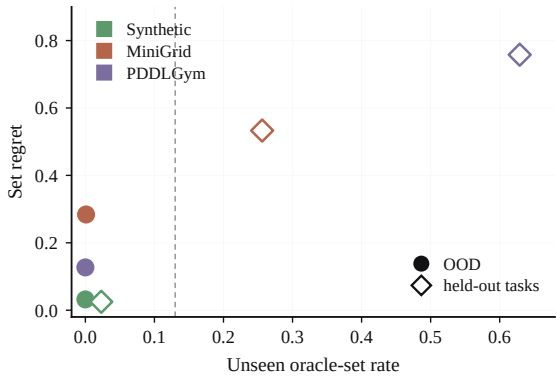

Figure 6: Held-out transfer as a diagnostic map. Color denotes suite (Synthetic, MiniGrid, PDDLGym); filled circles are OOD and open diamonds are held-out task types. Set regret rises on held-out task types that introduce new relevance structures.

Z Decision-Sufficient Memory: Theory

Setup. An observation is a finite set of facts $x = \{f_1, \ldots, f_n\}$. An objective ω from a family Ω assigns a value $Q_\omega(a \mid m)$ to each action $a \in \mathcal{A}$ when the solver re-decides from a retained memory $m \subseteq x$. The decision and regret are

$$a^\star_\omega(m) = \arg\max_{a \in \mathcal{A}} Q_\omega(a \mid m), \qquad \rho_\omega(m) = \max_a Q_\omega(a \mid x) - Q_\omega\big(a^\star_\omega(m) \mid x\big),$$

where $Q_\omega(\cdot \mid x)$ on the full observation is the ground-truth value. In the body, the same decision module, not knowing the full state, commits to the minimax-regret action a^{rob}_S over the uncertainty set $\Omega(S)$ (Appendix C). Here we fix ω and write $a^\star_\omega(m) = \arg\max_a Q_\omega(a \mid m)$ for the action the solver takes when the objective is known. The two coincide whenever the retained memory determines the optimal action across its whole uncertainty set. Zero realized regret, $\rho_\omega(m) = 0$, does not by itself imply this, so the synthetic oracle of Algorithm 1 is minimum-cardinality under its robust certificate, a criterion stronger than zero regret at the true state. Likewise, MuSiQue reports the budgeted set regret G_B of Section 2 and the synthetic suites report ρ_ω, regret against the full-information decision; the two agree when the budget can attain the full-information value. We make

two idealizing assumptions on the solver; the trained selector approximates a solver satisfying them, and the empirical sections quantify the approximation.

(A1) **Strict gap.** $\arg\max_a Q_\omega(a \mid x)$ is a singleton for every ω (generic position).

(A2) **Monotone sufficiency (no distraction).** For $m \subseteq m' \subseteq x$, $\rho_\omega(m) = 0 \Rightarrow \rho_\omega(m') = 0$: extra retained facts never break a decision that was already preserved.

Sufficiency is defined relative to this fixed solver: $Q_\omega(\cdot \mid x)$ on the full observation is the reference. A2 is the formal counterpart of "more context should not hurt"; it is exactly what fails for value-agnostic readers (§6, where more context lowers accuracy), so we treat it as an idealization and state every theorem under it explicitly.

Definition 3 (Decision-sufficient memory)**.** *A memory $m \subseteq x$ is* decision-sufficient *for ω at x if $\rho_\omega(m) = 0$, and for the family Ω if $\rho_\omega(m) = 0$ for all $\omega \in \Omega$. A* minimal *decision-sufficient set $M_\omega(x)$ is a $\subseteq$-minimal such m for ω; minimal sets need not be unique (Proposition 2).*

Counterfactual forgetting calls f *pivotal* for ω at x if $a^\star_\omega(x \setminus \{f\}) \neq a^\star_\omega(x)$. Pivotality is a *necessity* test, not a constructive one:

Lemma 1 (Pivotal facts are mandatory)**.** *Under A1–A2, if f is pivotal for ω at x then f belongs to* every *decision-sufficient set for ω. Hence every fact pivotal for some $\omega \in \Omega$ lies in every decision-sufficient memory for Ω.*

Proof. Suppose a decision-sufficient m omits a pivotal f, so $m \subseteq x \setminus \{f\}$. By A2, $x \setminus \{f\}$ is then decision-sufficient, so $\rho_\omega(x \setminus \{f\}) = 0$; under A1 the maximizer is unique, hence $a^\star_\omega(x \setminus \{f\}) = a^\star_\omega(x)$, contradicting pivotality of f. □

Z.1 Characterizing the decision-sufficient memory

Theorem 2 (Family sufficiency and bounds)**.** *Fix x and Ω, and assume A1–A2. Choose any minimal decision-sufficient set $M_\omega(x)$ per objective and let $P_\Omega(x) = \bigcup_{\omega\in\Omega} M_\omega(x)$.*

1. ***(Sufficiency / upper bound)*** *$P_\Omega(x)$ is decision-sufficient for Ω; hence the minimum size of a family-sufficient memory is at most $|P_\Omega(x)|$.*
2. ***(Lower bound)*** *Let $\mathrm{Piv}_\Omega(x)$ be the facts pivotal for some ω. Every decision-sufficient memory for Ω contains $\mathrm{Piv}_\Omega(x)$; in particular every minimal family-sufficient set $S^\star_\Omega(x)$ satisfies $\mathrm{Piv}_\Omega(x) \subseteq S^\star_\Omega(x)$.*
3. ***(Tightness)*** *If each $M_\omega(x)$ is unique, then every family-sufficient memory contains each $M_\omega(x)$, so $P_\Omega(x) \subseteq$ every family-sufficient set; since $P_\Omega(x)$ is itself family-sufficient by part (1), it is the unique minimum-cardinality family-sufficient set and we take $S^\star_\Omega(x) = P_\Omega(x)$. Without uniqueness, minimal sets may be non-unique and $P_\Omega(x)$ bounds only the minimum* size *(Proposition 2).*

Proof. (1) For each ω, $P_\Omega(x) \supseteq M_\omega(x)$ which is decision-sufficient for ω; by A2 (monotone sufficiency) $P_\Omega(x)$ is decision-sufficient for ω, and as ω was arbitrary, for Ω. *(2)* Each pivotal fact lies in every family-sufficient memory by Lemma 1, applied to its witnessing ω. *(3)* With each M_ω unique, any family-sufficient S has, for each ω, an inclusion-minimal ω-sufficient subset of S that must equal M_ω; hence $M_\omega \subseteq S$ for all ω and $P_\Omega = \bigcup_\omega M_\omega \subseteq S$. As P_Ω is family-sufficient by (1), it is the unique minimum-cardinality family-sufficient set. □

We call an inclusion-minimal family-sufficient set a *decision-sufficient memory*, and a least-cardinality one a *minimum decision-sufficient memory* $S^\star_\Omega(x)$ (MIN-DSM, Appendix Z.5): minimum-cardinality is strictly stronger than inclusion-minimal; Theorem 2(3) gives a sufficient condition under which the two coincide, and our empirical oracle targets the minimum-cardinality set. It is sufficient for the *decision*, not for reconstructing x: the distinction the rest of the section makes precise. (This is analogous in spirit to a minimal sufficient statistic, but it is a combinatorial property of one observation under a fixed solver, not a likelihood-based one; we use the term informally.)

Z.2 RELEVANCE IS COMBINATORIAL

Decision-sufficiency is combinatorial: single-fact pivotality neither composes nor suffices, and minimum sufficient sets need not be unique.

Proposition 2 (Non-additivity of decision-sufficiency)**.** *Single-fact pivotality neither composes into nor certifies a sufficient set:*

1. ***(Substitutable evidence)*** *There exist x, ω, f, g with neither individually pivotal yet the pair jointly pivotal: $a^\star_\omega(x \setminus \{f\}) = a^\star_\omega(x \setminus \{g\}) = a^\star_\omega(x)$ but $a^\star_\omega(x \setminus \{f, g\}) \neq a^\star_\omega(x)$. A one-shot rule that drops every fact non-pivotal at the full set (each of f, g has zero leave-one-out effect) then drops both and destroys the decision, although a minimal sufficient set ($\{f\}$ or $\{g\}$) exists. Under A2, sequential backward elimination starting from a sufficient set and deleting facts while preserving sufficiency until no further deletion is possible yields an inclusion-minimal sufficient set, but need not find a minimum-cardinality one.*

2. ***(Conjunctive evidence)*** *There exist multi-fact instances (e.g. a multi-hop chain) where the minimal sufficient set has size > 1 and every singleton is insufficient, so no single individually-relevant fact suffices.*

In both cases sufficiency is a property of subsets*. The general result, that no rule scoring facts only from their first-order (anchored singleton) signals can certify a minimal sufficient set, is Theorem 7.*

Constructions. (1) Substitutable: let $a^\star_\omega$ depend on x only through a quantity D that is independently derivable from f alone and from g alone (two redundant derivations). Removing f leaves g to supply D, so $a^\star_\omega$ is unchanged; likewise for g; removing both makes D unavailable and flips the decision. Neither is individually pivotal, the minimal sufficient sets are $\{f\}$ and $\{g\}$, and a one-shot non-pivotal deletion removes both. *(2) Conjunctive:* a two-hop chain f="A is in city C", g="C is the capital of D" with the decision a function of D: each of f, g is individually pivotal (by Lemma 1 both lie in every sufficient set), and the minimal sufficient set $\{f, g\}$ has size 2, so no singleton suffices. □

Z.3 SEPARATION FROM COMPRESSION AND SALIENCE

A *salience* policy keeps the top-k facts by an objective-independent score $s(f \mid x)$; a *δ-faithful compression* keeps a code from which x is reconstructable to tolerance δ. Decision-sufficiency separates from both.

Theorem 3 (Salience/compression: larger, and brittle)**.** *For every $n \geq 2$ there is an instance with $|x| = n$, a single objective ω (write $S^\star(x)$ for its minimal decision-sufficient memory), an objective-independent salience score s, and constants $\Delta > 0$, $V = O(1)$ with optimal value $\leq V$, such that:*

1. ***(Size)*** *$|S^\star(x)| = 1$, whereas (i) under s any top-k salience memory with $\rho_\omega = 0$ needs $k = n$, and (ii) any δ-faithful code for the constructed input family stores $\Omega(n)$ bits in expectation.*

2. ***(Regret)*** *The budget-matched top-1 salience memory has $\rho_\omega(m_s) = \Delta$ while $\rho_\omega(S^\star(x)) = 0$; choosing the penalty makes Δ/V arbitrarily large, so relative regret is unbounded.*

Construction. One *bridging* fact $f^\star$ is singly pivotal: flipping it changes $a^\star_\omega$, moving value by Δ (a high-penalty hazard on the otherwise-best route). Assign it the strictly lowest salience $s(f^\star \mid x) = \epsilon$, uniquely last under s. The other $n - 1$ facts are *decorative*: non-pivotal, all ranked above $f^\star$ by s, and drawn i.i.d. uniform from a constant-size alphabet of size $c \geq 2$. *Size:* $f^\star$ is singly pivotal so $S^\star(x) = \{f^\star\}$, $|S^\star(x)| = 1$. (i) To include $f^\star$ a top-k keeper under s must take $k = n$, since $f^\star$ is uniquely ranked last. (ii) A δ-faithful code with δ below the alphabet spacing must reproduce each decorative symbol exactly; by Shannon source coding, encoding $n - 1$ i.i.d. uniform symbols over an alphabet of size c requires at least $(n - 1) \log_2 c = \Omega(n)$ bits in expectation. *Regret:* top-1 salience keeps a decorative fact and omits $f^\star$, so the solver takes the dominated action and $\rho_\omega = \Delta$. Holding optimal value $\leq V$ while increasing the hazard penalty sends $\Delta/V \to \infty$. □

This is the regime of §6 (salience misses multi-hop bridging facts) and the synthetic suites (decorative high-prior facts): the separation is worst-case over salience orderings (there exists a salience score ranking the pivotal fact last). The size claim is stated in matched units (facts for selection, bits for compression), and the salience bound is for the score realized in the construction.

Z.4 Separation from action distillation under objective shift

A *direct action policy* distills the decision for a fixed ω_0, storing the action $a^\star_{\omega_0}(x)$ (or an ω-independent policy). On ω_0 it is optimal. The decision-sufficient *state* dominates it in *robustness* once the objective is revealed after the memory is formed.

Theorem 4 (State dominates action over an objective family)**.** *Let ω be drawn from Ω and revealed only after the memory is formed, and assume A1–A2. Then $S^\star_\Omega(x)$ attains $\rho_\omega = 0$ for every realized $\omega \in \Omega$. Any memory that stores a single action, or a policy that does not depend on the realized ω, attains $\rho_\omega > 0$ for some $\omega \in \Omega$ whenever the family's optima disagree ($\exists\, \omega_0, \omega_1$ with $a^\star_{\omega_0}(x) \neq a^\star_{\omega_1}(x)$). The dominance is in regret-coverage over Ω, not in size: $S^\star_\Omega(x)$ may be larger than one action, but a single ω_0-tied action cannot cover Ω.*

Proof. $S^\star_\Omega(x)$ is family-sufficient by definition, so $\rho_\omega = 0$ for all ω. A stored action a that ignores ω yields $\rho_{\omega_1} = Q_{\omega_1}(a^\star_{\omega_1} \mid x) - Q_{\omega_1}(a \mid x) > 0$ at any ω_1 whose optimum differs from a (such ω_1 exists when the family's optima disagree). (A lookup table $\omega \mapsto a^\star_\omega(x)$ would cover Ω but has size $|\Omega|$ and is no longer a single distilled action.) □

Theorem 4 formalizes the dichotomy of §X: a trained action policy ties decision-sufficient memory on a fixed objective yet cannot survive an objective revealed after compression, because an action is sufficient for one objective while the decision-sufficient state is sufficient for the whole family. Together, Theorems 2–4 characterize decision-sufficient memory as the minimal memory sufficient for the decision, and separate it from compression and salience (by size and by regret) and from action distillation (by robustness to objective shift).

Z.5 Computational hardness of the minimum decision-sufficient memory

Minimizing the decision-sufficient memory is not merely combinatorial (Proposition 2); it is NP-hard, with a tight approximation threshold, and the hardness arises intrinsically from the redundant-derivation structure the paper observes.

Definition 4 (Min-DSM)**.** *An instance of* minimum decision-sufficient memory *is $(x, \mathcal{A}, \omega)$ with a fact set x, actions $\mathcal{A}$, and an objective inducing $Q_\omega(a \mid m)$ satisfying A1–A2. The task is a decision-sufficient $m^\star$ ($\rho_\omega(m^\star) = 0$) of minimum cardinality; the decision version $\mathrm{DSM}_{\le k}$ asks whether some decision-sufficient m has $|m| \le k$.*

Under A1 the full set is ground truth with a unique maximizer, so for every m,

$$\rho_\omega(m) = 0 \iff a^\star_\omega(m) = a^\star_\omega(x). \tag{3}$$

Lemma 2 (Set Cover embeds as an objective)**.** *Fix a Set Cover instance: universe $U = \{u_1, \dots, u_N\}$, sets $\mathcal{S} = \{S_1, \dots, S_p\}$ with $\bigcup_j S_j = U$. Let $x = \{f_1, \dots, f_p\}$ with $f_j \leftrightarrow S_j$, $\mathcal{A} = \{a_0, a_1\}$, $\mathrm{cov}(m) = |\bigcup_{f_j \in m} S_j|$, and $Q_\omega(a_1 \mid m) = \mathrm{cov}(m)$, $Q_\omega(a_0 \mid m) = N - \frac{1}{2}$. This satisfies A1–A2, and for every m, $\rho_\omega(m) = 0 \iff \mathrm{cov}(m) = N \iff \{S_j : f_j \in m\}$ covers U; moreover $\rho_\omega(m) \in \{0, \frac{1}{2}\}$.*

Proof. A1: $Q_\omega(a_1 \mid m) = \mathrm{cov}(m) \in \mathbb{Z}$ while $Q_\omega(a_0 \mid m) = N - \frac{1}{2} \notin \mathbb{Z}$, so the argmax is always unique (gap $\ge \frac{1}{2}$); $\mathrm{cov}(m) = N \Rightarrow a^\star = a_1$, else $a^\star = a_0$. *Ground truth:* $\mathrm{cov}(x) = N$ so $a^\star_\omega(x) = a_1$, $\max_a Q_\omega(a \mid x) = N$; hence $\rho_\omega(m) = 0$ if $a^\star_\omega(m) = a_1$ (i.e. $\mathrm{cov}(m) = N$) and $\rho_\omega(m) = N - (N - \frac{1}{2}) = \frac{1}{2}$ otherwise, giving the stated equivalence via (3). *A2:* cov is monotone, so $\mathrm{cov}(m) = N$ and $m \subseteq m'$ force $\mathrm{cov}(m') = N$. □

Theorem 5 (Hardness and inapproximability)**.** $\mathrm{DSM}_{\le k}$ *is NP-complete when Q_ω is polynomial-time evaluable (in particular for the coverage subclass of Lemma 2), and* Min-DSM *is NP-hard in general. The map $\Phi(m) = \{S_j : f_j \in m\}$ of Lemma 2 is a cardinality-preserving bijection between*

decision-sufficient memories and set covers, so $\mathrm{OPT}_{\text{MIN-DSM}} = \mathrm{OPT}_{\text{SETCOVER}}$. *Consequently, for every* $\varepsilon > 0$ *it is NP-hard to approximate* MIN-DSM *within* $(1-\varepsilon)\ln N$ *(with* N *the number of derivation targets), unless* $\mathrm{P} = \mathrm{NP}$.

Proof. Membership (for poly-time-evaluable Q_ω, e.g. the coverage instances): checking $\rho_\omega(m) = 0$ is computing $\mathrm{cov}(m)$ and comparing to N, in polynomial time. By Lemma 2, m is decision-sufficient iff $\Phi(m)$ covers U, and Φ preserves cardinality on feasible solutions, so α-approximations transfer both ways. Set Cover decision is NP-complete (Karp, 1972) and, by Dinur & Steurer (2014); Feige (1998), admits no polynomial $(1-\varepsilon)\ln N$ approximation unless $\mathrm{P} = \mathrm{NP}$; the same bounds transfer to MIN-DSM. □

Proposition 3 (Value-graded greedy attains $\ln N + 1$)**.** *The greedy that repeatedly adds the fact maximizing the marginal coverage* $\mathrm{cov}(m \cup \{f\}) - \mathrm{cov}(m)$ *returns a decision-sufficient memory of size at most* $H_N \,\mathrm{OPT} \le (\ln N + 1)\,\mathrm{OPT}$*, where* $H_N = \sum_{j=1}^{N} 1/j$*, matching Theorem 5.*

Proof. cov is monotone submodular with $\mathrm{cov}(\varnothing) = 0$; the rule is the classical greedy set cover under Φ, whose output is a cover (hence decision-sufficient) with Chvátal's H_N guarantee (Chvátal, 1979) against $\mathrm{OPT} = \mathrm{OPT}_{\text{SETCOVER}}$, which equals the minimum decision-sufficient size by Theorem 5. □

Backward forgetting inherits minimality, not $\ln N$**.** The $\ln N + 1$ guarantee needs the *graded* coverage: raw regret is flat ($\rho_\omega \in \{0, \frac{1}{2}\}$), so a heuristic that deletes any fact whose removal preserves $\rho_\omega = 0$ returns only a $\subseteq$-minimal memory. Take the N singleton sets $\{u_i\}$ together with $D_1 = \{u_i : i \text{ odd}\}$ and $D_2 = \{u_i : i \text{ even}\}$: $\{D_1, D_2\}$ covers so $\mathrm{OPT} = 2$, yet backward forgetting may delete D_1 then D_2 (each redundant while all singletons remain), halting at all N singletons, a ratio $N/2$ that exceeds $\ln N + 1$ for every $N \ge 6$. The matching bound holds only for forward greedy on graded value; minimality of cardinality is strictly stronger than $\subseteq$-minimality.

Z.6 THE ANCHORED COLLISION IS MARGIN-ROBUST, NOT KNIFE-EDGE

The witness family of Theorem 1, $V_T(S) = |S| + \delta\,\mathbf{1}[S = T]$ with $0 < \delta < 1$, places its single distinguishing bump on one B-set T. The separation survives arbitrary small perturbations of *every* value the anchored signals do not observe. Let $a = \mathcal{A}(V_T)$ be the anchored observation shared across the family and $\mathcal{F}(a) = \{V : \mathcal{A}(V) = a\}$ its fiber; the anchored coordinates ($V(\varnothing)$, $V(F)$, the singletons, and the leave-one-out sets) are pinned to a, while the interior coordinates $\{V(S) : 2 \le |S| \le n-2\}$ are free.

Corollary 1 (Margin-robustness within the fiber)**.** *For each budget-*B *set* T *and any* $\varepsilon \in \big(0, \frac{1}{2}\min(\delta, 1-\delta)\big)$*, the relative-open neighborhood* $U_T = \{V' \in \mathcal{F}(a) : |V'(S) - V_T(S)| < \varepsilon \text{ for all } S \text{ with } 2 \le |S| \le n-2\}$ *is nonempty and consists entirely of monotone value functions whose unique budget-*B *optimum is* T*, with margin at least* $\delta - 2\varepsilon > 0$*. Consequently the conclusion of Theorem 1 holds not for isolated points but for every selection* $(V'_T)_T \in \prod_T U_T$*: the non-identification is stable, not knife-edge.*

Proof. Every $V' \in U_T$ agrees with V_T on the anchored coordinates by construction, so $U_T \subseteq \mathcal{F}(a)$ and is open in the interior coordinates; it contains V_T, hence is nonempty. *Monotonicity.* For $S \subsetneq S'$ with $|S'| = |S| + 1$, the unperturbed gap $V_T(S') - V_T(S)$ is at least $1 - \delta$ (it is 1 unless exactly one of S, S' equals T, where it is $1 \pm \delta$); since each perturbed coordinate moves by less than $\varepsilon < \frac{1}{2}\min(\delta, 1-\delta)$ and the anchored coordinates are fixed, the gap stays $\ge 1 - \delta - 2\varepsilon > 0$, and monotonicity extends to all nested pairs by transitivity. *Optimum.* Among B-sets, $V'(T) \ge B + \delta - \varepsilon$ while $V'(S) \le B + \varepsilon$ for $S \ne T$, so $V'(T) - V'(S) \ge \delta - 2\varepsilon > 0$; T is the unique budget-B optimum with that margin. Distinct T have disjoint optima, so an $\mathcal{A}$-measurable rule, seeing the same a, is optimal for at most one factor of $\prod_T U_T$; the worst-case success probability $\le 1/\binom{n}{B}$ of Theorem 1 is unchanged, and its expected-regret bound holds with the margin $\delta - 2\varepsilon$ in place of δ. □

Z.7 Relevance is combinatorial: no pointwise score recovers a minimal decision-sufficient set

Proposition 2 exhibited non-additive instances; here we prove the structural consequence: the value function is neither submodular nor supermodular, the decision-sufficient sets form no matroid, and *no* pointwise rule that scores facts independently from their anchored first-order profiles can certify a minimal decision-sufficient set. Write $v(m) = -\rho_\omega(m)$, with $v(x) = 0$, and the marginal $\Delta_f(m) = v(m \cup \{f\}) - v(m)$. Use a binary action set with $Q_\omega(1 \mid x) = 1, Q_\omega(0 \mid x) = 0$ so $\rho_\omega(m) = \mathbf{1}[m \notin \mathcal{S}_\omega]$ where $\mathcal{S}_\omega = \{m : \rho_\omega(m) = 0\}$ is the (upward-closed) family of decision-sufficient sets; specifying $\mathcal{S}_\omega$ specifies the instance and A1–A2 hold automatically. In this two-valued regime, A2 makes v monotone; it does not imply monotonicity of v for general regret values.

Proposition 4 (v is in general neither submodular nor supermodular)**.** *There is an instance satisfying A1–A2 whose v is neither submodular nor supermodular.*

Proof. Substitutable pair $\{a, b\}$, $\mathcal{S} = \{\{a\}, \{b\}, \{a, b\}\}$: $v(\varnothing) = -1$, else 0. With $S = \varnothing \subseteq T = \{a\}$, $\Delta_b(\varnothing) = 1$ but $\Delta_b(\{a\}) = 0$, so $\Delta_b(\varnothing) \not\leq \Delta_b(\{a\})$: supermodularity fails (both facts non-pivotal). *Conjunctive pair* $\{c, d\}$, $\mathcal{S} = \{\{c, d\}\}$: $v = -1$ except $v(\{c, d\}) = 0$. With $S = \varnothing \subseteq T = \{d\}$, $\Delta_c(\varnothing) = 0$ but $\Delta_c(\{d\}) = 1$, so $\Delta_c(\varnothing) \not\geq \Delta_c(\{d\})$: submodularity fails. Placing both pairs on disjoint coordinates, with product actions (a_1, a_2) and Q_ω additive across them (so $v(m) = v_1(m \cap \{a, b\}) + v_2(m \cap \{c, d\})$ by separability of the argmax) yields one instance where both inequalities fail; the sum remains monotone because each component is monotone. □

Proposition 5 (Exact modularity boundary)**.** *In the two-valued regime: if $\rho_\omega(m) = 0 \iff m \cap G \neq \varnothing$ for a set G of decisive facts (purely substitutable), v is submodular (and not supermodular unless $|G| \leq 1$); if $\rho_\omega(m) = 0 \iff S^\star \subseteq m$ for a unique minimal set $S^\star$ (purely conjunctive), v is supermodular (and not submodular unless $|S^\star| \leq 1$). Mixed structure is neither.*

Proof. Let $R > 0$ be the nonzero regret. Substitutable: $v(m) = R(\mathbf{1}[m \cap G \neq \varnothing] - 1)$; for $f \in G$ the marginal is $R\mathbf{1}[m \cap G = \varnothing]$, nonincreasing in the base set, so v is submodular; $|G| \geq 2$ gives $\Delta_f(\varnothing) = R > 0 = \Delta_f(\{f'\})$, not supermodular. Conjunctive: $v(m) = R(\prod_{f \in S^\star} \mathbf{1}[f \in m] - 1)$, a product of monotone indicators, supermodular; $|S^\star| \geq 2$ gives $\Delta_f(\varnothing) = 0 < R = \Delta_f(S^\star \setminus \{f\})$, not submodular. □

Theorem 6 (No matroid; unbounded greedy)**.** *Under A1–A2 with $\varnothing \notin \mathcal{S}_\omega$, the decision-sufficient sets are upward closed, hence not the independent sets of any matroid; and minimal decision-sufficient sets are in general not equicardinal, so no matroid has them as bases. Moreover, forward marginal-gain greedy and anonymous pointwise threshold selection have unbounded approximation ratios. With $M \geq 2$ disjoint chains $C_i = \{c_{i,1}, \ldots, c_{i,k}\}$ ($k \geq 2$) and $\mathcal{S} = \{m : \exists i,\ C_i \subseteq m\}$, the minimum certificate has size k, but identical first-order profiles force a common score threshold to select $\varnothing$ or all Mk facts (feasible ratio M). For $k = 3$, greedy with adversarial ties can require $M + 2$ facts, giving ratio $(M + 2)/3 \to \infty$.*

Proof. Upward closure is A2; the instance $\{h, p, q\}$ with $\mathcal{S} = \{m : h \in m \text{ or } \{p, q\} \subseteq m\}$ has minimal sets $\{h\}$ (size 1) and $\{p, q\}$ (size 2), and matroid bases are equicardinal, so no matroid represents it. For the chains: each singleton is insufficient ($k \geq 2$) and each fact non-pivotal ($M \geq 2$ leaves an intact chain), so all facts have identical profile $(\rho(\{f\}), \rho(x \setminus \{f\})) = (1, 0)$; an anonymous score assigns them one value, and a threshold yields $\varnothing$ or x, only the latter sufficient, of size Mk vs. optimum k. For greedy, set $k = 3$ and start from $\varnothing$. While every chain contains at most one selected fact, every available marginal is 0. Adversarial ties can therefore select $c_{1,1}, \ldots, c_{M,1}$, then $c_{1,2}$. Now $c_{1,3}$ has marginal 1 and all other available facts have marginal 0, so greedy selects it and stops at the first sufficient set. It has used $M + 2$ facts, whereas C_1 alone uses 3; the ratio is $(M + 2)/3$, unbounded as $M \to \infty$. □

Theorem 7 (No pointwise selector recovers a minimal decision-sufficient set)**.** *Call a selector* pointwise *if each fact's inclusion is a threshold on a score $s(f \mid x) = \Phi(\pi_\omega(f))$ depending on ω only through the first-order data $\pi_\omega(f) = (\rho_\omega(\varnothing), \rho_\omega(x), \rho_\omega(\{f\}), \rho_\omega(x \setminus \{f\}))$. No pointwise selector outputs a minimal decision-sufficient set on all A1–A2 instances: there exist two instances, indistinguishable to every pointwise selector, with disjoint minimal-certificate families.*

Proof. On $x = \{u_1, u_2, w_1, w_2\}$ take (I) $\mathcal{S} = \langle\{u_1, u_2\}, \{w_1, w_2\}\rangle$ and (II) $\mathcal{S}' = \langle\{u_1, w_1\}, \{u_2, w_2\}\rangle$ (upward closures). Minimal sets are $\{\{u_1, u_2\}, \{w_1, w_2\}\}$ and $\{\{u_1, w_1\}, \{u_2, w_2\}\}$, disjoint. Both have identical first-order data: $\rho(\varnothing) = 1$, $\rho(x) = 0$, $\rho(\{f\}) = 1$ and $\rho(x \setminus \{f\}) = 0$ for every f (each 3-element complement still contains a full generating pair). Thus $\pi_\omega = \pi'_\omega$ for all facts, so every pointwise selector produces the same output O on both; O minimal for both would require O in two disjoint families, impossible. The instances differ only in which *pairs* are jointly sufficient, a second-order property, so the certificate is irreducibly a property of subsets. □

Z.8 An information-theoretic identifiability limit

We formalize the surface-decorrelated setting of Section 7: when the pivotal index and its value are independent of observable surface features, a surface-only selector has a capacity-independent regret floor.

Model. Fix n facts, budget $k < n$, actions $|\mathcal{A}| = K \geq 2$, error cost $\Delta > 0$, and baseline payoff V. An instance draws a hidden pivotal index $J \in [n]$ and value $Y \in \mathcal{A}$; fact $f_i = (\varphi_i, y_i)$ has observable surface feature φ_i and a hidden value slot with $y_J = Y$, $y_i = \perp$ (no information) otherwise. Observing a fact reveals both. With $Q_\omega(a \mid x) = V - \Delta\mathbf{1}[a \neq Y]$ and $Q_\omega(a \mid m) = V - \Delta \Pr(Y \neq a \mid m)$, the memory-optimal action is the maximum a posteriori (MAP) estimate and

$$\rho_\omega(m) = \Delta\, \mathbf{1}[a^\star_\omega(m) \neq Y]. \tag{4}$$

A *surface-only selector* is any (randomized) map $\mathrm{Sel}(\Phi, U) \to S$, $|S| = k$, measurable in the surface profile $\Phi = (\varphi_1, \ldots, \varphi_n)$ and internal randomness U only.

Assumption 1 (Surface-uninformative pivot)**.** *$J \perp Y$, $(J, Y) \perp \Phi$, $J \sim \mathrm{Unif}[n]$, $Y \sim \mathrm{Unif}(\mathcal{A})$, and $U \perp (J, Y, \Phi)$. The surface law is exchangeable across facts and carries zero information about which fact is pivotal, i.e. $I(J; \Phi) = 0$.*

Theorem 8 (Identifiability lower bound)**.** *Under Assumption 1, for every surface-only selector with budget $k < n$,*

$$\mathbb{E}\,\rho_\omega(m) = \Delta\Big(1 - \tfrac{1}{K}\Big)\Big(1 - \tfrac{k}{n}\Big),$$

an exact identity independent of Sel*; hence the minimax regret over the class is at least this value. A value-supervised selector that observes the* expected *counterfactual-forgetting label $\bar{\ell}_i(j) = \mathbb{E}[\rho_\omega(x \setminus \{f_i\}) \mid J = j] = \Delta(1 - \frac{1}{K})\mathbf{1}[i = j]$, with the pivot index held fixed and the counterfactual value Y redrawn uniformly, identifies the pivot ($\bar{\ell}_i(j) \neq 0 \iff i = j$) and attains $\rho_\omega = 0$ by keeping f_J. The single-draw label $\rho_\omega(x \setminus \{f_J\})$ equals Δ only with probability $1 - \frac{1}{K}$: removing the pivot leaves a K-way tie broken correctly with probability $1/K$, so averaging over the counterfactual draw is what removes this noise.*

Proof. Since S is $\sigma(\Phi, U)$-measurable and $J \perp (\Phi, U)$ is uniform, $\Pr(J \in S) = \mathbb{E}[|S|/n] = k/n$. On $\{J \in S\}$ the value slot $y_J = Y$ is observed, so $a^\star_\omega(m) = Y$ and the indicator in (4) is 0. Since $Y \perp (J, \Phi, U)$, conditioning on $J \notin S$ and observing the selected empty value slots leaves Y uniform. Thus any estimator on this event errs with probability $1 - 1/K$. Averaging, $\mathbb{E}\,\rho_\omega(m) = \Delta(1 - \frac{1}{K})\Pr(J \notin S) = \Delta(1 - \frac{1}{K})(1 - \frac{k}{n})$. The minimax bound follows since a supremum over instances dominates the uniform-mixture average. Value supervision reveals J, so keeping f_J gives regret 0. □

Why scale cannot help. The bound contains no term for Sel: by the data-processing inequality any statistic of Φ has $I(J; \mathrm{Sel}(\Phi, U)) = 0$, so no surface-only rule retains the pivot above the prior rate k/n regardless of capacity or compute. The factor $1 - \frac{1}{K}$ is the probability of guessing wrong on a K-ary decision ($\frac{1}{2}$ for binary; as $K \to \infty$ the guessing success $\frac{1}{K}$ vanishes, so the factor $\to 1$ and the bound approaches $(1 - \frac{k}{n})\Delta$). The bound isolates the information limitation tested by the surface-decorrelated suites. The bound concerns the limiting case $I(J; \Phi) = 0$, where the pivot is independent of *all* observable features; it lower-bounds salience-only and prompt-only selectors, not value-supervised learning. In the empirical suites, task structure carries learnable information about pivotality beyond the generic salience cues; value-supervised training uses this

information to predict retained facts. The theorem thus states an oracle-information separation; amortized generalization is a separate, empirical claim.

Z.9 THE COUNTERFACTUAL-FORGETTING LABEL IS A FUNCTIONAL OF REGRET

The counterfactual-forgetting signal measures realized regret. We relate the label to this metric and bound the effect of perturbations to action values. Let A1 hold with margin $\gamma > 0$ ($Q_\omega(a^\star \mid x) - Q_\omega(a \mid x) \geq \gamma$ for $a \neq a^\star$, $a^\star = a^\star_\omega(x)$) and write $\Delta(m) = \max_a |Q_\omega(a \mid x) - Q_\omega(a \mid m)|$.

Lemma 3 (Margin–distortion stability)**.** *Under A1, for any m with $\varepsilon = \Delta(m)$: (i) $\rho_\omega(m) \leq 2\varepsilon$; (ii) $\rho_\omega(m) \in \{0\} \cup [\gamma, \infty)$; (iii) $\varepsilon < \gamma/2 \Rightarrow \rho_\omega(m) = 0$.*

Proof. (i) With $b = a^\star_\omega(m)$, $Q_\omega(b \mid x) \geq Q_\omega(b \mid m) - \varepsilon \geq Q_\omega(a^\star \mid m) - \varepsilon \geq Q_\omega(a^\star \mid x) - 2\varepsilon$, so $\rho_\omega(m) \leq 2\varepsilon$. (ii) If $b \neq a^\star$, A1 gives $\rho_\omega(m) \geq \gamma$. (iii) $\rho_\omega(m) \leq 2\varepsilon < \gamma$ forces $\rho_\omega(m) = 0$ by (ii). □

Theorem 9 (Grounding and graceful degradation)**.** *Under A1 (and A2 where noted):*

1. ***(Label = regret functional)** A fact is pivotal iff $\rho_\omega(x \setminus \{f\}) > 0$ iff $\rho_\omega(x \setminus \{f\}) \geq \gamma$; the pivotality label is $\ell(f) = \mathbf{1}[\rho_\omega(x \setminus \{f\}) > 0]$, read off realized regret rather than stipulated.*

2. ***(Soundness)** The subset oracle returns a set $m^\star$ of minimum cardinality among those its robust criterion certifies decision-sufficient ($\rho_\omega(m^\star) = 0$ by definition), so $\rho_\omega(m^\star) = 0 = \min_m \rho_\omega(m)$ since $\rho_\omega \geq 0$: the oracle target is a regret minimizer (minimum under the robust certificate; the primary experiments use tolerance $\varepsilon = 0$). Value-exactness $\Delta(m) = 0$ is a sufficient but not necessary condition for $\rho_\omega(m) = 0$ (Lemma 3(iii)); the oracle certifies the weaker $\rho_\omega(m^\star) = 0$, not $\Delta(m^\star) = 0$. Under A2, every pivotal fact lies in $m^\star$ (a necessary condition), but pivotal facts need not be jointly sufficient (Proposition 2), which is why sufficiency is a property of subsets and the oracle enumerates subsets rather than scoring facts one at a time.*

3. ***(Robustness, no A2)** Let $\ell_f = \sup_{m' \subseteq x} \max_a |Q_\omega(a \mid m') - Q_\omega(a \mid m' \setminus \{f\})|$ bound the one-action-value change from dropping f over* any *retained set. If $\mathrm{cf}(m) = \sum_{f \notin m} \ell_f \leq \varepsilon$, then $\rho_\omega(m) \leq 2\varepsilon \cdot \mathbf{1}[\varepsilon \geq \gamma/2]$: exactly 0 below $\gamma/2$, and bounded linearly beyond.*

Proof. (1) By (3) f pivotal $\iff$ $a^\star_\omega(x \setminus \{f\}) \neq a^\star$ $\iff$ $\rho_\omega(x \setminus \{f\}) > 0$; the dichotomy is Lemma 3(ii). (2) By definition the oracle returns $m^\star$ with $\rho_\omega(m^\star) = 0$, so $\rho_\omega(m^\star) = 0 = \min_m \rho_\omega(m)$ since $\rho_\omega \geq 0$; separately, $\Delta(m) = 0 \Rightarrow \rho_\omega(m) = 0$ by Lemma 3(iii), a sufficient condition only ($\rho_\omega(m) = 0$ does not require $\Delta(m) = 0$, since preserving the argmax gap does not fix every action value). Under A2, if a pivotal f had $f \notin m^\star$, then $m^\star \subseteq x \setminus \{f\}$ and monotone sufficiency would give $\rho_\omega(x \setminus \{f\}) = 0$, contradicting pivotality; hence every pivotal fact lies in $m^\star$. Pivotal facts are thus necessary but, by Proposition 2, not jointly sufficient. (3) Telescoping the dropped set, $\Delta(m) \leq \sum_{f \notin m} \ell_f = \mathrm{cf}(m) \leq \varepsilon$, then apply Lemma 3(i),(iii). □

Consequences. Part (1) identifies the label with the indicator of strictly positive leave-one-out regret. Part (3) applies without monotone sufficiency: for a non-monotone reader, the regret bound grows linearly with the certified perturbation budget, and the optimal action is preserved when that budget lies below half the full-observation gap.