

SEPO: Evidence-Grounded Prompt Optimization via Structural Editing

Xiaoyu Ma^{♣†} Haoyue Liu^{♣†} Yiwen Li[♣] Jionghao Zhu[♣]

Zhichao Wang[♣] Ye Chen[♡] Xiaoying Tang^{♣♣◇*}

[♣]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen

[♣]Shenzhen Future Network of Intelligence Institute (FNI-IL-Shenzhen)

[◇]Guangdong Provincial Key Laboratory of Future Networks of Intelligence, CUHK(SZ)

[♡]Xi'an Jiaotong University

Email: xiaoyuma@link.cuhk.edu.cn, tangxiaoying@cuhk.edu.cn

[†]Equal contribution. ^{*}Corresponding author.

Abstract

Existing API-only prompt optimisers are often described as *interpretable*, but in practice, this usually means only post-hoc inspectability: each iteration still rewrites the prompt as one opaque string, leaving a trace of full-prompt diffs rather than localisable, machine-readable edits. This paper introduces SEPO (Structural, Evidence-grounded Prompt Optimization), a multi-trajectory prompt optimiser centred on *edit-effect lineage feedback*. Rather than treating each iteration as an isolated whole-prompt rewrite, SEPO locally edits stable, typed units in a two-layer prompt schema, links the target and realised structural operations of each edit to the examples it newly fixes or breaks, and carries this edit–effect record forward to guide later architect calls on the same search branch. This makes prompt optimisation *ad-dressable*, *attributable*, and *actionable*. Across a 14-task held-out suite, SEPO improves over the strongest baseline, GEPA, by 3.1 pp on Llama-3.1-8B-Instruct and 2.2 pp on Qwen3-8B, reaching 61.9% and 73.3% macro accuracy. SEPO also lies on both the optimisation-time and test-time Pareto frontiers, spending 2.9M optimisation tokens versus 4.1M for GEPA and producing prompts over $5\times$ shorter.

1 Introduction

Prompt optimisation has become a de-facto recipe for adapting frozen LLMs to new tasks without weight updates: an architect LLM proposes a rewrite, the candidate is scored on a training set, and the cycle repeats (Ramnath et al., 2025; Yang et al., 2024; Pryzant et al., 2023; Agrawal et al., 2026; Yuksekgonul et al., 2025). Despite being commonly described as *interpretable*, most API-only prompt optimisers still expose the same underlying interface: each iteration rewrites the prompt

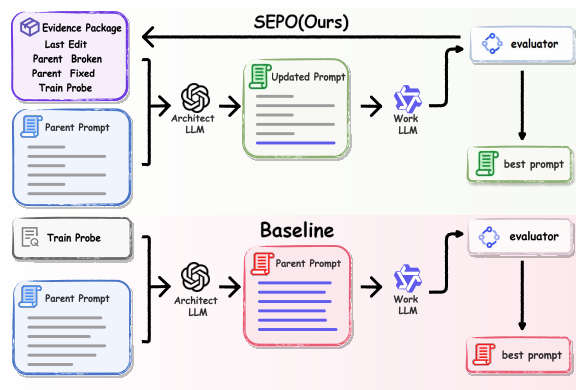


Figure 1: **SEPO at a glance.** SEPO makes localised, evidence-grounded edits (blue) to a structured prompt (top), each traceable to examples it fixed or broke, whereas monolithic rewriters rewrite the whole prompt (blue), opaquely (bottom).

as one opaque string. The resulting trace is therefore a sequence of full-prompt diffs: it shows *that* the prompt changed, but not *where* the relevant change occurred, *why* it was made, or *which* examples it helped or harmed. This limitation persists whether the rewriter is driven by text-gradient feedback or by LLM-as-optimiser sampling: the unit of edit remains the entire prompt.

A handful of recent structured optimisers begin to challenge this single-string interface. Some expose local edit targets, such as factor-, section-, token-, or tree-level units (Schnabel and Neville, 2024; Liu et al., 2026b; Sharma and Henley, 2026; Jain and Chowdhary, 2025; Kumar et al., 2025); others attribute failures to semantic units or agent contributions (Chen et al., 2026; Xia et al., 2026; Zhang et al., 2026); and reflective or self-evolving optimisers improve search with archives, critique, and synthesis (Agrawal et al., 2026; Agarwal et al., 2025). Existing prompt optimisers typically update a prompt based on failures surfaced in the current

round. However, rules added to address those failures can also change the worker’s behaviour on examples that it handled correctly under the parent, thereby introducing new regressions. Without this parent-to-child transition, later iterations may conflate persistent failures with regressions introduced by the preceding edit, leading to repeated rule accretion or another broad prompt rewrite. This can cause prompt bloat and make effective improvements harder to preserve across iterations.

In this paper, we propose SEPO, a multi-trajectory prompt optimiser that addresses this problem through *edit-effect lineage feedback*. Each edit targets a stable typed unit in a two-layer prompt schema; after the candidate is evaluated, SEPO links the target and realised operations of the edit to the examples it newly fixes or breaks and reuses this record in later architect calls on the same search branch. To preserve trajectory diversity, SEPO combines example-level admission with Lexicase-based parent selection (Spector, 2012), keeping specialists alive for rare training examples.

We evaluate SEPO on a 14-task held-out suite using Llama-3.1-8B-Instruct and Qwen3-8B as worker models. We cap optimisation at 2000 *metric calls* per task and seed—one metric call evaluates μ on one training example, consuming one Worker LLM forward pass. Under this matched budget, SEPO reaches 61.9/73.3% macro accuracy on Llama/Qwen—+3.1/ + 2.2 pp over the best of six API-only baselines (GEPA)—while landing on both optimisation- and test-time Pareto frontiers.

Our contributions are four-fold.

(i) We introduce SEPO and its core mechanism, *edit-effect lineage feedback*. This mechanism links the target and realised operations of each local edit to the examples it newly fixes or breaks and reuses this record in subsequent edits on the same search branch, making prompt optimisation *addressable*, *attributable*, and *actionable*.

(ii) We provide a controlled API-only comparison on a 14-task held-out suite covering reasoning (BBH, GSM-Hard), coding (MBPP+), multi-hop question answering (HotpotQA), and multidomain tasks from MMLU-Pro. Under the matched 2000-metric-call budget, SEPO consistently matches or exceeds six API-only baselines across both worker models, winning at least 11/14 tasks against every baseline.

(iii) We show that the gains from structured, lineage-aware editing are strongest on procedu-

ral, multi-step tasks—most clearly Word Sorting (+18.9 pp over GEPA on Llama) and MBPP+ (+8.7 pp on Qwen).

(iv) We show that SEPO sits on both the optimisation-time and test-time Pareto frontiers, spending 2.9M optimisation tokens and 203k deployment tokens per cell—compared with 4.1M and 424k for GEPA—with prompts over $5\times$ shorter than GEPA’s and remaining 6.9 pp above APSF, the only other method on both frontiers.

2 Related Work

Monolithic whole-prompt rewriting. A first family rewrites the prompt as an opaque string under a scalar validation reward (Ye et al., 2024): APE (Zhou et al., 2023) samples and re-scores instructions, OPRO (Yang et al., 2024) casts the LM as a numerical-feedback optimiser, GrIPS (Prasad et al., 2023) searches with gradient-free operators, and ProTeGi (Pryzant et al., 2023) introduces “textual gradient” beam search, extended by TextGrad (Yuksekgonul et al., 2025) to text-level backpropagation. Lacking component-level addresses, a rewrite that breaks one example yields no localisable change to attribute, and its trace collapses to untyped full-prompt diffs.

Structure-aware factorised optimisers. A second family exposes components as first-class units: *scaffold-based* methods optimise per-module instructions and demonstrations over a user-specified program (DSPy+MIPROv2 (Khatab et al., 2024; Opsahl-Ong et al., 2024)); *fixed-template* methods refine each section of a cross-task schema (Modular Prompt Optimisation (Sharma and Henley, 2026)); *type-aware* methods allocate repair operators by failure type (TARA (Liu et al., 2026a)); *decomposition-based* methods use a predefined taxonomy (Task-Facet (Juneja et al., 2025)) or self-discovered factors (aPSF (Liu et al., 2026b)); and *principle-aware* variants refine via multi-aspect critique (CriSPO (He et al., 2025)) or generalisable criteria (ZERA (Yi et al., 2025)). Although these factorisations expose local editable units, they typically do not explicitly link each realised edit to the examples it newly fixes or breaks, or carry this edit-effect relation forward along the search branch.

Reflective and evolutionary search with diversity. A third family uses natural-language reflection (Shinn et al., 2023), population diversity, or prompt-aware evaluation scheduling (Ma et al.,

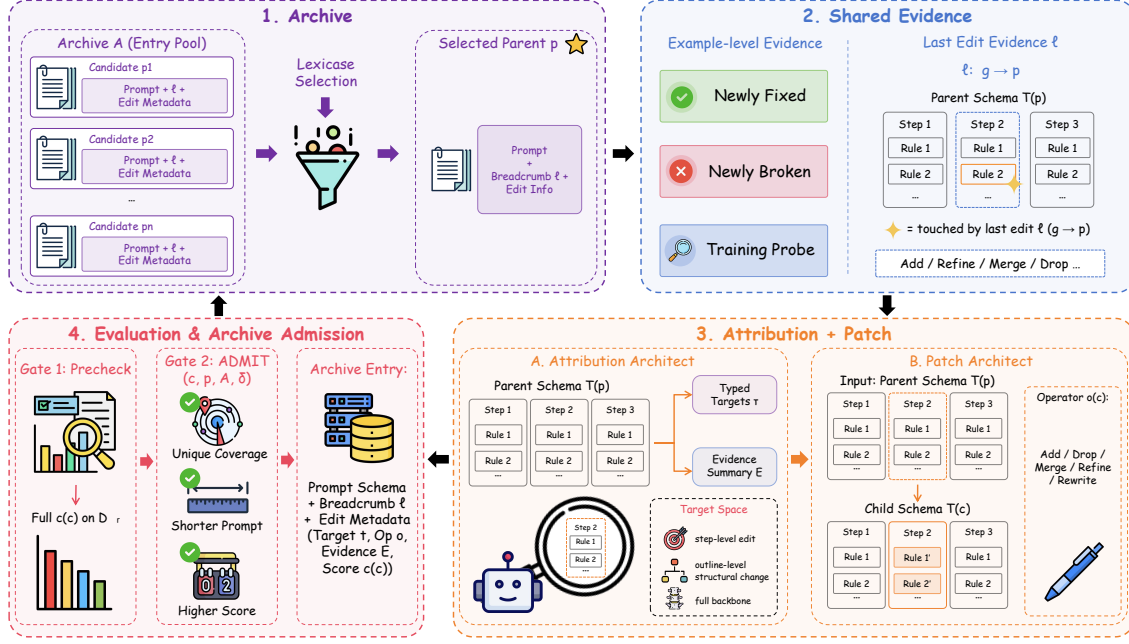


Figure 2: **Per-iteration overview of SEPO.** Parent selection (§3.4) draws from the archive $\mathcal{A}$; the lineage-aware minibatch and breadcrumb (§3.3) feed the attribution and patch architects; the resulting child is gated by Gate 1 (subsample precheck) and admitted under the three example-level rules of Gate 2 (§3.4).

2026). Self-Refine (Madaan et al., 2023) and ProMST (Chen et al., 2024) show that iterative feedback can substitute for an explicit scalar score, and population-based methods add diversity: Evo-Prompt (Guo et al., 2024) ports Genetic Algorithm and Differential Evolution operators onto a prompt population, Promptbreeder (Fernando et al., 2024) co-evolves mutation prompts alongside the task prompts they edit, GEPA (Agrawal et al., 2026) (Genetic-Pareto) couples per-module reflective mutation with a Pareto-frontier archive, and PromptAgent (Wang et al., 2024a) runs Monte-Carlo tree search with reflective error feedback. Outside prompt optimisation, verifier-grounded planning has likewise paired typed structural patches with deterministic constraint checking (Wang et al., 2026). Yet even when archives or search trees preserve multiple candidates, reflection often remains centred on the current candidate’s traces, without linking an edit’s target and realised operations to the examples newly fixed or broken in the parent-to-child transition, or carrying this record forward to subsequent edits on the same search branch.

3 Method

SEPO is a multi-trajectory prompt optimiser: each iteration applies one local edit to a structured prompt and records the edit together with its

example-level outcomes as an edit-effect lineage state, making the optimisation process auditable and providing reusable feedback for later edits on the same search branch. SEPO rests on three principles. **(P1) Localised edits** over stable, addressable prompt units give every edit an explicit target (§3.2). **(P2) Edit-effect lineage feedback** uses two architect calls, *attribution* and *patch*. Both calls share an evidence packet comprising the parent’s preceding edit, the examples that edit newly fixed or broke, and the associated breadcrumb (§3.3). **(P3) Trajectory-diverse archiving** uses example-level admission and Lexicase-based parent selection to determine which candidate branches merit further search (§3.4). Figure 2 and Algorithm 1 detail one iteration.

3.1 Problem Setup

We assume a *frozen* worker LLM W for downstream inference and a *frozen* architect LLM A used in the two roles described in §3.3. Given a task $\mathcal{T}$ with a training set $\mathcal{D}_{\text{tr}}$ used during optimisation and a disjoint held-out test split, we optimise a schema s according to

$$s^* = \arg \max_s \mathbb{E}_{(x,y) \sim \mathcal{T}} [\mu(W(s, x), y)], \quad (1)$$

where μ is a binary task metric and $W(s, x)$ is the worker’s output under the prompt rendered from

s , with the expectation estimated by the empirical mean over $\mathcal{D}_{\text{tr}}$. For each s we record its score vector $\mathbf{c}(s) \in \{0, 1\}^{|\mathcal{D}_{\text{tr}}|}$ and correct set $C(s) = \{i : \mathbf{c}(s)_i = 1\}$ under a budget of B metric calls.

3.2 Prompt representation

Unlike whole-prompt rewriters without local failure addresses, we represent each candidate with a task-agnostic two-layer schema:

$$s = (\alpha, \{\rho_j\}_{j=1}^m), \quad (2)$$

A top-level *outline* $\alpha = (\alpha_1, \dots, \alpha_K)$ preserves global intent, while each rule ρ_j is attached to one step and constrains local edits without fragmenting context. The attribution architect targets one step or the whole outline, and the patch architect revises it (§3.3). Linking each addressable edit to its example-level effects provides the interface for edit-effect lineage feedback; other representations can provide the same interface. The schema below instantiates Eq. 2 for BBH Word Sorting with $K=3$ outline steps and step-local rules.

Operator label space. After each patch call, a deterministic parser extracts the child’s outline steps and step-local rules; a structural diff against the parent, together with the attributed target, then labels each changed step or rule with one operator from $\mathcal{O}$,

$$o(c) \triangleq \text{INFEROP}(\text{DIFF}(p, c), \tau(c)), \quad (3)$$

where INFEROP is deterministic and $o(c)$ collects one label per changed location; $\tau(c)=\tau$ is the attributed target assigned to the child. These labels are recorded on the child for use in the next iteration’s breadcrumb. The vocabulary $\mathcal{O}$, the structural diff $\text{DIFF}(p, c)$ (Algorithm 3), and the per-location classification rule are in Appendix C.2.

Example schema (BBH Word Sorting, seed 45, test acc. 58%)

Step 1. Understand the problem — read carefully, parse all entities, and decompose into sub-questions if needed.

Step 2. Reason step by step.

- Copy the words EXACTLY as they appear after the marker “List:”; do not substitute, invent, or import any words from training data or memory.
- Count the input words once, then sort in strict A–Z order in a single pass.
- CRITICAL: after sorting, do *not* re-process the list (no re-reading or side-by-side checks)—both trigger hallucination loops on this worker; output immediately.

Step 3. Output the sorted list.

- Output on the final line as “Answer: word1 word2 . . .” — exactly the extracted words, only re-ordered, separated by single spaces with no commas, newlines, numbering, or bullets.

3.3 Lineage-guided reflective edit generation

Each iteration of SEPO proceeds in five steps. We (i) draw a parent p from the archive $\mathcal{A}$ (§3.4); (ii) build an evidence packet from p and its grandparent g , comprising training examples that p ’s last edit broke or newly fixed, train probes, and a summary ℓ of what the last edit changed; (iii) run an *attribution* architect call returning a typed target τ (Eq. 4)—identifying either an edit within one outline step or a structural change to the outline—together with the failure reason; (iv) run a *patch* architect call returning a full revised schema $\tilde{s}$, which is parsed into the child c ; (v) infer the realised operators from the structural diff. After the child is evaluated (§3.4), this completes its edit-effect lineage tuple $(\tau, \ell, o(c), \text{effect})$ for later architect calls on the same search branch.

Balanced evidence packet. Both architect calls consume the same minibatch $M = (M_{\text{focus}}, M_{\text{anchor}}, M_{\text{train}})$, where for p ’s last edit

$$\text{NEWLY_BROKEN}(p) \triangleq \{i : \mathbf{c}(g)_i=1, \mathbf{c}(p)_i=0\},$$

$$\text{NEWLY_FIXED}(p) \triangleq \{i : \mathbf{c}(g)_i=0, \mathbf{c}(p)_i=1\},$$

$$a_i \triangleq |\mathcal{A}|^{-1} \sum_{s \in \mathcal{A}} \mathbf{c}(s)_i \in [0, 1].$$

M_{focus} samples $\text{NEWLY_BROKEN}(p)$ with weights $w_i \propto a_i$ (the archive agreement on example i), over-sampling failures on examples the archive shows the worker can solve rather than failures likely caused by capability limits. The architect uses these newly broken examples together with train probes from M_{train} to infer why the breakages occurred and to steer the next edit away from the same failure mode; recovering in-capability failures directly supports generalisation. M_{anchor} draws from $\text{NEWLY_FIXED}(p)$, preserving these recent gains. Slot sizes and expressiveness arguments are in Appendix D.1.

Fallback without regressions. When NEWLY_BROKEN is empty—at cold start ($g=\emptyset$) or when the parent’s last edit introduced no regression on $\mathcal{D}_{\text{tr}}$ —the iteration drops M_{focus} and uses M_{anchor} together with M_{train} . In this case, no training regression is available to explain, so the iteration continues improving on the train probes while preserving one anchor.

Attribution and patch. The attribution architect consumes (p, M, ℓ) and diagnoses where each error arises in the schema and why it occurred, returning a typed target τ and failure rationale r ,

$$(\tau, r) = A_{\text{attr}}(p, M, \ell), \quad (4)$$

where τ takes one of five typed forms:

- **step <N>**: edit confined to step N (the patch architect may add, drop, refine, or merge rules inside that step, or rewrite the step’s top-level text);
- **ADD step at <pos>**: insert a missing reasoning stage at outline position <pos>;
- **SPLIT step <N>**: split two conflated responsibilities inside step N ;
- **DROP step <N>**: remove step N ;
- **backbone**: rewrite the entire outline.

The attribution call returns only the edit location τ and failure rationale r , not a new schema. The patch architect takes (p, τ, r, M, ℓ) and emits a full revised schema text

$$\begin{aligned} \tilde{s} &= A_{\text{patch}}(p, \tau, r, M, \ell), \\ c &= \text{PARSESHEMA}(\tilde{s}), \end{aligned} \quad (5)$$

Carry-forward lineage record. After producing and evaluating child c , we record its attributed target τ , diff-inferred realised operators $o(c)$, breadcrumb ℓ , and newly-fixed/newly-broken examples as the edit-effect lineage tuple $(\tau, \ell, o(c), \text{effect})$; effect stores the example sets, while reporting uses only their counts. If c is later selected as a parent, ℓ and these transition examples jointly condition the next attribution and patch calls. The **—LINEAGE** and **—BREADCRUMB** ablations (§4.3) remove the lineage evidence and the breadcrumb summary ℓ respectively.

3.4 Example admission and Lexicase sampling

To control evaluation cost, SEPO first evaluates each child on the M_{train} probe slot and discards candidates that do not strictly improve over the parent on those probes. A child that passes this precheck is evaluated on the remaining training examples, reusing the probe results to complete its full-training score vector, and is admitted into the archive when at least one of the following holds:

$$\begin{cases} \text{(a)} & C(c) \setminus \bigcup_{s \in \mathcal{A}} C(s) \neq \emptyset \\ \text{(b)} & |c| < |p| \\ \text{(c)} & \sum c(c) > \sum c(p) \end{cases} \quad (6)$$

That is, the child adds unique coverage, is shorter than the parent, or achieves a higher score than the parent. The unique-coverage criterion preserves specialists that solve rare training examples.

We select the next parent from the archive using Lexicase selection (Spector, 2012) with probability 0.8 and uniform random sampling with probability 0.2. Lexicase filters candidates using a randomly ordered sequence of training examples, preserving specialists for difficult cases, while random sampling retains additional exploration. Full budget accounting and implementation details are given in Appendix C.

4 Experiments

Experimental questions. We evaluate SEPO through four questions:

- **Q1: Held-out effectiveness.** Under a matched API-only optimisation and deployment protocol, does SEPO produce stronger final prompts than strong prompt optimisers?
- **Q2: Edit-diagnostic mechanism.** Are SEPO’s gains driven by its edit-aware diagnostic loop, rather than by generic reflection, static factorisation, or unlocalised prompt rewriting?
- **Q3: Behavioral scope and auditability.** Where does SEPO help or fail, and do its edit traces make these outcomes inspectable at evidence, target, operator, and example-level effect?
- **Q4: Cost-effectiveness.** What optimisation-time and deployment-time cost do SEPO’s gains require, and does SEPO remain competitive on accuracy–cost trade-offs?

4.1 Experimental Setup and Protocol

Benchmarks. Our primary suite contains 14 held-out tasks spanning four families:

- **BBH (8 tasks)** (Suzgun et al., 2023; Srivastava et al., 2023): causal judgement, date understanding, disambiguation QA, formal fallacies, logical deduction over seven objects, reasoning about colored objects, snarks, and word sorting, chosen to cover BBH’s distinct capability types (world-knowledge causal inference, temporal reasoning, linguistic disambiguation, formal logic, multi-step deduction, symbolic manipulation, sarcasm understanding, and algorithmic sorting).
- **MMLU-Pro (3 tasks)** (Wang et al., 2024b): computer science, economics, and psychology, span-

Method	BBH								Code	MMLU-Pro			Reasoning		Avg.
	Caus.	Date	Disamb.	Fall.	Deduct.	Obj.	Snarks	Word.	MBPP+	CS	Econ.	Psych.	GSM-H	HQA	
Worker LLM: Llama-3.1-8B-Instruct															
Seed	37.5 $\pm$ 0.0	66.7 $\pm$ 0.0	61.7 $\pm$ 0.0	44.5 $\pm$ 0.0	45.0 $\pm$ 0.0	82.5 $\pm$ 0.0	72.4 $\pm$ 0.0	0.0 $\pm$ 0.0	57.3 $\pm$ 0.0	44.7 $\pm$ 0.0	57.5 $\pm$ 0.0	56.7 $\pm$ 0.0	35.8 $\pm$ 0.0	25.0 $\pm$ 0.0	49.1 $\pm$ 0.0
OPRO	59.9 $\pm$ 3.2	65.2 $\pm$ 2.3	65.8 $\pm$ 4.1	42.8 $\pm$ 3.8	45.7 $\pm$ 5.2	81.8 $\pm$ 9.0	82.3 $\pm$ 3.2	5.2 $\pm$ 0.6	60.2 $\pm$ 1.4	41.0 $\pm$ 2.6	52.2 $\pm$ 2.6	59.8 $\pm$ 2.6	30.2 $\pm$ 4.9	69.6 $\pm$ 8.7	54.4 $\pm$ 3.9
APSF	52.6 $\pm$ 8.3	68.2 $\pm$ 1.2	62.8 $\pm$ 2.4	45.8 $\pm$ 7.3	45.2 $\pm$ 4.3	83.5 $\pm$ 1.3	76.8 $\pm$ 4.0	25.2 $\pm$ 12.0	60.7 $\pm$ 2.3	48.7 $\pm$ 4.5	57.8 $\pm$ 2.8	60.3 $\pm$ 2.3	30.1 $\pm$ 2.5	51.9 $\pm$ 11.2	55.0 $\pm$ 4.7
MIPROv2	63.3 $\pm$ 2.9	67.8 $\pm$ 2.3	65.2 $\pm$ 4.7	49.3 $\pm$ 8.6	44.5 $\pm$ 2.2	79.0 $\pm$ 1.7	76.6 $\pm$ 1.4	3.3 $\pm$ 0.3	56.0 $\pm$ 2.2	47.3 $\pm$ 3.0	52.7 $\pm$ 1.9	59.3 $\pm$ 2.6	37.2 $\pm$ 2.1	62.5 $\pm$ 1.1	54.6 $\pm$ 2.6
MPO	52.6 $\pm$ 3.3	73.2 $\pm$ 6.9	67.0 $\pm$ 2.2	44.8 $\pm$ 4.0	47.7 $\pm$ 4.4	83.7 $\pm$ 5.1	75.0 $\pm$ 2.1	56.5 $\pm$ 2.2	59.0 $\pm$ 1.7	47.7 $\pm$ 1.4	55.5 $\pm$ 4.6	56.0 $\pm$ 0.9	33.8 $\pm$ 2.0	67.6 $\pm$ 0.9	58.6 $\pm$ 3.0
GEPA	59.9 $\pm$ 3.8	75.3 $\pm$ 6.1	64.5 $\pm$ 6.6	47.5 $\pm$ 2.2	47.0 $\pm$ 3.9	86.0 $\pm$ 3.1	77.3 $\pm$ 2.1	39.3 $\pm$ 18.3	60.0 $\pm$ 1.3	46.5 $\pm$ 3.0	54.5 $\pm$ 1.3	59.3 $\pm$ 4.2	37.2 $\pm$ 2.1	68.4 $\pm$ 1.9	58.8 $\pm$ 4.3
SEPO	61.1 $\pm$ 4.6	76.5 $\pm$ 4.3	68.8 $\pm$ 4.6	50.8 $\pm$ 3.6	47.3 $\pm$ 2.1	87.5 $\pm$ 0.9	82.3 $\pm$ 1.6	58.2 $\pm$ 1.3	62.6 $\pm$ 1.8	46.7 $\pm$ 1.8	58.8 $\pm$ 2.8	58.7 $\pm$ 1.8	37.6 $\pm$ 1.8	70.2 $\pm$ 3.8	61.9 $\pm$ 2.6
Worker LLM: Qwen3-8B (HQA = HotpotQA Hard)															
Seed	56.9 $\pm$ 0.0	83.8 $\pm$ 0.0	44.3 $\pm$ 0.0	46.0 $\pm$ 0.0	84.7 $\pm$ 0.0	94.5 $\pm$ 0.0	76.0 $\pm$ 0.0	5.3 $\pm$ 0.0	15.7 $\pm$ 0.0	68.7 $\pm$ 0.0	73.8 $\pm$ 0.0	66.2 $\pm$ 0.0	48.6 $\pm$ 0.0	35.5 $\pm$ 0.0	57.1 $\pm$ 0.0
OPRO	60.6 $\pm$ 3.2	78.7 $\pm$ 1.8	66.0 $\pm$ 7.0	47.3 $\pm$ 3.6	86.7 $\pm$ 4.2	97.0 $\pm$ 2.2	82.3 $\pm$ 1.6	7.7 $\pm$ 1.3	66.7 $\pm$ 1.0	69.3 $\pm$ 1.2	75.0 $\pm$ 2.2	67.7 $\pm$ 3.0	69.0 $\pm$ 0.9	32.8 $\pm$ 2.3	64.8 $\pm$ 2.5
APSF	55.7 $\pm$ 4.5	84.3 $\pm$ 2.3	65.0 $\pm$ 1.3	64.2 $\pm$ 2.0	79.5 $\pm$ 1.3	94.3 $\pm$ 2.3	77.6 $\pm$ 5.5	31.5 $\pm$ 2.6	68.7 $\pm$ 1.2	66.7 $\pm$ 2.8	75.2 $\pm$ 2.8	66.7 $\pm$ 3.1	53.8 $\pm$ 7.5	38.9 $\pm$ 1.1	65.9 $\pm$ 2.9
MIPROv2	57.7 $\pm$ 1.3	86.0 $\pm$ 2.2	77.0 $\pm$ 2.3	56.7 $\pm$ 14.4	83.5 $\pm$ 3.0	97.3 $\pm$ 0.8	81.2 $\pm$ 2.8	6.7 $\pm$ 6.7	68.3 $\pm$ 1.2	69.5 $\pm$ 2.2	76.0 $\pm$ 0.9	68.3 $\pm$ 1.2	62.8 $\pm$ 0.8	32.7 $\pm$ 1.1	66.0 $\pm$ 2.9
MPO	56.7 $\pm$ 3.0	86.0 $\pm$ 4.3	70.5 $\pm$ 7.2	67.7 $\pm$ 9.0	80.5 $\pm$ 3.1	97.7 $\pm$ 0.8	78.1 $\pm$ 3.6	50.5 $\pm$ 0.9	67.5 $\pm$ 1.7	65.0 $\pm$ 1.7	77.5 $\pm$ 1.3	64.0 $\pm$ 2.3	62.6 $\pm$ 0.9	35.3 $\pm$ 2.9	68.5 $\pm$ 3.1
GEPA	61.1 $\pm$ 1.5	86.3 $\pm$ 1.2	74.5 $\pm$ 2.6	78.5 $\pm$ 7.1	88.7 $\pm$ 2.5	98.5 $\pm$ 1.3	78.6 $\pm$ 3.9	51.7 $\pm$ 5.3	64.0 $\pm$ 1.3	69.0 $\pm$ 2.2	73.3 $\pm$ 1.2	67.5 $\pm$ 1.3	67.5 $\pm$ 1.1	36.5 $\pm$ 0.7	71.1 $\pm$ 2.4
SEPO	58.9 $\pm$ 3.7	90.0 $\pm$ 1.7	77.7 $\pm$ 1.4	75.7 $\pm$ 4.9	92.3 $\pm$ 3.0	99.0 $\pm$ 0.9	82.6 $\pm$ 3.0	54.7 $\pm$ 3.9	72.7 $\pm$ 1.1	70.8 $\pm$ 1.3	76.2 $\pm$ 0.8	66.7 $\pm$ 2.0	70.0 $\pm$ 8.0	39.1 $\pm$ 0.6	73.3 $\pm$ 2.6

Table 1: **Held-out test accuracy (%)** on the 14-task suite; mean over three seeds (subscript: std). Tinted bold cells mark the best per Worker LLM block. HotpotQA: Medium (Llama), Hard/HQA (Qwen).

ning a technical, an economic–quantitative, and a behavioural–social domain.

- **Code:** MBPP+ (Liu et al., 2023), scored by pass@1 (Chen et al., 2021).
- **Reasoning:** GSM-Hard (Gao et al., 2023) and HotpotQA (Yang et al., 2018) (Medium split under Llama, Hard split under Qwen).

Models. We run our pipeline with two frozen Workers of different architecture and capability to test cross-family generalisation:

- **Worker LLMs:** Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen3-8B (Yang et al., 2025), temperature 0.0 for deterministic metric evaluation.
- **Architect LLM:** Qwen3.5-397B-A17B (Qwen Team, 2026), temperature 0.7 to encourage diverse candidate edits.

Baselines and protocol. We compare SEPO against six API-only baselines—an unoptimised SEED prompt, APSF, OPRO, MIPROv2, MPO, and GEPA—under a shared budget of 2000 *metric calls* per task and seed (architect LLM calls are *not* counted toward this budget). This matches DSPy / MIPROv2’s max_metric_calls=2000 (Khatab et al., 2024; Opsahl-Ong et al., 2024). All methods start from the same task-agnostic understand–reason–answer seed on every benchmark. The cross-worker HotpotQA column is not comparable: Llama uses Medium because limited worker headroom compresses performance on Hard, whereas Qwen uses Hard (Appendix G.2). The minibatch’s

three slots ($M_{\text{focus}}, M_{\text{anchor}}, M_{\text{train}}$) have sizes $(n_f, n_a, n_{tr}) = (2, 1, 3)$. We report macro accuracy (unweighted 14-task mean), per-baseline win rate, optimisation/deployment tokens, and final prompt length; full configs in Appendix A.2.

4.2 Q1: SEPO Beats All Baselines on Both Workers in Macro Accuracy

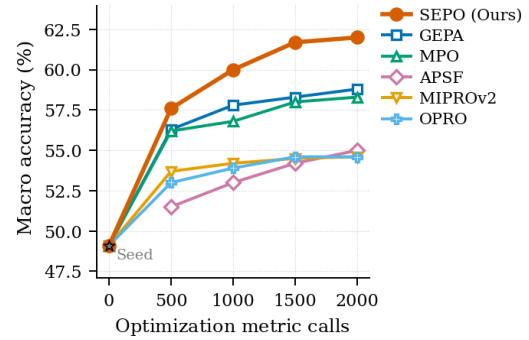


Figure 3: **Budget-matched accuracy curves** (Llama-3.1-8B-Instruct). Test accuracy at metric-call budgets 0/500/1000/1500/2000.

We evaluate whether structured, lineage-aware editing yields stronger held-out prompts than competing optimisers under a matched API-only protocol; per-task results are in Table 1.

1. SEPO reaches 61.9% on Llama (+3.1 pp over GEPA; ≥ 3.3 pp over the rest) and 73.3% on Qwen (+2.2 pp over GEPA; ≥ 4.8 pp over the rest), beating GEPA on 13/14 (Llama) and 11/14 (Qwen), and $\geq 11/14$ against every other baseline under both workers.

2. The gains span procedural and reasoning families—**BBH** (+4.5/ + 1.6 pp), **Code** (MBPP+, +2.6/ +8.7 pp), and **Reasoning** (HotpotQA + GSM-Hard, +1.1/ + 2.5 pp)—ruling out a single-domain artefact, and concentrate on procedural tasks where one failed local sub-step can invalidate the final answer. Word Sorting is the clearest case: SEPO is best under both workers (Llama 58.2% / Qwen 54.7%, +18.9 pp over GEPA on Llama; the seed collapses to 0.0% on Llama), and MBPP+ shows the same pattern (62.6% / 72.7%). On the MMLU-Pro multidomain tasks, SEPO remains within the leading cluster but with thinner per-task margins, illustrating that the magnitude of prompt-optimization gains varies across task families.
3. SEPO also reaches a given accuracy with fewer metric calls: at every metric-call budget in Figure 3 it stays at or above every baseline, so its lead holds across the budget axis and is not an artefact of the 2000-metric-call cap.

4.3 Q2: Each Principle (P1/P2/P3) of SEPO is Load-Bearing

We ablate three principles to test whether SEPO’s gains are driven by its edit-aware diagnostic loop, not by generic reflection or unlocalised rewriting. We isolate four single-removal variants—**–ATTRIBUTION** (drops P1), **–BREADCRUMB** and **–LINEAGE** (drop P2 at two strengths), and **–MULTITRAJ** (drops P3)—and analyse their accuracy, length, and acceptance signatures.

We ablate on five tasks spanning code, math, temporal, linguistic, and multidomain tasks (MBPP+, GSM-Hard, BBH Date, BBH Snarks, and MMLU-Pro CS). Every single-removal variant costs 1.8–4.7 pp on the five-benchmark diagnostic average, with the largest damage on procedural multi-sub-skill tasks. Matched proposal-level controls further isolate these roles: random targets reduce net gain from +2.5 to –1.7 examples per 100 proposals, while replacing newly broken cases with random current errors reduces it to +0.8 (Appendix H.3).

Each variant leaves a distinct signature in Table 2 and Figure 4; P1–P3 below.

P1 — typed targeting localises edits. Removing the attribution architect (**–ATTRIBUTION**) inflates prompt length $1.97\times$ (228→450 tokens), drops accuracy 3.1 pp (Table 2), grows each edit $\sim 1.7\times$ in diff size (2.23→3.77), and lowers acceptance from 0.354 to 0.305 at matched compute (318 vs 320

Variant	P	Acc.	σ	ΔAcc	Len.
Full SEPO	—	61.1	2.2	—	228
–ATTRIBUTION	P1	58.0	2.5	–3.1	450
–LINEAGE	P2	58.2	1.5	–2.9	292
–BREADCRUMB	P2	59.3	2.0	–1.8	235
–MULTITRAJ	P3	56.4	2.3	–4.7	198

Table 2: **Aggregate ablation** on the five-benchmark diagnostic subset (Llama-3.1-8B-Instruct; 3 seeds; σ : std). Acc. %, Len. tokens; **bold/italic**: cells discussed in P1/P3. Breakdowns: Appendices H.1, H.2.

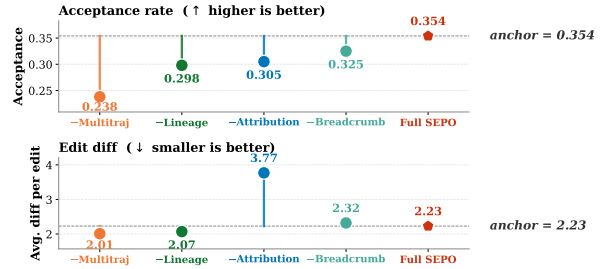


Figure 4: **Edit-quality signature.** Stem length = each variant’s drift from the Full SEPO anchor (dashed) in acceptance rate (top) and average edit size (bottom), sorted by acceptance; colour = principle (P1/P2/P3).

candidate edits; Figure 4). Both signatures share one mechanism: SEPO edits one step per iteration and accumulates, with τ enforcing the per-iteration target. Without it, the patch architect lacks a single target, so it touches multiple steps at once — larger edits inflate length and clear admission less often. The damage concentrates on **linguistic-reasoning tasks** (§4.2), whose compositional rule scaffolds are most fragile under multi-region rewrites (BBH-Snarks –8.5 pp, Table 3).

	P1	P2		P3
Task	−ATTR	−LIN	−BRD	−MT
BBH Date	−2.5	−4.3	−0.5	−4.7
BBH Snarks	−8.5	−3.8	−5.7	−7.3
MBPP+	−1.7	−4.1	−2.8	−7.0
GSM-Hard	−1.2	−2.6	−0.6	−2.5
MMLU-Pro CS	−1.7	+0.1	+0.3	−2.2

Table 3: **Per-task ΔAcc** (five benchmarks, Llama-3.1-8B-Instruct). **Bold**: largest drop per variant; *italic* row: scope-negative case. Full heatmap: Appendix H.2.

P2 — edit-effect lineage feedback supports net-positive edits. **–BREADCRUMB** (removes the prior-edit trace ℓ) costs 1.8 pp, while **–LINEAGE** (removes ℓ and the newly-fixed/newly-broken transition information) costs 2.9 pp (Table 3). These

nested ablations indicate complementary roles: ℓ summarises what changed, while the transition examples identify which cases were newly fixed or broken; together they guide the next edit. The effect is most pronounced on procedural, multi-step tasks that require sequencing local rules, whereas MMLU-Pro CS is roughly indifferent, illustrating that the benefit of lineage evidence varies with task structure.

P3 — archive escapes local optima of greedy editing. —MULTITRAJ disables the archive and Lexicase selection (best-only, greedy single-trajectory editing) and drops 4.7 pp at a *shorter* prompt ($\times 0.87$, 228 $\rightarrow$ 198 tokens) with acceptance falling from 0.354 to 0.238 (Table 2, Figure 4); the result is consistent with a reduced exploration radius rather than prompt bloat. SEPO’s archive mitigates this by keeping diverse candidates alive: **Admission** (Eq. 6) retains unique-coverage, shorter, or higher-scoring children rather than ranking on a single scalar, and **Lexicase** parent selection keeps rare-example solvers as future parents (Appendix H.4). The auxiliary controls show that retaining and revisiting multiple candidate branches provides the dominant benefit, while Lexicase selection and unique-coverage admission each add a modest average gain.

4.4 Q3: Behavioral Scope and Auditability

Auditability. SEPO logs a typed tuple $(\tau, \ell, o(c), \text{effect})$ for every child (§3.3); Figure 5 reads one packet end to end.

Audit packet for one SEPO edit (BBH Word Sorting)	
Local target: τ = step 3, the output formatting rule .	
Parent breadcrumb ℓ : previous edit added count/sort rules under step 2:	
Count the input words once, then sort in strict A–Z order in a single pass.	
Observed side effect: +15 newly fixed on sorting, but with sorting now reliable the worker renders output freely without format constraints , newly breaking 4 correctly-sorted cases on format.	
– Newly broken (focus): Ex #13 is sorted correctly but renders with commas ; Ex #16 breaks a word onto a new line — both fail the single-line answer.	
– Anchor kept: Ex #17 — already correct sort and correct format , preserved by adding rather than overwriting.	
– Train probe: an independent comma-joined case is likewise fixed , so the rule is not confined to focus examples.	
Patch on Step 3: one operation per failure mode.	
○ REWRITESTEP: rename step 3 to “Output the sorted list.”	
+ ADDRULE: output on the final line as “Answer: word1 word2 ...” — exactly the extracted words, only reordered, separated by single spaces with no commas, newlines, numbering, or bullets (closes Ex#13/Ex#16).	
Operator: $o(c) = \{\text{REWRITESTEP, ADDRULE}\}$ Effect: +17 fixed , –6 broken , 20 $\rightarrow$ 31, kept .	

Figure 5: A **local audit packet** (representative; full edit tree in Figure 6, Appendix E).

The breadcrumb summarises the parent’s preceding edit, the focus examples show which cases broke, the patch addresses those cases, and the anchor plus train probe test whether the benefit extends beyond the focus cases. Together, these signals make the edit packet both retrospectively inspectable and useful for guiding subsequent edits. The walkthrough, the edit tree (Figure 6), and example final prompts are in Appendix E.

Rule-level op	#	share	Structural op	#	share
ADDRULE	423	75.4%	REWRITESTEP	34	7.3%
DROPRULE	147	31.7%	ADDSTEP	14	3.0%
REFINERULE	126	37.2%	SPLITSTEP	18	4.0%
MERGERULES	3	0.6%	BACKBONE	1	0.3%

Table 4: **Typed edit operations** (14-task suite, 3 seeds; $N \approx 338$ edits). # = total occurrences; *share* = % of edits containing the op (one edit may use several ops and repeat one, so shares sum to $>100\%$).

Aggregated over the suite, the same log also characterises *how* SEPO edits (Table 4). **Edits stay overwhelmingly step-local:** structural rewrites that add, split, or replace outline steps remain rare, so localised editing is SEPO’s dominant editing mode, in contrast to monolithic whole-prompt rewriting. Each edit is also composite—most add a rule while many also drop or refine one—so SEPO actively prunes and revises rather than merely accreting.

4.5 Q4: SEPO Dominates Cost-Effectiveness across Architect and Worker Channels

We compare SEPO’s optimisation- and deployment-time costs in Table 5.

SEPO achieves the best ρ under a shared budget by trimming both the architect and worker

Method	Opt. (M tok)				Dep.	Len.	ρ	\$/cell	ρ_s
	Acc.	Arch.	Work.	Total	(k)	(tok)		(USD)	(pp/\$)
OPRO	54.4	6.84	3.11	9.95	244.0	358	0.53	1.66	3.2
MIPROV2	54.6	3.52	2.60	6.12	208.4	230	0.90	0.89	6.2
APSF	55.0	0.69	2.13	2.82	157.8	65	2.09	0.19	31.1
MPO	58.6	0.19	3.40	3.59	261.0	416	2.65	0.15	63.3
GEPA	58.8	0.23	3.88	4.11	423.7	1,324	2.36	0.18	53.9
SEPO	61.9	0.34	2.59	2.93	203.2	236	4.37	0.17	75.3

Table 5: **Efficiency decomposition** (Llama-3.1-8B-Instruct, 14-task macro). ΔAcc is each method’s gain over the Seed prompt (49.1%). $\rho = \Delta\text{Acc}/\text{Total Opt. (pp/M tok)}$; $\rho_s = \Delta\text{Acc}/(\text{\$/cell})$ (pp/USD); list prices, exchange rate, and per-benchmark breakdowns in Appendix G.3.

channels, while every baseline overspends on at least one. Table 5 decomposes the cost drivers and reports the cost-effectiveness ratio ρ . SEPO reaches $\rho=4.37$ — $1.6\times$ the next-best MPO (2.65), $1.9\times$ GEPA (2.36), $2.1\times$ APSF (2.09; lowest total opt. but 6.9 pp weaker), and $5\text{--}8\times$ MIPROv2 (0.90) / OPRO (0.53). SEPO trims *both* channels at once: localised edits hold the architect side to 0.34M tokens per cell (10–20 $\times$ less than OPRO and MIPROv2), and a 236-token prompt holds worker optimisation to 2.59M (vs. GEPA 3.88M, MPO 3.40M) and deployment to 203k per cell (half of GEPA’s 424k). Each baseline overspends on one channel — OPRO/MIPROv2 on the architect, GEPA/MPO on worker length. At list prices, SEPO costs \$0.17/cell ($\rho_{\$}=75.3$ pp/USD) — $\sim 10\times$ cheaper than OPRO (\$1.66), and within \$0.02 of MPO/GEPA (\$0.15/\$0.18) while +3.1 pp more accurate.

5 Conclusion

We presented SEPO and its core mechanism, *edit-effect lineage feedback*, which links each local edit’s target and realised operations to the examples it newly fixes or breaks and reuses this record in later architect calls on the same search branch. A two-layer prompt schema provides a stable, addressable implementation. Across 14 tasks and two workers, SEPO beats six API-only baselines (+3.1/ +2.2 pp over the strongest, GEPA) and lies on both the optimisation- and test-time Pareto frontiers, yielding a stronger accuracy–cost trade-off.

Limitations

Instruction-space scalability. Our evaluation covers prompt lengths and instruction spaces representative of existing prompt-optimization work. Whether SEPO’s efficiency advantage extends to substantially longer prompts or more compositional agentic skills involving multiple steps, tool calls, or state tracking remains open. **Task-tailored schemas.** By design, SEPO tailors a schema per task, yielding a task-specialised artefact rather than a universal prompt; cross-task transfer is left to future work. **Evaluation setting.** We evaluate SEPO only on single-turn, single-call, text-based tasks; its structured, attributable edits may suit broader domains—agentic pipelines and multi-modal prompts—which we leave to future work.

Acknowledgments

This work is supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2025A1515012968, in part by the Shenzhen Science and Technology Program under Grant No. JCYJ20240813113502004, in part by the National Natural Science Foundation of China under Grant No. 62001412, in part by Shenzhen Stability Science Program 2023, in part by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), and in part by the Shenzhen Key Lab of Crowd Intelligence Empowered Low-Carbon Energy Network (Grant No. ZDSYS20220606100601002).

References

- Eshaan Agarwal, Raghav Magazine, Joykirat Singh, Vivek Dani, Tanuja Ganu, and Akshay Nambi. 2025. [PromptWizard: Optimizing prompts via task-aware, feedback-driven self-evolution](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19974–20003.
- Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziemis, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alex Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. 2026. [GEPA: Reflective prompt evolution can outperform reinforcement learning](#). In *International Conference on Learning Representations*, volume 2026, pages 8479–8565.
- Dongyu Chen, Jian Ma, Xianpeng Zhang, Lei Zhang, Haonan Lu, Chen Chen, Chuangchuang Wang, and Kai Tang. 2026. [Learning from prompt itself: the hierarchical attribution prompt optimization](#). *Preprint*, arXiv:2601.02683.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. [Evaluating large language models trained on code](#). *Preprint*, arXiv:2107.03374.
- Yongchao Chen, Jacob Arkin, Yilun Hao, Yang Zhang, Nicholas Roy, and Chuchu Fan. 2024. [PRompt optimization in multi-step tasks \(PROMST\): Integrating human feedback and heuristic-based sampling](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3859–3920, Miami, Florida, USA. Association for Computational Linguistics.

- Pavlos S. Efrimidis and Paul G. Spirakis. 2006. [Weighted random sampling with a reservoir](#). *Information Processing Letters*, 97(5):181–185.
- Chrisantha Fernando, Dylan Sunil Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2024. [Promptbreeder: Self-referential self-improvement via prompt evolution](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 13481–13544. PMLR.
- Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. [PAL: Program-aided language models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799. PMLR.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2024. [Connecting large language models with evolutionary algorithms yields powerful prompt optimizers](#). In *International Conference on Learning Representations*, volume 2024, pages 34133–34156.
- Han He, Qianchu Liu, Lei Xu, Chaitanya Shivade, Yi Zhang, Sundararajan Srinivasan, and Katrin Kirchhoff. 2025. [CriSPO: Multi-aspect critique-suggestion-guided automatic prompt optimization for text generation](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24014–24022.
- Yash Jain and Vishal Chowdhary. 2025. [Local prompt optimization](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 75–81, Albuquerque, New Mexico. Association for Computational Linguistics.
- Gurusha Juneja, Gautam Jajoo, Hua Li, Jian Jiao, Nagarajan Natarajan, and Amit Sharma. 2025. [Task facet learning: A structured approach to prompt optimization](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23473–23496, Vienna, Austria. Association for Computational Linguistics.
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan A, Saiful Haq, Ashutosh Sharma, Thomas Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. [DSPy: Compiling declarative language model calls into state-of-the-art pipelines](#). In *International Conference on Learning Representations*, volume 2024, pages 54928–54958.
- Shanu Kumar, Akhila Yesantarao Venkata, Shubhanshu Khandelwal, Bishal Santra, Parag Agrawal, and Manish Gupta. 2025. [SCULPT: Systematic tuning of long prompts](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14996–15029, Vienna, Austria. Association for Computational Linguistics.
- Haoyue Liu, Xiaoyu Ma, Ye Chen, Shuguang Cui, and Xiaoying Tang. 2026a. [One rewrite to fix them all? type-aware repair allocation for text-to-image prompt optimization](#). *Preprint*, arXiv:2607.18724.
- Haoyue Liu, Zhichao Wang, Yongxin Guo, Hao-ran Shou, and Xiaoying Tang. 2026b. [Adaptive prompt structure factorization: A framework for self-discovering and optimizing compositional prompt programs](#). *Preprint*, arXiv:2604.06699.
- Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. [Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 21558–21572. Curran Associates, Inc.
- Xiaoyu Ma, Yiwen Li, Haoyue Liu, Zhichao Wang, Ye Chen, Yongxin Guo, and Xiaoying Tang. 2026. [Select smarter, not more: Prompt-aware evaluation scheduling with submodular guarantees](#). *Preprint*, arXiv:2604.11328.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-Refine: Iterative refinement with self-feedback](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.
- Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. [Optimizing instructions and demonstrations for multi-stage language model programs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9340–9366, Miami, Florida, USA. Association for Computational Linguistics.
- Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. 2023. [GrIPS: Gradient-free, edit-based instruction search for prompting large language models](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3845–3864, Dubrovnik, Croatia. Association for Computational Linguistics.
- Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. 2023. [Automatic prompt optimization with “gradient descent” and beam search](#).

- In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968, Singapore. Association for Computational Linguistics.
- Qwen Team. 2026. [Qwen3.5: Towards native multi-modal agents](#).
- Kiran Ramnath, Kang Zhou, Sheng Guan, Soumya Smruti Mishra, Xuan Qi, Zhengyuan Shen, Shuai Wang, Sangmin Woo, Sullam Jeoung, Yawei Wang, Haozhu Wang, Han Ding, Yuzhe Lu, Zhichao Xu, Yun Zhou, Balasubramaniam Srinivasan, Qiaojing Yan, Yueyan Chen, Haibo Ding, and 2 others. 2025. [A systematic survey of automatic prompt optimization techniques](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33078–33110.
- Tobias Schnabel and Jennifer Neville. 2024. [Symbolic prompt program search: A structure-aware approach to efficient compile-time prompt optimization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 670–686, Miami, Florida, USA. Association for Computational Linguistics.
- Prith Sharma and Austin Z. Henley. 2026. [Modular prompt optimization: Optimizing structured prompts with section-local textual gradients](#). *Preprint*, arXiv:2601.04055.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. [Reflexion: language agents with verbal reinforcement learning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc.
- Lee Spector. 2012. [Assessment of problem modality by differential performance of lexibase selection in genetic programming: a preliminary report](#). In *Proceedings of the 14th annual conference companion on Genetic and evolutionary computation*, pages 401–408.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. [Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models](#). *Transactions on Machine Learning Research*.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. [Challenging BIG-Bench tasks and whether chain-of-thought can solve them](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.
- Haoyu Wang, Cheng Feng, Liuyang Bian, Ruiyang Huang, Lei Wei, Yafei Wen, Xiaoxin Chen, and Xiaoying Tang. 2026. [Plans you can check: Verifier-grounded learning of an open-weight planner for executable video-editing](#). *Preprint*, arXiv:2608.25622.
- Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric Xing, and Zhiting Hu. 2024a. [PromptAgent: Strategic planning with language models enables expert-level prompt optimization](#). In *International Conference on Learning Representations*, volume 2024, pages 23967–24001.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. 2024b. [MMLU-Pro: A more robust and challenging multi-task language understanding benchmark](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 95266–95290. Curran Associates, Inc.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Yihan Xia, Taotao Wang, Shengli Zhang, Zhangyuhua Weng, Bin Cao, and Soung Chang Liew. 2026. [HiveMind: Contribution-guided online prompt optimization of LLM multi-agent systems](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29767–29774.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024. [Large language models as optimizers](#). In *International Conference on Learning Representations*, volume 2024, pages 12028–12068.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.
- Qinyuan Ye, Mohamed Ahmed, Reid Pryzant, and Fereshte Khani. 2024. [Prompt engineering a prompt engineer](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 355–385,

Bangkok, Thailand. Association for Computational Linguistics.

Seungyoun Yi, Minsoo Khang, and Sungrae Park. 2025. [ZERA: Zero-init instruction evolving refinement agent – from zero instructions to structured prompts via principle-based optimization](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23323–23337, Suzhou, China. Association for Computational Linguistics.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Pan Lu, Zhi Huang, Carlos Guestrin, and James Zou. 2025. [Optimizing generative AI by backpropagating language model feedback](#). *Nature*, 639(8055):609–616.

Guibin Zhang, Junhao Wang, Junjie Chen, Wangchunshu Zhou, Kun Wang, and Shuicheng YAN. 2026. [AgentTracer: Who is inducing failure in the LLM agentic systems?](#) In *The Fourteenth International Conference on Learning Representations*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. [Large language models are human-level prompt engineers](#). In *The Eleventh International Conference on Learning Representations*.

A Reproducibility Statement

This appendix consolidates the information needed to reproduce every number in §4. Sections A.1–A.3 cover data, models, compute, and reporting conventions; Sections B–C cover the SEPO loop and architect prompts; Section F covers each baseline; the remaining appendices report extended quantitative results, ablations, and qualitative analysis.

A.1 Datasets, Splits, and Licences

Table 6 summarises the 14 held-out tasks used in §4, drawn from five public benchmarks: **BBH** (8 reasoning sub-tasks), **MMLU-Pro** (3 subject sub-tasks), **MBPP+**, **GSM-Hard**, and **HotpotQA**. For each task we report the source benchmark, the number of optimisation-time training examples ($|\mathcal{D}_{\text{tr}}|$), the size of the held-out test split, the task metric, and the source licence. All splits are fixed across optimisers and seeds; the same indices are used for every method. No example in $\mathcal{D}_{\text{tr}}$ overlaps with the held-out test split.

Evidence-packet probes. Every task uses a fixed training set $\mathcal{D}_{\text{tr}}$ and a disjoint held-out test split. Evidence packets, including the M_{train} slot in Appendix D.1, are constructed from this training set after the focus and anchor slots have been selected. This convention is used uniformly across BBH, MMLU-Pro, MBPP+, HotpotQA, and GSM-Hard (Table 6).

All datasets are used in accordance with their licences and intended research-evaluation purpose. None contains personally identifying information; HotpotQA passages are drawn from Wikipedia under the CC BY-SA licence.

A.2 Models and Decoding Configuration

We keep the evaluation harness fixed across optimisers and random seeds. The frozen worker LLMs are **Llama-3.1-8B-Instruct** (Meta, 8.0B parameters) and **Qwen3-8B** (Alibaba, 7.6B parameters); each worker is used for both optimisation-time scoring and held-out deployment. Worker inference uses temperature 0.0, top- p 1.0, and a maximum generation length of 2,048 tokens, so that metric calls are deterministic under a fixed prompt and example.

SEPO uses **Qwen3.5-397B-A17B** (a Mixture-of-Experts model with 17B active parameters out of a 397B total) as the architect LLM for the attribution and patch calls, with temperature 0.7 and a

Task	$ \mathcal{D}_{\text{tr}} $	Test	Metric	Licence
BBH Causal Judgement	50	137	exact match	MIT
BBH Date Understanding	50	200	exact match	MIT
BBH Disambiguation QA	50	200	exact match	MIT
BBH Formal Fallacies	50	200	exact match	MIT
BBH Logical Deduction (7)	50	200	exact match	MIT
BBH Colored Objects	50	200	exact match	MIT
BBH Snarks	50	128	exact match	MIT
BBH Word Sorting	50	200	exact match	MIT
MMLU-Pro CS	100	200	multi-choice	MIT
MMLU-Pro Economics	100	200	multi-choice	MIT
MMLU-Pro Psychology	100	200	multi-choice	MIT
MBPP+	100	228	pass@1	Apache-2.0
GSM-Hard	200	400	exact match	MIT
HotpotQA [†]	200	400	EM (string)	CC BY-SA

Table 6: Dataset statistics and licences. All BBH and MMLU-Pro splits follow each benchmark’s original release. Every task uses the same evidence-packet convention described above. [†]**HotpotQA split swap.** We use the *Medium* split under the Llama worker and the *Hard* split under the Qwen worker. The two splits share the same distractor-setting passages but differ in extraction difficulty: the Hard split requires the worker to resolve a distractor entity before extracting the answer, while the Medium split has cleaner gold-passage alignment. Under the Llama worker, limited answer accessibility on Hard compresses performance across optimisers; a matched Hard-split sensitivity test preserves their ranking (Appendix G.2). We therefore use Medium under Llama and Hard under Qwen. The implication is that the HotpotQA column in Table 1 (column “HQA”) is not directly comparable between the Llama and Qwen blocks; the gap across blocks reflects both the worker change and the split swap.

maximum generation length of 16,000 tokens to encourage diverse candidate edits. Baselines are run through their corresponding API-only optimiser interfaces while sharing the same worker, data split, and metric-call budget; baseline-specific architect choices are reported in Appendix F.

A.3 Random Seeds and Reporting Conventions

For each (method, benchmark) cell we run three independent optimisation seeds; the deployment held-out evaluation reuses the optimiser’s selected prompt and is itself deterministic (temperature 0). Tables report the seed mean with the standard deviation as the subscript (e.g., $62.0_{\pm 2.6}$). All accuracies are reported in percent on the held-out test split, not the training split used for optimiser feedback.

A.4 Use of AI Assistants

We used an AI assistant solely for English proofreading and stylistic polishing. All technical content, experiments, and conclusions were produced

and verified by the authors.

B Prompts: Seed Schema and Architect Templates

This appendix lists the three text artefacts that ground every SEPO iteration: the cold-start seed schema s_0 , the attribution architect’s prompt template, and the patch architect’s prompt template. The schema parser at the end of the section is the deterministic counterpart that converts the architect’s JSON output back into a two-layer schema.

B.1 Seed Schema and Initialisation

The cold-start seed schema s_0 is *task-agnostic*: every benchmark in our suite starts from the same generic three-step backbone with no step-local rules. This puts every optimiser on an equal footing — no task-specific priors are injected into the seed — and means the first iteration’s NEWLY_BROKEN is empty by construction, so the loop enters the no-regression fallback (Appendix D.2) on iteration 1. The exact seed used in all runs is shown below.

Seed schema s_0 (task-agnostic)

1. Understand the problem – read carefully, parse all entities, and decompose into sub-questions if needed.
2. Reason step by step.
3. Output the final answer on the last line as: Answer: (X).

The seed has no step-local rules: all sub-rules in the converged schema (e.g. the BBH Word Sorting example in §3.2) are introduced by the patch architect during optimisation; no rule is copied from $\mathcal{D}_{tr}$ or any benchmark-specific prompt template.

B.2 Attribution Prompt Template

SEPO’s first architect call returns a typed target τ and a failure-pattern rationale; it does not rewrite the schema. The prompt is identical across benchmarks — only the schema text, the evidence packet, and the breadcrumb vary per call. Below is the full block structure of the attribution prompt; $\{\}$ -braced fields are populated at render time.

Attribution architect prompt

Your task: diagnose why the current schema is failing, by reading parent’s behavior on the train cases shown below. Decide which single schema component to fix, and explain why in your own words.

The schema is the worker’s system prompt across MANY cases of this task family – not just the few cases shown here. Your fix

should generalize to unseen cases of this task family. Different evidence cases may fail for materially different reasons; if so, say that explicitly in your rationale instead of forcing them into a single narrative.

The schema has been refined across prior optimization rounds via discrete, traceable patches. The block below shows what the LAST patch did, so you can analyse its observed example-level effects.

```
{lineage breadcrumb  $\ell$  - op-specific before-state diff}
```

=== CURRENT SCHEMA ===

(Numbered lines '1.', '2.', '3.' are TOP-LEVEL STEPS. Indented '-' lines are SUB-RULES belonging to the most recent numbered step above them.)

```
{p}
```

=== END SCHEMA ===

=== Parent on train ===

{focus + anchor + train cases, each rendered as Train Case #i with transition tag, Input, Gold, Worker output}

=== ANALYSIS GUIDANCE ===

1. Read each evidence case yourself – input, gold, worker output . . .
2. If evidence cases span materially different failure subtypes, say so explicitly and pick the SINGLE dominant subtype to target.

=== TARGET PREFERENCE ===

Choose the smallest target that explains the dominant failure:

1. Use "step <N>" when the failure can be repaired inside step N.
2. Use explicit structure targets only when the top-level numbered step structure itself is the problem:
 - "ADD step at <pos>" - "SPLIT step <N>"
 - "DROP step <N>"
3. Use "backbone" only when the whole 1./2./3. macro structure is wrong.

=== OUTPUT ===

Return ONLY a JSON object:

```
{"target": "step <N>" | "ADD step at <pos>" | "SPLIT step <N>" | "DROP step <N>" | "backbone",
"rationale": "1-2 sentences, MAX 300 chars",
"supporting_observations": [2-4 short observations]}
```

B.3 Patch Prompt Template

The second architect call consumes the same evidence packet plus the attribution output $(\tau, \text{rationale})$ and emits the full revised schema $\tilde{s}$. We list the full block structure of the patch prompt below; the style block uses the balanced variant in every run of our launch scripts.

Patch architect prompt

Your task: produce ONE revised version of the schema below so the worker stops failing on the diagnosed pattern.

The schema is the worker’s system prompt across MANY cases of this task family; the cases below are a small sample it will NOT see again. Your patch must generalize.

A diagnosis has already been performed (see THIS ROUND’S DIAGNOSIS below). Use it together with the train evidence and the lineage breadcrumb to decide what to change. {lineage breadcrumb ℓ }

=== THIS ROUND’S DIAGNOSIS ===

target: { τ }

rationale: {rationale from attribution}

=== SCOPE RULE ===

For target=’step <N>’, target is the step at the locus of the dominant failure. Choose the smallest applicable operation below:

- edit an existing sub-rule under the target step
- drop a sub-rule under the target step
- merge two overlapping sub-rules under the target step into one
- add a new sub-rule under the target step
- rewrite the target step’s numbered top-level text

If target is explicitly structural, execute that structure edit:

- target=’ADD step at <pos>’: insert a new top-level numbered step at position <pos>.
- target=’SPLIT step <N>’: split step N into two consecutive top-level steps.
- target=’DROP step <N>’: remove step N as a top-level step.
- target=’backbone’: rewrite the global top-level structure.

Sub-rules under non-target steps are preserved unless they become irrelevant after the change.

=== CURRENT SCHEMA ===

{ p }

=== END SCHEMA ===

=== Parent on train ===

{focus + anchor + train cases with transition tags}

=== STYLE GUIDANCE ===

From the worker responses and gold answers above, identify the niche, task-specific procedures, constraints, and edge cases the worker is getting wrong, and encode each as a directly executable rule with a worked example. Avoid jargon; every rule must be directly executable.

=== OUTPUT SCHEMA FORMAT ===

Schema: numbered points (1./2./3./...) with “-” indented elaborations. Step text can be a short imperative OR a longer paragraph with persona/role context.

Example output for target=’step 2’ (adding two task-specific sub-rules under step 2):

1. Understand the problem.
2. Reason step by step.

- When the option uses positive words to describe an objectively bad situation, classify as sarcasm.
- Compare literal meaning against real-world facts; if mismatched, the option is sarcastic.

3. Output the answer on the last line as: Answer: (X).

In this example the backbone (1./2./3.) is unchanged; only elaborations were added under step 2.

=== OUTPUT FORMAT ===

Return ONLY a JSON object:

```
{"schema": "<complete outline-tree markdown>",  
  "rationale": "one or two sentences explaining what changed and why"}
```

B.4 Schema Parser

A deterministic parser extracts the outline step list and per-step rule lists from the patch architect’s JSON output. Across the attribution and patch stages, our runs made 2,520 architect calls in total. Of these, 58 patch calls (2.3% of all architect calls) returned missing JSON or schema text that did not match the “<N>.” regex. The corresponding children were discarded with the audit tuple recorded as parse_failure, and all consumed tokens were included in the reported optimisation cost.

C Algorithm Details

We give two views of SEPO’s per-iteration loop. Algorithm 1 is the slim, audit-tuple-first description of one iteration, with explicit cross-references to attribution (Eq. 4), patch (Eq. 5), operator inference (Eq. 3), and example-level admission (Eq. 6). Algorithm 2 is the engineering view that spells out the helpers (parent selection, evidence-packet assembly, breadcrumb rendering, subsample precheck, archive capacity). The two algorithms describe the same loop.

Metric-call accounting. The focus and anchor slots provide cached parent outputs as architect evidence; Gate 1 evaluates the new child only on the n_{tr} examples in M_{train} . A candidate rejected at Gate 1 therefore incurs n_{tr} new metric calls. For a candidate that passes, these n_{tr} child results are reused and only the remaining $|\mathcal{D}_{tr}| - n_{tr}$ examples are evaluated, for $|\mathcal{D}_{tr}|$ new metric calls in total. Below, COMPLETEEVAL evaluates only the uncached training examples and then assembles the full score vector with the cached Gate 1 results.

Algorithm 1: SEPO: one iteration of lineage-aware structural editing.

Input : seed schema s_0 ; training set $\mathcal{D}_{tr}$; worker W ;
architects A_{attr} , A_{patch} ; probe size n_{tr} ;
metric-call budget B .

Output : archive $\mathcal{A}$.

```

1  $c(s_0) \leftarrow \text{EVAL}(W, s_0, \mathcal{D}_{tr})$ ;  $\mathcal{A} \leftarrow \{s_0\}$ ;
   $u \leftarrow |\mathcal{D}_{tr}|$ 
2 while  $u + |\mathcal{D}_{tr}| \leq B$  do
3    $p \leftarrow \text{SELECTPARENT}(\mathcal{A})$ ;  $g \leftarrow \pi(p)$ 
  // Lexicase w.p. 0.8; uniform w.p. 0.2
4   if  $g = \emptyset$  then
5      $\text{NB}, \text{NF} \leftarrow \emptyset, \emptyset$  // cold start
6   else
7      $\text{NB} \leftarrow \{i : c(g)_i=1, c(p)_i=0\}$ ;
   $\text{NF} \leftarrow \{i : c(g)_i=0, c(p)_i=1\}$ 
8    $M \leftarrow (M_{\text{focus}}, M_{\text{anchor}}, M_{\text{train}})$  // sampled
  from  $(\text{nb}, \text{nf}, \mathcal{D}_{tr})$ ; fallback if  $\text{nb}=\emptyset$ 
9    $\ell \leftarrow \text{BREADCRUMB}(o(p), \tau(p), +|\text{NF}|, -|\text{NB}|)$ 
10   $(\tau, r) \leftarrow A_{attr}(p, M, \ell)$  // Eq. 4
11   $\tilde{s} \leftarrow A_{patch}(p, \tau, r, M, \ell)$  // Eq. 5
12   $c \leftarrow \text{PARSESHEMA}(\tilde{s})$ ; if  $c = \perp$  then
  continue // parse failure
13   $\pi(c) \leftarrow p$ ;  $\tau(c) \leftarrow \tau$ ;
   $o(c) \leftarrow \text{INFEROP}(\text{DIFF}(p, c), \tau(c))$ 
  // Eq. 3
14   $c_{M_{\text{train}}}(c) \leftarrow \text{EVAL}(W, c, M_{\text{train}})$ 
15  if  $\sum_{i \in M_{\text{train}}} c_{M_{\text{train}}}(c)_i \leq \sum_{i \in M_{\text{train}}} c(p)_i$ 
  then  $u + = n_{tr}$ ; continue // Gate 1:
  precheck
16   $c(c) \leftarrow$ 
   $\text{COMPLETEEVAL}(W, c, \mathcal{D}_{tr}; c_{M_{\text{train}}}(c))$ 
  // reuse Gate 1 results
17   $u + = |\mathcal{D}_{tr}|$ 
18  if Eq. 6 holds then  $\mathcal{A} \leftarrow \mathcal{A} \cup \{c\}$  // Gate 2:
  admit
19 return  $\mathcal{A}$ 

```

C.1 Expanded Engineering View

Where each module appears. The five-module decomposition of §3 maps onto Algorithm 2 as follows. Parent selection (§3.4) is the LEXICASEORUNIFORM call; the counterfactual minibatch (§3.3 Step 1) is BUILDMINIBATCH, which contains the no-regression fallback when NEWLY_BROKEN(p) is empty; the lineage breadcrumb (§3.3 Step 2) is RENDERBREADCRUMB; the attribution and patch architects (§3.3 Steps 3–4) are the A_{attr} and A_{patch} calls; the two-layer schema (§3.2) is materialised by PARSESHEMA and the assignment of $\pi(c)$, $\tau(c)$ on the next line; operator inference (§3.2) is the INFEROP call; the subsample precheck and example-level admission (§3.4) are SUBSAMPLEPRECHECK and ADMIT; archive capacity is MANAGECAPACITY.

C.2 Operator Label Space

The schema of Eq. 2 is two-layer (a top-level *outline* plus step-local *rules*), so the operator vocabulary $\mathcal{O}$ in Eq. 3 is naturally indexed by which layer it edits. Table 7 lists the eight sub-operations in $\mathcal{O} = \mathcal{O}_{\text{rule}} \cup \mathcal{O}_{\text{outline}}$ together with the architect’s target form τ that elicits each one (the architect chooses “the smallest applicable operation” from these alternatives; see Appendix B.3). $\mathcal{O}_{\text{rule}}$ contains four rule-layer sub-ops that act on the step-local rule list, leaving the top-level outline unchanged; $\mathcal{O}_{\text{outline}}$ contains four outline-layer sub-ops that touch the top-level outline. Each sub-op is its own $o(c)$ label, so the post-hoc operator takes eight distinct values; these are the eight labels counted in Table 4 of the main paper.

The architect’s *target* τ has a slightly richer surface vocabulary (step <N>, ADD step at <pos>, SPLIT step <N>, DROP step <N>, backbone; see Appendix B.2) than $o(c)$ because τ pre-locates the edit before the patch is written. The realised $o(c)$ is recovered post-hoc from the structural diff (Alg. 3); the DROP step <N> target was never

Algorithm 2: SEPO expanded loop with helper subroutines.

Input: seed s_0 ;;
training set $\mathcal{D}_{tr}$;;
 W, A ;;
budget B ;;
iter cap T ;;
archive cap C ;;
(n_f, n_a, n_{tr}).
Output: archive $\mathcal{A}$.

```

1  $c(s_0) \leftarrow \text{EVAL}(W, s_0, \mathcal{D}_{tr})$ ;;
2  $\mathcal{A} \leftarrow \{s_0\}$ ;;
3  $u \leftarrow |\mathcal{D}_{tr}|$ ;;
4 for  $t = 1, \dots, T$  do
5   if  $u + |\mathcal{D}_{tr}| > B$  then break;
6    $p \leftarrow \text{LEXICASEORUNIFORM}(\mathcal{A})$ ;
  // Lexicase w.p. 0.8; uniform w.p. 0.2
7    $M \leftarrow \text{BUILDMINIBATCH}(p, \mathcal{A}, \mathcal{D}_{tr})$ ;
  // fallback when  $\text{nb}(p)=\emptyset$ 
8    $\ell \leftarrow \text{RENDERBREADCRUMB}(p, \pi(p))$ ;
9    $(\tau, r) \leftarrow A_{attr}(p, M, \ell)$ ; // Eq. 4
10   $\tilde{s} \leftarrow A_{patch}(p, \tau, r, M, \ell)$ ; // Eq. 5
11  if  $\text{PARSESHEMA}(\tilde{s})$  fails then continue;
12   $c \leftarrow$  schema with  $\pi(c)=p, \tau(c)=\tau$ ;
13   $o(c) \leftarrow \text{INFEROP}(\text{DIFF}(p, c), \tau(c))$ ; // Eq. 3
14  if  $\text{SUBSAMPLEPRECHECK}(p, c, M_{\text{train}})$  rejects
  then  $u + = n_{tr}$ ; continue;
15   $c(c) \leftarrow \text{COMPLETEEVAL}(W, c, \mathcal{D}_{tr})$ ;
  // reuse Gate 1 results
16   $u + = |\mathcal{D}_{tr}|$ ;
17   $\text{ADMIT}(c, p, \mathcal{A})$ ; // Gate 2, Eq. 6
18  if  $|\mathcal{A}| > C$  then  $\text{MANAGECAPACITY}(\mathcal{A}; C)$ ;
19 return  $\mathcal{A}$ 

```

Sub-op	Target τ	Effect on schema
$\mathcal{O}_{\text{rule}}$ (rule-layer): 4 sub-ops, all under step $\langle N \rangle$		
ADDRULE	step $\langle N \rangle$	insert sub-rule in step N
DROPRULE	step $\langle N \rangle$	remove sub-rule from step N
REFINERULE	step $\langle N \rangle$	rewrite one sub-rule in step N
MERGERULES	step $\langle N \rangle$	merge two sub-rules in step N
$\mathcal{O}_{\text{outline}}$ (outline-layer): 4 sub-ops, one $o(c)$ label each		
REWRITESTEP	step $\langle N \rangle$	rewrite step N 's top-level text
ADDSTEP	ADD step at $\langle \text{pos} \rangle$	insert new top-level step at $\langle \text{pos} \rangle$
SPLITSTEP	SPLIT step $\langle N \rangle$	split step N into two steps
BACKBONE	backbone	rewrite entire 1./2./3. structure

Table 7: Operator vocabulary $\mathcal{O}$ decomposed by schema layer. Each row is a sub-operation together with the architect’s target form τ that elicits it; each row also maps to its own post-hoc $o(c)$ label. All four rule-layer sub-ops live inside one step’s rule list. All four outline-layer sub-ops change the top-level outline structure or text and are distinguished post-hoc from the diff together with the typed target τ (Alg. 3); the architect never declares the operator label itself. The architect’s DROP step $\langle N \rangle$ target form is supported in the prompt template (Appendix B.2) but yielded no admitted children in our 14-task $\times$ 3-seed runs, so it does not appear as a separate $o(c)$ label here. The verbatim patch-prompt phrase eliciting each sub-op is given in Appendix B.3 (SCOPE RULE block).

admitted in our runs and therefore has no corresponding $o(c)$ label in Table 7.

Why DIFF is non-trivial. The patch architect returns a free-form schema text $\tilde{s}$; it does not annotate which child rule corresponds to which parent rule, nor does it self-declare an edit operator. Recovering $o(c) \in \mathcal{O}$ therefore reduces to first computing a *structural* parent–child alignment, after which the operator label follows by counting non-empty buckets. Because rules are short natural-language strings, we align by sequence similarity rather than by edit distance over tokens, and we use two thresholds to separate three cases that the patch prompt actually distinguishes (Appendix B.3, SCOPE RULE): a refinement of one parent rule, a merge of two parent rules into one, and the introduction of a new rule.

Algorithm 3 computes the diff at two levels. It first aligns the outline steps in order, recording added, dropped, refined, and copied step locations. Within each aligned step pair, each child rule is greedily matched against the top-two unused parent rules from that same step under the similarity σ . The three branches (REFINED, MERGED, ADDED) are gated by the two thresholds θ_R, θ_M together with an exact-match shortcut (COPIED) and

Algorithm 3: DIFF(p, c) — deterministic structural diff between a parent and child schema. Used by INFEROP (Eq. 3) and the breadcrumb renderer (Appendix D.3).

Input : parent schema $p = (\alpha_p, \{\rho_{p,i}\}_{i=1}^{m_p})$ and child schema $c = (\alpha_c, \{\rho_{c,j}\}_{j=1}^{m_c})$, where $R_{p,k} = \{\rho_{p,k,i}\}$ and $R_{c,k'} = \{\rho_{c,k',j}\}$ denote the rule lists attached to parsed steps k and k' ; similarity $\sigma(s_1, s_2)$ (Python difflib.SequenceMatcher ratio, $\in [0, 1]$); thresholds $\theta_R = 0.55$ (refine), $\theta_M = 0.40$ (merge).

Output : hierarchical diff $\mathcal{D} = (\mathcal{D}_{\text{outline}}, \mathcal{D}_{\text{rule}})$.

```

1  $H, \mathcal{D}_{\text{outline}} \leftarrow \text{ALIGNOUTLINE}(\alpha_p, \alpha_c; \sigma)$ 
  // order-preserving step alignment
2 ADDED, DROPPED, REFINED, MERGED, COPIED  $\leftarrow \emptyset$ 
3 foreach aligned step pair  $(k, k') \in H$  do
4   USED $[k, i] \leftarrow \text{FALSE}$  for  $i = 1, \dots, |R_{p,k}|$ 
5   for  $j = 1, \dots, |R_{c,k'}|$  do
6      $S \leftarrow$  list of  $(i, \sigma(\rho_{p,k,i}, \rho_{c,k',j}))$  for
7        $\neg \text{USED}[k, i]$ , sorted by  $\sigma$  descending
8     if  $S = \emptyset$  or  $S_1.\sigma < \theta_M$  then
9       ADDED  $+= \{(k', j)\}$ ; continue
10     $(i^*, s^*) \leftarrow S_1$ 
11    if  $\text{STRIP}(\rho_{p,k,i^*}) = \text{STRIP}(\rho_{c,k',j})$  then
12      COPIED  $+= \{(k, i^*, k', j)\}$ ;
13      USED $[k, i^*] \leftarrow \text{TRUE}$ ; continue
14    if  $|S| \geq 2 \wedge S_2.\sigma \geq \theta_M \wedge s^* < 0.95 \wedge$ 
15      LOOKSLIKEMERGE $(\rho_{p,k,i^*}, \rho_{p,k,S_2.i}, \rho_{c,k',j})$ 
16      then
17        MERGED  $+= \{(k, i^*, S_2.i, k', j)\}$ ;
18        USED $[k, i^*], \text{USED}[k, S_2.i] \leftarrow$ 
19          TRUE, TRUE
20      else if  $s^* \geq \theta_R$  then
21        REFINED  $+= \{(k, i^*, k', j)\}$ ;
22        USED $[k, i^*] \leftarrow \text{TRUE}$ 
23      else ADDED  $+= \{(k', j)\}$ 
24    DROPPED  $+= \{(k, i) : \neg \text{USED}[k, i]\}$ 
25  $\mathcal{D}_{\text{rule}} \leftarrow (\text{ADDED}, \text{DROPPED},$ 
26   REFINED, MERGED, COPIED)
27  $\mathcal{D} \leftarrow (\mathcal{D}_{\text{outline}}, \mathcal{D}_{\text{rule}})$ 
28 return  $\mathcal{D}$ 
```

a merge-evidence predicate LOOKSLIKEMERGE defined below. Any parent rule to which no child rule was assigned is reported as DROPPED. The resulting hierarchical diff record, together with the typed target τ , is the input to INFEROP; its per-step outline locations are also used for breadcrumb rendering (Appendix D.3). Here ALIGNOUTLINE computes a maximum-similarity monotone one-to-one alignment of outline steps under σ (ties are resolved by the lower parent index); unmatched child and parent steps are marked as added and dropped, while matched steps are copied if their stripped texts are identical and refined otherwise.

Merge-evidence

LOOKSLIKEMERGE(a, b, c) checks that the child string c inherits distinctive vocabulary from *both* candidate parents, not just from one. Let $W(s)$ denote the set of lower-cased whitespace-separated tokens of length at least four in string s (the length filter removes function words), and define the “unique-from- x tokens that appear in c ” shorthand

$$U_{x|y}(c) \triangleq (W(x) \setminus W(y)) \cap W(c).$$

LOOKSLIKEMERGE is then

$$\text{LOOKSLIKEMERGE}(a, b, c) \triangleq [U_{a|b}(c) \neq \emptyset] \wedge [U_{b|a}(c) \neq \emptyset].$$

The companion guard $s^* < 0.95$ on the best similarity inside Alg. 3 prevents the merge branch from firing when one parent rule already covers nearly the entire child — in that case the edit is more naturally a refinement of that parent with auxiliary phrasing borrowed from a sibling.

InferOp classification rule. With the hierarchical diff record $\mathcal{D}$ and typed target τ in hand, **INFEROP** emits one label per changed location, so $o(c)$ is a multiset over $\mathcal{O}$ rather than a single dominant label. This is why a single edit in Table 4 can carry several labels (e.g. **REFINERULE**+**ADDRULE** when the patch refines one sub-rule and adds another) and the per-op shares sum above 100%. **INFEROP** classifies each location independently by the following rules:

1. Each *merged rule alignment*—two parent rules replaced by one child rule that shares distinctive vocabulary with both— contributes a **MERGERULES** label.
2. Each remaining added rule contributes an **ADDRULE** label; each refined rule contributes a **REFINERULE** label; each dropped rule contributes a **DROPRULE** label.
3. Outline-layer changes are classified jointly with the typed target τ , each contributing one label: $\tau=\text{ADD}$ step at $\langle \text{pos} \rangle$ with $|\alpha_c|=|\alpha_p|+1 \rightarrow \text{ADDSTEP}$; $\tau=\text{SPLIT}$ step $\langle N \rangle$ aligning one parent step to two adjacent child steps $\rightarrow \text{SPLITSTEP}$; $\tau=\text{step}$ $\langle N \rangle$ with exactly one step’s top-level text replaced and step count preserved $\rightarrow \text{REWRITESTEP}$; otherwise (e.g. $\tau=\text{backbone}$ or the diff spans multiple steps) $\rightarrow \text{BACKBONE}$.

The procedure is deterministic given the parent-child schema pair and typed target τ ; the architect never self-declares any of these operator labels.

predicate.

D Evidence Packet Construction

This appendix complements §3.3 (*Balanced evidence packet*) by specifying the slot sizes, weights, and breadcrumb-rendering template.

D.1 Slot Sizes and Weights

The minibatch $M = (M_{\text{focus}}, M_{\text{anchor}}, M_{\text{train}})$ uses three slots with sizes $n_f=2, n_a=1, n_{tr}=3$, totalling up to six examples; in fallback iterations without newly broken examples, M_{focus} is empty. M_{focus} samples without replacement from **NEWLY_BROKEN**(p) with weights $w_i \propto a_i$ (Efremidis and Spirakis, 2006), where

$$a_i \triangleq |\mathcal{A}|^{-1} \sum_{s \in \mathcal{A}} \mathbf{c}(s)_i \in [0, 1]$$

is the archive agreement on example i . M_{anchor} samples the parent’s **NEWLY_FIXED** examples, weighted by $1 - a_i$ toward rare cases few archive members solve; when no **NEWLY_FIXED** case exists it falls back to specialist solves (correct in the parent but in none of its ancestors), then to any other case in $C(p)$, and M_{train} samples uniformly from training-set examples not already selected into M_{focus} or M_{anchor} . None of the three slots is recoverable from $\sum_i \mathbf{c}(p)_i$ alone, which motivates explicit edit-effect lineage conditioning rather than scalar-reward conditioning.

D.2 Fallback without Regressions

When **NEWLY_BROKEN**(p) is empty (cold start $g=\emptyset$, or the parent’s last edit caused no regression on $\mathcal{D}_{\text{tr}}$), M_{focus} is empty and we fall back to $M = (\emptyset, M_{\text{anchor}}, M_{\text{train}})$; the anchor and train slot sizes are unchanged ($n_a=1, n_{tr}=3$). The architect prompt is told there is no regression focus so that diagnoses target train probes rather than hallucinating a non-existent training regression.

D.3 Breadcrumb Rendering

The breadcrumb ℓ is rendered as a short multiline block titled `=== PARENT’S LAST PATCH ===`. It is *not* a single per-operator template: instead the renderer iterates over the grandparent $\rightarrow$ parent structural diff and emits one bullet section per non-empty diff bucket. A single breadcrumb can therefore carry multiple operator types if the last patch made several changes at once (e.g. an **ADDRULE** together with a **REFINERULE**). Each change is anchored to the parent schema by a content *snippet* (the first ~ 25 chars of the after-text, or a ****Bold**

Title** if present), so the architect can locate the change by substring search rather than by a numeric rule id, which keeps the CURRENT_SCHEMA block free of inline markers.

Section structure. The renderer composes the block from the following pieces, in order:

1. Title line `=== PARENT’S LAST PATCH ===`.
2. Optional Last rationale `→ . . .` line (truncated to 160 chars) carrying the previous round’s attribution rationale.
3. Outline-layer sections (only if the outline diff is non-empty): Rewritten outline step(s), New outline step(s), or a fallback Backbone text changed (no per-step anchors).
4. Rule-layer sections, one per non-empty diff bucket: Refined sub-rule(s), New sub-rule(s), Merged-into sub-rule(s), and a tally line Dropped <K> sub-rule(s) (content removed).
5. A final summary line Effect on train: `+<f> newly_fixed, -<b> newly_broken`.

Cold-start branch. When the parent is the seed schema (no prior patch to learn from), the renderer emits the fixed message (parent is the seed schema; no prior patch to learn from) under the title and skips all diff sections.

Breadcrumb example (REFINERULE + ADDRULE in the same edit)

```
=== PARENT’S LAST PATCH ===
Last rationale → Worker conflated sarcastic
and literal options on step 2; adding an
explicit sentiment-mismatch check.
Refined sub-rule(s):
- in step 2 → When the option uses
positive. . .
New sub-rule(s):
- in step 2 → Compare literal meaning
agai. . .
Effect on train: +2 newly_fixed, -1
newly_broken
```

Both architect calls (attribution and patch) consume this same rendered block as the {lineage_block} placeholder in Appendix B.2 and B.3. The —BREADCRUMB ablation (§4.3) replaces the entire block with a single line that strips everything except the Effect on train summary; the —LINEAGE ablation removes the block entirely and uniformises M_{focus} over the parent’s current errors.

E Qualitative Analysis

This appendix illustrates SEPO’s schemas across tasks (§E.1) and then grounds the main paper’s auditability claims in the actual optimisation log of one run (§E.2–§E.4; BBH Word Sorting, seed 45, Llama-3.1-8B-Instruct worker). All numbers and text are reconstructed from the single result.json each run writes by default; no extra instrumentation is required.

E.1 Cross-Task Schema Examples

The schemas SEPO converges to look different across tasks because their failure modes differ. Together with the BBH Word Sorting schema in §3.2, the two examples below show that the two-layer schema architecture supports task-tailored rule distributions without imposing one: on MBPP+, rules concentrate on extraction (function-name fidelity, type parsing, constraint mapping) and implementation; on HotpotQA Medium, rules concentrate on grounding and entity verification. Rule text is taken verbatim from SEPO’s final prompts (light line-wrapping only).

Example schema (MBPP+, seed 45, test acc. 63%)

Step 1. Understand the problem and extract constraints.

- Identify the required function name from the problem description or assert statement (e.g., `assert func_name(. . .)`) and ensure your solution defines a function with this exact name.
- Parse the input/output types and specific logic rules (e.g., “remove characters at odd indices”, “sum of digits equal to length”) before planning.
- Explicitly map the problem’s specific logical constraints to a concrete algorithm before coding (e.g., for a magic square: verify all rows, columns, and diagonals sum to the same value; for character replacement: iterate through the string and swap specific target characters).

Step 2. Plan and implement the solution in code.

- Write a complete Python function that implements the parsed logic exactly as described, ignoring any unrelated patterns from training data.
- Include the function definition, the logic body, and a test call using the provided assert statement to verify correctness.

Step 3. Output the final answer.

- If the task asks for a function, output the full Python code block containing the function and the test.
- Output the final result on the last line as “Answer: (X)”, where X is the return value of the function or the code itself if requested.

Example schema (HotpotQA Medium, seed 45, test acc. 68%)

Step 1. Parse the question and locate the relevant context.

- Identify the key entities and constraints in the question (e.g., specific dates, titles, roles, locations), then scan the provided passages for sentences mentioning those exact entities.

Step 2. Verify, link, and answer.

- Strictly ground your answer in the provided text; do not use outside knowledge or hallucinate details not present in the context.
- Verify entity matches before extracting attributes: if the question asks about Entity A, do not extract facts about Entity B even if they appear in the same passage or share a category.
- For multi-hop questions, explicitly trace the chain across passages (e.g., Entity A in passage 1 to Property B in passage 2); once the answer is reached, stop reasoning immediately and do not repeat the same logical loop.

Step 3. Output the final answer.

- The final line must be exactly “Answer: (X)” where X is the concise value with no introductory phrases, explanations, or repetitions of the question.

E.2 Worked Iteration Trace: Discovering an Anti-Verification Rule

The log records a two-iteration sequence in which SEPO *adds* a self-verification rule, observes it triggering worker repetition loops, and *removes* it—replacing it with the opposite instruction. The final schema is the winner shown in Figure 6.

Iteration 26 — adding self-verification. The breadcrumb from the parent (fa5513, 31/50 correct on $\mathcal{D}_{tr}$) reports a REWRITESTEP+ADDRULE edit on step 3 with example-level effect (+17, −6). The evidence packet’s focus slot draws two newly broken training examples on which the worker *has* the right sorting logic but emits a typo: **Ex #33** sorts skat before scandalous in its reasoning yet writes the reversed order on the answer line; **Ex #32** shows the same on castrate/cancelled. The attribution architect returns τ =step2 with the reason “the worker ignores its own verification rule, leading to sorting typos despite correct logic.” The patch architect adds one CRITICAL bullet under step 2:

Before finalizing, perform a strict side-by-side verification: compare your sorted list against the original input word-for-word; if the order of any two adjacent words differs from their alphabetical order, immediately re-sort.

The post-hoc operator is ADDRULE; the child (aebcc2, 32/50) is admitted by the score-above-parent rule of Eq. 6.

Iteration 32 — removing self-verification. The breadcrumb from aebcc2 reports the iter-26 ADDRULE. The focus slot now draws two examples that the iter-26 child *newly broke*: **Ex #40** enters an infinite repetition loop, hallucinating the word anaheim hundreds of times in place of the sort, and **Ex #7** outputs an incorrectly sorted list (wiley before whimper) while falsely claiming the verification passed. The attribution architect again localises to τ =step 2 with the reason “the CRITICAL verification sub-rule is ineffective and triggering loops; the model cannot execute self-verification without hallucinating.” The patch architect *removes* both the side-by-side verification and the confirmation bullets—both are forms of post-sort re-processing—and adds an anti-re-processing rule:

CRITICAL: after sorting, do not re-process the list (no re-reading or side-by-side checks)—both trigger hallucination loops on this worker; output immediately.

Crucially the edit is *not* a narrow patch on one signal: its evidence packet draws *both* the focus regressions (Ex #40, #7, fixed on $\mathcal{D}_{tr}$) and the train probe (Tr #44, #49, #35, the M_{train} slot), so the same “stop re-processing” principle addresses the parent’s regression and the train errors together. The aggregate training score stays at 32/50, but the child (7b0643) solves a training example no other archive member solves; it is admitted by the **unique-contribution** branch (a) of Eq. 6—a typed, Pareto-style admission a scalar reward cannot express. This rule (“do not re-process”) contradicts standard chain-of-thought (Wei et al., 2022) practice and is reproducibly discoverable *only* because the per-example (NEWLY_BROKEN, NEWLY_FIXED) transitions in the evidence packet expose it; whole-prompt rewriters with reward-only feedback have no comparable mechanism.

SEPO edit trajectory to the final prompt — BBH Word Sorting (winning lineage teal).
Each edge: [operators] + rule. No parent regression $\Rightarrow$ train-only fallback (Tr-driven); a regression $\Rightarrow$ repair the newly-broken Ex cases.

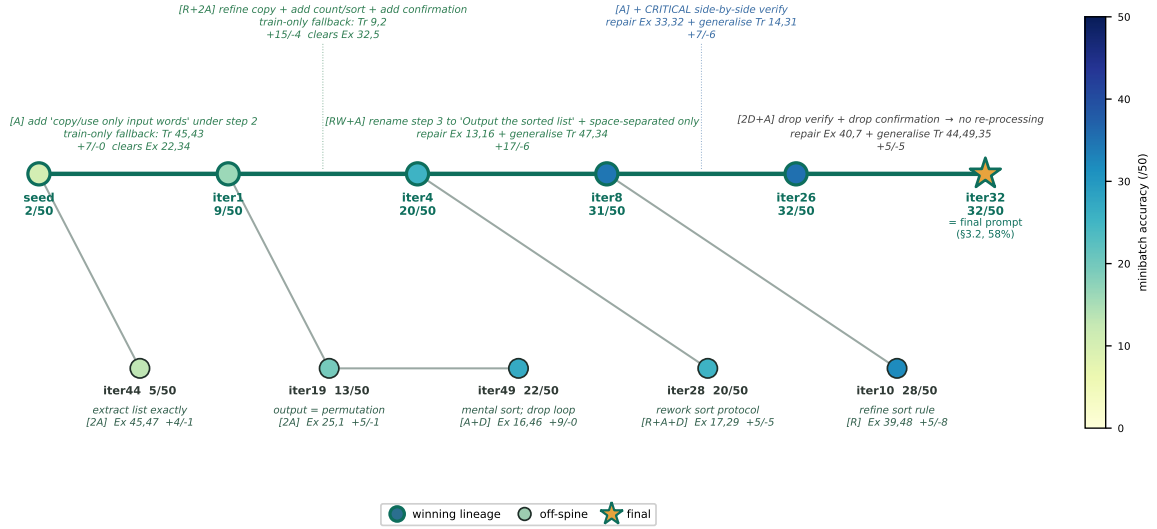


Figure 6: Edit trajectory to the final prompt (BBH Word Sorting, seed 45). The teal spine is the winning lineage to the §3.2 prompt. Each edge shows its operator(s) (A=Add, R=Refine, D=Drop, RW=RewriteStep) and the rule changed; the accumulated rules at iter 32 are exactly the §3.2 prompt. The train probe (Tr, the M_{train} slot) is in *every* packet, giving two real modes: when the parent left *no* regression the edit is a *no-regression fallback* driven by Tr (and still lifts $\mathcal{D}_{\text{tr}}$); when the parent *broke* cases the edit *repairs* those newly-broken training cases (Ex) *and* lets Tr steer generalisation. Lower branches are off-spine accepted schemas. Iter 4 $\rightarrow$ 8’s RW renames step 3 to §3.2’s task-specific “Output the sorted list” heading; the modes, the train probe, every other operator, and the Ex repairs are grounded in the log. Node colour = accuracy; +fixed/–broken on $\mathcal{D}_{\text{tr}}$. Per-edit detail is in the text.

E.3 Edit Tree from Seed to Final Winner (BBH Word Sorting)

Figure 6 shows the path to SEPO’s final Word Sorting prompt (§3.2). We walk through each edit below: its operation, the failing cases that motivated it, and—from the per-example scores—whether it repaired the parent edit’s regression or a train-probe failure.

Per-edit walkthrough (winning lineage). For each accepted edit we give the operator(s) (A=AddRule, R=RefineRule, D=DropRule, RW=RewriteStep) and the rule changed; the rules accumulated at iter 32 are exactly the §3.2 prompt. The train probe (Tr, the M_{train} slot) is in *every* packet (§3.3), giving two modes: when the parent left *no* $\mathcal{D}_{\text{tr}}$ regression the edit is a *no-regression fallback* driven by Tr; when the parent *broke* cases the edit *repairs* those newly-broken Ex cases *and* is steered by Tr to generalise. Ex repairs are verified on $\mathcal{D}_{\text{tr}}$, and the modes and train probe are grounded in the log; iter 4 $\rightarrow$ 8’s RW renames step 3 to §3.2’s task-specific “Output the sorted list” heading.

- **seed $\rightarrow$ iter 1** (2 $\rightarrow$ 9), **A**: add “copy/use only the input words” rule under step 2. *Train-only*

fallback (seed broke nothing): driven by the train probe Tr #45, #43; the rule generalises to clear the persistent hallucination failures Ex #22, #34. *Effect*: +7/–0.

- **iter 1 $\rightarrow$ iter 4** (9 $\rightarrow$ 20), **R+2A**: refine the copy rule (“do not import from memory”), add “count the input words once, then sort in strict A–Z order in a single pass,” and add a confirmation rule (“re-state the full sorted list to check that no word is dropped”). *Train-only fallback* (iter 1 broke nothing): Tr #9, #2 drives it; the confirmation catches the hallucinated / dropped-word failures Ex #32, #5. *Effect*: +15/–4.
- **iter 4 $\rightarrow$ iter 8** (20 $\rightarrow$ 31), **RW+A** on step 3: rename step 3 heading to “Output the sorted list” and add “output one space-separated line, no commas/newlines” under step 3. *Repairs* iter 4’s regression Ex #16, #13 (correct sort, wrong format); anchor Ex #17 (already correct sort + correct format) is preserved by the rule being additive rather than overwriting, with Tr #47, #34 steering generalisation. *Effect*: +17/–6.
- **iter 8 $\rightarrow$ iter 26** (31 $\rightarrow$ 32), **A**: add “CRITICAL:

side-by-side verify” (an *attempt* to fix sort typos). *Repairs* iter 8’s regression Ex #33, #32 (reversed output), with Tr #14, #31 steering generalisation. *Effect*: +7/−6.

- **iter 26 → iter 32** (32 → 32, final), **2D+A**: drop both the side-by-side verification rule and the confirmation rule (both are forms of post-sort re-processing) and add “CRITICAL: do not re-process; trust the single-pass sort.” *Repairs* iter 26’s own regression Ex #40, #7 (the verify rule’s loop / false-verify damage; the confirmation rule also re-reads the list and is subsumed by the same anti-re-process rule), with Tr #44, #49, #35 steering generalisation. *Effect*: +5/−5; removes the loop pathology (final test 58%).

So SEPO never stalls: with a regression it surgically repairs the newly-broken Ex cases while the train probe steers generalisation; with none it keeps improving in no-regression fallback—a dual mode whole-prompt rewriters lack. The operators span RW/R/A/D, and the accumulated rules at iter 32 reproduce the §3.2 prompt exactly; the modes, the train probe, the Ex repairs, and the RW/R/A/D operators are grounded in the log, with iter 4→8’s RW renaming step 3 to the task-specific “Output the sorted list” heading.

E.4 Failure-Mode Analysis

SEPO’s Word Sorting schema repairs the dominant failure modes (entity hallucination, sort-method loops, output-format collapse) but its effect saturates with input length. Word Sorting uses the same 50-train / held-out-test convention as the other BBH tasks, with a disjoint 200-example held-out test split. Table 8 bins a 100-example diagnostic subset of that held-out test split by the number of words to be sorted: short lists (≤ 5 words) reach 79% while long lists (≥ 16 words) collapse to 31%. The degradation is monotone and held across every schema variant we examined. The cause is that the schema’s discipline is *instructional*, not *algorithmic*: rules such as “copy the words exactly” and “trust the initial sort” raise the probability of a correct single pass but do not give the worker an external scratchpad, so the per-token error of an n -word sort still compounds with n . This delineates the scope of prompt-only structural editing: it repairs systematic, length-independent failure modes (format, hallucination, verification loops)

but cannot substitute for an algorithmic execution mechanism on long inputs.

List length (words)	n	Diagnostic acc.
≤ 5	24	0.79
6–10	30	0.70
11–15	20	0.40
≥ 16	26	0.31
Diagnostic subset	100	0.56

Table 8: Held-out accuracy of SEPO’s BBH Word Sorting schema (§3.2) binned by sort length on a 100-example diagnostic subset drawn from the 200-example held-out test split with the Llama-3.1-8B-Instruct worker. Accuracy degrades monotonically with list length, the signature of instructional (not algorithmic) repair.

F Baseline Configurations

For each (method, benchmark) cell in the main comparison experiments, we run three independent optimisation seeds. All baselines share the worker LLM, dataset splits, three-seed schedule, and the `max_metric_calls=2000` metric-call budget. Architect choices, hyperparameters, and budget-normalisation details follow below.

SEED (unoptimised). The cold-start seed schema s_0 of Appendix B.1, deployed without any optimisation. Reported for reference.

OPRO. Architect: Qwen3.5-397B-A17B. Architect temperature 1.0, candidates per step 5, optimiser steps 40, optimisation trajectory window 20. Architect prompts follow the public reference implementation.

APSF. Architect: Qwen3.5-397B-A17B. Number of factors 5, candidates per factor 3, factor selection greedy on the training slice. We use the public reference implementation with no other changes.

MIPROv2. DSPy v2.5; architect Qwen3.5-397B-A17B; `auto="light"` configuration ($n_{\text{init}}=8$ bootstrap demonstrations, $n_{\text{trials}}=20$). Bayesian optimiser uses the default prior.

MPO (Sharma and Henley, 2026). Architect: Qwen3.5-397B-A17B. We use the released hyperparameters ($T=40$ optimisation iterations, $b=5$ candidates per iteration, trajectory window 10).

GEPA. Architect: Qwen3.5-397B-A17B. Pareto-front archive with capacity 32; reflection

prompt and reflective sample size 3 follow the released configuration. We rerun the released code with the same metric-call budget as SEPO.

G Extended Experimental Results

This appendix supplements §4 with additional Llama runs on high-variance tasks, a matched HotpotQA-Hard sensitivity test, per-method and per-benchmark token-cost decompositions, and

budget-matched accuracy curves.

G.1 Additional Runs

We also run three additional paired Llama optimisations on five tasks with high run-to-run variance. Across these 15 added comparisons, SEPO wins 12, and its six-run mean remains higher on all five tasks (Table 9).

Task	Reported 3 runs mean $\pm$ SD (S/G)	Seed 46 (S/G)	Seed 47 (S/G)	Seed 48 (S/G)	All 6 runs mean $\pm$ SD (S/G)	6-run Δ	Added W/T/L
Causal Judgement	61.1 $\pm$ 4.6/59.9 $\pm$ 3.8	64.2/62.8	55.5/56.9	57.7/56.9	60.1 $\pm$ 4.2/59.4 $\pm$ 3.3	+0.7	2/0/1
Date Understanding	76.5 $\pm$ 4.3/75.3 $\pm$ 6.1	82.0/80.0	76.5/82.0	77.5/74.0	77.6 $\pm$ 3.5/77.0 $\pm$ 5.0	+0.6	2/0/1
Disambiguation QA	68.8 $\pm$ 4.6/64.5 $\pm$ 6.6	74.5/74.0	68.5/63.5	67.5/61.0	69.5 $\pm$ 3.8/65.3 $\pm$ 6.1	+4.2	3/0/0
Logical Deduction	47.3 $\pm$ 2.1/47.0 $\pm$ 3.9	49.5/45.5	44.0/51.0	46.5/41.5	47.0 $\pm$ 2.2/46.5 $\pm$ 3.9	+0.5	2/0/1
Word Sorting	58.2 $\pm$ 1.3/39.3 $\pm$ 18.3	59.5/56.0	57.0/31.0	58.0/31.5	58.2 $\pm$ 1.1/39.4 $\pm$ 14.7	+18.8	3/0/0

Table 9: Additional paired Llama runs on five high-variance tasks. S/G denotes SEPO/GEPA; the W/T/L column counts only the three added seeds. Accuracies and Δ are in percentage points.

G.2 Matched HotpotQA-Hard Sensitivity

To test whether the Llama HotpotQA split choice changes the method ranking, we evaluate the compared methods below on the same 400 held-out Hard examples.

Method	HQA-Hard EM (%)
Seed	8.00
OPRO	24.00
APSF	26.00
MPO	28.25
GEPA	28.75
SEPO	29.25

Table 10: Matched HotpotQA-Hard sensitivity test with the Llama-3.1-8B-Instruct worker on 400 held-out examples.

Replacing HotpotQA-Medium with Hard in the Llama 14-task macro changes SEPO from 61.9% to 59.0% and GEPA from 58.8% to 55.9%; SEPO therefore remains first, with the same 3.1 pp lead. The compressed 24.0–29.25% range reflects limited answer accessibility: with the fixed seed prompt, pass@1 is 8.0%, oracle pass@64 is 31.0%, and 69.0% of examples remain incorrect in all 64 samples. We use Medium under Llama to retain a more discriminative optimiser comparison.

G.3 Per-Benchmark Token Cost

List prices used in Table 5. The \$/cell and $\rho_{\$}$ columns of Table 5 are computed from public list

prices as of 2026-05-26. The architect (Qwen3.5-397B-A17B) is priced at the SiliconFlow public rate of ¥1.20/M input and ¥7.20/M output tokens (0–128k context tier); the worker (Llama-3.1-8B-Instruct) is priced at the OpenRouter listing of \$0.02/M input and \$0.05/M output tokens. CNY amounts are converted at the 2026-05-26 spot rate of 1 USD=6.79 CNY, yielding architect prices of \$0.177/M input and \$1.060/M output. For method m , we compute

$$C_m = \sum_{x \in \{A, W_{\text{opt}}, W_{\text{dep}}\}} (p_x^{\text{in}} T_{x,m}^{\text{in}} + p_x^{\text{out}} T_{x,m}^{\text{out}}), \quad (7)$$

Here x indexes the architect, worker-optimisation, and worker-deployment channels; p and T denote the corresponding list price and measured token count. Each cell includes one held-out deployment of the selected prompt; we average cell costs across seeds and tasks. Input/output separation matters because architect output is approximately six times as expensive as input at this snapshot. For MPO, totals are measured but the unavailable I/O split is estimated from GEPA, so only its dollar estimate depends on this approximation. The token-based ρ is provider-independent, whereas $\rho_{\$} = \Delta \text{Acc} / C_m$ covers all three billed channels; dollar results are therefore a dated price snapshot.

Table 11 reports the aggregate input/output split of each cost axis per method (one row per method, no per-benchmark breakdown). Tables 12–14 then decompose each axis benchmark by benchmark.

Method	Architect (M tok)		Worker-Opt (M tok)		Worker-Dep (k tok)	
	Input	Output	Input	Output	Input	Output
OPRO	6.44	0.40	2.04	1.07	134.7	109.3
MIPROv2	3.32	0.20	1.64	0.96	105.0	103.4
APSF	0.69	0.00	1.33	0.80	78.9	78.9
MPO [†]	0.17	0.02	2.68	0.72	198.3	62.7
GEPA	0.20	0.03	3.06	0.82	321.9	101.8
SEPO	0.32	0.02	1.61	0.98	108.8	94.4

Table 11: Per-method input/output token decomposition (M / k tokens per cell, mean over 3 seeds, Llama-3.1-8B-Instruct 14-task suite). Row sums match the corresponding columns of Table 5. [†]MPO’s I/O split is estimated from GEPA; only its row totals are measured.

Method	BBH								Code	MMLU-Pro			Reasoning		Mean
	Caus.	Date	Disamb.	Fall.	Deduct.	Obj.	Snarks	Word.		MBPP+	CS	Econ.	Psych.	GSM-H	HQA
OPRO	21.22	4.91	9.68	19.47	13.91	19.56	3.14	0.46	0.13	0.71	0.63	1.76	0.04	0.08	6.84
MIPROv2	12.05	5.71	5.23	1.07	1.44	5.38	3.02	0.53	1.34	5.46	2.92	4.61	0.18	0.28	3.52
APSF	1.57	1.67	0.28	0.95	0.89	0.34	0.14	0.91	0.28	0.61	0.57	0.24	0.61	0.58	0.69
MPO	0.32	0.30	0.25	0.27	0.23	0.30	0.22	0.21	0.09	0.10	0.13	0.10	0.04	0.07	0.19
GEPA	0.24	0.23	0.15	0.54	0.52	0.18	0.13	0.37	0.68	0.07	0.07	0.04	0.03	0.00	0.23
SEPO	0.38	0.50	0.28	0.58	0.89	0.32	0.28	0.59	0.10	0.18	0.17	0.13	0.07	0.30	0.34

Table 12: Per-benchmark *architect* token cost (millions of tokens per cell, mean over 3 seeds). The MEAN row matches the *Arch.* column of Table 5.

Method	BBH								Code	MMLU-Pro			Reasoning		Mean
	Caus.	Date	Disamb.	Fall.	Deduct.	Obj.	Snarks	Word.	MBPP+	CS	Econ.	Psych.	GSM-H	HQA	Macro
OPRO	3.14	3.78	2.19	5.40	4.93	3.09	1.48	1.70	1.35	3.68	2.81	2.92	2.47	4.65	3.11
MIPROv2	3.10	3.11	1.42	4.37	4.96	2.23	1.20	1.17	1.14	2.96	2.36	2.08	2.05	4.28	2.60
APSF	1.08	2.02	0.49	3.76	2.48	1.08	0.30	4.56	1.16	2.68	2.20	2.05	1.25	4.69	2.13
MPO	3.98	3.76	2.56	5.20	5.71	2.94	2.13	4.33	1.52	2.60	3.16	2.39	2.28	4.99	3.40
GEPA	5.62	4.11	4.28	6.95	5.59	3.94	3.08	6.06	1.64	3.80	3.17	2.54	2.16	1.32	3.88
SEPO	3.03	2.84	1.50	3.76	3.44	1.93	1.48	2.64	0.95	2.76	2.24	2.14	3.26	4.26	2.59

Table 13: Per-benchmark *worker-optimisation* token cost (millions of tokens per cell, mean over 3 seeds). The MEAN row matches the *Work.* column of Table 5.

Method	BBH								Code	MMLU-Pro			Reasoning		Mean
	Caus.	Date	Disamb.	Fall.	Deduct.	Obj.	Snarks	Word.	MBPP+	CS	Econ.	Psych.	GSM-H	HQA	Macro
OPRO	84.7	268.0	180.8	226.4	319.0	209.2	86.0	65.7	117.0	287.6	234.5	241.7	380.6	714.7	244.0
MIPROv2	83.5	233.4	123.6	186.3	324.0	182.2	67.7	46.7	98.9	225.6	195.3	175.9	316.1	659.0	208.4
APSF	36.0	137.9	36.9	144.4	164.8	77.6	14.9	177.7	87.3	183.5	157.8	154.2	175.8	660.6	157.8
MPO	108.1	277.2	223.1	216.4	359.1	251.9	123.5	168.1	132.4	206.8	264.7	203.7	350.9	767.7	261.0
GEPA	208.0	413.4	535.7	337.9	447.8	468.3	236.9	295.6	153.1	506.8	424.1	500.8	741.6	661.3	423.7
SEPO	92.7	207.0	128.4	161.4	260.7	160.6	83.0	117.6	81.3	219.8	185.9	181.4	263.7	701.0	203.2

Table 14: Per-benchmark *worker-deployment* token cost (thousands of tokens per cell, mean over 3 seeds). Deployment cost is the average tokens per worker call multiplied by the test-set size. HotpotQA’s seed cost is dominated by retrieval context length, which inflates all methods uniformly. The MEAN row matches the *Dep.* column of Table 5.

Together the four tables let any downstream dollar-cost computation be done at the level of (method, axis, I/O channel), and let reviewers verify that the per-method aggregates match the main-paper Table 5 both in total (cross-checked in the MEAN rows of the per-benchmark tables) and in I/O composition. Worker-optimisation and deployment tokens are reported per evaluation run; architect tokens are reported as measured.

G.4 Per-Benchmark Budget-Matched Curves

Figure 7 extends the Q1 macro headline of the main paper to the budget axis: for every benchmark family we re-evaluate each method’s stored *best-so-far* prompt at four metric-call checkpoints (500, 1000, 1500, 2000), starting from the shared seed prompt at metric call 0. All baselines, including SEPO, are evaluated under identical conditions; in particular, APSF’s intermediate snapshots correspond to the *step-aligned best-so-far* candidate at the end of each optimisation step (each step consumes ≈ 200 metric calls in the $\{4 \text{ candidates}\} \times \{50 \text{ val examples}\}$ schedule).

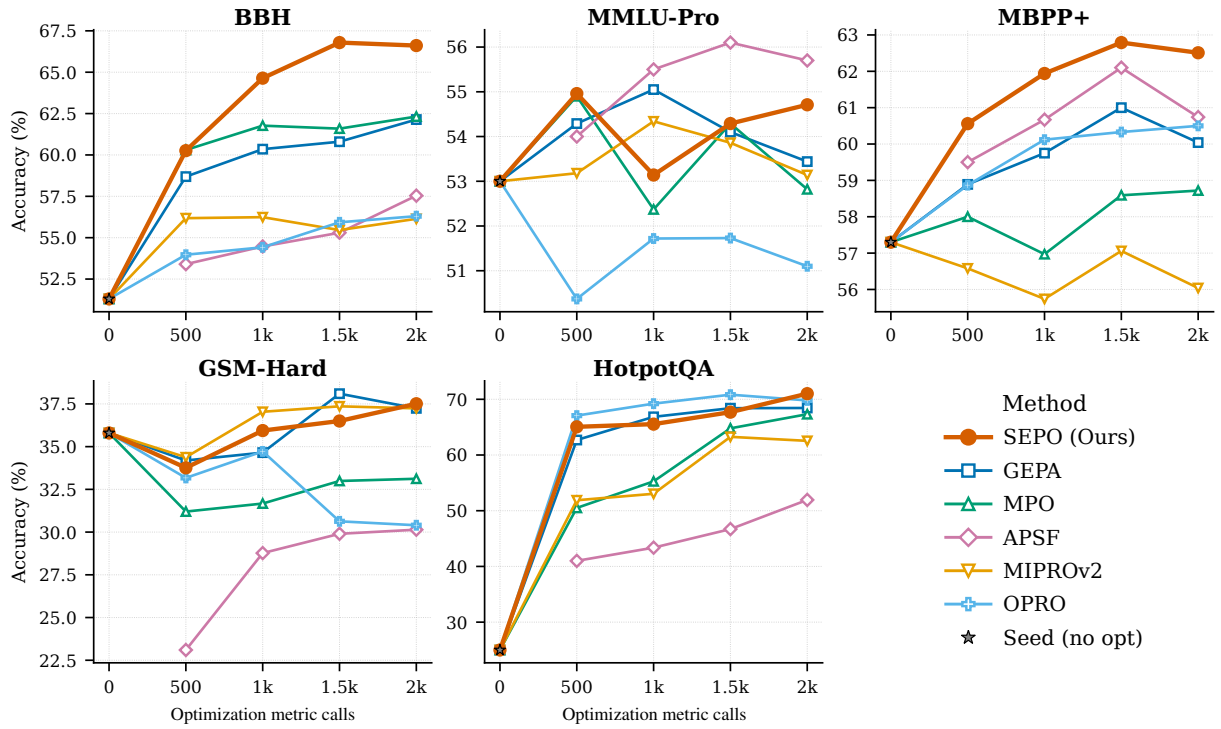


Figure 7: **Per-benchmark budget-matched accuracy curves** (Llama-3.1-8B-Instruct, held-out). For each benchmark family — BBH and MMLU-Pro shown as the mean over their sub-tasks, and MBPP+, GSM-Hard, and HotpotQA shown individually — we plot the held-out accuracy of every method’s stored best-so-far prompt at metric-call checkpoints 0 (seed), 500, 1000, 1500, and 2000, under a shared evaluation protocol. SEPO (highlighted) leads consistently on BBH and MBPP+, remains competitive on MMLU-Pro, GSM-Hard, and HotpotQA, and yields the strongest macro-level trajectory.

H Ablation Studies (Extended)

This appendix expands the Q2 ablation section (§4.3) with the design map, the five-benchmark diagnostic score table, proposal-level target/error controls, and two archive controls. We separate the four destructive removals in the main text from the auxiliary controls PARETOFREQPARENT

and —UNIQUEADM: the former asks whether the loop’s main modules are load-bearing, while the latter asks which archive sub-decisions account for the P3 gain.

H.1 Ablation Design Map

Variant	Principle removed	Falsifiable prediction	Primary readout
-ATTRIBUTION	P1: typed target on the two-layer schema	edits drift toward whole-prompt rewrites; final prompts grow longer	prompt length and length variance
-LINEAGE	P2: example-level lineage evidence (strong form)	gain falls back to generic error-reflection level	held-out accuracy
-BREADCRUMB	P2: operator/locator summary ℓ only	small or null drop if ℓ is redundant with example deltas	held-out accuracy
-MULTITRAJ	P3: multi-trajectory archive + Lexicase	greedy best-only editing shrinks exploration; the largest single-removal accuracy drop	held-out accuracy and acceptance rate
PARETOFREQPARENT	P3: Lexicase parent selection only	frequency-weighted sampling recovers much, but not all, of the archive benefit	five-task diagnostic accuracy
-UNIQUEADM	P3: unique-solve admission branch only	archive remains useful, but specialist admission should contribute a small margin	five-task diagnostic accuracy

Table 15: Ablation design map for Q2. The first four rows are the main destructive removals; the final two rows are auxiliary archive controls that isolate parent sampling and admission.

Table 15 maps the main removal and archive-control variants to the SEPO principle it removes, the falsifiable prediction it tests, and the primary readout that operationalises that principle. The proposal-level controls are specified separately in §H.3. The protocol of Table 1 is held fixed across all variants (worker LLM, architect, seed schemas, metric-call budget, held-out split). The

—BREADCRUMB row is included explicitly so the design admits a null result for the operator-summary sub-claim. The final two rows are auxiliary P3 controls: they preserve the archive, so their expected drops are smaller than the best-only —MULTITRAJ removal.

H.2 Five-Benchmark Diagnostic Scores

Task	Full	—ATTR	—LIN	—BRD	—MT	PARETOFREQ	—UNIQUEADM
BBH Date	76.5	74.0	72.2	76.0	71.8	76.6	75.8
BBH Snarks	82.3	73.8	78.5	76.6	75.0	81.1	81.5
MBPP+	62.6	60.9	58.5	59.8	55.6	61.2	61.5
GSM-Hard	37.6	36.4	35.0	37.0	35.1	35.9	36.3
MMLU-Pro CS	46.7	45.0	46.8	47.0	44.5	46.3	47.6
Avg.	61.1	58.0	58.2	59.3	56.4	60.2	60.5

Table 16: Absolute scores for the five-benchmark diagnostic ablation subset (Llama-3.1-8B-Instruct, held-out accuracy in %); **bold** marks the best entry in each row.

Table 16 reports the absolute held-out scores for the five-benchmark diagnostic subset, not a separate 14-task aggregate: the average is the arithmetic mean of BBH Date, BBH Snarks, MBPP+, GSM-Hard, and MMLU-Pro CS. The subset is chosen

to expose the three mechanisms discussed in §4.3: structured procedural tasks for P1/P2, a code task for transfer, and a knowledge-leaning negative case for scope.

The absolute table sharpens the P3 interpre-

tation. Removing multi-trajectory search entirely is the most damaging intervention on the subset (−MT: 56.4%, −4.7 pp), whereas the two archive controls remain close to Full (PARETOFREQPARENT: 60.2%, −0.9 pp; −UNIQUEADM: 60.5%, −0.6 pp). Thus the main mechanistic contrast is not that every archive sub-decision is monotone on every task: Pareto-frequency sampling slightly improves BBH Date, and removing unique-admission slightly improves MMLU-Pro CS. The robust conclusion is instead that preserving multiple specialised candidates matters much more than either auxiliary archive choice in isolation, while Lexicase and unique-admission add a small average margin.

H.3 Proposal-Level Target and Error Controls

We add two budget-matched controls on the same five-benchmark subset. RANDOMTARGET receives the same evidence as Full SEPO but replaces the attributed target with a random legal structural target. RANDOMERROR replaces newly broken focus examples with the same number of examples sampled uniformly from the parent’s current error set, regardless of whether they were already incorrect or newly broken by the preceding edit.

Variant	Test Acc.	Fixes /100	Breaks /100	Net gain /100	Net-positive edits
Full SEPO	61.1	7.5	5.0	+2.5	61%
−ATTRIBUTION	58.0	8.0	7.3	+0.7	47%
RANDOMTARGET	57.4	4.8	6.5	−1.7	34%
RANDOMERROR	59.9	6.3	5.5	+0.8	51%

Table 17: Proposal-level diagnostics on the five-benchmark subset (Llama-3.1-8B-Instruct). Fixes and breaks are corrected and newly broken training examples per 100 evaluated proposals; net gain is their difference. Net-positive edits are proposals with more fixes than breaks.

Despite receiving identical evidence, RANDOMTARGET produces fewer fixes and a negative net gain. Replacing newly broken cases with random current errors also reduces net gain from +2.5 to +0.8 and net-positive edits from 61% to 51%. These controls support separate roles for target attribution, which routes evidence to an appropriate structural unit, and transition-aware sampling, which identifies regressions introduced by the preceding edit.

H.4 Archive Sub-Decision Controls

Two auxiliary controls isolate which archive sub-decision drives the P3 gain. Both preserve the archive, admission rule, and metric-call budget of Full SEPO, changing only one decision at a time (Table 16). PARETOFREQPARENT replaces Lexicase parent selection (§3.4) with the frequency-weighted Pareto-front sampling used by GEPA. It averages 60.2% on the five-benchmark subset (0.9 pp below Full), recovering most of the multi-trajectory benefit and even edging out Full on BBH Date (76.6 vs. 76.5) while trailing on the other four tasks; we read this as evidence that Lexicase is a safer default for keeping specialist parents, not that Pareto-front sampling is uniformly weak. −UNIQUEADM removes only the “unique-solve” admission branch of Eq. 6—children can still enter through the higher-score and shorter-prompt branches—and averages 60.5% (0.6 pp below Full, 4.1 pp above −MULTITRAJ). Both drops are small and in the expected direction, so P3 is primarily the effect of retaining and revisiting multiple candidates, with Lexicase parent selection and unique-solve admission each adding a modest average margin.