LEARNING WHEN AND WHAT TO TEACH: SELF-EVOLVING PRIVILEGED SEARCH AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Training long-horizon search agents is difficult because useful recovery trajectories are scarce, while stronger teachers, guides, or judges make improvement depend on external capability. As base models grow stronger, progress should instead turn their own capabilities and task-internal signals into self-evolution. We observe that search answers are often structured entities or relations from which the same model can reason backward to propose missing recovery actions, although direct answer conditioning risks privilege leakage. We introduce `MIST-OPD`, which instantiates the same base model as an answer-blind Student and a training-only privileged Guide. At each current Student state, the Guide proposes six levels of answer-free intervention. Matched continuations compare each candidate with no guidance, its adjacent weaker level, and a same-level placebo from an unrelated same-source answer; the first candidate in the ladder that improves over every applicable control is retained. Accepted boundaries teach the Guide when, how much, and what to teach; controlled distillation transfers each accepted repair at a state visited by the current Student. Thus the Student determines where to learn, while the matched public-execution gate determines which proposed correction is admitted. The updated pair processes fresh failures, and deployment removes the Guide and answer. Using Qwen3.5-9B, `MIST-OPD` reaches 52.8% macro accuracy across six benchmarks and exceeds the strongest distillation baseline by 4.4 points; gains transfer to 4B and code.

1 INTRODUCTION

Training long-horizon search agents on difficult tasks is fundamentally a problem of discovering recovery behavior. A final answer may depend on many interdependent decisions—which query to issue, which evidence to retain, when to backtrack, and when to stop (Nakano et al., 2021; Yao et al., 2022)—yet feedback arrives only after the trajectory. On the hardest instances, every sampled trajectory can fail, removing the reward contrast required by policy-gradient and group-relative objectives (Schulman et al., 2017; Shao et al., 2024). Process supervision and hierarchical credit provide denser feedback on annotated steps or visited states, but do not by themselves expose an unvisited recovery branch (Lightman et al., 2024; Feng et al., 2025). A stronger teacher, guide, or judge can instead supply a missing action through distillation (Zhao et al., 2026; Penaloza et al., 2026; Lazaridis et al., 2026); this solution, however, makes improvement depend on an external model more capable than the Student.

As base models become stronger, continued progress should not require an ever-stronger external supervisor. This motivates *self-evolution*: turning the model's own interactions and task-internal signals into new capability (Zelikman et al., 2022; Gulcehre et al., 2023; Madaan et al., 2023; Shinn et al., 2023). Existing approaches nevertheless inherit a discovery bottleneck. Outcome-driven methods such as GRPO and GiGPO need successful rollouts or reward variation (Shao et al., 2024; Feng et al., 2025); self-training and reflection recycle the model's own solutions or critiques but may remain within its current support (Zelikman et al., 2022; Madaan et al., 2023; Shinn et al., 2023). On-policy distillation supplies full-vocabulary feedback at Student-visited prefixes, yet its coverage remains tied to the states and prefixes that the current policy reaches; it does not by itself validate a missing recovery branch. A stronger teacher can expand that coverage but forfeits the self-evolution premise (Zhao et al., 2026; Yu et al., 2026; Zhu et al., 2026; Yang et al., 2026). Thus, in the all-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

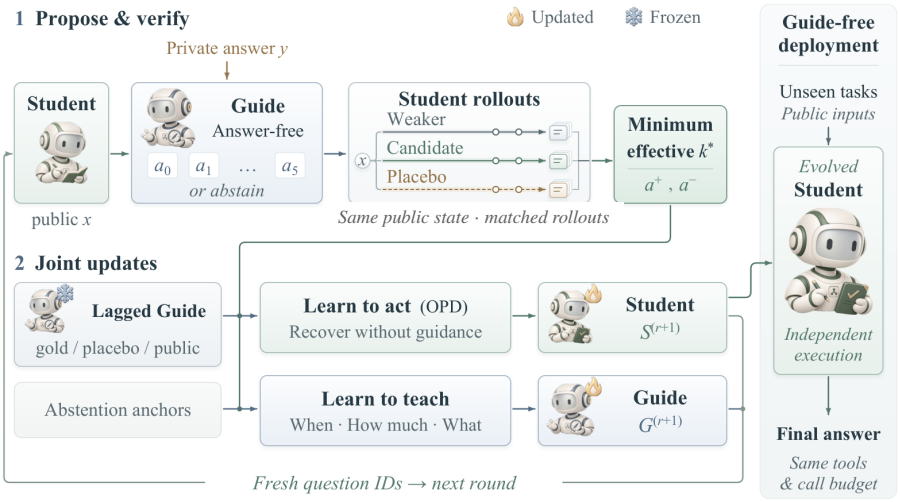


Figure 1: **One `MIST-OPD` round.** Matched rollouts screen proposals; deployment retains only the Student.

failure regime, an agent may recognize that its search failed without having a better action from which to learn.

We identify an overlooked task-internal source of such actions. In search, reference answers are often structured entities or relations. Given an answer, the same model can reason backward to an alias, a bridge entity, an inverse relation, or a branch that should have been pruned, thereby proposing a recovery that its answer-blind rollout did not find. This makes the answer useful not as a deployment input, but as *privileged information for behavior discovery*. The opportunity creates a corresponding risk: an answer-aware policy can appear stronger by copying the answer, inserting an answer-specific query, or exploiting another shortcut unavailable at test time. Distilling its full distribution therefore conflates transferable search strategy with privilege leakage.

We introduce `MIST-OPD` (Minimum-Intervention Self-Teaching with On-Policy Distillation), which turns reference answers into transferable search behavior without a separate learned teacher or learned judge/reward model. Here, on-policy refers to current Student states; screened Guide proposals expand the actions available for learning. As Figure 1 shows, an answer-blind Student and a training-only Guide start from the same base model. The Student exposes a public state; using the answer, the Guide either abstains or proposes six ordered, answer-free interventions from a general operation to an executable recovery route. We do not imitate these proposals directly. From the identical state, matched continuations compare a candidate with no guidance, its adjacent weaker level, and a same-level placebo, retaining the first tier in the prespecified order that passes every applicable control. The accepted boundary trains the Guide's future proposals, while controlled distillation, selected-action supervision, and replay transfer the accepted repair to the Student. State collection remains on-policy although the repair extends beyond Student-sampled prefixes. The updated pair then generates fresh states, closing the loop. Deployment removes every privileged component and retains only the answer-blind Student under the original tool and context budget.

With Qwen3.5-9B (Qwen Team, 2026), `MIST-OPD` reaches 52.8% macro accuracy across BrowseComp, FRAMES, GAIA, HLE, HotpotQA, and BrowseComp-Plus (Wei et al., 2025; Krishna et al., 2024; Mialon et al., 2024; Phan et al., 2025; Yang et al., 2018), improves every benchmark, and exceeds the strongest distillation baseline by 4.4 points. The improvement persists on a 4B backbone and transfers to two repository-level code benchmarks; removing the hierarchy, controlled residual, or Guide updates consistently weakens performance. Our contributions are threefold. First, structured answers become a task-internal source of recovery actions beyond those discovered by conventional self-training or reflection (Zelikman et al., 2022; Madaan et al., 2023; Shinn et al., 2023). Second, graded matched-state tests identify the first-passing tier and filter shortcut effects detected by the controls before distillation, rather than transferring an answer-aware distribution wholesale (Zhao et al., 2026; Yu et al., 2026; Zhu et al., 2026). Third, alternating updates jointly evolve a same-backbone Guide and Student over fresh current Student states, while deployment remains strictly answer-blind.

2 MIST-OPD

2.1 LEARNING SETTING

We study an interactive search agent whose public history at step t is $h_t = (q, a_{<t}, o_{<t})$. The deployed Student $S_\theta(a_t \mid h_t)$ observes only this history and the ordinary tool schema. During training, a separately optimized Guide G_ϕ may also observe private information y, consisting of the reference answer and verified aliases. Guide and Student start from the same base model, so their asymmetry is information rather than initial capability; only the Student is deployed.

The answer can reveal a missing recovery operation, but it can also produce an unavailable shortcut. We therefore do not imitate the Guide directly. For each public state $x_i = h_t$, we identify an accepted teaching boundary

$$e_i = (k_i^\star, a_i^+, a_i^-, a_i^\varnothing, w_i),$$

where $k_i^\star$ is the first passing tier in the predefined order, a_i^+ its action, a_i^- its designated weaker reference, $a_i^\varnothing$ the unguided action, and w_i its matched-evidence reliability weight. The answer, Guide text, intervention label, outcome, and Guide logits never enter the Student input. One round therefore has three stages: the Guide proposes answer-free repairs, public continuations select the first passing repair in the predefined order, and the accepted boundary updates both policies.

2.2 SCREENED INTERVENTIONS FROM PRIVILEGED ANSWERS

Given (x_i, y_i), the Guide either abstains or proposes the ordered ladder in Table 1. Each tier renders to a legal answer-free action; malformed, leaking, or non-ordered ladders are rejected. The ordering is operational: $k_i^\star$ is the first tier in this prespecified ladder that passes every applicable control, not an information-theoretic minimum.

Table 1: The ordered teaching ladder. Each tier represents a prespecified degree of answer-free specificity and renders to an executable next action.

Level	Role	Information added by the Guide
$\perp$	Abstain	No intervention; the public policy is already sufficient.
L_0	Diagnose	Identify the failed branch without proposing a direction.
L_1	Operate	Select a generic operation: rewrite, broaden, backtrack, or reverse.
L_2	Type	Specify the relation or entity type that the next query should test.
L_3	Attribute	Add a disambiguating role or attribute of an intermediate entity.
L_4	Bridge	Specify an answer-free bridge that determines the query skeleton.
L_5	Route	Provide an executable recovery route while withholding the answer.

We screen proposals from the identical serialized state. A candidate a_i^k is compared with the unguided action $a_i^\varnothing$, a same-level placebo $a_i^{p,k}$ generated from an unrelated same-source answer, and, for $k > 0$, the adjacent weaker action a_i^{k-1}. All arms use the same continuation policy, environment, retrieval backend, and tool budget. Action generation and environment interaction are answer-blind; during construction only, the reference answer scores terminal correctness after each continuation. Thus the mechanism is a matched public-execution gate, not an answer-free scorer. For a matched repetition, u defeats v iff u ends correct and v does not; v defeats u symmetrically, and all other outcomes are ties. Let $W_i(u, v)$ and $L_i(u, v)$ count these two outcomes across repetitions. Define $C_i(u, v) = \mathbf{1}[W_i(u, v) \geq m \wedge L_i(u, v) = 0]$. We retain

$$k_i^\star = \min\big\{k : C_i(a_i^k, a_i^\varnothing) = 1 \ \wedge\ [k = 0 \ \vee\ C_i(a_i^k, a_i^{k-1}) = 1] \ \wedge\ C_i(a_i^k, a_i^{p,k}) = 1\big\}. \quad (1)$$

Thus every accepted level must beat no guidance and its placebo, while levels above L_0 must also beat the adjacent weaker level; we evaluate the full ladder before retaining a state, and any reversal at a stronger tier rejects the entire state rather than only that tier. For the training statistics, let $v_i^{\mathrm{ref}} = a_i^\varnothing$ when $k_i^\star = 0$ and $v_i^{\mathrm{ref}} = a_i^{k_i^\star - 1}$ otherwise. We set $W_i = W_i(a_i^+, v_i^{\mathrm{ref}})$ and $L_i = L_i(a_i^+, v_i^{\mathrm{ref}})$, without pooling the three controls, and use $w_i = (W_i + L_i)/R \leq 1$ as the matched-evidence

reliability weight. A state is labelled ABSTAIN only if the unguided arm succeeds in at least one matched repetition and no assisted or placebo arm defeats it in any repetition; states satisfying neither rule are rejected. This is an empirical screening rule rather than a significance test; we use $R = 2, m = 1$ for construction efficiency and audit retained tiers with held-out continuations.

2.3 LEARNING FROM THE ACCEPTED BOUNDARY

The Guide proposes abstention, intervention level, answer-free hint, and action; matched public continuations assign the first-passing tier $k_i^\star$ and action a_i^+. Accepted boundaries train later proposals but never bypass the public-execution gate. The Guide loss is:

$$\begin{aligned} \mathcal{L}_G &= \lambda_H \mathcal{L}_{\text{level+hint}} + \lambda_A \mathcal{L}_{\text{rank}} + \lambda_C \mathcal{L}_{\text{ctrl}} + \lambda_{GR} \mathcal{L}_{\text{Guide-replay}}, \\ \mathcal{L}_{\text{rank}} &= -\log \sigma(\beta m_i^{\mathsf{g}}), \\ \mathcal{L}_{\text{ctrl}} &= \sum_{z \in \{\mathsf{p},\mathsf{0}\}} [m_i^z]_+. \end{aligned} \tag{2}$$

Here, $a_i^- = v_i^{\text{ref}}$ and $m_i^z = |a_i^+|^{-1} \log G_\phi^z(a_i^+ \mid x_i) - |a_i^-|^{-1} \log G_\phi^z(a_i^- \mid x_i)$ is the length-normalized action-sequence log-likelihood margin in view z; we use $\beta = 1$. The level-and-hint and Guide-replay terms are token cross-entropies, and $[u]_+ = \max(u, 0)$. A lagged Guide $G_{\bar{\phi}}$ scores the accepted action at one public prefix under the true answer (g), source-matched placebo (p), and no private value (0), rather than transferring its full distribution directly:

$$r_{it}(v) = \log G_{\bar{\phi},t}^{\mathsf{g}}(v) - \tfrac{1}{2} \log G_{\bar{\phi},t}^{\mathsf{p}}(v) - \tfrac{1}{2} \log G_{\bar{\phi},t}^{\mathsf{0}}(v). \tag{3}$$

This contrast discounts components shared across views but does not rule out every semantic shortcut; the matched gate is therefore an empirical screening mechanism rather than a guarantee of privilege removal. We zero residuals for every tokenization of the answer, aliases, placebo, and donor markers, then center and clip the rest. The anchored Student target is

$$q_{it}(v) = \begin{cases} S_{\text{ref},t}(v), & v \in F_i, \\ (1 - b_{it}) \dfrac{S_{\text{ref},t}(v) e^{\eta w_i \widetilde{r}_{it}(v)}}{\sum_{u \notin F_i} S_{\text{ref},t}(u) e^{\eta w_i \widetilde{r}_{it}(u)}}, & v \notin F_i, \end{cases} \tag{4}$$

where $b_{it} = \sum_{v \in F_i} S_{\text{ref},t}(v)$ and S_{ref} is the Student checkpoint frozen at the beginning of the current round and held fixed throughout that round's Student update. The target preserves every forbidden coordinate and the total allowed mass, while the Student objective adds selected-action supervision, a training-only ordinal help head, and behavior replay:

$$\mathcal{L}_S = \lambda_{\text{KD}} D_{\text{KL}}(q_i \| S_\theta) + \lambda_{\text{CE}} \text{CE}(a_i^+ \mid x_i) + \lambda_{\text{ord}} \mathcal{L}_{\text{ord}}(x_i, k_i^\star) + \lambda_{\text{help}} \mathcal{L}_{\text{help}}(x_i, h_i) + \lambda_R \mathcal{L}_{\text{replay}}. \tag{5}$$

Here, $h_i = 1$ for accepted boundaries and $h_i = 0$ for verified abstention anchors. $\mathcal{L}_{\text{help}}$ is binary cross-entropy for predicting from the public state whether intervention is required; $\mathcal{L}_{\text{ord}}$ and $\mathcal{L}_{\text{replay}}$ are cross-entropies on the tier label and public replay, respectively. The KL is evaluated only at accepted action positions. The ordinal head asks the public representation to predict $\{\text{ABSTAIN}, L_0, \ldots, L_5\}$, but it does not route deployment-time actions and is removed with the Guide. The Student never receives the answer, Guide instruction, or rollout outcome as an input; scored execution outcomes only select and weight the accepted correction used for training. The main Student objective also contains no action-DPO term, so Student-side preference optimization cannot silently explain the controlled-distillation result.

2.4 ALTERNATING SELF-EVOLUTION AND DEPLOYMENT

For each state, the lagged Guide constructs q_i and remains fixed while the Student is updated. We then update the live Guide and refresh the Guide EMA exactly once; the resulting EMA checkpoint is used for the next state. Across ID-disjoint rounds, Student updates change visited failures and Guide updates change later candidate ladders, while the public-execution gate remains the selector.

Unlike trajectory-on-policy distillation, our *state-on-policy repair distillation* keeps public states on-policy while screened Guide proposals expand action support; accepted repair prefixes are therefore Guide-proposed rather than sampled by the Student.

At inference, the answer, Guide, intervention ladder, placebo donor, teacher logits, and ordinal head are removed. Deployment is simply $S_\theta(a_t \mid h_t)$ under the original context, tools, and call budget.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Backbone models. Our primary experiments use Qwen3.5-9B. The Guide and Student are separately trainable copies initialized from the same backbone, so their only initial asymmetry is access to the training answer. We repeat the complete pipeline with Qwen3.5-4B to test whether the method depends on a large privileged model. Only the answer-blind Student is evaluated in either setting.

Training data. We construct teaching states from public questions drawn equally from 2WikiQA and MuSiQue (Ho et al., 2020; Trivedi et al., 2022). Student rollouts are collected before any answer is exposed; the Guide sees the answer only after a public state has been serialized. Construction, replay, and evaluation question identifiers are disjoint. A frozen 384-question construction cohort yields 296 technically valid public states after checking tool legality, serialization, source identity, and overlap with the evaluation manifests. The controlled intervention test, rather than answer access alone, determines which states enter optimization. Later rounds draw fresh, ID-disjoint cohorts from the same 12,000-question inventory; full round-wise counts remain in Appendix E.

Evaluation benchmarks. We evaluate full-agent search on BrowseComp, BrowseComp-Plus, FRAMES, GAIA-text, HLE-text, and HotpotQA (Wei et al., 2025; Krishna et al., 2024; Mialon et al., 2024; Phan et al., 2025; Yang et al., 2018). Together these benchmarks cover open-web retrieval, multi-page evidence aggregation, tool-using assistance, specialized knowledge, and structured multi-hop reasoning. We use the complete released text split of each benchmark: 1,266, 830, 824, 127, 2,158, and 7,405 questions, respectively. GAIA-text and HLE-text exclude 38 and 342 attachment-dependent or multimodal examples; we make no claim about those instances. Accuracy is computed from the final answer after the complete trajectory, and the reported Total is the unweighted mean over the six benchmarks. Macro averages, differences, and drops are computed from exact correct counts before values are rounded to one decimal.

Baselines. We first report two untrained reference conditions. *Question-only* asks Qwen3.5-9B to answer directly from the original question, without search calls or an accumulated tool trajectory. *Qwen3.5-9B search agent* gives the same model the evaluation-time search tools, prompt, context limit, and call budget, without task-specific optimization. We then compare with SFT on accepted actions and answer-blind GRPO, trajectory-on-policy distillation (OPSD), and privileged π-Distill (Shao et al., 2024; Zhao et al., 2026; Penaloza et al., 2026). When the objective permits, these learned baselines reuse the same backbone, public-state distribution, replay mixture, and update-token budget. OPSD obtains teacher targets along Student-sampled continuations, whereas `MIST-OPD` distills only Guide-proposed repairs admitted by matched public-execution gating. These comparisons align public states and Student update-token budget; the best-of-N control additionally matches continuation calls. They do not match Guide proposal, three-view scoring, or Guide-update compute.

Implementation and evaluation protocol. The Student and Guide use learning rates 2×10^{-7} and 1×10^{-7}, respectively, with 8,192-token sequences and eight source-balanced replay anchors per accepted state. We run one deterministic unshuffled pass within each of four rounds, in the order Student, live Guide, then Guide EMA; the EMA decay is 0.9. The main Student loss weights are 0.25 for controlled distillation, 0.2 for the selected action, 1.0 for replay, 0.5 for the ordinal hierarchy, and 0.1 for public help prediction. Hyperparameters and checkpoint selection are fixed before reading the complete evaluation results. At test time, every method in the primary 9B and 4B comparisons receives the same question, tool schema, context limit, model-call budget, search backend, and answer normalization. No Guide, answer, hint, intervention label, outcome, or teacher logit is available.

3.2 MAIN RESULTS

Table 2 compares `MIST-OPD` with prompting, outcome optimization, and distillation baselines under the same answer-blind agent interface. We highlight three findings.

Table 2: Main results (accuracy, %). BC: BrowseComp; BC+: BrowseComp-Plus; HPQA: HotpotQA. Avg. is the unweighted mean over all six benchmarks. Every method is answer-blind at evaluation.

Method	BC	BC+	FRAMES	GAIA	HLE	HPQA	Avg.
Qwen3.5-9B search agent	20.9	30.1	60.8	50.4	24.5	69.2	42.7
Qwen3.5-9B question-only	11.5	20.4	51.2	40.2	15.8	59.4	33.1
SFT on accepted a^+	22.6	38.0	67.2	54.3	25.9	77.4	47.6
GRPO, no answer	23.8	34.0	66.4	49.6	27.1	75.1	46.0
OPSD	24.6	37.5	69.2	52.0	27.0	76.3	47.7
π-Distill	25.1	38.2	68.7	53.5	27.4	77.0	48.3
MIST-OPD Student	**29.6**	**42.7**	**72.9**	**58.3**	**31.7**	**81.3**	**52.8**

Table 3: Scale transfer with Qwen3.5-4B (accuracy, %). Avg. is the macro-average over the five reported benchmarks.

Method	BC	BC+	FRAMES	GAIA	HPQA	Avg.
Qwen3.5-4B search agent	9.0	18.2	48.2	38.6	58.4	34.5
OPSD	12.2	20.8	49.3	40.2	59.4	36.4
π-Distill	12.9	21.6	49.6	40.2	59.8	36.8
MIST-OPD Student	**17.5**	**26.3**	**53.4**	**44.1**	**63.4**	**40.9**

MIST-OPD improves guide-free search across all six benchmarks. The evolved Student reaches 52.8% macro accuracy and outperforms the strongest distillation baseline on every benchmark. The gain ranges from 4.2 to 4.7 points and is therefore not driven by one favorable domain. The same Student also outperforms the untrained search agent by 10.1 points and the question-only condition by 19.7 points.

Controlled teaching outperforms ordinary distillation. The strongest distillation baseline, π-Distill, obtains 48.3%, while MIST-OPD reaches 52.8%. The advantage is remarkably stable across domains, ranging from 4.2 to 4.7 points. This comparison uses the same public states and Student update-token budget, testing the combined effect of first-passing-tier selection and controlled residual projection rather than simply exposing the Student to a privileged teacher distribution.

Accepted-action replay alone does not reproduce the gain. SFT uses the same full-parameter setting and accepted recovery actions with standard supervised learning, but omits intervention ranking, controlled residual learning, and Guide–Student co-evolution. Its 47.6% macro accuracy, compared with 52.8% for MIST-OPD, shows that the gain does not follow from replaying accepted trajectories alone. Answer-blind GRPO, OPSD, and π-Distill reach 46.0%, 47.7%, and 48.3%, respectively.

Technical failures remain in the fixed evaluation denominator. A continuation-call-matched best-of-N control reaches 50.3% macro accuracy, 2.5 points below MIST-OPD (Appendix Table 9).

3.3 Generalization across scale and domain

Scaling to a 4B backbone. We repeat the complete training pipeline with Guide and Student initialized from the same Qwen3.5-4B checkpoint. The intervention ladder, acceptance rule, replay sources, and answer-blind evaluation interface remain unchanged. MIST-OPD exceeds π-Distill across all five reported benchmarks and by 4.1 points on their macro-average (Table 3).

Transfer to repository-level code agents. We next replace the search answer with a verified reference patch and test evidence. The Student observes only the issue, repository state, public test output, and tool history; the Guide uses the patch to propose answer-free localization, repair, and test actions. We train on SWE-smith-style tasks and evaluate on SWE-bench Verified and SWE-

Table 4: Repository-level code-agent transfer (resolve rate, %). All methods use the same guide-free coding-agent scaffold at evaluation.

Method	SWE-bench Verified	Multilingual	Avg.
Base, Qwen3.5-9B	53.2	39.7	46.4
GRPO, no answer	53.8	41.3	47.6
π-Distill	54.6	42.0	48.3
MIST-OPD Student	**58.0**	**46.3**	**52.2**

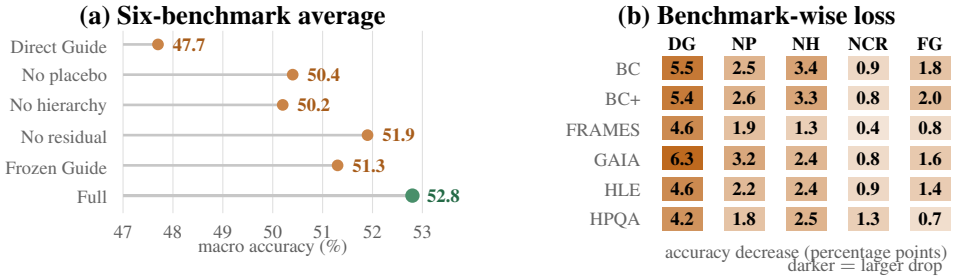


Figure 2: Component ablations. DG omits public gating; NP the placebo; NH the hierarchy; NCR the controlled residual; and FG Guide updates.

bench Multilingual (Yang et al., 2025; Jimenez et al., 2024). At deployment the patch, Guide, and intervention ladder are removed. MIST-OPD reaches 52.2% average resolve rate, exceeding π-Distill by 3.9 points (Table 4).

4 Further Analysis

4.1 Ablation analysis

We ablate public gating, the placebo, hierarchy, controlled residual, and Guide updates under the same answer-blind evaluation. Figure 2 shows macro and per-benchmark effects; Appendix D reports exact scores.

Screening and controlled distillation are complementary. Direct Guide distillation without matched screening reaches 47.7%; removing the placebo reaches 50.4%, gold-only distillation reaches 51.9%, and the full MIST-OPD reaches 52.8%. These results indicate that the gain arises from combining matched repair selection with anchored controlled distillation, rather than from either component alone.

The remaining components also improve the point estimate. The binary Abstain/Help variant and a Frozen Guide reach 50.2% and 51.3%, respectively. Removing the ladder costs 3.4/3.3 points on the two BrowseComp sets, while freezing the Guide hurts most where later failures require new proposals.

4.2 Does the Guide learn a less specific intervention?

We audit 45 accepted construction states and nine separate abstention anchors. Each arm uses two matched held-out continuations for each of 45 audit states (90 outcomes); none is used to choose the tier.

The selected boundary is graded and nearly maximally effective. Levels $L_0, \ldots, L_5$ occur $1, 11, 11, 9, 6$, and 7 times. The selected first-passing tier reaches 32/90 (35.6%), versus 11/90 with no guidance (12.2%), 14/90 with placebo (15.6%), 16/90 with the adjacent weaker tier (17.8%), and 35/90 with L_5 (38.9%). Thus selection improves over all controls while retaining most of maximal help. These 45 states come only from 2WikiQA and MuSiQue; evaluation questions are excluded. Repeat-count sensitivity at fixed $m = 1$ gives acceptance-set Jaccard scores of 65.3/61.7% for

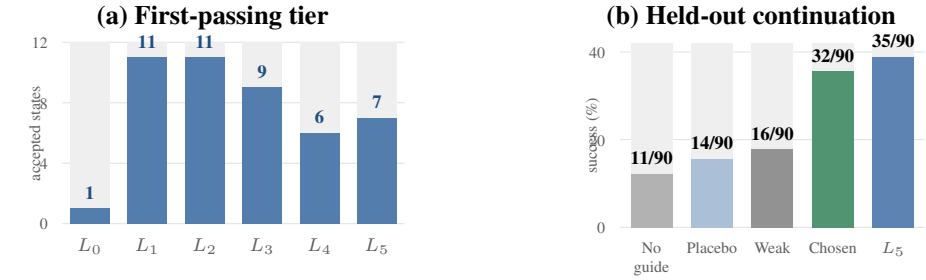


Figure 3: Held-out intervention diagnostics on 45 construction-only states (two matched continuations per state and arm). Chosen substantially outperforms the three controls while retaining most of the effectiveness of L_5.

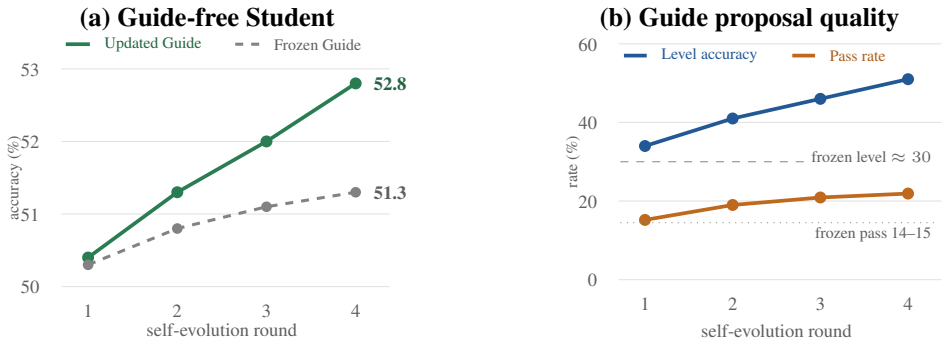


Figure 4: Four-round self-evolution. Only the updated Guide keeps improving the guide-free Student, proposal acceptance, and level accuracy; Frozen Guide plateaus at 51.3%, 14–15%, and about 30%, respectively.

$R = 5/10$; the same tier is selected on 57.1/55.3% of the corresponding unions (Appendix Table 8). The acceptance sets are not nested.

4.3 Does adapting the Guide improve the alternating system?

A Frozen Guide retains fresh Student states, the matched test, replay, and update budget, but not an adapting proposal policy. The arms are nearly equal after round 1 (50.4% versus 50.3%); by round 4, full reaches 52.8% while frozen stops at 51.3%. The live Guide's proposal pass rate rises 15.2%–21.9%. Its tier-prediction accuracy—the fraction of accepted states whose predicted intervention level matches the gate-selected first-passing tier $k^\star$—rises 34%–51%, whereas the frozen Guide remains at 14–15% and about 30%, respectively. Round-wise denominators and accepted counts appear in Appendix E. Fixed-state continuation measures immediate proposal utility, while the matched one-update comparison measures the end-to-end value of each Guide's accepted set (Appendix Table 10). The latter includes both supervision quality and accepted-set size.

The Guide learns both when and what to teach. Among Guide proposals before public gating, L_5 falls from $69/150$ (46.0%) in round 1 to $49/169$ (29.0%) in round 4. Among the 45 accepted Round-1 states, $7/45$ (15.6%) use L_5. The evolving Guide therefore becomes more selective, not merely more interventionist, and relies less often on the maximally explicit route as its level prediction improves.

The gain is not an artifact of changing cohort size. Table 5 separates the num-

Table 5: Round-wise self-evolution. Pass is measured over valid states; accuracy is the guide-free six-benchmark macro.

Round	Questions	Valid	Accepted	Pass	Full	Frozen
1	384	296	45	15.2%	50.4	50.3
2	360	268	51	19.0%	51.3	50.8
3	340	244	51	20.9%	52.0	51.1
4	320	228	50	21.9%	52.8	51.3
Total	1,404	1,036	197	–	–	–

ber of questions, technically valid states, and accepted boundaries. Although the fresh cohort decreases from 384 to 320 questions, the absolute number of accepted boundaries remains between 45 and 51 while the pass rate rises from 15.2% to 21.9%. The Frozen Guide sees the same changing Student states and update budget but remains near 14–15% acceptance and 30% level accuracy. The divergence therefore supports the end-to-end benefit of adapting the proposal policy under the alternating procedure. The fixed-state controls distinguish immediate proposal utility from end-to-end downstream updating; the latter retains each Guide's own accepted-set size. Construction costs appear in Appendix E.3; deployment adds no privileged calls.

4.4 What happens when no intervention is accepted?

Rejection is not equivalent to abstention. An abstention example requires the unguided continuation to succeed and remain non-inferior to every assisted arm; it therefore teaches the Guide that help was unnecessary. A state is instead discarded when no level beats both no guidance and its placebo, when a stronger level reverses an earlier gain, or when a proposed action is malformed or leaks an answer alias. Treating these failures as abstention would confound "the Student does not need help" with "the Guide did not produce valid help."

This conservative distinction reduces the number of usable boundaries but keeps the learning signal interpretable. Accepted states provide evidence for both the action and the first-passing tier; verified unguided successes supervise when to abstain; all other cases contribute neither privileged targets nor negative labels. The method therefore improves by accumulating accepted repairs, rather than by forcing every answer-conditioned proposal into the Student. Appendix F details the rejection taxonomy and remaining failure modes.

5 Related Work

Search-agent learning and self-improvement. Outcome optimization and process rewards improve behavior represented in sampled trajectories, while tool-using agents enlarge the space of trajectories that can be collected (Schulman et al., 2017; Shao et al., 2024; Lightman et al., 2024; Feng et al., 2025; Nakano et al., 2021; Yao et al., 2022). Self-training and reflection reuse the model's own solutions, rationales, or critiques (Zelikman et al., 2022; Madaan et al., 2023; Shinn et al., 2023); they improve execution within the current policy's support but do not directly identify a missing recovery action. Hindsight and counterfactual methods instead diagnose failures from subsequent observations or outcomes (Liu et al., 2026a; Bonagiri et al., 2026). `MIST-OPD` shares this repair-oriented goal, but uses the private answer only during construction to propose candidate actions and admits them through matched public execution.

Privileged and on-policy distillation. Learning using privileged information separates richer training signals from the inputs available at deployment (Vapnik & Vashist, 2009). In language models, OPSD and π-Distill transfer training-only information at Student-relevant states (Zhao et al., 2026; Penaloza et al., 2026). Closely related variants refine this transfer through evidence guidance, dual routing, outcome-guided steering, anchored residuals, or token-wise reliability weighting (Lazaridis et al., 2026; Yu et al., 2026; Yang et al., 2026; Zhang, 2026; Liu et al., 2026b). These methods primarily determine how a privileged signal is transferred once it is available. `MIST-OPD` focuses on the preceding discovery and selection problem: action-level matched screening identifies a graded first-passing repair, controlled residual projection transfers it, and alternating Guide–Student updates revisit fresh Student states. Deployment retains only the answer-blind Student.

6 Conclusion

`MIST-OPD` converts training-only answers into answer-free recovery actions. A same-backbone Guide proposes graded interventions; matched public execution screens them against no guidance, a weaker tier, and a placebo, while controlled residual projection transfers the first-passing repair. Alternating updates revisit fresh Student states; deployment retains only the answer-blind Student. The 9B Student reaches 52.8% macro accuracy across six benchmarks, exceeding the strongest distillation baseline by 4.4 points, with gains transferring to 4B and code agents.

AI USE DISCLOSURE

Generative AI tools were used to assist with language editing, document organization, and LaTeX formatting. All AI-assisted text was reviewed and revised by the authors, who verified the claims, citations, experimental results, and reported numbers and take responsibility for the final content.

REFERENCES

Akash Bonagiri, Devang Borkar, Gerard Janno Anderias, Setareh Rafatirad, and Houman Homayoun. CausalFlow: Causal attribution and counterfactual repair for LLM agent failures. *arXiv preprint arXiv:2605.25338*, 2026. URL https://arxiv.org/abs/2605.25338.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.10978.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. Reinforced self-training (ReST) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023. URL https://arxiv.org/abs/2308.08998.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, 2020. URL https://arxiv.org/abs/2011.01060.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.06770.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. *arXiv preprint arXiv:2409.12941*, 2024. URL https://arxiv.org/abs/2409.12941.

Aristotelis Lazaridis, Dylan Bates, Aman Sharma, Brian King, Vincent Lu, and Jack FitzGerald. EDGE-OPD: Internalizing privileged context with evidence guided on-policy distillation. *arXiv preprint arXiv:2605.23493*, 2026. URL https://arxiv.org/abs/2605.23493.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Haoran Liu, Yuwei Zhang, Xiyao Li, Bohan Lyu, and Jingbo Shang. HERO: Hindsight-enhanced reflection from environment observations for agentic self-distillation. *arXiv preprint arXiv:2606.11559*, 2026a. URL https://arxiv.org/abs/2606.11559.

Xiaogeng Liu, Xinyan Wang, Yingzi Ma, Yechao Zhang, and Chaowei Xiao. When are teacher tokens reliable? position-weighted on-policy self-distillation for reasoning. *arXiv preprint arXiv:2605.21606*, 2026b. URL https://arxiv.org/abs/2605.21606.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2303.17651.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2311.12983.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021. URL https://arxiv.org/abs/2112.09332.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. *arXiv preprint arXiv:2602.04942*, 2026. URL https://arxiv.org/abs/2602.04942.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025. URL https://arxiv.org/abs/2501.14249.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2303.11366.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10: 539–554, 2022. URL https://arxiv.org/abs/2108.00573.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009. doi: 10.1016/j.neunet.2009.06.042.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. BrowseComp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025. URL https://arxiv.org/abs/2504.12516.

John Yang, Kilian Lieret, Carlos E. Jimenez, Alexander Wettig, Kabir Khandpur, Yanzhe Zhang, Binyuan Hui, Ofir Press, Ludwig Schmidt, and Diyi Yang. SWE-smith: Scaling data for software engineering agents. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://arxiv.org/abs/2504.21798.

Yuxiao Yang, Xiaoyun Wang, and Weitong Zhang. OGLS-SD: On-policy self-distillation with outcome-guided logit steering for LLM reasoning. *arXiv preprint arXiv:2605.12400*, 2026. URL https://arxiv.org/abs/2605.12400.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. URL https://arxiv.org/abs/1809.09600.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022. URL https://arxiv.org/abs/2210.03629.

Xinlei Yu, Gen Li, Qingyi Si, Guibin Zhang, Yuqi Xu, Congcong Wang, Shuai Dong, Kaiwen Tuo, Xiangyu Zeng, Kaituo Feng, Qunzhong Wang, Yang Shi, Xiaobin Hu, Xiangyu Yue, Jiaqi Wang, and Shuicheng Yan. DOPD: Dual on-policy distillation. *arXiv preprint arXiv:2606.30626*, 2026. URL https://arxiv.org/abs/2606.30626.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2203.14465.

Wenhao Zhang. Beyond absolute imitation: Anchored residual guidance for privileged on-policy distillation. *arXiv preprint arXiv:2606.10385*, 2026. URL https://arxiv.org/abs/2606.10385.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026. URL https://arxiv.org/abs/2601.18734.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes. *arXiv preprint arXiv:2605.11182*, 2026. URL https://arxiv.org/abs/2605.11182.

A TRAINING ALGORITHM

1. Freeze source-balanced matched states, unguided, adjacent, and placebo actions, the repeat plan, reference-policy replay anchors, and leakage masks.
2. For each state, compute the matched-evidence reliability weight w_i.
3. Run the lagged Guide EMA under gold, placebo, and public conditions.
4. Construct the projected, answer-blind anchored target in Eqs. 3–4.
5. Update the Student with controlled KL, chosen-action CE, the ordinal help policy, and eight replay anchors.
6. Update the live Guide with privileged action contrast, first-passing-tier supervision, abstention, and anchoring.
7. Update the Guide EMA exactly once and continue to the next state.
8. Publish both role checkpoints for fresh, ID-disjoint construction; evaluate only the Student with Guide mode disabled.

B INFORMATION-FLOW CONTRACT

Table 6: Private information is restricted to construction and teacher views.

Stage	Public state	Private answer	Outcomes
Intervention proposal	Read	Read	Forbidden
Repeated execution	Read	Forbidden	Write
Boundary selection	Read	Scoring only	Read
Teacher update	Read	Read	Evidence weight only
Student update	Read	Forbidden	Evidence weight only
Evaluation	Read	Forbidden	Write for final scoring

C PROMPT AND INTERACTION PROTOCOL

Serialized public state. At each construction decision, both policies share $x_t = (q, \tau_t, z_t, o_t)$, comprising the question, last public tool, current tool argument, and latest public observation. The serialized argument and observation are capped at 500 and 4,200 characters, respectively. Only the Guide additionally receives the private target answer and verified aliases.

Guide proposal prompt. The fixed construction prompt is:

```
You are a privileged proposal policy for a
search agent.  Given the question, current
public search state, latest public observation,
and private target answer, output strict
JSON: {"decision":"ABSTAIN|L0|L1|L2|L3|L4|L5",
"hint":"..."}.  Choose the least informative candidate
tier expected to enable a useful next action.  Never
output the final answer or a verified alias.
```

For a forced counterfactual arm, the decision is fixed to L_k and the Guide generates only a tier-conforming hint. Malformed JSON, a mismatched tier, an empty non-abstention hint, an overlength hint, or an answer/alias match invalidates the proposal.

Public tier prediction and action rendering. A construction-only public prompt requests exactly one label from $\{\text{ABSTAIN}, L_0, \ldots, L_5\}$; it sees no answer, Guide text, or outcome. This target trains the ordinal head but does not route deployment. For search tasks, an accepted answer-free hint is rendered through the ordinary agent interface into one legal next call, `search(query)` or `batch_search(queries)`. The renderer sees the public state and hint and emits a query of at

Table 7: Complete ablation scores (accuracy, %). DG directly distills Guide proposals without public gating; NP removes the placebo; NH removes the hierarchy; NCR uses gold-only distillation; and FG freezes the Guide.

Dataset	DG	NP	NH	NCR	FG	Full
BrowseComp	24.1	27.1	26.2	28.8	27.8	**29.6**
BrowseComp-Plus	37.3	40.1	39.4	41.8	40.6	**42.7**
FRAMES	68.3	71.0	71.6	72.6	72.1	**72.9**
GAIA-text	52.0	55.1	55.9	57.5	56.7	**58.3**
HLE-text	27.1	29.5	29.4	30.9	30.4	**31.7**
HotpotQA	77.1	79.5	78.8	80.0	80.6	**81.3**
Macro	47.7	50.4	50.2	51.9	51.3	**52.8**
Drop from full	-5.1	-2.4	-2.5	-0.8	-1.4	**0.0**

Table 8: Repeat-count sensitivity on the frozen Round-1 construction slice with $m = 1$. Because additional repetitions can reveal either a win or a reversal, the accepted sets are not nested. The final column rebuilds supervision and retrains.

R	Accepted / valid	Pass (%)	Jaccard vs. $R = 2$	Same tier / union	Macro (%)
2	45 / 296	15.2	–	–	52.8
5	36 / 296	12.2	32 / 49 (65.3%)	28 / 49 (57.1%)	52.4
10	31 / 296	10.5	29 / 47 (61.7%)	26 / 47 (55.3%)	52.1

Table 9: Continuation-call-matched best-of-N control (accuracy, %). For each frozen public state, BoN samples $N = 26$ continuations from the same Student policy, matching `MIST-OPD`'s 26 continuation calls per state. This does not match proposal generation, lagged-Guide scoring, or Guide-update compute. The same privileged construction scorer selects one continuation, which is transferred with the same Student update and replay; evaluation is guide-free.

Dataset	π-Distill	BoN	`MIST-OPD`
BrowseComp	25.1	27.0	**29.6**
BrowseComp-Plus	38.2	40.1	**42.7**
FRAMES	68.7	70.6	**72.9**
GAIA-text	53.5	55.1	**58.3**
HLE-text	27.4	29.4	**31.7**
HotpotQA	77.0	79.3	**81.3**
Macro	**48.3**	**50.3**	**52.8**

most 16 words, without an answer string. The concrete tool call, rather than the Guide prose, is the Student target. Repository-level code tasks use the same separation with the public coding scaffold and test interface.

Scoring. Each continuation receives only a terminal bit $c(u) \in \{0, 1\}$. Search tasks use the benchmark's normalized-answer scorer; code tasks use public-test success. Timeouts, malformed outputs, and tool or provider failures remain in the frozen denominator and receive zero. In a paired comparison, u beats v iff $c(u) = 1$ and $c(v) = 0$; equal outcomes are ties. No answer-evidence heuristic enters the gate.

D Guide-Behavior and Safety Diagnostics

Changing the repeat count without retraining leaves the already trained Student at 52.8%; the last column rebuilds supervision and retrains. Terminal performance is more stable than boundary membership.

Table 10: Fixed-state Guide controls. Both panels start from the same Round-3 Student and the same 228 Round-4 public states. Panel (b) uses each Guide's complete accepted set (50 versus 34 states), so it measures the combined effect of supervision quality and quantity rather than per-example quality.

(a) Guided continuation without Student retraining		
Condition	Success	Rate (%)
No guide	49 / 228	21.5
Frozen Guide	56 / 228	24.6
Updated Guide	71 / 228	**31.1**
(b) One matched Student update, then guide-free evaluation		
Update source	Six-domain macro	Change from 52.0
Frozen Guide	51.9	-0.1
Updated Guide	**52.5**	$+0.5$

Table 11: Frozen construction funnel. Counts before private conditioning are reported separately from the accepted-boundary audit.

Stage	Total
Public training inventory	12,000
Round-1 frozen cohort	384
Round-1 technically valid states	296
Round-1 accepted audit states	45
Independent abstention anchors	9

The gate accepts 34/228 and 50/228 states for the Frozen and Updated Guides, respectively. Among the 28 states accepted by both, their first-passing tiers agree on 19 (67.9%).

E Data Construction and Reproducibility

E.1 Construction funnel

Table 11 separates question inventory, public-state collection, and the mechanism audit. This distinction prevents the number of accepted boundaries from being confused with the number of evaluation questions. In every round, answers are read only after that round's disjoint question cohort and public states have been frozen.

The construction ledger stores a normalized question hash, a token multiset, the source split, the serialized public-prefix hash, and frozen branch identifiers. Exact matches and token-Jaccard overlap of at least 0.8 against the evaluation universe are rejected. Near-duplicate public histories are assigned as a group so that two serializations of the same trajectory cannot cross a construction/evaluation boundary. Failed technical executions remain in the frozen denominator and are never silently replaced.

E.2 Optimization configuration

One pass traverses the round-specific accepted set; it does not denote a single optimizer step. With microbatch one, the four round-specific accepted sets yield 197 cumulative Student updates. Each update applies token-level losses to one accepted repair and eight source-balanced replay records (with action/replay caps of 256/2,048 tokens); 197 is therefore the number of accepted states, not the number of supervised tokens.

The Guide and Student are separately trainable copies initialized from the same checkpoint. All reported 9B, 4B, and code experiments optimize both copies with full-parameter updates; no LoRA or other parameter-efficient adapters are used. The stated losses and update order apply throughout. Only the Student checkpoint is eligible for deployment or task evaluation.

Table 12: Optimization and runtime configuration used by the joint Guide–Student implementation. DPO is disabled in the main method.

Quantity	Value
Student / Guide learning rate	2×10^{-7} / 1×10^{-7}
Self-evolution rounds	4
Parameterization	Full-parameter Student and Guide updates; no LoRA or adapters
Training passes per round and ordering	1 complete pass, no shuffle; Student, live Guide, then Guide EMA
Guide EMA decay	0.9
Maximum sequence length	8,192 tokens
Action / replay supervision caps	256 / 2,048 tokens
Microbatch and accumulation	1; eight replay terms per accepted state
Precision / gradient control	bf16 autocast; fp32 parameters and AdamW states; checkpointing; norm clip 1.0
Residual scale / coordinate clip	0.25 / 2.0
Student distillation / chosen CE / replay weights	0.25 / 0.2 / 1.0
Guide level+hint / rank / control / replay weights	0.5 / 1.0 / 1.0 / 0.05
Student hierarchy / public-help weights	0.5 / 0.1
Student action-DPO weight	0.0
Repeated intervention test	2 matched repeats, 13 branches per state
Acceptance threshold	At least one win and zero reversals against each control

E.3 Construction cost and inference efficiency

Privileged teaching is paid for during construction, not deployment. With six levels, one placebo per level, and one shared no-guide arm, a selected state has 13 branches. The frozen online configuration uses $R = 2$ matched repetitions, so the worst-case interaction budget is 26 continuation episodes per state. A reference round with three states from each of the two construction sources therefore contains $2 \times 3 \times 13 \times 2 = 156$ episodes. After gating, each accepted state requires three lagged-Guide views (gold, placebo, and public), one Student training record, one live-Guide training record, and eight replay anchors. At inference, the deployed Student retains the evaluation protocol's public inputs, search API, context limit, model-call count, and tool budget; the Guide, answer, and counterfactual branches are removed.

F Failure Modes and Scope of the Mechanism

F.1 Limitations

Matched intervention effects are local to the continuation policy and search environment used during construction; a gate-admitted next action need not remain optimal after a large policy shift. Repeated rollouts and conservative updates reduce this mismatch but increase construction cost.

The intervention ladder imposes an ordinal structure, but neighboring levels need not be monotone at every state. Rejecting directional reversals favors precision over data efficiency, and a source-matched placebo cannot remove every semantic confound. Likewise, residual and token projection suppress direct token-level privilege transfer but do not formally exclude every semantic encoding of an answer; our evidence is therefore answer-blind, disjoint evaluation rather than a privacy guarantee. Entity-centric search is especially suitable for backward proposal. Code repair establishes one form of domain transfer, while mathematical and multimodal settings remain outside the present evidence. The 4B analysis also lacks HLE-text, so its evidence is limited to the five reported benchmarks and their matched learned baselines.

F.2 REJECTED STATES

The gate can reject a state for four scientifically different reasons. First, the public policy may already solve it; only independently verified cases of this kind become abstention anchors. Second, no level may improve over the unguided action, its adjacent weaker action, and the placebo. Third, a stronger level may reverse a previous gain, violating the ordinal teaching assumption. Fourth, a proposal may be invalid because its action is malformed or contains the answer or a verified alias. Only the first case is a positive abstention label; treating the other three as abstention would teach the Guide that failed construction means help was unnecessary. When no stable boundary exists, `MIST-OPD` forgoes the example rather than distilling an unverified privileged preference.