REASONFORK: ROUTING SMALL AND LARGE MODELS BY FUTURE CONSEQUENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Collaboration between small and large language models can make reasoning more efficient by letting a small draft model handle routine steps and consulting a larger teacher selectively. The core challenge is to identify *reasoning forks* where replacing a draft action with a teacher action can improve task success within a limited step budget while controlling teacher costs. Existing routers based on draft confidence or process scores evaluate the draft's proposed action, but do not explicitly estimate how replacing it with a teacher action would affect subsequent task completion. To address this limitation, we introduce **ReasonFork**, a lightweight router that learns when to intervene by comparing the *future consequences* of draft and teacher actions. During training, the draft and teacher each propose an action from the same decision state. An offline evaluator compares the two successors under the same remaining step budget, and the difference in task outcomes provides supervision for the router. At inference, ReasonFork estimates the benefit of replacement using only information available before the teacher call and weighs it against the cost of calling the teacher. At each intervention, the teacher provides one action, after which the draft resumes the task. Experiments on five benchmarks for symbolic reasoning and interactive tasks demonstrate the effectiveness of ReasonFork. On reasoning tasks, the reported ReasonFork configurations match or exceed the accuracy of Always Teacher with approximately 72% lower teacher cost, and achieve absolute accuracy gains of 5.4–6.4% over strong baselines at comparable teacher cost.

1 INTRODUCTION

Letting a small draft model handle most reasoning steps and seek help from a larger teacher at key moments offers a way to balance task performance and computational efficiency. The timing of intervention also affects downstream cost: different actions can induce trajectories of different lengths, and each additional step requires model generation (Kwon et al., 2023) and may involve tool calls (Gao et al., 2025) or environment interactions (Yao et al., 2023). Therefore, efficient collaboration requires choosing when to intervene so that a teacher action helps the draft complete the task in fewer steps while keeping teacher costs low. The core challenge is to identify intervention points at which replacing a draft action can enable the draft to reach the goal within a limited step budget.

Existing methods commonly use confidence or process scores to decide whether the draft needs help at the current step, while learned sequential policies account for task outcomes and calling costs over a trajectory (Lee et al., 2026; Kapoor et al., 2026). However, a score for the draft's proposed action does not directly measure the benefit of replacing it with a teacher action. An uncertain draft action may have no better teacher alternative, while a confident action may lead to unnecessary steps or prevent task completion. Effective teacher routing therefore requires identifying forks where a teacher replacement can improve task completion within the remaining budget.

Figure 1 illustrates how intervention timing affects whether a task can be completed within the remaining budget. The draft confidently proposes a legal action, but accepting it leaves too few steps to reach the goal. A confidence-based router calls the teacher only at the next uncertain step, when it is already too late. ReasonFork instead replaces the draft action with a teacher action at the fork, leaving enough steps for the draft to reach the goal. The example raises a central question:

Which fork is worth a teacher call?

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.

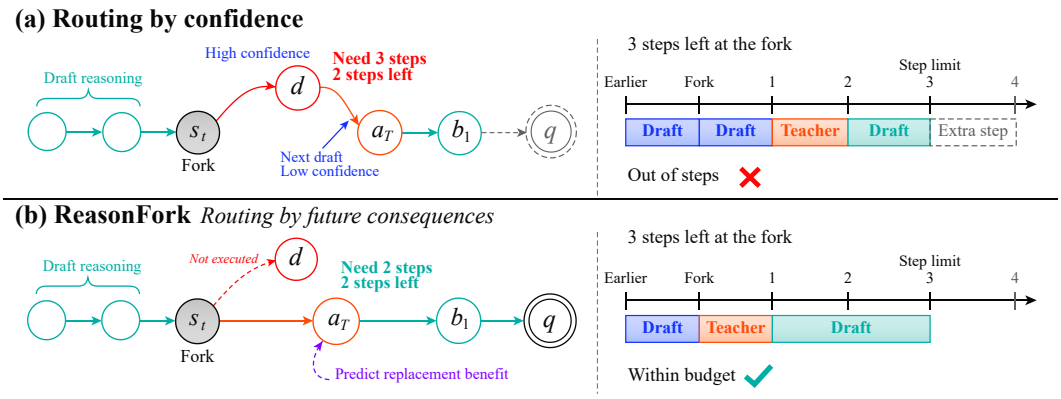


Figure 1: **Route at the consequential fork.** **(a)** A confidence-based router accepts a high-confidence draft action that causes a detour. Intervention at the next uncertain step is too late to finish within budget. **(b)** ReasonFork replaces the action at the fork, allowing the draft to finish within budget.

We introduce ReasonFork, which learns to identify *reasoning forks* where a single teacher action can change the outcome of a task. To construct offline supervision, the draft and teacher each propose an action from the same legal action set at the same decision state. An offline evaluator then compares the outcomes of the two successors under the same remaining action budget. This paired comparison yields repair, tie, or harm labels based on the task's evaluation criterion and available offline feedback. We use these labels to train a lightweight router that estimates expected replacement benefit using only inputs available at the current decision point.

At inference, the router calls the teacher when the estimated benefit justifies the calling cost. The teacher supplies one action, which is verified before control returns to the draft. Otherwise, the draft's proposal is retained. Since future paths are evaluated only to construct offline supervision, online routing requires no search over subsequent action sequences.

We derive a local routing rule that weighs the expected gains and losses from replacing the draft action against calling cost, using only information available before the teacher call. We further characterize when legal actions differ in whether the goal remains reachable within the remaining budget, and give sufficient conditions for a teacher replacement to preserve a feasible path. Experiments cover symbolic reasoning on RuleTaker, ProofWriter, and ProverQA, alongside interactive tasks in ALFWorld and WebShop. Across the five proof settings, the ReasonFork configurations achieve an average accuracy of 95.4%, reducing the average error rate by 27.2% compared with Always Teacher. Teacher cost is approximately 72% lower on average relative to Always Teacher under the corresponding settings.

Our main contributions are:

- We introduce ReasonFork, a lightweight router that learns which reasoning forks warrant intervention from differences in the future outcomes of draft and teacher actions.
- We derive an intervention rule that maximizes expected local utility given pre-call information, and establish sufficient conditions under which legal replacements preserve finite-horizon feasibility.
- We evaluate ReasonFork across five benchmarks and test router transfer to other benchmarks without retraining. Experiments demonstrate the effectiveness and efficiency of our method.

2 Related Work

Model allocation. Language-model cascades compose model calls through structured control flow (Dohan et al., 2022). FrugalGPT and AutoMix use response quality estimates to decide whether to escalate a request to another model (Chen et al., 2024; Aggarwal et al., 2024). Hybrid LLM and RouteLLM select a model using predicted difficulty or learned preferences (Ding et al., 2024; Ong et al., 2025), while RouterEval and LLMRouterBench evaluate routing across models and tasks (Huang et al., 2025; Li et al., 2026). Stepwise routing adjusts model allocation during reasoning. STEER uses draft confidence (Lee et al., 2026). TRIM uses process scores and studies threshold and reinforcement learning policies under partial observation (Kapoor et al., 2026). Other methods learn sequential allocation policies that balance task success and calling cost over the full reasoning trajectory under budget constraints (Si et al., 2026; Zhang et al., 2026).

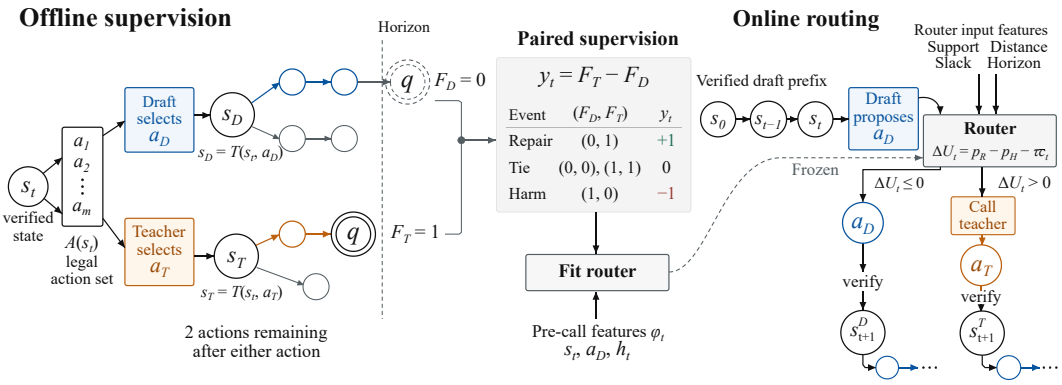


Figure 2: **Overview of ReasonFork with exact feasibility and structural features. Left:** Compare draft and teacher successors under the same remaining budget. **Center:** Use feasibility differences as paired training labels. **Right:** Balance replacement benefit against calling cost; after a teacher action, the draft resumes.

Search and future evaluation. Self-consistency samples complete reasoning paths and aggregates their answers (Wang et al., 2023). Tree of Thoughts and Graph of Thoughts organize intermediate reasoning into branching or interconnected structures (Yao et al., 2023; Besta et al., 2024). RAP and LATS use simulated continuations and feedback to guide online search (Hao et al., 2023; Zhou et al., 2024), while TS-LLM learns a value function for tree search (Wan et al., 2024). Local Branch Routing limits lookahead to a small tree and learns which branch to retain (Yin et al., 2026). For external retrieval, Kim (2026) compares outcomes with and without search to learn whether retrieval is needed.

Executable and verifiable reasoning. PAL, Faithful CoT, Logic-LM, and LINC use interpreters or symbolic solvers (Gao et al., 2023; Lyu et al., 2023; Pan et al., 2023; Olausson et al., 2023). TacticZero, HyperTree Proof Search, and LeanDojo verify proof steps in formal environments (Wu et al., 2021; Lample et al., 2022; Yang et al., 2023). PlanBench evaluates action sequences in planning domains (Valmeekam et al., 2023). Learned verifiers assess complete solutions (Cobbe et al., 2021) or intermediate steps (Uesato et al., 2022; Lightman et al., 2024; Wang et al., 2024).

Unlike prior work, ReasonFork bases routing on how teacher replacement changes task completion, identifying *which reasoning fork warrants a teacher call*. The router learns to predict this replacement benefit from offline outcomes and weighs it against the calling cost without online search.

3 REASONFORK

REASONFORK has two stages. **Offline supervision** compares draft and teacher successors from the same state: improvements are REPAIR, deteriorations HARM, and equivalent outcomes TIE under the task's outcome criterion. **Online routing** uses pre-call inputs to estimate replacement benefit and allocate single-action teacher interventions, after which the draft resumes. Figure 2 and Algorithms 1–2 illustrate the exact-feasibility instantiation.

3.1 ACTIONS, OUTCOMES, AND THE TEACHER BUDGET

For proof tasks, the executor exposes a verified state s_t, remaining horizon h_t, legal action set $\mathcal{A}(s_t)$, and targets $q, \bar{q}$ (query and complement), without revealing the correct polarity. The draft policy π_{S} and teacher policy π_{L} select a_D, a_T from the same ordered action set. Executing a yields $T(s_t, a)$.

Let $D^\star_\cup(s)$ denote the minimum number of legal actions required to derive either target, with value ∞ when neither target is reachable. The exact post-action feasibility indicator is

$$F^\star_t(a) = \mathbf{1}[D^\star_\cup(T(s_t, a)) \le h_t - 1]. \tag{1}$$

For a complete unit-cost executor, $F^\star_t(a) = 1$ holds if and only if a legal completion exists within $h_t - 1$ further actions (Proposition 2, Appendix A). An *exact reasoning fork* admits two legal actions with different feasibility indicators. Exact feasibility is evaluated offline. Final success depends on the trajectory executed by the routed policy.

For routing policy ρ, define expected teacher cost $C(\rho) = \mathbb{E}_\rho[\sum_t z_t c_t]$. We seek

$$\max_\rho \ \mathbb{E}_\rho[\mathbf{1}\{\text{solve}\}] \quad \text{s.t.} \quad C(\rho) \le \beta C_{\mathrm{AT}}, \tag{2}$$

where z_t indicates a teacher call, c_t counts its realized prompt and completion tokens, and C_{AT} is the expected cost of Always Teacher on the same task distribution. Expectations cover tasks and policy-induced trajectories, and β specifies the relative budget. The trained router and threshold grid remain fixed during evaluation.

3.2 WHICH TEACHER CALLS HAVE VALUE?

Let $V_t(a)$ be the offline outcome value of successor $T(s_t, a)$. With a fixed tie tolerance $\eta \geq 0$, define

$$\Delta_t = V_t(a_T) - V_t(a_D), \qquad y_t = \mathbf{1}[\Delta_t > \eta] - \mathbf{1}[\Delta_t < -\eta]. \tag{3}$$

The labels $+1, -1, 0$ denote REPAIR, HARM, and TIE. A fixed utility mapping $u(\Delta_t)$ assigns positive value to repairs, negative value to harms, and zero to ties; it may retain outcome magnitude. Given pre-call information ϕ_t, the expected replacement value is

$$\mu(\phi_t) = \mathbb{E}[u(\Delta_t) \mid \phi_t]. \tag{4}$$

Outcome criteria, tie tolerances, and utilities are fixed before evaluation (Appendix B.2). For exact feasibility, $V_t(a) = F_t^\star(a)$, $\eta = 0$, and $u(\Delta_t) = \Delta_t$, so $y_t = F_T - F_D$ and $\mu(\phi_t) = p_R(\phi_t) - p_H(\phi_t)$, where $F_D = F_t^\star(a_D)$, $F_T = F_t^\star(a_T)$, and p_R, p_H are the conditional repair and harm probabilities. Draft risk $\Pr(F_D = 0 \mid \phi_t)$ cannot distinguish repairable failures from pairs of infeasible successors; paired outcomes also account for harm.

Proposition 1 (Conditional intervention rule)**.** *Let the cost proxy $\bar{c}_t = \max(\widehat{c}_t, \epsilon)$, $\epsilon > 0$, and its price $\tau_t \geq 0$ be determined by ϕ_t. Relative to retaining the draft, define local utility $U_t(z) = z\, u(\Delta_t) - \tau_t z \bar{c}_t$. Its conditional expectation is maximized by*

$$z_t^\star = \mathbf{1}\{\mu(\phi_t) > \tau_t \bar{c}_t\}, \tag{5}$$

with the keep-draft convention at equality.

Proof. Since $\bar{c}_t$ and τ_t are fixed given ϕ_t,

$$\Delta U_t = \mathbb{E}[U_t(1) - U_t(0) \mid \phi_t] = \mu(\phi_t) - \tau_t \bar{c}_t.$$

Calling is optimal when $\Delta U_t > 0$. Otherwise, the draft is retained. □

The proposition characterizes the optimal single-step choice for conditional expected utility, while Equation 2 defines the trajectory objective used for evaluation. To estimate replacement value, offline supervision pairs draft and teacher actions at states visited by the draft policy.

3.3 PREDICTING INTERVENTION VALUE BEFORE A CALL

ReasonFork separates value estimation from budget allocation. Learned estimators use paired outcomes, whereas structural estimators use fixed support-based scores. Both select single-action interventions from pre-call inputs.

For proof states, a support hypergraph provides features for predicting $F_t^\star$ differences. Reverse traversal from q and $\bar{q}$ identifies relevant literals in the union support cone $\mathcal{C}_\cup$ and assigns relaxed distance $\widetilde{d}_\cup(v)$. Approximate post-action slack is

$$\widetilde{\sigma}_t(a) = h_t - 1 - \widetilde{d}_\cup(\mathrm{head}(a)). \tag{6}$$

Cone membership measures relevance to a goal. Negative slack indicates that the relaxed distance exceeds the steps available after the current action. This slack is a feature rather than an exact feasibility certificate. The following theorem characterizes when exact feasibility differs across legal actions.

Theorem 1 (Reachability forks)**.** *Assume a sound, complete, monotone, unit-cost action system with upward-persistent executability. Let $d = D_\cup^\star(s_t)$ satisfy $0 < d < \infty$, and write $d_a = D_\cup^\star(T(s_t, a))$. For integer horizon $h_t \geq 1$, every legal action satisfies $d_a \in \{d-1, d\}$, and*

$$F_t^\star(a_D) \neq F_t^\star(a_T) \quad \Longleftrightarrow \quad h_t = d \ \text{ and } \ \{d_{a_D}, d_{a_T}\} = \{d-1, d\}. \tag{7}$$

Proof. Replay a shortest completion after a. Monotonicity and upward persistence preserve its steps, omitting conclusions already present, so $d_a \leq d$. Prefixing a to a shortest successor completion gives $d \leq 1 + d_a$. Integrality yields $d_a \in \{d-1, d\}$. In $F_t^\star(a) = \mathbf{1}[d_a \leq h_t - 1]$, all actions have label one when $h_t > d$ and zero when $h_t < d$. At $h_t = d$, the label is one exactly when $d_a = d - 1$. □

Algorithm 1: Offline paired supervision

Require: Calibration states, development tasks
Initialize paired dataset $\mathcal{D} \leftarrow \varnothing$;
for each draft state (s_t, h_t, a_D) **do**
  Extract ϕ_t^{str} and $\widehat{c}_t$;
  Teacher selects $a_T \in \mathcal{A}(s_t)$;
  $F_D \leftarrow F_t^\star(a_D)$;
  $F_T \leftarrow F_t^\star(a_T)$;
  $\boldsymbol{y_t \leftarrow F_T - F_D}$;
  Add $(\phi_t^{\mathrm{str}}, y_t)$ to $\mathcal{D}$;
end for
Fit $p_\theta(y \mid \phi^{\mathrm{str}})$ on $\mathcal{D}$;
Construct S using Eq. 8;
Select τ from the calibrated grid;
return Frozen S, τ;

Algorithm 2: Online teacher routing

Require: State s_0, horizon H, frozen S, τ
$t \leftarrow 0$, $h_0 \leftarrow H$;
while $h_t > 0$, unsolved, $\mathcal{A}(s_t) \neq \varnothing$ **do**
  Draft proposes $a_D \in \mathcal{A}(s_t)$;
  Read pre-call features ϕ_t;
  $a \leftarrow a_D$;
  if $\boldsymbol{S(\phi_t) > \tau}$ **then**
    Request teacher action a_T;
    $a \leftarrow a_T$ if valid, else a_D;
  end if
  Verify a, set $s_{t+1} \leftarrow T(s_t, a)$;
  $h_{t+1} \leftarrow h_t - 1$, $t \leftarrow t + 1$;
end while
return Final state s_t;

Within these proof systems, Theorem 1 generalizes Figure 1: a feasibility fork requires a tight remaining action horizon and exactly one of the paired actions to advance a shortest completion. The result characterizes exact feasibility. Relaxed slack remains a feature, and TIE need not imply equally easy continuations.

Appendix C presents training-free structural rules and their comparison with learned intervention values.

Outcome-value estimation. The router combines support relevance, distance, slack, and action-set statistics into ϕ_t^{str}. On RuleTaker, a multinomial logistic model predicts paired outcome probabilities and computes

$$S_t^{\mathrm{str}} = \frac{p_\theta(y = +1 \mid \phi_t^{\mathrm{str}}) - \lambda_{\mathrm{harm}} p_\theta(y = -1 \mid \phi_t^{\mathrm{str}})}{\max(\widehat{c}_t, \epsilon)}. \tag{8}$$

The numerator estimates $\mu(\phi_t)$ with repair utility 1 and harm utility $-\lambda_{\mathrm{harm}}$; $\lambda_{\mathrm{harm}} = 1$ recovers unit feasibility utility. For fixed price $\tau_t = \tau$, $S_t^{\mathrm{str}} > \tau$ applies Proposition 1 using these estimates. Pre-call information ϕ_t includes the structural features and cost proxy. Appendix B.2 gives the estimators and allocation rules for each task.

Offline and online execution. Algorithm 1 trains the router using paired feasibility labels. Algorithm 2 then applies the trained router and returns control to the draft after each accepted action. For cost normalization, the pre-call proxy $\widehat{c}_t$ is the total number of draft input and output tokens at the current step. Budget accounting uses realized teacher tokens, including those spent on invalid responses. Such responses fall back to the draft action.

No online successor search. Reverse support takes $O(|V| + |E|)$ preprocessing time and memory on an ordinary grounded graph. Each routing step reads the legal action set, extracts features, and evaluates the predictor once.

Task observations and feedback. Without an exposed graph, let $b_t = e(o_{\le t}, a_{<t})$ represent public observation and action history. Pre-call inputs combine this representation with the draft proposal, remaining horizon, and public interface information. Offline planning or paired continuation returns supply outcome labels. Event predictors trained on these labels estimate repair and harm, with fixed utilities mapping their outputs to replacement value. Online routing uses these predictions without hidden state or successor search. Fitting and budget allocation are specified in Appendix B.2.

4 Experiments

4.1 Experimental setup

Tasks and models. We evaluate RuleTaker, ProofWriter, and ProverQA (Clark et al., 2020; Tafjord et al., 2021; Qi et al., 2025), alongside 134 ALFWorld episodes and 500 WebShop tasks (Shridhar et al., 2021; Yao et al., 2022). Proof confirmation uses 300 RuleTaker and 148 ProofWriter questions per model family, and 439 ProverQA questions. Model pairs appear in Table 1.

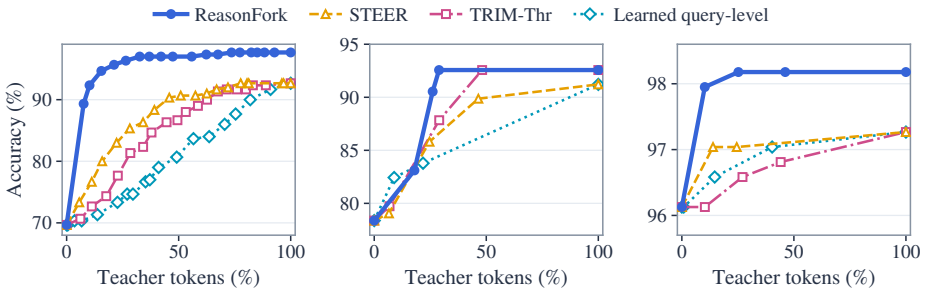
Protocol and comparisons. Learned query-level and stepwise routers share models, prompts, legal actions, calibration splits, and token accounting. Each frozen policy follows its own trajectory. T-tok counts teacher prompt and completion tokens relative to Always Teacher, while T-act measures teacher-action fractions. Appendix B details baseline training, allocation, and cost accounting.

Table 1: **Proof accuracy and teacher allocation.** Accuracy (Acc.) and teacher-token share (T-tok) are percentages. Model headers denote draft→teacher pairs. Bold and underlined values indicate the best and second-best accuracy, respectively.

	RuleTaker				ProofWriter QDep3				ProverQA	
	Qwen2.5-1.5B → Qwen2.5-7B		Phi-4-mini → Phi-4		Qwen2.5-1.5B → Qwen2.5-7B		Phi-4-mini → Phi-4		Qwen2.5-7B → Qwen2.5-14B	
Policy	Acc.↑	T-tok↓	Acc.↑	T-tok↓	Acc.↑	T-tok↓	Acc.↑	T-tok↓	Acc.↑	T-tok↓
Always Draft	69.67	0.00	77.33	0.00	78.38	0.00	83.78	0.00	96.13	0.00
Always Teacher	<u>92.67</u>	100.00	**94.67**	100.00	<u>91.22</u>	100.00	<u>92.57</u>	100.00	<u>97.27</u>	100.00
Learned query-level	73.33	22.69	85.67	22.41	82.43	8.88	87.16	14.32	96.58	14.66
RouteLLM-adapt	75.67	18.73	85.33	28.28	79.05	3.79	84.46	3.30	96.36	22.19
Confidence	86.33	34.54	<u>92.00</u>	36.83	85.81	24.83	89.86	21.45	96.13	16.82
History-only	88.33	31.80	91.00	32.97	81.76	12.08	85.81	9.39	95.44	15.78
STEER	86.33	34.11	91.00	34.03	85.81	24.66	89.86	22.42	97.04	24.50
TRIM-Thr	82.33	34.14	89.00	32.15	87.84	29.00	89.19	20.60	96.58	27.06
ReasonFork	**97.00**	32.61	**94.67**	31.39	**92.57**	28.93	**94.59**	21.91	**98.18**	25.23

4.2 Results on proof tasks

Accuracy with selective teacher use. Across five proof settings, the ReasonFork configurations in Table 1 average 95.4% accuracy with approximately 72% fewer teacher tokens than Always Teacher. On Qwen RuleTaker, ReasonFork reaches 97.00% accuracy at 32.61% T-tok, while learned query-level reaches 77.00% at 37.13% T-tok (Figure 3).



(a) RuleTaker **(b) ProofWriter** **(c) ProverQA**

Figure 3: **Accuracy across teacher budgets.** Frozen policies on three Qwen proof tasks, using the structural value estimator on ProofWriter. T-tok includes teacher prompt and completion tokens. Full curves, including Phi, appear in Figure 13.

Performance across budgets. On 360 independent RuleTaker theories per pair, ReasonFork exceeds the aggregate-state PPO router TRIM-Agg by 17.66 accuracy-pAUC points on Qwen and 12.71 on Phi. Table 2 compares common token ranges. Appendix B gives the 54-policy training grid.

Table 2: **Comparison with learned sequential routing.** Accuracy pAUC on 360 independent RuleTaker theories per pair, over common teacher-token ranges. Gains include paired 95% intervals.

Pair	T-tok range	ReasonFork	TRIM-Agg	Gain	95% CI
Qwen	2.27–23.84	**89.28**	71.62	+17.66	14.92–20.38
Phi	6.09–54.02	**92.96**	80.25	+12.71	10.24–15.19

Learning intervention value. On the same independent theories, ReasonFork gains 1.18 and 1.59 accuracy-pAUC points over draft risk on Qwen and Phi, respectively. Both paired 95% intervals exclude zero. Appendix C reports the full comparisons, including training-free structural rules. Section 4.4 evaluates the routing targets on interactive tasks.

4.3 TEACHER ALLOCATION UNDER PARTIAL OBSERVATION

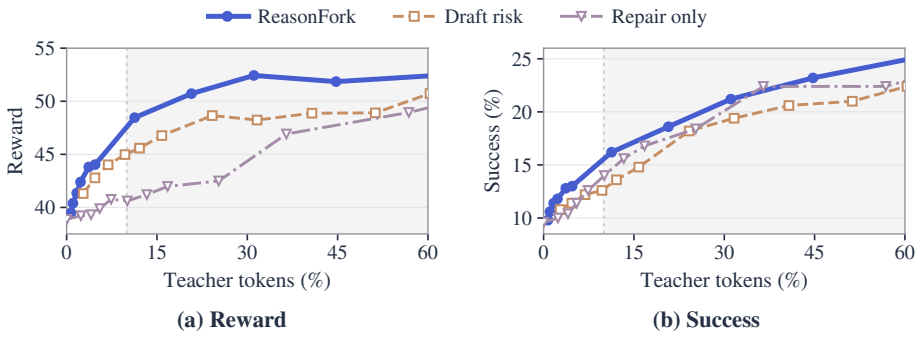
ALFWorld without an exposed proof graph. The router uses public observations, action history, and a local action-schema graph, without hidden simulator state or online search. The Phi-trained router is frozen before Qwen evaluation on all 134 unseen episodes. Teacher calls are capped at 10 or 25 per episode, with the cutoff adapting to the remaining calls and steps.

Table 3: **ALFWorld across call budgets.** Results on all 134 unseen episodes. Success and T-tok are percentages, and token counts are in millions. Always Draft and Always Teacher are endpoint references. Best and second-best routing success rates within each call budget are bolded and underlined.

	10-call cap		25-call cap		Token cost at 25 calls	
Policy	Success↑	T-tok↓	Success↑	T-tok↓	Total↓	Per success↓
Always Draft	3.73	0.00	3.73	0.00	9.202	1.840
Always Teacher	26.12	100.00	26.12	100.00	7.511	0.215
Direct action	1.49	0.00	1.49	0.00	8.492	4.246
Confidence	9.70	36.69	12.69	72.47	14.028	0.825
History	11.19	30.10	18.66	56.92	12.350	0.494
TRIM-Thr	<u>11.94</u>	30.44	14.18	62.30	13.243	0.697
Teacher takeover	11.19	23.73	<u>20.15</u>	37.57	10.271	0.380
ReasonFork	**12.69**	28.01	**21.64**	49.34	11.213	0.387

More completions from the available calls. At the 25-call cap, ReasonFork completes 29 episodes, compared with 17 for Confidence and 19 for TRIM-Thr. Its success gains are 8.96 and 7.46 points, with paired 95% intervals of 2.24–16.42 and 0.75–14.18. ReasonFork also uses 31.9% fewer teacher tokens than Confidence and 20.8% fewer than TRIM. The same call cap therefore yields more completions with less teacher usage.

Success across budgets. ReasonFork improves success with fewer teacher tokens than Confidence and TRIM at both evaluated caps. Figure 18 reports the paired gains over every stepwise routing baseline at the 25-call cap.



(a) Reward **(b) Success**

Figure 4: **The contribution of signed supervision on WebShop.** Inputs and predictor capacity are fixed. Shading marks the 10–60% pAUC range. Lines join frozen selectors. Full curves appear in Figure 9.

Total model usage. ReasonFork uses 11.21M total tokens versus Confidence's 14.03M, a 20.06% reduction. Dividing by completed tasks gives 0.387M tokens per success, compared with 0.825M. Appendix E.2 details the scheduler and cost accounting.

4.4 WHAT PAIRED SUPERVISION ADDS

Isolating the value of repair and harm. We compare draft-risk, repair-only, and signed scores. ALFWorld reuses the same fitted heads and scheduler for 1,072 own trajectories on 134 episodes. WebShop fixes inputs, capacity, training data, and allocation grid for 19,000 trajectories on 500 tasks. Complete protocols appear in Appendices E.4 and C.3.

Table 4: **Supervision-target controls on interactive tasks.** ALFWorld reports normalized success AUC from B10/B25 and shared endpoints over 0–100% teacher tokens; WebShop reports success and reward pAUC over 10–60%. The final row gives ReasonFork gains over draft risk and repair only (top/bottom).

	ALFWorld	WebShop	
Supervision target	Success AUC	Success pAUC	Reward pAUC
Draft risk	13.20	18.98	48.37
Repair only	14.71	20.22	45.41
ReasonFork	**17.32**	**21.22**	**51.43**
ReasonFork gain	+4.13	+2.24	+3.06
vs. draft risk / repair only	+2.62	+1.00	+6.01

ReasonFork gains 4.13 success-AUC points over draft risk on ALFWorld. On WebShop, it gains 2.24 success-pAUC points and 3.06 reward-pAUC points over draft risk, and 6.01 reward-pAUC points over repair-only. All four paired intervals exclude zero.

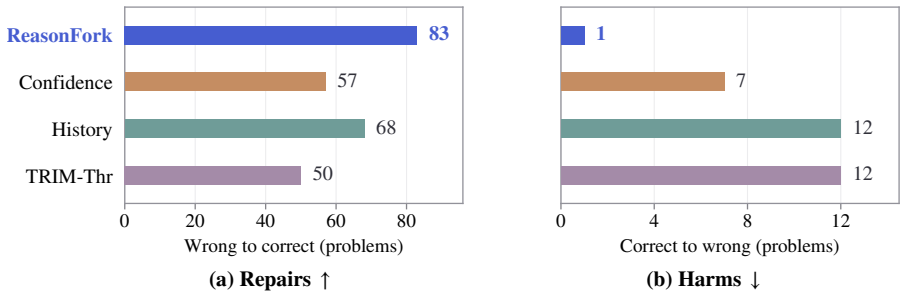
Table 5: **Actual completion after a paired intervention.** Both branches use identical selected states and frozen draft continuations. We report teacher-first minus draft-first completion gains in percentage points, with paired 95% intervals in brackets. Proxy groups report realized outcomes at H. Bold marks the primary H-horizon gains.

Paired states			**Completion gain (pp)**		**Proxy REPAIR at H**		**Proxy TIE at H**		
Split	**Pair**	N	H	$2H$	N	Repair (%)	N	Repair (%)	Harm (%)
Calibration	Qwen	714	**+16.53** [10.89, 21.92]	+2.66 [0.29, 5.16]	72	68.06	619	16.64	3.23
	Phi	415	**+17.59** [11.52, 23.82]	+2.89 [0.91, 5.39]	46	69.57	366	14.75	3.28
Confirmation	Qwen	471	**+18.26** [11.97, 24.54]	+3.61 [1.04, 6.60]	46	65.22	410	20.49	4.15
	Phi	332	**+18.98** [12.11, 25.94]	+3.01 [−0.31, 6.98]	32	71.88	297	17.85	3.37

From feasibility to realized completion. At selected action-disagreement states, we force each paired action and continue with the frozen draft policy (Table 5).

Teacher-first completion gains 16.53–18.98 points at H, with paired intervals above zero, and 2.66–3.61 points at $2H$. Among proxy-REPAIR states, 65.22–71.88% realize a repair. No non-TIE state reverses direction. Proxy-TIE states can yield repairs or harms under the draft policy, so equal feasibility labels need not imply equally easy continuations. Appendix D.13 details these selected disagreement states.

Against repair-only, WebShop combines higher reward with success noninferiority at the prespecified one-point margin. Paired success intervals include zero on both interactive tasks. ReasonFork yields more repairs and fewer harms at similar teacher usage (Figure 5).



(a) Repairs ↑ **(b) Harms ↓**

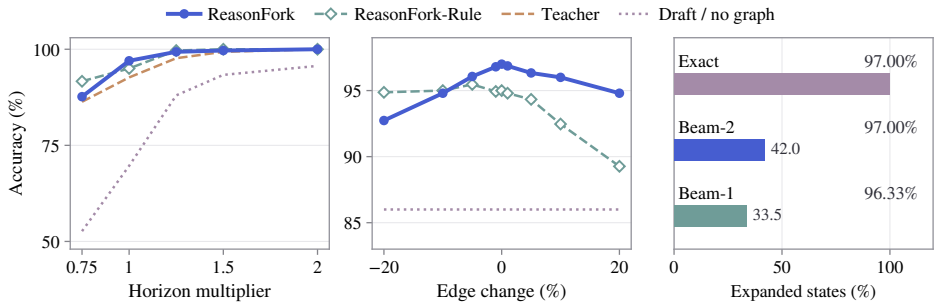
Figure 5: **Repairs and harms after routing.** Paired final outcomes relative to Always Draft on 300 Qwen RuleTaker problems. The four policies use 31.80–34.54% teacher tokens.

4.5 TRANSFER AND ROBUSTNESS

Table 6: **Transfer without target-outcome fitting.** Source weights and thresholds are frozen. RT: RuleTaker; PW: ProofWriter. Acc. and T-tok are percentages. Target sets are fixed; additional analysis is in Appendix F.

Transfer	Policy / source cutoff	Acc.↑	T-tok↓
Qwen → Phi	Frozen / 30%	93.67	27.02
	Frozen / 50%	95.00	50.82
	Phi local reference / 30%	94.67	31.39
RT → PW QDep2/3/5	Always Teacher	74.33	100.00
	Frozen / 30%	75.00	38.93
RT → PW QDep3	Always Teacher	91.22	100.00
	Frozen / 30%	94.59	29.75

Transfer without outcome refitting. Table 6 reuses source weights and thresholds. The Qwen-trained router reaches 93.67% on Phi at 27.02% T-tok. Across task interfaces, the RuleTaker classifier reaches 94.59% on ProofWriter QDep3 at 29.75% T-tok. The source classifier, scaling, and threshold remain unchanged throughout target evaluation. Frozen ALFWorld transfer further tests reuse under partial observation.



(a) Reasoning budget **(b) Support perturbations** **(c) Offline supervision**

Figure 6: **Changing budgets, structure, and offline supervision.** (a,b) Policies and thresholds remain frozen. Negative edge changes denote deletion. (c) Expanded states relative to exact evaluation, with final accuracy above each bar. Complete horizon costs and perturbations appear in Appendix D.

Budget and structure sensitivity. We compare ReasonFork with ReasonFork-Rule, a training-free variant that calls the teacher when the draft action falls outside the goal-support cone (Appendix C). The horizon sweep traces how additional steps improve completion. ReasonFork reaches full accuracy at twice the original budget. At 20% support-edge deletion or addition, ReasonFork retains 92.73% or 94.80% accuracy, compared with 86.00% without graph features. Removing graph features on Phi lowers accuracy by 6.00 points, with a paired 95% interval of 3.33–9.00. These tests probe the contribution of structure without choosing new thresholds from test outcomes.

Approximating offline supervision. Beam-1 uses 33.54% of exact state expansions with 99.06% signed-label agreement and 96.33% final accuracy. Beam-2 reaches 100% signed-label agreement using 41.98% of expansions and reproduces the exact-label deployed policy at 97.00% accuracy. Features, grouped folds, classifier, cost proxy, and calibration-only threshold selection remain fixed. Only offline labels change, while online routing remains free of successor search. On 448 states visited only after teacher intervention, repair AUROC is 0.978, with a 95% interval of 0.956–0.993. Appendix D gives the full diagnostics.

5 CONCLUSION

REASONFORK learns the value of a teacher replacement from paired offline outcomes and intervenes for one action without online search. Independent RuleTaker controls support this target over draft-risk prediction and the structural rule. Transfer and interactive-task evaluations extend the decision framework across model families and task interfaces.

AI USE STATEMENT

We used generative AI tools to assist with language editing of the manuscript and to write portions of the research code. Language editing focused on grammar, wording, and readability. The AI-assisted code was reviewed and tested by the authors to check its correctness and consistency with the methods described in the paper. We take responsibility for the final content of this work, including all text, code, claims, and results produced with the assistance of generative AI.

REFERENCES

Marah I. Abdin, Jyoti Aneja, Harkirat S. Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report. Technical Report MSR-TR-2024-57, Microsoft Research, 2024. arXiv:2412.08905.

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-Mini technical report: Compact yet powerful multimodal language models via Mixture-of-LoRAs. arXiv:2503.01743, 2025.

Pranjal Aggarwal, Aman Madaan, Ankit Anand, Srividya Pranavi Potharaju, Swaroop Mishra, Pei Zhou, Aditya Gupta, Dheeraj Rajagopal, Karthik Kappaganthu, Yiming Yang, Shyam Upadhyay, Manaal Faruqui, and Mausam. AutoMix: Automatically mixing language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi:10.52202/079017-4164.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of Thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, 2024. doi:10.1609/aaai.v38i16.29720.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. In *Proceedings of IJCAI*, pages 3882–3890, 2020.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. arXiv:2110.14168, 2021.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. Hybrid LLM: Cost-efficient and quality-aware query routing. In *International Conference on Learning Representations*, 2024.

David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A. Saurous, Jascha Sohl-Dickstein, Kevin Murphy, and Charles Sutton. Language model cascades. arXiv:2207.10342, 2022.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 10764–10799, 2023.

Silin Gao, Jane Dwivedi-Yu, Ping Yu, Xiaoqing Ellen Tan, Ramakanth Pasunuru, Olga Golovneva, Koustuv Sinha, Asli Celikyilmaz, Antoine Bosselut, and Tianlu Wang. Efficient tool use with chain-of-abstraction reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2727–2743, 2025.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *Proceedings of EMNLP*, pages 8154–8173, 2023. doi:10.18653/v1/2023.emnlp-main.507.

Zhongzhan Huang, Guoming Ling, Yupei Lin, Yandong Chen, Shanshan Zhong, Hefeng Wu, and Liang Lin. RouterEval: A comprehensive benchmark for routing LLMs to explore model-level scaling up in LLMs. In *Findings of EMNLP*, pages 3860–3887, 2025. doi:10.18653/v1/2025.findings-emnlp.208.

Vansh Kapoor, Aman Gupta, Hao Chen, Anurag Beniwal, Jing Huang, and Aviral Kumar. TRIM: Hybrid inference via targeted stepwise routing in multi-step reasoning tasks. In *International Conference on Learning Representations*, 2026.

Minho Kim. When should LLMs search? Counterfactual supervision for search routing. arXiv:2607.05752, 2026. Accepted at the FAGEN Workshop at ICML 2026.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023.

Guillaume Lample, Timothee Lacroix, Marie-Anne Lachaux, Aurelien Rodriguez, Amaury Hayat, Thibaut Lavril, Gabriel Ebner, and Xavier Martinet. HyperTree proof search for neural theorem proving. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Sangmook Lee, Dohyung Kim, Hyukhun Koh, Nakyeong Yang, and Kyomin Jung. Confidence-guided stepwise model routing for cost-efficient reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):31483–31491, 2026. doi:10.1609/aaai.v40i37.40413.

Hao Li, Yiqun Zhang, Zhaoyan Guo, Chenxu Wang, Shengji Tang, Qiaosheng Zhang, Yang Chen, Biqing Qi, Peng Ye, Lei Bai, Zhen Wang, and Shuyue Hu. LLMRouterBench: A massive benchmark and unified framework for LLM routing. In *Findings of ACL*, pages 37733–37754, 2026. doi:10.18653/v1/2026.findings-acl.1881.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison 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.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *Proceedings of IJCNLP-AACL*, pages 305–329, 2023. doi:10.18653/v1/2023.ijcnlp-main.20.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of EMNLP*, pages 5153–5176, 2023. doi:10.18653/v1/2023.emnlp-main.313.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *International Conference on Learning Representations*, 2025.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of EMNLP*, pages 3806–3824, 2023. doi:10.18653/v1/2023.findings-emnlp.248.

Chengwen Qi, Ren Ma, Bowen Li, He Du, Binyuan Hui, Jinwang Wu, Yuanjun Laili, and Conghui He. Large language models meet symbolic provers for logical reasoning evaluation. In *International Conference on Learning Representations*, 2025.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021.

Wenwen Si, Insup Lee, and Osbert Bastani. Policy-guided stepwise model routing for cost-effective reasoning. arXiv:2605.06116, 2026.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of ACL-IJCNLP*, pages 3621–3634, 2021. doi:10.18653/v1/2021.findings-acl.317.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. arXiv:2211.14275, 2022.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. PlanBench: An extensible benchmark for evaluating large language models on planning and reasoning about change. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Allen Van Gelder, Kenneth A. Ross, and John S. Schlipf. The well-founded semantics for general logic programs. *Journal of the ACM*, 38(3):620–650, 1991.

Ziyu Wan, Xidong Feng, Muning Wen, Stephen Marcus Mcaleer, Ying Wen, Weinan Zhang, and Jun Wang. AlphaZero-like tree-search can guide large language model decoding and training. In *Proceedings of the 41st International Conference on Machine Learning*, pages 49890–49920, 2024.

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

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of ACL*, pages 9426–9439, 2024. doi:10.18653/v1/2024.acl-long.510.

Minchao Wu, Michael Norrish, Christian Walder, and Amir Dezfouli. TacticZero: Learning to prove theorems from scratch with deep reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Kaiyu Yang, Aidan Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan J. Prenger, and Animashree Anandkumar. LeanDojo: Theorem proving with retrieval-augmented language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. arXiv:2412.15115, 2024.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, pages 20744–20757, 2022. doi:10.52202/068431-1508.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, 2023.

Yutong Yin, Mingyu Jin, Jin Pan, Changyi Yang, Zijie Xia, Dhruv Pai, Shuming Hu, Zhen Zhang, Chenyang Zhao, Jinman Zhao, Wujiang Xu, Raymond Li, Xin Eric Wang, Julian McAuley, and Zhaoran Wang. Efficient and trainable language model test-time scaling via local branch routing. arXiv:2606.25354, 2026.

Caiqi Zhang, Menglin Xia, Xuchao Zhang, Daniel Madrigal, Ankur Mallick, Samuel Kessler, Victor Ruehle, and Saravan Rajmohan. Budget-aware agentic routing via boundary-guided training. arXiv:2602.21227, 2026.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language Agent Tree Search unifies reasoning, acting, and planning in language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 62138–62160, 2024.

APPENDIX CONTENTS

A Proof Semantics and Offline Evaluation **14**
- A.1 Action systems . . . 14
- A.2 Exact state distance . . . 14
- A.3 Support hypergraph and relaxed distance . . . 14
- A.4 Validity and the main-text guarantees . . . 15
- A.5 Exact evaluator and reverse support . . . 15
- A.6 Boundary cases . . . 16

B Experimental Setup and Reproducibility **16**
- B.1 Dataset adapters . . . 16
- B.2 Outcome labels, score fitting, and allocation . . . 17
- B.3 Model interfaces and decoding . . . 18
- B.4 Baseline definitions . . . 18
- B.5 Evaluation and statistical protocol . . . 19

C Paired-Supervision Controls **20**
- C.1 Independent supervision-target comparison . . . 20
- C.2 Original-set target and capacity diagnostics . . . 22
- C.3 WebShop supervision targets . . . 23

D Proof Frontiers and Mechanism Diagnostics **23**
- D.1 ReasonFork variants . . . 24
- D.2 A finer threshold grid . . . 25
- D.3 Intervention outcomes . . . 25
- D.4 Calibration signal quality . . . 25
- D.5 Complete frozen teacher-budget frontiers . . . 26
- D.6 Horizon construction and sensitivity . . . 26
- D.7 Target polarity and states visited after intervention . . . 27
- D.8 Support-graph perturbations . . . 28
- D.9 Bounded-search supervision . . . 29
- D.10 Complete on-policy evaluation . . . 29
- D.11 Component ablation . . . 29
- D.12 Intervention anatomy . . . 29
- D.13 Feasibility labels and fixed-policy continuation . . . 30

E Partially Observed Agent Tasks **31**
- E.1 ALFWorld partial-observation adapter . . . 31
- E.2 ALFWorld cross-model confirmation . . . 31
- E.3 WebShop without an exposed task graph . . . 32
- E.4 ALFWorld controls with shared fitted heads . . . 34

F Frozen Transfer **34**
- F.1 Transfer details . . . 34
- F.2 RuleTaker-to-ProofWriter frozen transfer . . . 35

A Proof Semantics and Offline Evaluation

A.1 Action systems

Let $\mathcal{L}$ be a finite set of signed ground literals. A state contains a set $K \subseteq \mathcal{L}$, a visible query q, its explicit complement $\bar{q}$, and the remaining action horizon h. A grounded action is a tuple

$$a = (P_a, N_a, c_a, \chi_a), \tag{9}$$

where P_a is a finite set of positive prerequisites and N_a contains task-specific negative or guarded prerequisites. The declared conclusion is c_a. Semantic certificates such as well-founded falsity are recorded in χ_a (Van Gelder et al., 1991). The action is executable at K when $P_a \subseteq K$, every guarded prerequisite is licensed by χ_a, and $c_a \notin K$.

The evaluated proof domains are monotone:

$$T(K, a) = K \cup \{c_a\}. \tag{10}$$

No previous literal is retracted, and each committed action costs one symbolic step. The off-cone result also assumes *upward-persistent executability*: if an action is executable at K, it remains executable after adding other sound conclusions unless its own conclusion is already present, in which case the action can be omitted. Negative or guarded prerequisites therefore require immutable semantic certificates over the derivation. Section A.6 treats non-monotone and weighted extensions separately.

Soundness. Every exposed action agrees with the benchmark's symbolic semantics. Model explanations do not change the state. Only a validated identifier and its declared conclusion are committed.

Completeness. A grounded action graph is complete when it represents every benchmark-valid executable action and every represented edge is executable. Parser correctness alone is insufficient because a faithful parser may still omit a proof rule or valid grounding. Exact infeasibility and the off-cone guarantee require this stronger condition.

A.2 Exact state distance

For $z \in \{q, \bar{q}\}$, define

$$d_z^\star(K) = \min\left\{m : \exists a_{1:m} \text{ executable from } K,\ z \in T^m(K, a_{1:m})\right\}, \tag{11}$$

with $d_z^\star(K) = \infty$ if no sequence exists. The union distance is

$$D_\cup^\star(K) = \min\{d_q^\star(K), d_{\bar{q}}^\star(K)\}. \tag{12}$$

This quantity is computed by the offline evaluator, not by the deployed router.

A.3 Support hypergraph and relaxed distance

A grounded rule with prerequisite set P_a and conclusion c_a defines a directed hyperedge $P_a \to c_a$. Reverse indices map each conclusion to the rules that can establish it. Starting from the visible target set $\mathcal{G} = \{q, \bar{q}\}$, reverse traversal marks every literal on a prerequisite chain ending in $\mathcal{G}$. These marked literals form the union support cone $\mathcal{C}_\cup$.

For ordinary unary edges, reverse breadth-first search gives exact shortest graph depth. General hyperedges instead use a relaxed support depth that is monotone under reverse propagation. One useful recursion is

$$\widetilde{d}_\cup(c) = 0\ (c \in \mathcal{G}), \qquad \widetilde{d}_\cup(p) = \min_{a:p\in P_a}\left[1 + \widetilde{d}_\cup(c_a)\right], \tag{13}$$

subject to task-specific treatment of conjunction and alternative rules. This literal-level quantity is a feature rather than a state-level distance unless the adapter provides an equivalence proof.

Proposition 2 (Finite-horizon feasibility)**.** *For a complete unit-cost executor, action $a \in \mathcal{A}(s_t)$ preserves a route to a visible target within $h_t - 1$ further actions if and only if $D_\cup^\star(T(s_t, a)) \le h_t - 1$.*

Proof. A minimizing sequence witnesses sufficiency. Conversely, any feasible sequence of at most $h_t - 1$ actions bounds the exact minimum by the same number. □

Algorithm 3 Exact union distance used for offline labels

Require: state fact set K, visible targets $\mathcal{G}$, grounded executor
if $K \cap \mathcal{G} \neq \varnothing$ **then**
  return 0
end if
initialize queue $Q \leftarrow [(K, 0)]$ and visited set $V \leftarrow \{K\}$
while Q is nonempty **do**
  pop (K', d) from the front of Q
  for $a \in \mathcal{A}(K')$ **do**
    $K'' \leftarrow T(K', a)$
    if $K'' \cap \mathcal{G} \neq \varnothing$ **then**
      return $d + 1$
    end if
    if $K'' \notin V$ **then**
      add K'' to V and append $(K'', d + 1)$ to Q
    end if
  end for
end while
return ∞

A.4 Validity and the main-text guarantees

If the initial fact set is valid and each committed identifier belongs to the current legal action set, induction on the trajectory shows that every successor is executor-reachable. This invariant guarantees executability, not eventual success. Proposition 2 applies exact state distance to the post-action horizon. Proposition 1 and Theorem 1 are proved in the main text.

For Corollary 1, $\mathcal{C}(z)$ contains every literal, including z, on a prerequisite path of at least one complete proof of z, and $\mathcal{C}_\cup = \mathcal{C}(q) \cup \mathcal{C}(\bar{q})$. A reverse traversal that includes additional syntactic dependencies may over-approximate this set. A conclusion outside such an over-approximation is still outside every complete proof. Missing a valid prerequisite, in contrast, invalidates the off-cone guarantee. The persistence assumption also requires immutable certificates for any negative or guarded premises. It does not follow from monotone fact storage alone.

Proof of Corollary 1. No complete proof of z uses the off-cone conclusion $c_D = \mathrm{head}(a_D)$, so adding it leaves $d^\star_z(K)$ unchanged. After the teacher action, replay a shortest proof of z from K. Monotonicity and upward-persistent executability preserve every remaining step; omit any conclusion already present. Hence $d^\star_z(T(K, a_T)) \leq d^\star_z(K) = d^\star_z(T(K, a_D))$. The inequality is immediate when z is unreachable. Taking the minimum over $q, \bar{q}$ and comparing the successors at the common horizon $h_t - 1$ gives the feasibility claim.

If $D^\star_\cup(s_t) = \infty$, every legal successor also has infinite distance: a finite successor completion, prefixed by the current action, would complete the original state. Theorem 1 therefore isolates all nontrivial feasibility forks in unfinished states with finite completion distance. Two feasible successors can still yield different outcomes under a fixed draft policy; Section 4.4 examines that distinction empirically.

A.5 Exact evaluator and reverse support

State-level breadth-first evaluation. The exact evaluator represents each search node by its complete set of established facts. Applying one currently executable action produces a successor. The first state containing either visible polarity determines the union distance.

The evaluator returns the successor distance and feasibility label. Router supervision requires neither a recovered path nor parent-pointer storage.

State versus literal distance. Consider two rules $A \wedge B \to Q$ and $C \to B$. A literal graph may assign a short distance from A to Q without accounting for the status of B or the actions needed to establish it. State-level BFS instead tracks the complete fact set. It therefore respects prerequisite conjunction, duplicate conclusions, and action availability. Online support depth remains a low-cost relevance and slack feature, while the exact label uses state-level distance.

Algorithm 4 Reverse union-support construction

Require: grounded actions, targets $q, \bar{q}$
build conclusion-to-producer index
$d(q) \leftarrow 0$, $d(\bar{q}) \leftarrow 0$; initialize worklist with both targets
while worklist is nonempty **do**
  pop conclusion c
  for each producer a of c **do**
    for each prerequisite $p \in \mathrm{pre}(a)$ **do**
      if p is unseen or $d(p) > d(c) + 1$ **then**
        $d(p) \leftarrow d(c) + 1$; push p
      end if
    end for
  end for
end while
$\mathcal{C}_\cup \leftarrow \{v : d(v) < \infty\}$
return $\mathcal{C}_\cup, d$

Reverse support. The reverse module first constructs a mapping

$$\mathrm{producers}(v) = \{a : \mathrm{head}(a) = v\}. \tag{14}$$

The algorithm initializes a worklist with q and $\bar{q}$. For each popped conclusion, it visits every producer and marks the prerequisites as support literals. Alternative producers create multiple reverse branches. Duplicate visits retain the minimum relaxed depth. Any literal never reached by this traversal is off-cone.

Complexity. For an ordinary grounded graph, reverse traversal costs $O(|V|+|E|)$ time and memory. Hypergraph grounding depends on rule arity and domain size. Exact state evaluation is used only to construct calibration labels. Appendix D evaluates bounded label construction.

A.6 Boundary cases

The exact feasibility definition extends to positive action costs $w(a)$ by using the shortest executable weighted completion,

$$D^\star_w(s) = \min_{a_{1:m} \text{ reaching } \mathcal{G}} \sum_{i=1}^{m} w(a_i), \qquad F^\star_t(a) = \mathbf{1}\{D^\star_w(T(s_t, a)) \le H_t - w(a)\}. \tag{15}$$

A lower bound below the remaining budget is only a necessary condition. A sufficient certificate requires a feasible continuation, or an upper bound attained by such a continuation. Symbolic action cost remains separate from teacher-token allocation.

In non-monotone environments, actions may remove facts or consume resources, so Corollary 1 does not apply. Exact state-level feasibility is still meaningful under the full transition model, while literal support is only a routing feature. With multiple visible terminal classes, the evaluator can use their union without exposing the gold class.

For a concrete on-cone detour, let $K = \{A\}$ and two actions remain. The rules $A \to B$, $B \to q$ give a two-action proof, while $A \to C$, $C \to D$, $D \to q$ provide a longer alternative. After choosing $A \to C$, only one action remains and the target needs two. ReasonFork-Rule does not trigger because C is on-cone; slack distinguishes this deadline risk. Adding $A \to X$ with no path from X to either target gives the off-cone case. Undefined negative cycles are a separate semantic issue: $p \leftarrow \mathrm{not}\, q$ and $q \leftarrow \mathrm{not}\, p$ license neither rule without a falsity certificate.

B Experimental Setup and Reproducibility

B.1 Dataset adapters

RuleTaker. RuleTaker pairs natural-language facts and rules with queries at controlled depths. The adapter normalizes entities and predicates, then grounds rule variables over the instance domain. Each grounded application becomes one action. Positive premises must already be established. A

default-negative premise is exposed only when well-founded evaluation marks its atom false. Atoms left undefined by negative cycles do not license a rule.

The reported Qwen calibration contains 300 records and 1,487 same-state paired events. These events comprise 72 repairs, 23 harms, and 1,392 ties. The frozen adapter audit records 4,045 actions across 271 fork-bearing targets, with 399 structural certificates. The main Qwen and Phi confirmations each contain 300 problems separated from calibration and development by theory grouping.

ProofWriter. ProofWriter supplies grounded multi-step theories with implications, explicit negative facts, and proofs. Multiple grounded rules may share a conclusion. Stable action identifiers therefore include the rule and substitution rather than only the head literal. Alternative proofs are retained, and exact distance minimizes over their executable successor states.

The strict QDep3 confirmation contains 148 examples with zero overlap with calibration material at either the theory or problem level. A separate 900-example mixed QDep2/3/5 anchor is theory-disjoint and was inspected during development. It serves as an additional analysis.

ProverQA. ProverQA is created through an LLM–symbolic-prover generation pipeline. The adapter normalizes clauses and literal signs, then exposes sound CNF unit-resolution actions. An action is legal when the current closure falsifies all but one literal of a clause and the remaining literal has not been derived. Tautologies and duplicate literals require explicit normalization, as do contradictory closure states.

The ProverQA evaluation set contains 439 examples. The reported table combines the original held-out runs with additional completions generated under the same frozen configuration. Only aggregate measurements are redistributed.

B.2 Outcome labels, score fitting, and allocation

Paired outcome labels. The learned routers use REPAIR/TIE/HARM supervision from same-state draft and teacher comparisons. RuleTaker evaluates exact feasibility; ProverQA compares signed horizon-and-slack utility and retains its magnitude. ALFWorld uses offline planner feasibility, while WebShop uses paired continuation returns. A TIE means that the outcome criterion does not distinguish an improvement or deterioration. For the WebShop target controls, reward differences within $[-0.05, 0.05]$ are ties (Appendix C.3).

Outcome estimators. On RuleTaker, ReasonFork predicts paired feasibility outcomes with a standardized multinomial logistic model. It uses $C = 0.25$, lbfgs, at most 5,000 iterations, and seed 20260719 on the structural features listed below. Five-fold cross-validation groups events by theory, so each out-of-fold prediction comes from a model trained on other theories. Each training fold learns its own mean and scale. The deployed model then refits the unchanged specification on the complete calibration bank. Missing or infinite distances receive explicit indicators and fixed numeric fill values. Candidate list position is excluded, and structural-distance ties share a dense rank. The estimator predicts $p(\text{REPAIR}), p(\text{TIE}), p(\text{HARM})$. For the exact-feasibility differences $\Delta_t \in \{-1, 0, +1\}$, $u(\Delta_t) = \Delta_t$ maps these probabilities to expected signed value. Utility values differ from loss class weights. Primary budgeted operating points divide the value by a pre-call length proxy. Each pair shares a tokenizer and strict candidate-ID schema, so this proxy is the observed draft prompt-plus-completion length at the current state. Realized teacher tokens are available only after a call. The proxy schedules calls but is excluded from the ReasonFork outcome classifier and is not treated as a calibrated predictor of teacher tokens.

ProverQA predicts REPAIR/TIE/HARM probabilities and uses the conditional expectation of its magnitude-aware utility. ALFWorld estimates repair through fork and conditional-repair probabilities, alongside a harm predictor, and applies a remaining-call scheduler (Appendix E.2). WebShop fits separate repair and harm heads. Its target controls assign utility 1 to repairs, -2 to harms, and 0 to ties, yielding $p_R - 2p_H$ in Equation 4, with development-calibrated selectors (Appendix C.3). All estimator inputs are pre-call; utilities and allocation rules are fixed before evaluation.

ProofWriter instantiation. ProofWriter uses a frozen structural value estimator based on whether the draft action lies outside the support cone, under a three-call receding-horizon budget (Table 1). Table 6 additionally evaluates direct transfer of the RuleTaker classifier, scaling, and threshold without target-task refitting.

Table 7: Features used on RuleTaker. Exact future labels and gold answer polarity are unavailable online.

Feature family	REASONFORK-RULE	REASONFORK	REASONFORK-HV
Off-cone indicator	✓	✓	✓
Support distance and rank	✗	✓	✓
Approximate slack / horizon	✗	✓	✓
Branch and cone statistics	✗	✓	✓
Semantic certificates	✗	✓	✓
Draft confidence	✗	✗	✓
Accepted-action history	✗	✗	✓
Pre-call draft-length proxy (scheduling)	✗	✓	✓
Exact successor BFS at deployment	✗	✗	✗
Gold label / official polarity	✗	✗	✗

ReasonFork-HV. ReasonFork-HV (Horizon-Value) is our analysis variant. In RuleTaker, it augments the outcome classifier with draft minimum and mean token probability. It also includes recent accepted-action history, prior off-cone rate, and cost-related predictors. Scheduling follows the same signed-utility-per-length-proxy rule as the primary ReasonFork operating points. This analysis variant tests whether confidence, history, and cost add information beyond the primary structural representation.

Thresholds. For RuleTaker, calibration OOF predictions fix the 10/30/50% threshold grid before development or confirmation outcomes are read. Development evaluates only these declared points and selects the primary displayed policy. The Qwen ReasonFork-HV-positive point uses a separately declared zero expected-value threshold and reaches 57.35% realized cost. Every threshold is locked before confirmation. These fractions target expected teacher use over a data split. They do not impose a separate token cap on each problem.

B.3 MODEL INTERFACES AND DECODING

The proof model pairs are Qwen2.5-1.5B/7B, Phi-4-mini/Phi-4, and Qwen2.5-7B/14B for ProverQA. The frozen ALFWorld cross-model test applies Phi-trained routers to Qwen2.5-7B/14B. Endpoint and tokenizer revisions, temperature, maximum completion length, and guided-decoding settings are fixed per suite. An unfavorable legal action is never retried. Invalid teacher outputs fall back to the draft and retain their incurred token charge. Returned prose cannot modify the verified state.

Both endpoints receive the same visible state and ordered action identifiers. The shared proof prompt has the following structure:

> Select the next executable reasoning action.
> Verified facts: [STATE]. Query: [QUERY]. Complement: [COMPLEMENT].
> Legal actions: [ID] [PREMISES] ⇒ [CONCLUSION].
> Return exactly one listed identifier. Do not invent a rule or conclusion.

Task adapters specialize this rendering and the required output schema. Inference terminates on a derived target, horizon exhaustion, or an empty legal action set. The benchmark-specific rule for dual polarity or unknown answers is fixed before evaluation. Each policy maintains its own successor trajectory; fixed-draft replay is used only for paired calibration labels.

B.4 BASELINE DEFINITIONS

Endpoints and generic controls. Always Draft commits the small-model action and incurs zero teacher tokens. Always Teacher commits the teacher action at every step and defines 100% T-tok. Raw confidence thresholds draft token probability, while History uses previous confidences, calls, or position features without the current support graph. All cutoffs use calibration or development data rather than confirmation outcomes.

Learned query-level routing. A multinomial logistic classifier predicts REPAIR, TIE, and HARM from non-graph pre-call features, with grouped out-of-fold calibration. At the first live step, its expected signed value determines whether the instance is assigned to the draft or teacher; this choice is fixed for the full trajectory. Thresholds target the same 10/30/50% calibration teacher-token regions

as the stepwise policies and remain frozen during evaluation. Model pairs, prompts, executable actions, calibration splits, and realized teacher-token accounting are shared with the stepwise controls.

No-graph learned routing. A standardized three-class logistic model uses $C = 0.25$, lbfgs, 5,000 iterations, and seed 20260720. Its inputs are horizon fraction, history, two confidence values, candidate count, and the pre-call length proxy. Support distance, cone membership, and certificates are excluded. The model predicts signed REPAIR/TIE/HARM utility and divides expected utility by the same length proxy. This control tests the contribution of the structural inputs.

STEER. Our executable-action implementation retains confidence-based step routing. It fits an unlabeled two-component diagonal Gaussian mixture to draft mean-token log probability with 10 initializations, regularization 10^{-6}, and seed 20260720. The posterior probability of the low-confidence component is thresholded at development teacher-token regions.

TRIM-Thr. Our threshold implementation retains process-score routing and one-step return. It uses a 4,096-dimensional hashed character 3–5-gram view of the visible state and draft action. Seven non-graph scalars complete the input. A balanced logistic classifier with $C = 0.25$, liblinear, and 5,000 iterations predicts draft-deadline safety. Development data select the safety threshold.

TRIM-Agg. The sequential comparator trains an aggregate-state PPO router with a task process-reward model and rollout dropout. Public inputs include the accepted action history, draft confidence, current position, and remaining budget; the support graph is excluded. Nine cost penalties and seeds 10, 20, and 30 yield 27 policies per model pair. All 54 policies are evaluated on their own trajectories over 360 theory-disjoint problems per pair, giving 19,440 trajectories. Independent outcomes do not select the penalty or seed. Table 2 compares normalized accuracy pAUC over each pair's common teacher-token range. Intervals use 5,000 paired theory-group bootstrap draws, recomputing both curves, their support, and the teacher normalizer.

Teacher takeover. After the frozen no-graph trigger first calls the teacher, takeover assigns the remaining suffix to it. It tests persistent teacher control with the no-graph trigger.

All policies share the executable interface and token accounting. Their thresholds are fixed without confirmation outcomes, and every reported point uses its own successor trajectory.

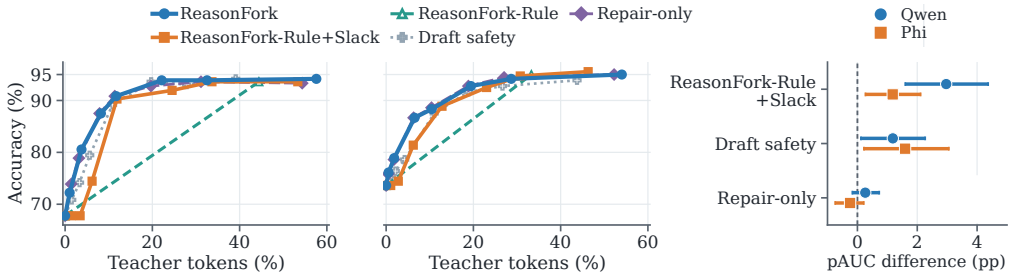
B.5 Evaluation and statistical protocol

Split roles. Calibration supplies paired outcome labels and fits the learned scores. Development chooses score thresholds and primary operating points. Models, prompts, features, code, and splits are locked before confirmation evaluation. RuleTaker groups examples at the theory level. Strict ProofWriter groups at both the theory and problem levels. ProverQA uses the frozen configuration described in Appendix B for every reported completion. ALFWorld groups state events by training episode during out-of-fold selection. Its complete 134-episode valid-unseen split is evaluated only after the features, fitted model, and thresholds are fixed.

Final accuracy and repairs. Task accuracy is the proportion of final predictions matching the benchmark label. A final *repair* is a confirmation problem that Always Draft gets wrong and the evaluated policy gets right. A final *harm* reverses those outcomes. These task-level counts differ from the event-level calibration label but retain the same signed interpretation.

McNemar test. For two policies evaluated on the same problems, let n_{10} count cases correct only for policy A and n_{01} count cases correct only for policy B. The exact two-sided McNemar test conditions on $n_{10} + n_{01}$ under a binomial null with probability one half. We use this test for locked operating-point comparisons.

Bootstrap intervals. The paired problem-level bootstrap resamples confirmation problems with replacement and recomputes accuracy differences. Calibration AUROC analysis resamples theory groups to preserve within-theory dependence. Confidence intervals use the 2.5th and 97.5th percentiles. The release script records the replicate count and random seed.



(a) Qwen 1.5B → 7B **(b) Phi 3.8B → 14B** **(c) Paired gain**

Figure 7: Independent target controls and paired pAUC differences. Every curve follows its own trajectory; comparator-specific common supports are reported below.

Accuracy–teacher-budget AUC. Let a method provide observed operating points (c_j, a_j) with normalized teacher budget $c_j \in [0, 1]$. Interpolation on a common grid gives

$$\mathrm{AUC}(\pi) = \int_0^1 \widetilde{a}_\pi(c)\, dc. \tag{16}$$

The Qwen comparison envelope takes the pointwise maximum of the prespecified Learned query-level, No-graph, STEER, and TRIM-Thr curves before integration. It may switch among these four families across cost levels. Each bootstrap resample reconstructs the envelope. Other diagnostic routers are reported separately from this declared comparison set.

Cost definitions. Teacher allocation counts logical prompt-plus-completion tokens for every called endpoint on each policy's trajectory, even when deterministic calls are cached across policies. Always Teacher defines 100% within the split. Total neural work additionally counts draft inference. We report teacher allocation and total unweighted model tokens separately. Isolated end-to-end latency comparisons were outside the measurement protocol. Teacher-token savings alone do not imply lower total model usage or latency.

C Paired-Supervision Controls

We first report controls on an independent RuleTaker test set, followed by diagnostics on the original test sets.

C.1 Independent supervision-target comparison

A rule variant. ReasonFork-Rule is our training-free variant: it calls when the draft conclusion lies outside the support cone: $S_t^{\mathrm{rule}} = \mathbf{1}[\mathrm{head}(a_D) \notin \mathcal{C}_\cup]$. Let $d_z^\star(K)$ denote the minimum number of legal actions needed to derive target z from fact set K.

Corollary 1 (Conditional off-cone non-harm)**.** *Under Theorem 1's action-system assumptions, let $\mathcal{C}_\cup$ contain all literals, including the goals, on every complete proof of q or $\bar{q}$. If $\mathrm{head}(a_D) \notin \mathcal{C}_\cup$, every sound legal replacement a_T satisfies, for $z \in \{q, \bar{q}\}$,*

$$d_z^\star(T(K, a_T)) \le d_z^\star(K) = d_z^\star(T(K, a_D)), \qquad F_t^\star(a_T) \ge F_t^\star(a_D). \tag{17}$$

Appendix A.4 proves the corollary and specifies the cone and persistence assumptions. Non-harm need not mean positive gain. ReasonFork-Rule+Slack adds deadline prioritization (Appendix C.1). Both variants use fixed structural rules; learned estimators fit replacement value from paired outcomes.

This experiment tests the learned intervention target against a stronger training-free rule on new RuleTaker problems. The source is the official depth-5 meta-test partition. We first removed every theory group used by the earlier Qwen and Phi confirmation suites. Deterministic ordering then selected 60 theories from each depth-by-answer cell, using query depths 2, 3, and 5. This produced 360 distinct theory groups. Seven eligible groups were left unused. The split has no theory-group overlap with either earlier suite, and none of the selected group or problem identifiers appeared in the previous result records.

The original protocol was fixed before constructing the split, running the models, or accessing test answers. We added ReasonFork-Rule+Slack after constructing the split but before either model ran,

without changing the selected theories. All policies and thresholds were fixed before evaluation. Reference answers were used only to score the completed trajectories.

Policies and allocation grid. The Qwen pair is Qwen2.5 1.5B to 7B. The Phi pair is Phi-4-mini to Phi-4. Generation uses temperature 0 and at most 32 completion tokens. Candidate output is constrained to strict JSON with the `outlines` backend. The request seed is 20260720. Retry and output repair are disabled. The learned controls share the 22 pre-call inputs used by ReasonFork, five theory-group folds, and a StandardScaler–LogisticRegression pipeline with $C = 0.25$. Draft safety predicts whether the draft successor misses the deadline. Repair-only predicts whether the paired teacher action changes a failure into a success. The signed target distinguishes REPAIR, HARM, and TIE. Their respective scores are $p(\text{unsafe})/\hat{c}_t$, $p(\text{REPAIR})/\hat{c}_t$, and $(p(\text{REPAIR}) - p(\text{HARM}))/\hat{c}_t$. Grouped out-of-fold calibration sets the seven cutoffs at teacher-token fractions of 1, 2.5, 5, 10, 20, 30, and 50 percent. Score ties are ordered by problem identifier and then turn. Ridge controls use the same folds, inputs, three outputs, and cutoff grid, with $\alpha = 4$.

ReasonFork-Rule+Slack requires no fitted outcome model or coefficients. For the draft action at turn t, let o_t indicate that it lies outside the union-support cone. Let $\delta_t = \max\{0, -\text{slack}_t\}$. Slack subtracts one plus the draft support distance from the remaining action count. A fixed hash of problem identity and turn supplies the label-independent tie-breaker u_t. The policy orders events by

$$K_t = (o_t, \delta_t, u_t) \tag{18}$$

in descending lexicographic order. Every off-cone event therefore precedes every on-cone event. Within either group, the policy ranks the more severe deadline deficit first. The hash acts only on exact ties. For an off-cone action, the same precomputed structural feature supplies the best finite support distance used in the slack calculation. Each cutoff is the largest prefix that stays within the requested calibration token fraction. ReasonFork-Rule is evaluated separately as a fixed policy that calls on every off-cone draft action and nowhere else.

Every policy follows its own trajectory. A teacher call replaces one draft action from the same legal action set, after which control returns to the draft. The matrix contains 52 policies per model family: two endpoints, seven cutoffs for each of three Logistic and three Ridge targets, ReasonFork-Rule, and seven ReasonFork-Rule+Slack cutoffs. Teacher tokens include the teacher prompt and completion and are normalized by Always Teacher within the same family.

Matched-allocation curves. We linearly interpolate each frozen curve without a monotone hull. Duplicate allocation coordinates receive their mean accuracy. Normalized partial AUC is the mean interpolated accuracy on the common interior range. This range runs from 1.52 to 39.10 percent teacher tokens for Qwen and from 1.22 to 43.75 percent for Phi. Each of 5,000 paired theory-group bootstrap draws recomputes the token coordinates, Always-Teacher denominator, common range, accuracy, and partial AUC.

ReasonFork exceeds ReasonFork-Rule+Slack by 2.97 partial-AUC points on Qwen. Its 95 percent confidence interval runs from 1.60 to 4.37 points. The Phi gain is 1.18 points, with an interval from 0.27 to 2.11. In both model families, the learned curve has higher matched-allocation partial AUC than deadline-aware prioritization. Relative to draft safety, signed supervision gains 1.18 points on Qwen and 1.59 on Phi. The corresponding intervals run from 0.12 to 2.27 and from 0.22 to 3.06. Against repair-only, the Qwen difference is 0.26 point, with an interval from -0.17 to 0.73. The Phi difference is -0.25 point, with an interval from -0.74 to 0.21.

Deadline-sensitive subgroup. The subgroup is fixed from the initial graph and public horizon, before model output. Its legal action set contains one action that can finish within the remaining horizon and another action that is off-cone or exceeds that horizon. This rule selects 316 of the 360 theories. ReasonFork again exceeds ReasonFork-Rule+Slack. The Qwen gain is 3.09 partial-AUC points, with a confidence interval from 1.62 to 4.61. The Phi gain is 1.50 points, with an interval from 0.55 to 2.51. The shared allocation ranges are 1.70 to 42.25 percent for Qwen and 1.37 to 47.57 percent for Phi.

Estimator and fixed-point checks. For the equal-parameter Ridge control, signed supervision exceeds draft safety for Phi by 1.68 partial-AUC points, with an interval from 0.46 to 2.97. The Qwen difference is 0.63 point, with an interval from -0.36 to 1.56. Against repair-only, signed Ridge is lower by 0.45 point on Qwen, with an interval from -0.89 to -0.03. The Phi difference is -0.16 point, with an interval from -0.60 to 0.25. Table 9 reports every prespecified comparison at the 30

Table 8: Normalized pAUC and paired contrasts on the independent RuleTaker split. Full-set ranges are 1.52–39.10% (Qwen) and 1.22–43.75% (Phi); deadline ranges are 1.70–42.25% and 1.37–47.57%.

	Full set		Deadline subgroup	
Curve	Qwen	Phi	Qwen	Phi
Logistic draft safety	89.68	89.65	88.99	88.45
Logistic repair only	90.60	91.48	90.00	90.71
Logistic signed value (ReasonFork)	90.87	91.24	90.28	90.57
Ridge draft safety	89.61	89.28	88.87	88.10
Ridge repair only	90.68	91.11	90.14	90.30
Ridge signed value	90.24	90.96	89.66	90.14
ReasonFork-Rule	79.59	87.21	77.65	85.90
ReasonFork-Rule+Slack	87.90	90.06	87.19	89.07

Paired differences and 95% intervals

		Qwen			Phi		
Scope	Contrast	Δ	CI low	CI high	Δ	CI low	CI high
Full	ReasonFork − ReasonFork-Rule+Slack	+2.97	+1.60	+4.37	+1.18	+0.27	+2.11
Full	ReasonFork − ReasonFork-Rule	+11.28	+9.25	+13.34	+4.03	+2.75	+5.39
Full	ReasonFork − draft safety	+1.18	+0.12	+2.27	+1.59	+0.22	+3.06
Full	ReasonFork − repair only	+0.26	-0.17	+0.73	-0.25	-0.74	+0.21
Full	Ridge signed − draft safety	+0.63	-0.36	+1.56	+1.68	+0.46	+2.97
Full	Ridge signed − repair only	-0.45	-0.89	-0.03	-0.16	-0.60	+0.25
Deadline	ReasonFork − ReasonFork-Rule+Slack	+3.09	+1.62	+4.61	+1.50	+0.55	+2.51
Deadline	ReasonFork − ReasonFork-Rule	+12.63	+10.43	+14.80	+4.68	+3.32	+6.13
Deadline	ReasonFork − draft safety	+1.29	-0.0004	+2.53	+2.12	+0.62	+3.70
Deadline	ReasonFork − repair only	+0.28	-0.21	+0.81	-0.13	-0.64	+0.34
Deadline	Ridge signed − draft safety	+0.80	-0.35	+1.89	+2.04	+0.65	+3.46
Deadline	Ridge signed − repair only	-0.48	-1.01	-0.01	-0.16	-0.66	+0.29

Table 9: Paired comparisons at the frozen 30% calibration cutoffs.

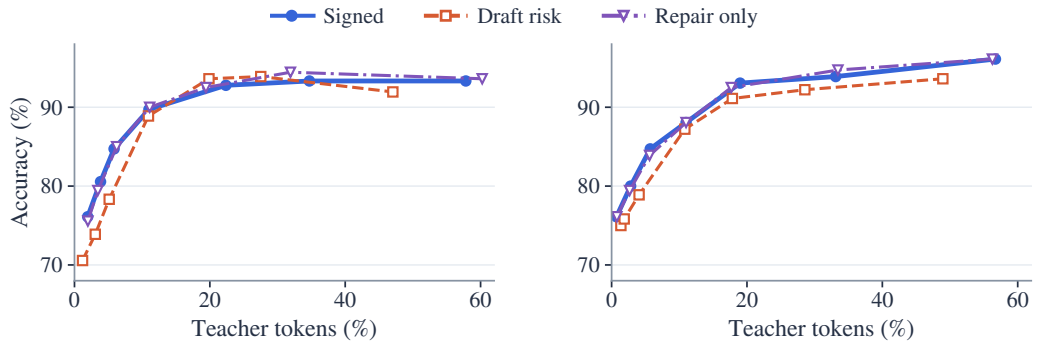
Family	Estimator	Comparator	Δ	CI low	CI high	S/C	Exact p
Qwen	Logistic	ReasonFork-Rule	+0.28	-1.11	+1.67	4/3	1.0000
Qwen	Logistic	ReasonFork-Rule+Slack	+0.28	-1.39	+1.94	5/4	1.0000
Qwen	Logistic	draft safety	+0.83	-1.39	+3.06	10/7	0.6291
Qwen	Logistic	repair only	+0.28	-0.83	+1.39	3/2	1.0000
Qwen	Ridge	draft safety	-0.56	-2.78	+1.67	8/10	0.8145
Qwen	Ridge	repair only	-1.11	-2.22	-0.28	0/4	0.1250
Phi	Logistic	ReasonFork-Rule	-0.83	-2.22	+0.28	1/4	0.3750
Phi	Logistic	ReasonFork-Rule+Slack	-0.56	-1.94	+0.83	2/4	0.6875
Phi	Logistic	draft safety	+1.39	-1.11	+3.61	12/7	0.3593
Phi	Logistic	repair only	-0.28	-1.67	+0.83	2/3	1.0000
Phi	Ridge	draft safety	+1.67	-0.56	+3.89	11/5	0.2101
Phi	Ridge	repair only	-0.83	-1.94	+0.00	0/3	0.2500

percent calibration cutoff. These pointwise comparisons have different realized allocations and are not treated as matched-budget tests. Their intervals use the same 5,000 paired theory-group draws. Exact two-sided McNemar tests provide the p values. S/C gives the numbers correct only under signed routing and only under its comparator.

Operating points and estimator sensitivity. Figure 7 shows the Logistic targets and ReasonFork-Rule+Slack. Figure 8 gives the equal-parameter Ridge controls at all seven frozen cutoffs. Their realized token allocations are plotted directly; the complete full-set and deadline-subgroup values are provided in the accompanying data file. Table 8 retains both scopes and their paired uncertainty.

C.2 Original-set target and capacity diagnostics

These additional diagnostics use the existing RuleTaker confirmation splits ($n = 300$ each). Calibration locks every estimator and protocol before collecting the new model outputs. Thresholds, the curve grid, and the overlap interval are fixed without confirmation correctness.



(a) Qwen / Ridge **(b) Phi / Ridge**

Figure 8: **Independent Ridge target controls.** All seven frozen cutoffs on 360 theories per model pair. Inputs and parameter counts are held fixed; lines join measured accuracy and teacher-token pairs.

Matched-input target comparisons. These controls use the targets, 22 inputs, five theory-group folds, and seven-cutoff calibration grid specified in Appendix C.1. Logistic uses $C = 0.25$; the Ridge model with the same parameter count uses three outputs and $\alpha = 4$. Only target coding and the fixed output contrast change. Table 10 reports all six 30%-calibration operating points.

Matched-allocation curves linearly interpolate the frozen thresholds and endpoints without a monotone hull. Table 10 gives pAUC levels and 5,000-draw paired intervals over the shared interior range. Signed Logistic exceeds draft safety for both families; its difference from repair-only is not resolved. Ridge results provide a separate capacity sensitivity check.

At a common expected budget $B = 30\%$, deterministic Ridge mixtures are formed from token counts recorded without consulting test answers. Qwen expected accuracies are 95.15 for draft safety, 96.34 for repair-only, and 96.14 for signed value. Phi gives 92.70 for draft safety and 92.97 for signed value. The predeclared repair-only policy pair for Phi does not bracket B. These point estimates are not used as matched-budget tests because the predeclared 95% bracketing criterion is not met.

Comparison with the rule variant. On Qwen, REASONFORK and REASONFORK-RULE have seven versus one discordant correct outcomes. The difference is $+2.00$ points, with exact McNemar $p = 0.0703$. Every REASONFORK-only correct trajectory includes a teacher call that replaces an on-cone negative-slack draft action by a nonnegative-slack action. On Phi, the two discordant correct outcomes both favor REASONFORK-RULE. The paired difference is -0.67 point, with $p = 0.500$.

C.3 WEBSHOP SUPERVISION TARGETS

The study uses 500 fixed tasks, with the training and development banks, public inputs, capacity, and allocation grid held fixed. All three logistic heads are evaluated at each routed state, including in single-target controls. They use 20,038 inputs, class balancing, and $C = 0.25$. Paired reward gains above 0.05 define repair and losses below -0.05 define harm. Draft risk predicts draft completion failure; repair-only uses p_R, and ReasonFork uses $p_R - 2p_H$.

Twelve development-calibrated selectors per target and two reference policies yield 19,000 trajectories. This study is separate from the earlier operating points in Table 19. Normalized pAUC covers 10–60% teacher tokens, using the same-run Always-Teacher total of 3,153,270 tokens. Figure 9 retains all measured points. Intervals use 10,000 paired task-bootstrap draws with seed 20260903. Success against repair-only meets the frozen one-point noninferiority margin; its difference interval includes zero.

The accompanying data file records draft, scorer, and teacher tokens, total usage, and cost per success for every selector.

D PROOF FRONTIERS AND MECHANISM DIAGNOSTICS

Table 1 reports the primary operating points. Here we give additional thresholds, paired outcomes, and analyses of the routing decisions.

Table 10: RuleTaker target-supervision controls ($n = 300$ per family). Cells report confirmation accuracy / teacher tokens normalized by Always Teacher (%). Thresholds use 30% calibration teacher tokens. LogReg and Ridge are calibration-locked additional controls. Paired intervals use 5,000 bootstraps; p values are exact two-sided McNemar tests.

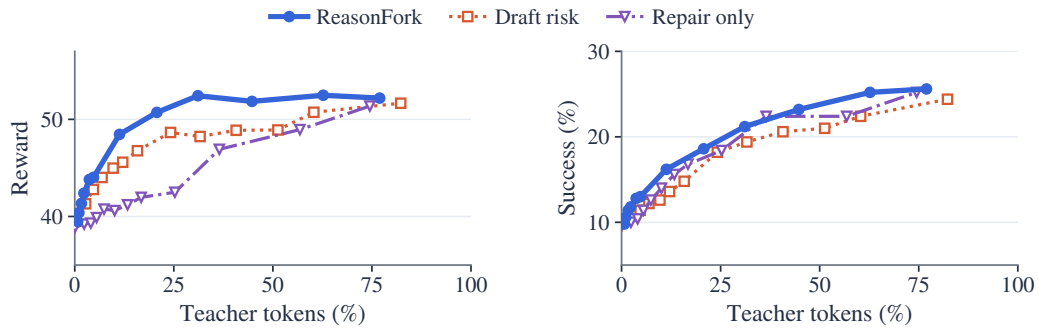
Estimator	Family	Draft safety	Repair-only	Signed value
LogReg	Qwen	94.00 / 27.00	96.00 / 29.86	97.00 / 32.61
LogReg	Phi	92.00 / 28.23	94.33 / 29.54	94.67 / 31.39
Ridge	Qwen	95.33 / 27.69	96.33 / 29.91	96.33 / 33.82
Ridge	Phi	92.67 / 29.24	93.67 / 34.03	93.33 / 33.80

(a) Matched-allocation LogReg curves

Family	Common T-tok	pAUC: safety / repair / signed	Signed − safety Δ (95% CI)	Signed − repair Δ (95% CI)
Qwen	1.67–38.70	90.69 / 93.13 / 93.21	+2.52 (1.43–3.69)	+0.08 (−0.21–0.45)
Phi	1.22–44.97	90.40 / 91.54 / 91.47	+1.07 (0.04–2.21)	−0.07 (−0.36–0.25)

(b) Paired accuracy differences

Estimator	Family	Contrast	Δ accuracy (95% CI)	p
LogReg	Qwen	Signed − safety	+3.00 (0.67–5.33)	0.0225
LogReg	Qwen	Signed − repair	+1.00 (0.00–2.33)	0.250
LogReg	Phi	Signed − safety	+2.67 (0.67–4.67)	0.0215
LogReg	Phi	Signed − repair	+0.33 (0.00–1.00)	1.000
Ridge	Qwen	Signed − safety	+1.00 (−1.00–3.00)	0.508
Ridge	Qwen	Signed − repair	0.00 (−1.00–1.00)	1.000
Ridge	Phi	Signed − safety	+0.67 (−2.00–3.33)	0.804
Ridge	Phi	Signed − repair	−0.33 (−1.00–0.00)	1.000



(a) Reward **(b) Success**

Figure 9: Complete WebShop target curves. All twelve measured points per target are shown; lines connect evaluated selectors.

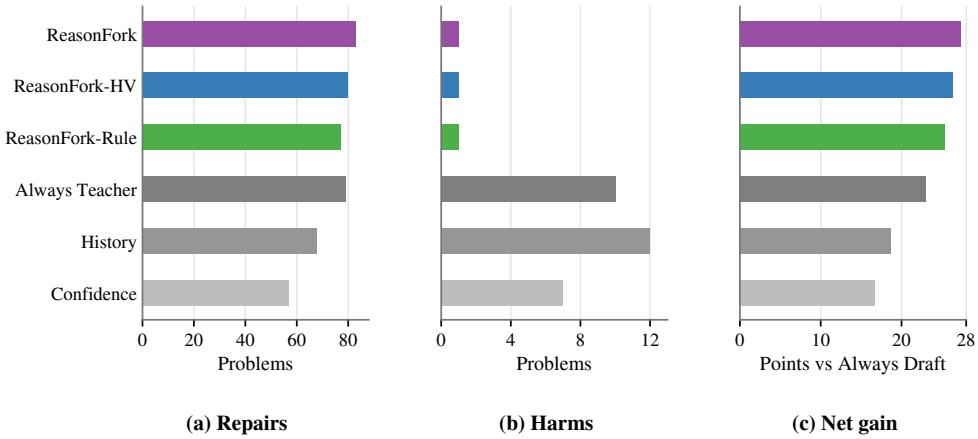
D.1 REASONFORK VARIANTS

Table 11: **ReasonFork and its variants.** Model pairs, settings, units, and highlighting follow Table 1. RT: RuleTaker. PW: ProofWriter QDep3. ProverQA includes the same additional completions.

	RT / Qwen		RT / Phi		PW / Qwen		PW / Phi		ProverQA	
Policy	Acc.↑	T-tok↓	Acc.↑	T-tok↓	Acc.↑	T-tok↓	Acc.↑	T-tok↓	Acc.↑	T-tok↓
ReasonFork-Rule	95.00	39.80	**95.33**	36.40	**92.57**	28.14	**94.59**	21.15	**98.18**	13.22
ReasonFork	**97.00**	32.61	94.67	31.39	**92.57**	28.93	**94.59**	21.91	**98.18**	25.23
ReasonFork-HV	96.00	33.35	93.67	31.94	**92.57**	28.93	**94.59**	21.91	**98.18**	13.26

Table 12: Qwen RuleTaker policy outcomes.

Method	Accuracy	Teacher tokens	Repair	Harm
Always Draft	69.67	0.00	0	0
Always Teacher	92.67	100.00	79	10
Learned query-level-30	73.33	22.69	13	2
Raw confidence-30	86.33	34.54	57	7
History-30	88.33	31.80	68	12
No-graph-30	86.00	32.96	60	11
STEER-30	86.33	34.11	57	7
TRIM-Thr-30	82.33	34.14	50	12
Teacher takeover-30	86.33	33.89	61	11
REASONFORK-RULE	95.00	39.80	77	1
REASONFORK-30	**97.00**	32.61	**83**	1
REASONFORK-HV-30	96.00	33.35	80	1
REASONFORK-HV-positive	96.67	57.35	82	1



(a) Repairs **(b) Harms** **(c) Net gain**

Figure 10: Qwen intervention decomposition. Repairs, harms, and net accuracy gains are paired against Always Draft on the same 300 problems.

D.2 A FINER THRESHOLD GRID

The 18-point Qwen RuleTaker grid is derived only from calibration scores and replayed on confirmation without label-based threshold fitting. All 27,900 trajectories for our methods have complete state records. ReasonFork has a normalized AUC of 0.953821, followed by ReasonFork-HV at 0.949707 and history at 0.876395. The remaining values are 0.872525 for STEER, 0.872356 for confidence, 0.864495 for no-graph, 0.848011 for TRIM-Thr, and 0.809774 for learned query-level routing. The prespecified pointwise envelope contains Learned query-level, No-graph, STEER, and TRIM-Thr. ReasonFork exceeds this envelope by 0.081296. A problem-paired bootstrap with 5,000 draws places the difference between 0.059026 and 0.096049.

Figure 13 gives the complete task-level frontiers for the structural variants and comparison policies.

D.3 INTERVENTION OUTCOMES

Figure 10 separates problem-level repairs from regressions relative to Always Draft.

D.4 CALIBRATION SIGNAL QUALITY

On 1,487 Qwen RuleTaker calibration events, ReasonFork-HV reaches a repair AUROC of 0.971918 and ReasonFork reaches 0.962274. The corresponding values are 0.866804 for history, 0.611744 for raw confidence, and 0.583952 for confidence logistic. Repair average precision is 0.701917 for ReasonFork-HV and 0.686904 for ReasonFork. Both structural signals separate repair events more sharply than history or confidence alone. Figure 11 compares these signals directly.

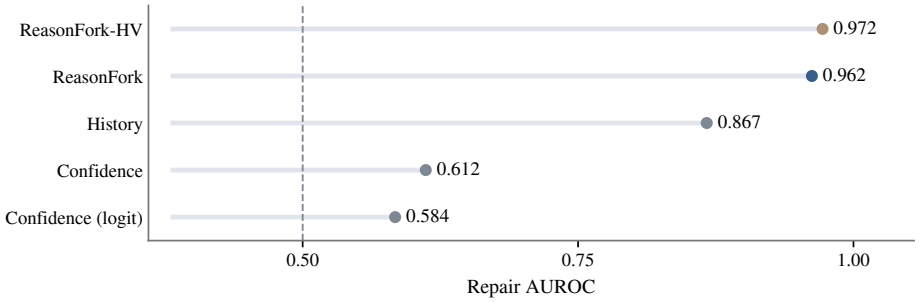


Figure 11: Repair discrimination across five signals on 1,487 Qwen RuleTaker calibration events. The dashed line marks chance.

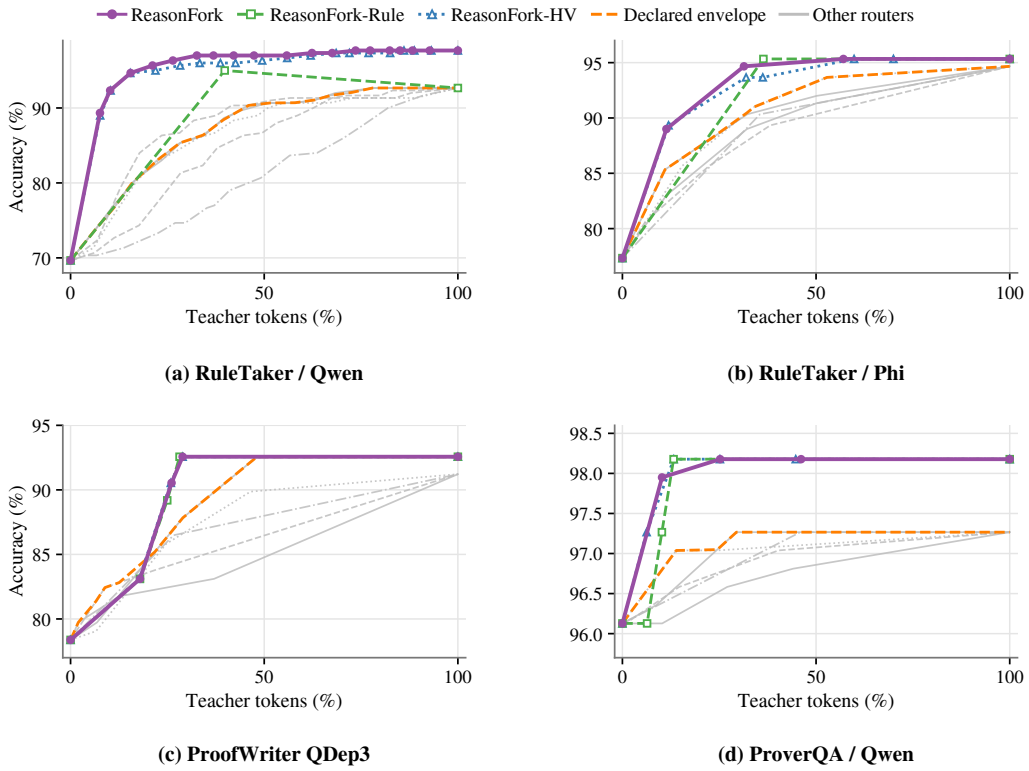


Figure 12: Four-setting summary, including the Phi RuleTaker frontier. Each point follows its own trajectory at a frozen calibration-selected threshold.

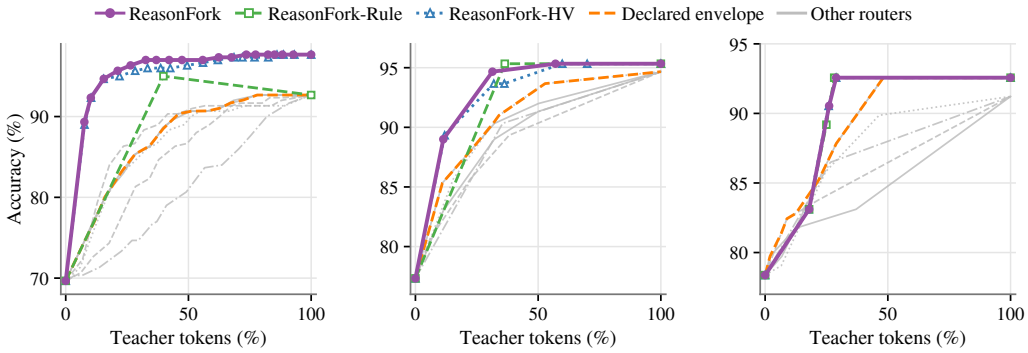
D.5 Complete frozen teacher-budget frontiers

Figure 13 shows the reported frozen operating points for Qwen RuleTaker, independent Phi, and held-out ProofWriter. It includes both endpoints and the baseline methods used to form each envelope. Every curve follows calibration-selected thresholds on its own successor trajectories. The primary operating points remain unchanged.

Across the frozen RuleTaker curves, normalized AUC gains over the declared envelope are 0.0813 for Qwen and 0.0215 for Phi. The paired 95% intervals are 0.0590–0.0960 and 0.0099–0.0325, respectively.

D.6 Horizon construction and sensitivity

For RuleTaker, the declared proof-depth stratum fixes the public action horizon at $H = 3, 6, 9$ for QDep $2, 3, 5$. This per-instance rule depends on the official depth stratum but not on answer polarity. It is identical across calibration, development, and confirmation. The 300-problem confirmation set



(a) RuleTaker / Qwen **(b) RuleTaker / Phi** **(c) ProofWriter QDep3**

Figure 13: Frozen accuracy–teacher-token frontiers for RuleTaker/Qwen, RuleTaker/Phi, and ProofWriter QDep3. ReasonFork is primary. ReasonFork-Rule is training-free. ReasonFork-HV is the analysis variant. Gray lines show external routers. Orange shows their envelope.

Table 13: Qwen RuleTaker horizon sensitivity ($n = 300$). Frozen policies follow their own trajectories. Method cells give accuracy / teacher-token fraction (%). R/H/T counts evaluator-only labels at called states. Anatomy categories overlap; low positive means slack in $[0, 1]$.

H'/H	Mean H'	Draft	Teacher	ReasonFork-Rule	ReasonFork	R/H/T	Fork %
0.75	5.00	52.67	86.33	91.67/40.64	87.67/33.97	62/0/337	72.83
1.00	6.00	69.67	92.67	95.00/39.80	97.00/32.61	32/0/364	61.65
1.25	8.00	88.00	97.67	99.67/39.48	99.33/33.60	5/0/410	42.88
1.50	9.33	93.33	99.33	100.00/39.33	99.67/35.81	3/0/439	41.09
2.00	12.00	95.67	100.00	100.00/39.27	100.00/39.41	0/0/483	41.34

Structural anatomy of the same rollouts

H'/H	Events	Off-cone (%)	Negative slack (%)	Low positive (%)	Union fork (%)
0.75	1,189	40.37	14.38	52.99	72.83
1.00	1,215	39.18	4.94	32.10	61.65
1.25	1,187	38.75	0.93	10.87	42.88
1.50	1,178	39.73	0.51	2.89	41.09
2.00	1,166	40.91	0.09	1.11	41.34

has minimum/mean/median/maximum $3/6/6/9$. After freezing every model, feature map, threshold, and failure rule, we set $H' = \lceil mH \rceil$ for $m \in \{0.75, 1, 1.25, 1.5, 2\}$. Each policy runs its own trajectory.

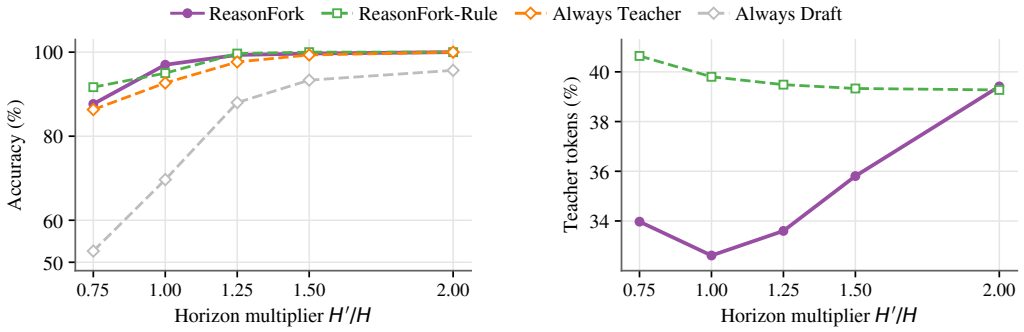
Figure 14 summarizes the same sweep on a shared horizontal scale.

Across $0.75H, H, 1.25H, 1.5H$, the margins over Always Teacher are 1.33, 4.33, 1.67, and 0.33 accuracy points. At $2H$, both reach 100%. ReasonFork leads ReasonFork-Rule by two points at the frozen primary horizon. ReasonFork-Rule performs best at the shortened and near-saturated longer horizons. As more steps become available, fewer actions prevent completion and both methods approach full accuracy.

D.7 Target polarity and states visited after intervention

The union target keeps official answer polarity hidden from the router. We check its agreement with gold-specific feasibility separately and report ambiguous or dual-reachable states. Table 14 uses every calibration pair and the frozen Qwen ReasonFork confirmation rollout. Exact successor checks remain evaluator-only. Online results in the table use the initial subset of 983 states with paired outcomes. The complete evaluation below scores all 1,215 deployed states.

Because the frozen state hash includes problem identity, exact cross-split state matches are zero by construction under theory-disjoint confirmation. We therefore compare states using standardized structural-feature distance. Within this subset, teacher-induced states comprise 38.05% of counterfactually scorable events, and repair discrimination remains stable after the first intervention. The analysis measures the frozen router under this state shift and performs no on-policy refitting.



(a) Accuracy **(b) Teacher use**

Figure 14: Frozen RuleTaker/Qwen policies under changes to the public reasoning horizon. Panel (a) reports accuracy and panel (b) realized teacher-token use. Each policy follows its own successor trajectory, with no threshold retuning.

Table 14: Union-target validity and deployed-state evaluation. AP is repair average precision. Online rows report the initial 983 states with paired outcomes, without imputation. The complete evaluation covers all 1,215 deployed states in Appendix D.

Audit	Events	Statistic	Value
Union vs gold	2,974 actions	feasibility agreement	100.00%
Union REPAIR	72	gold repair / wrong polarity	72 / 0
Union HARM	23	gold harm	23
Polarity state	2,974 actions	q only / $\bar{q}$ only / both / neither	43.98 / 44.89 / 0 / 11.13%
Before first teacher	609	repair AUROC / AP / prevalence	0.977 / 0.566 / 1.31%
Teacher-induced	374	repair AUROC / AP / prevalence	0.973 / 0.785 / 6.42%
Counterfactual coverage	1215 online states	observed / missing, no imputation	983 / 232
Exact state support	983	calibration hash matches	0
Feature support	983	median / p90 nearest RMS distance	0.170 / 0.415

Table 15: Robustness to support-graph corruption and bounded-search labels. Graph cells report accuracy / teacher-token fraction (%); label rows report final accuracy, teacher fraction, and event-level repairs and harms.

Condition	ReasonFork	ReasonFork-Rule	No-graph
Clean	97.00 / 32.61	95.00 / 39.80	86.00 / 32.96
Delete 10%	94.80 / 29.46	95.00 / 47.83	86.00 / 32.96
Delete 20%	92.73 / 26.78	94.87 / 53.29	86.00 / 32.96
Add 10%	96.00 / 30.33	92.47 / 36.50	86.00 / 32.96
Add 20%	94.80 / 28.50	89.27 / 32.47	86.00 / 32.96

Bounded-search labels

Labels	Accuracy	Teacher fraction	Event repair	Event harm
Exact	97.00	32.61	32	0
Beam-1	96.33	31.16	25	0

D.8 SUPPORT-GRAPH PERTURBATIONS

Reverse-support edges are deleted or augmented independently at 1/5/10/20% under five fixed seeds. Executable rules and legal actions remain fixed, as do prompts, horizons, and cached model outputs. All 36,900 trajectories complete.

At 20% deletion, node-cone recall is 72.04% and ReasonFork records 106/2 repair/harm events. At 20% addition, node-cone precision is 61.47% and it records 184/0 repairs/harms.

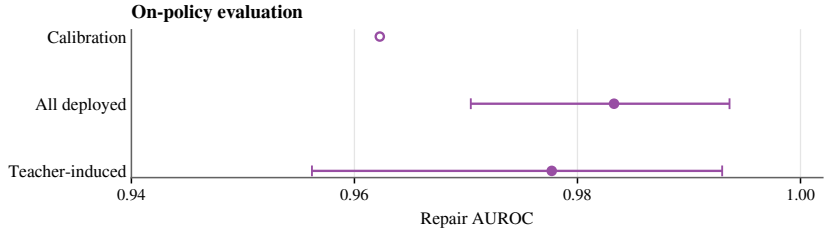


Figure 15: ReasonFork repair discrimination on calibration states and on states visited by the frozen deployed policy. Filled points show task-bootstrap 95% intervals. The open calibration marker reports the stored point estimate.

Table 16: Phi RuleTaker component ablation ($n = 300$). T-budget is normalized teacher tokens.

Variant	Accuracy	T-budget	Repair	Harm	Δ acc.
Full ReasonFork	94.67	31.39	54	2	0.00
Without horizon/slack features	94.67	32.57	55	3	0.00
Without all graph features	88.67	33.17	36	2	-6.00
Without signed harm utility	94.33	30.29	53	2	-0.33
Without cost normalization	95.33	32.33	56	2	+0.67

D.9 Bounded-search supervision

Beam-1 changes only the signed training labels. It retains the features, grouped folds, classifier, seed, cost proxy, and calibration-only threshold selection.

Beam-1 signed-label agreement is 99.06% at 33.54% of exact expansions. Its accuracy difference from exact is -0.67 points. The paired-bootstrap confidence interval runs from -2.00 to 0.67, and exact McNemar $p = 0.625$. Beam-2 reaches 100% signed agreement at 41.98% expansions and produces the same deployed policy as exact labels.

D.10 Complete on-policy evaluation

The frozen Qwen ReasonFork confirmation rollout contains 1,215 decision events. The completed same-state teacher evaluation provides paired outcomes for all 1,215 events. Across all states, repair AUROC/AP is 0.98328/0.74329. The 448 states visited only after an earlier teacher intervention include 24 repairs. Their repair AUROC/AP is 0.97769/0.78514, and the bootstrap AUROC interval runs from 0.95619 to 0.99298. The frozen router retains this discrimination on its shifted successor distribution. Figure 15 shows the same discrimination on the complete deployed and teacher-induced state subsets.

D.11 Component ablation

Five component variants fit their estimators and thresholds on the disjoint calibration split, then run on the original Phi confirmation set without access to test answers. This ablation was designed after the original test results were known.

Figure 16 complements the exact table with paired uncertainty, cost changes, and intervention outcomes.

The graph-free degradation is paired and significant. ReasonFork is correct on 19 problems that the graph-free variant misses; the reverse occurs once. The exact McNemar test gives $p = 4.0054 \times 10^{-5}$. A bootstrap 95% interval runs from 3.33 to 9.00 points. No-horizon has identical accuracy with two discordances in each direction. The no-signed difference involves one discordant problem, and the no-cost difference involves two. The paired ablation supports the overall graph contribution; the individual horizon, harm, and cost contrasts remain within statistical uncertainty.

D.12 Intervention anatomy

Among the 1,215 states visited by Qwen ReasonFork-30, 39.18% of draft proposals are off-cone and 4.94% have negative relaxed slack. Another 32.10% have slack zero or one, with overlap among

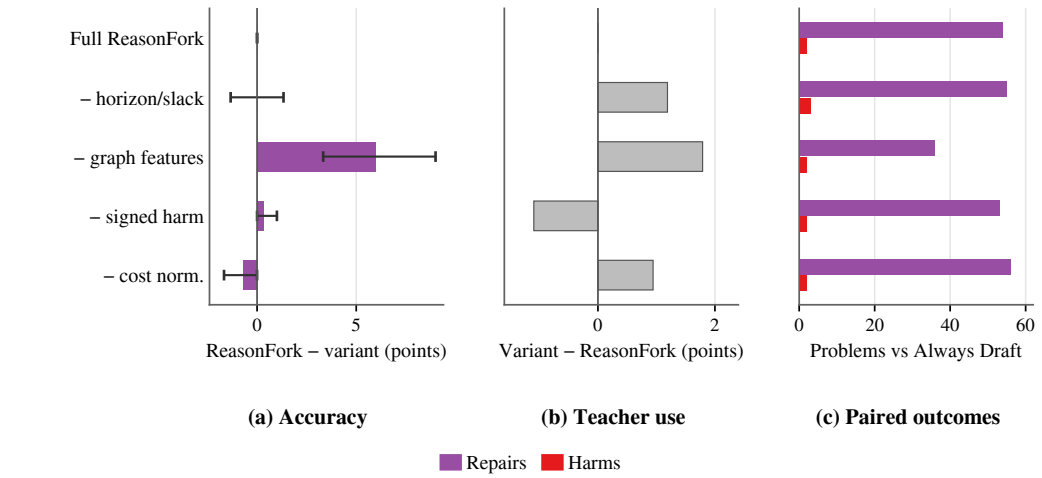


Figure 16: Calibration-locked Phi component ablation. Panels show paired accuracy differences with bootstrap 95% intervals and teacher-token changes. Repairs and harms are relative to Always Draft.

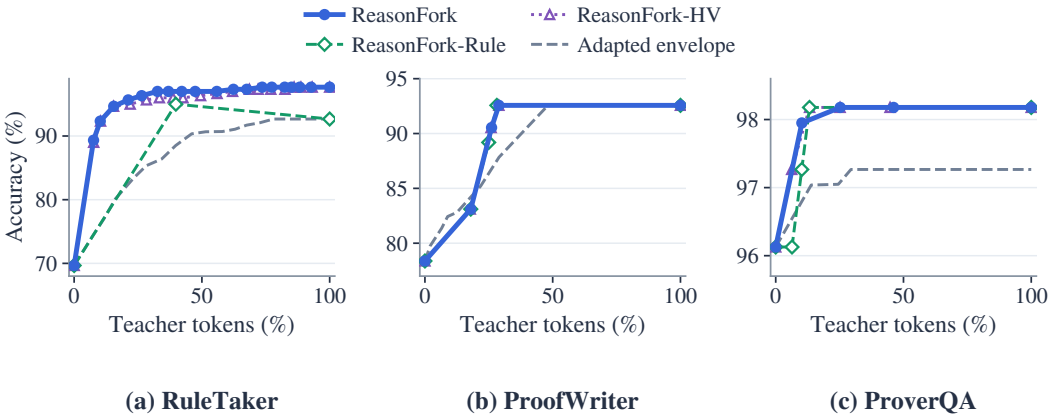


Figure 17: **Our variants across teacher budgets.** All three Qwen task panels shown in the main comparison are retained. ReasonFork is the primary method, ReasonFork-Rule is our training-free variant, and ReasonFork-HV is our analysis variant. The prespecified comparison envelope is defined above.

categories. Their union covers 61.65% of states, while the frozen router calls on 32.59%. The selected calls therefore differ from those of the off-cone rule. The component ablation above also separates the overall graph contribution from horizon and slack alone.

D.13 Feasibility labels and fixed-policy continuation

The paired experiment selects states where the stored draft and teacher propose different actions. After forcing either action, both branches use the same frozen draft policy under the same continuation conditions. The measured contrast concerns these selected local interventions, rather than an average effect over all states or the value of a full routing policy.

Across the four groups, 65.22–71.88% of proxy-repair states realize repairs. No non-TIE state reverses the predicted repair/harm direction. At the original horizon, teacher-first completion gains 16.53–18.98 points, with all paired cluster-bootstrap intervals above zero. At twice the horizon, gains fall to 2.66–3.61 points. This attenuation is consistent with differences in near-term recovery under a deadline.

Table 5 reports completion gains and the realized outcomes within proxy-REPAIR and proxy-TIE states. In calibration Phi, calibration Qwen, confirmation Qwen, and confirmation Phi order, the repair/TIE/harm counts are 46/366/3, 72/619/23, 46/410/15, and 32/297/3. At H, respectively 1/3, 14/23, 11/15, and 3/3 proxy-HARM states realize a harm; none realizes a repair. Proxy-TIE states

account for a substantial share of the gain, since equal reachability labels need not imply equal success under the draft policy. These results support local directional alignment within the selected action-disagreement states.

E Partially Observed Agent Tasks

E.1 ALFWorld partial-observation adapter

ALFWorld places household instruction following in an interactive TextWorld environment (Shridhar et al., 2021). Unlike the proof adapters, its state is partially observed and its transitions are non-monotone. Opening, closing, heating, cooling, cleaning, taking, or placing an object can change future preconditions. At each step, the environment exposes an observation and a finite `admissible_commands` set. Both models choose from that same action set, and the environment executes the selected command.

Here ReasonFork uses observable features in place of a complete support graph. They encode the task, observation and action history, current command set, draft proposal and log probability, remaining horizon, and a public local action-schema graph. Hidden PDDL state, expert plans, simulator predicates, and online search are excluded. During training only, a privileged planner evaluates counterfactual draft and teacher successor states to form exact finite-horizon REPAIR/HARM/TIE labels. The complete training bank contains 164 episodes and 8,090 state events. Its labels comprise 52 REPAIR, 32 HARM, and 8,006 TIE events.

Episode-grouped out-of-fold selection chooses a factorized histogram-gradient-boosting score. It models the fork probability separately from repair given a fork and from harm. Its one-step intervention AUC is 0.004258, compared with 0.002279 for the matched History-only representation. An episode-cluster bootstrap places the difference between 0.000893 and 0.002586. The representation and two thresholds are frozen before any valid-unseen outcome is evaluated. Confirmation evaluates every loadable valid-unseen episode ($n = 134$). The same frozen policy definitions are used for the cross-model test.

Partial-observation controls. The stepwise ALFWorld controls use the same hard 10- or 25-call scheduler and execute their own successor trajectories. History removes the action-schema features while retaining the public interaction history. Confidence ranks negative mean draft-token log probability against its frozen calibration distribution. Once triggered, Belief takeover uses the same ReasonFork score but assigns subsequent actions to the teacher until the call budget is exhausted. The no-teacher direct-action control is a frozen candidate ranker trained on all 8,090 calibration events. It identifies the same-state teacher action from the task and public observations. Its inputs also include action history, the legal action set, the draft proposal, confidence, and the public action schema. The control commits its highest-ranked legal action without a teacher call. The Phi and Qwen control suites use identical policy definitions. The Qwen configuration was fixed before any Qwen ALFWorld output.

E.2 ALFWorld cross-model confirmation

The Phi-trained router is frozen before evaluation on all 134 Qwen valid-unseen episodes. Each policy follows its own trajectory and uses only public observations, action history, and the environment-provided legal action set.

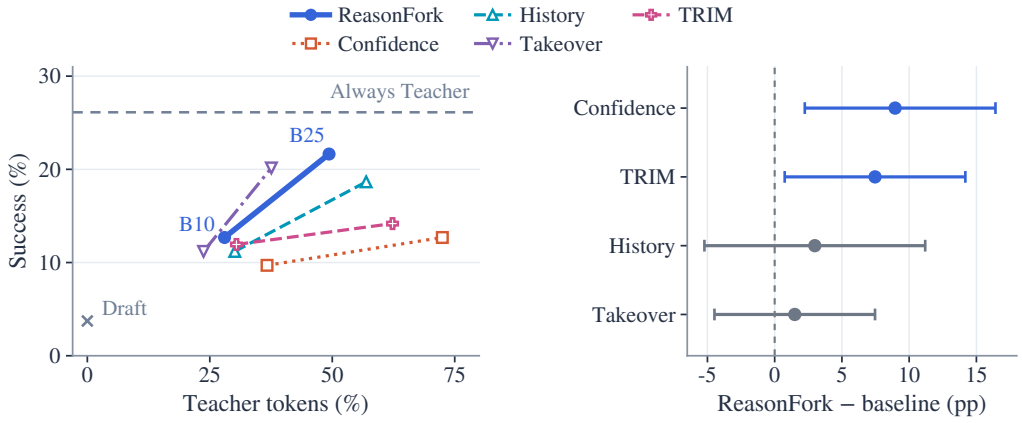
At the 25-call cap, ReasonFork attains the highest routed success. It improves success by 8.96 points over confidence while using about 23.1 percentage points fewer teacher tokens.

The added TRIM-Thr-25 control reaches 14.18% success at 62.30% teacher tokens. ReasonFork's success gain is 7.46 points, with a 95% interval of 0.75–14.18. The matched request-level operating point has 18.26% success at 49.24% teacher actions and 47.89% teacher tokens. It interpolates the executed request-50 and request-75 policies, with weight 0.62228 on the latter. Teacher-action allocation matches ReasonFork; teacher-token usage differs. The added evaluation covers 134 episodes and nine executed policies, totaling 1,206 trajectories with no generation failures, parse failures, or missing records.

Dynamic budgets and cost accounting. The frozen ALFWorld scheduler compares the score percentile to $1 - B_{\mathrm{rem}}/H_{\mathrm{rem}}$, where the remaining teacher calls and steps determine the current

Table 17: ALFWorld Qwen valid-unseen results. Success, teacher-action fraction, and teacher-token fraction are percentages. *Request-level operating point interpolated to ReasonFork-25's teacher-action fraction.

Policy	Success ↑	T-actions ↓	T-tokens ↓
Always Draft	3.73	0.00	0.00
Direct-action ranker	1.49	0.00	0.00
Always Teacher	26.12	100.00	100.00
Confidence-10	9.70	19.40	36.69
History-10	11.19	19.28	30.10
TRIM-Thr-10	11.94	19.46	30.44
Belief takeover-10	11.19	19.91	23.73
ReasonFork-10	12.69	19.46	28.01
Confidence-25	12.69	48.53	72.47
History-25	18.66	49.10	56.92
TRIM-Thr-25	14.18	48.77	62.30
Matched request*	18.26	49.24	47.89
Belief takeover-25	20.15	51.67	37.57
ReasonFork-25	**21.64**	49.24	**49.34**



(a) Budget allocation **(b) Success gains at B25**

Figure 18: **Allocating the teacher budget on ALFWorld.** (a) Measured 10- and 25-call policies; endpoints remain separate references. (b) Paired success gains at the 25-call cap with 95% intervals.

cutoff. The per-episode cap is 10 or 25 calls. Figure 18 connects the measured points and keeps the draft and teacher endpoints as separate references.

TRIM-Thr uses 2,286,259 and 4,679,901 teacher tokens at the 10- and 25-call caps, respectively. Both use the same 7,511,386-token Always-Teacher reference as the primary controls. At 25 calls, its 8,563,571 draft tokens bring total usage to 13,243,472. Request-level allocation and full controls appear in Table 17.

An aggregate summary linearly connects the two measured points and the shared draft/teacher endpoints over 0–100% teacher tokens. Normalized pAUC is 18.06 for ReasonFork and 11.81 for Confidence. The gain is 6.25 points, with a 95% interval of 2.84–9.96 from 20,000 paired episode-bootstrap draws that recompute costs and the teacher denominator. No additional budget points were evaluated.

Counts are unweighted model tokens, with measurement scope defined in Appendix B.5.

E.3 WebShop without an exposed task graph

WebShop evaluates language agents on a simulated shopping site with executable state transitions and graded task rewards (Yao et al., 2022). Unlike the proof domains, it does not expose a target-support graph. The online router receives the instruction and current page together with the committed action history and remaining horizon. It also sees the draft proposal, its forced log probability, and the

Table 18: **ALFWorld total token accounting.** Prompt and completion tokens are included. Counts are in millions; cost per success divides total tokens by the number of completed episodes. Stepwise policies use the 25-call cap.

Policy	Draft	Teacher	Total	Successes	Cost/success
Always Draft	9.202	0.000	9.202	5	1.840
Always Teacher	0.000	7.511	7.511	35	0.215
Confidence	8.585	5.443	14.028	17	0.825
ReasonFork	7.507	3.706	11.213	29	0.387

Table 19: Frozen WebShop results on 500 shared tasks. Reward measures partial credit, while exact success requires task completion. Teacher-action counts refer to executed actions. Total tokens include draft, teacher, and scorer tokens. *Request-level allocation interpolated to ReasonFork's 6.39% teacher actions; additional costs are not inferred.

Policy	Reward ↑	Exact ↑	Teacher act. ↓	Teacher tok. ↓	Total tok. ↓	Wall s ↓
Always Draft	39.50	10.2	0.00	0	3,472,705	7.58
Direct learned action	48.82	10.2	0.00	0	1,876,795	5.24
Request-level (10)	41.15	12.2	11.52	340,142	3,541,827	8.95
History	46.63	14.6	7.65	338,543	4,483,241	10.22
TRIM-Thr	48.12	14.6	6.85	292,843	4,428,199	9.56
Matched request*	40.14	11.16	6.39	—	—	—
ReasonFork	46.35	14.8	6.39	266,533	4,435,933	9.35
Teacher takeover	46.67	15.6	8.80	395,701	4,361,789	10.20
Confidence	51.45	17.4	14.49	453,058	4,479,805	10.99
Request-level (20)	43.31	14.4	25.30	760,881	3,434,517	11.40
Always Teacher	50.70	24.4	100.00	3,133,089	3,133,089	25.08

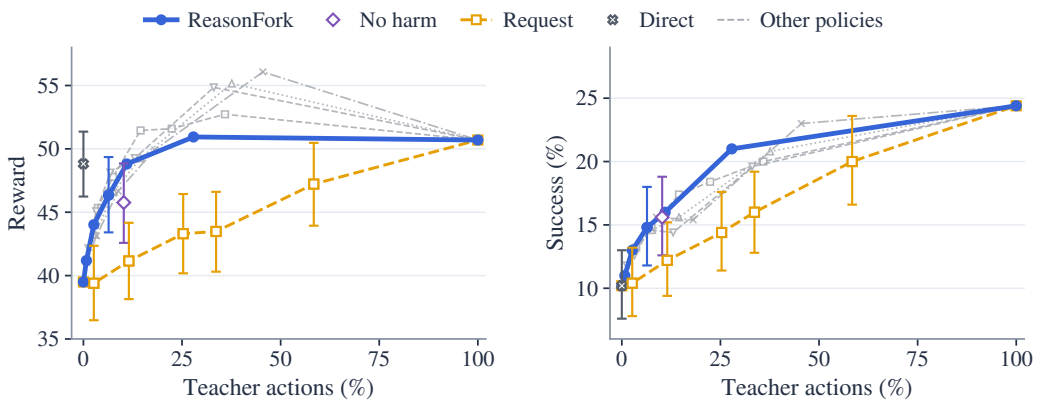
current public interface. The interface records whether search is available. It describes the type and number of visible clickable actions, including repetition and change. The router receives no hidden product metadata, future pages, expert trajectories, or simulator-side target checks.

Frozen protocol. The draft is Qwen2.5-1.5B-Instruct and the teacher is Phi-4. A disjoint bank of 600 training tasks contains 4,161 paired decision states. One-step counterfactual execution labels 1,435 as REPAIR, 310 as HARM, and 2,416 as TIE. We fit separate repair and harm heads. All selectors are calibrated on 120 development tasks before the 500 shared test tasks are evaluated. Every stepwise policy uses the same draft and teacher endpoints. The parser, executable transition interface, maximum 15-step horizon, and development-only threshold selection are also shared. Confidence is the teacher-forced log probability of the parsed draft action rather than a score of a separately regenerated sequence. The Direct control learns to rank visible click actions from 56,908 training candidates. It uses the draft for open-text search and makes no online teacher call.

Figure 19 shows every frozen selector threshold for both partial reward and exact completion.

Decision-level routing and matched request allocation. At its frozen operating point, ReasonFork obtains 46.35% reward and 14.8% exact success with 6.39% of actions executed by the teacher. Linear interpolation between the two adjacent frozen request thresholds at the same realized teacher-action budget gives a 6.21-point reward gain. Its 95% confidence interval runs from 4.02 to 8.46. Exact completion improves by 3.64 points, with a 95% confidence interval from 1.44 to 5.95. Both intervals use 20,000 paired task-level resamples. At the observed 10% selector, reward rises by 5.20 points. The 95% confidence interval runs from 2.87 to 7.62. Against the 20% request selector, which realizes 25.30% teacher actions, the gain is 3.04 points. Its 95% confidence interval runs from 0.29 to 5.80. The matched comparison holds the teacher-action budget fixed and changes where it is spent, from complete requests to selected decisions.

Operating-point comparisons for harm scoring and takeover. Removing the harm head raises the teacher-action fraction from 6.39% to 10.23%, while changing reward from 46.35% to 45.77%. Permanent takeover uses 8.80% teacher actions and reaches 46.67% reward. At these operating points, the signed selector uses fewer teacher actions than permanent takeover while giving nearly the same reward.



(a) Reward **(b) Success**

Figure 19: Complete frozen WebShop allocation frontiers. Reward measures partial progress; exact success requires task completion. Lines connect evaluated thresholds.

Partial reward versus exact completion. ReasonFork improves exact completion over Direct by 4.6 points. Its 95% confidence interval runs from 2.2 to 7.2. Direct retains a higher graded reward without teacher calls. The two metrics distinguish partial progress from completed execution.

E.4 ALFWorld controls with shared fitted heads

Table 4 fixes the public features, three fitted heads, Qwen2.5-7B/14B model pair, and dynamic scheduler. The scores are p_{fork}, $p_R = p_{\text{fork}} p_{R|\text{fork}}$, and $p_R - 8p_H$. Episode-grouped out-of-fold selection on 8,090 events from 164 calibration episodes chooses the harm weight from nine candidates before these evaluation runs. WebShop uses weight 2 in its separate control. Each ALFWorld score is evaluated at 10 and 25 calls, alongside two endpoints, on 134 fixed episodes, yielding 1,072 complete, unique trajectories. Normalized AUC integrates the four-point piecewise-linear curve over 0–100% teacher tokens. The same-run teacher denominator is 7,508,621 tokens; these endpoints are separate from the primary study above. At 25 calls, signed, draft-risk, and repair-only routing complete 28, 19, and 23 episodes with 50.61%, 61.86%, and 59.85% teacher tokens. Paired 95% intervals use 20,000 episode-bootstrap draws that recompute costs and the teacher denominator. The frozen joint criterion required a positive AUC-difference lower bound against draft risk and a nonnegative bound against repair-only. The observed lower bounds are 1.35 and -0.21 AUC points, respectively: the draft-risk bound meets its threshold, while the repair-only interval crosses zero.

F Frozen Transfer

F.1 Transfer details

The main transfer matrix appears in Table 6. Here we distinguish target cost recalibration from fully frozen deployment and report paired task-level uncertainty.

Qwen-to-Phi ReasonFork transfer. The source router is fit only on Qwen2.5-1.5B/7B RuleTaker calibration. Target replay uses Phi-4-mini/Phi-4 on all 300 confirmation problems. Table 6 reports frozen source thresholds; Table 20 adds thresholds recalibrated using only Phi calibration scores and teacher-token charges. Both variants keep the learned weights fixed and are constructed without Phi correctness, repair, or harm outcomes. All points use the corresponding policy's own trajectory.

Table 20: Additional Qwen-to-Phi transfer points with target cost-only recalibration. Source weights remain fixed; no Phi outcome labels select the threshold. Fully frozen thresholds are in Table 6.

Target cost cutoff	Accuracy	T-tok	Repairs	Harms
10%	89.00	11.86	35	0
30%	94.00	30.38	53	3
50%	95.00	54.60	55	2

F.2 RULETAKER-TO-PROOFWRITER FROZEN TRANSFER

The RuleTaker scaler, classifier, cost proxy, and 30% threshold are applied unchanged on the fixed ProofWriter target sets, with model selection completed on the source task. Table 6 gives the operating points.

On QDep3, frozen transfer improves over Always Draft by 16.22 points. The 95% paired-bootstrap interval runs from 10.14 to 22.97 points, and the McNemar test gives $p = 3.03 \times 10^{-6}$. Its 3.38-point advantage over Always Teacher has six paired wins and one reversal ($p = 0.125$). Across QDep2/3/5, the transferred router is 0.67 points above Always Teacher with 61.07% fewer teacher tokens. The paired difference interval, from -2.33 to 3.67 points, includes zero.