BEYOND REFUSAL: POLICY-GROUNDED ON-POLICY DISTILLATION FOR SAFETY DECISIONS AND RECOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety distillation can teach a language model to refuse without teaching it the policy distinction behind the refusal—or how to recover once generation has entered a harmful trajectory. We study this gap in on-policy distillation (OPD) through representation readouts, trajectory diagnostics, and budget-matched interventions. Rule and boundary generalization lag behind refusal, while recovery states become less frequent in student rollouts as refusals emerge; explicit policy supervision improves boundary decisions, and targeted state coverage further improves recovery. Motivated by these findings, we introduce DEEPEN, a policy-grounded OPD framework that makes safety judgments explicit and restores access to decision-critical states. DEEPEN predicts structured safety labels to condition generation, trains on boundary pairs and harmful-prefix starts, and uses brief teacher handoffs triggered by policy errors to supply a corrective segment before returning control to the student. Overlap-aware token-level distillation reinforces local corrections, while handoffs anneal to zero for student-only deployment. Across four preselected representatives from distinct model families under matched response-token budgets, DEEPEN improves mean boundary score (BDA) by 6.6 percentage points and useful recovery by 18.1 points over an adapted PolicyAlign baseline; the largest task-wise capability decrease is 4.5 points relative to the initial student. These results support treating safety distillation as a joint problem of policy-decision learning and safety-critical state coverage, rather than refusal imitation alone.

1 INTRODUCTION

On-policy distillation (OPD) trains student language models with teacher feedback on their own generated responses (Agarwal et al., 2024; Xu et al., 2026). For safety, the goal is not simply to reproduce refusals. Students must distinguish harmful requests from benign boundary cases and recover when a response has already begun to go wrong.

Existing safety-oriented OPD methods condition teachers on safety policies or privileged context and distill their token predictions (Fu et al., 2026; Wu et al., 2026). However, similar refusals can conceal different rules, leaving the judgments behind them implicit. Student rollouts also determine where feedback is available: early refusal can bypass states that require recovery. This raises a joint challenge: *How can OPD transfer the safety judgments behind responses and teach students to apply them on unfamiliar requests and harmful response prefixes?*

We investigate this challenge through representation readouts, activation interventions, and trajectory diagnostics (Section 3; Appendix E.3). Refusal improves faster than rule and boundary generalization, while recovery-state visits decline as refusal emerges. Budget-matched interventions show that explicit policy rationales improve boundary decisions and targeted state coverage further improves recovery. These findings motivate supervising both *what safety judgment to make* and *how to act on it along a continuation*.

Building on these findings, we propose DEEPEN (Section 4). Students predict structured safety judgments that condition response generation, and train on boundary cases and harmful-prefix continuations. When policy errors occur, brief teacher handoffs supply corrective segments before the

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.

student resumes. Overlap-aware token-level distillation reinforces corrective actions, while handoffs are gradually withdrawn so that deployment uses only the student.

Under matched response-token budgets across four preselected model-family representatives, DEEPEN improves mean boundary score (BDA) by 6.6 percentage points and useful recovery by 18.1 points over adapted PolicyAlign (PA*). The largest task-wise capability decrease is 4.5 points relative to the initial student. Section 5 evaluates these gains alongside benign answering, capability retention, and training and deployment costs.

Our contributions are:

- **A diagnosis of safety transfer.** We connect refusal–rule generalization gaps to decision supervision and corrective-state coverage.
- **A method that couples judgment and recovery.** DEEPEN combines structured policy supervision, targeted training states, and bounded teacher handoffs with local distillation.
- **A controlled evaluation.** Cross-family comparisons and budget-matched ablations assess boundary decisions, recovery, capability retention, and resource costs.

2 Related Work

Safety judgments and representations. Qi et al. (2025) identify alignment concentrated in early tokens and train recovery from harmful prefixes. Deliberative Alignment teaches explicit reasoning over safety specifications (Guan et al., 2024). Representation engineering reads and controls safety-relevant internal representations (Zou et al., 2023). We study how rule representations and student trajectory coverage jointly constrain safety knowledge transfer in OPD.

On-policy safety distillation. GKD trains students on self-generated sequences using teacher feedback (Agarwal et al., 2024). OPSA and PolicyAlign extend this paradigm to safety through privileged-context and policy-conditioned teachers, respectively (Fu et al., 2026; Wu et al., 2026). DEEPEN additionally supervises student-predicted risk, rules, and boundary conditions as explicit learning targets that condition response generation, coupling policy knowledge to its application.

Trajectory intervention and distribution matching. TrOPD combines trust-region supervision with teacher-prefix guidance, while Relay-OPD introduces brief teacher handoffs for prefix failures (Xing et al., 2026; Xu et al., 2026). Tail-Aware Top-k OPD preserves probability mass outside the teacher's top-k tokens during distillation (Huang & Wei, 2026). DEEPEN uses policy errors to trigger teacher corrections and trains students to apply the corrected judgments during their own continuations, while overlap-dependent distillation supplies local action supervision.

3 From Refusal to Safety Decisions: Examining OPD Supervision

Scope of the diagnostic analysis. The diagnostics in this section examine how token-level supervision, policy information, and state coverage affect safety transfer. Metric definitions and experimental protocols are given in Appendix B.

On-policy distillation (OPD) supplies token-level teacher feedback on states sampled from the current student policy (Agarwal et al., 2024). We ask whether improved refusal also transfers safety decisions. Learning dynamics and position ablations first show that short refusal prefixes capture most standard safety gains, whereas decision-boundary generalization and recovery from harmful prefixes improve less. Trajectory diagnostics then reveal limited coverage of teacher-provided policy rationales and declining visits to recovery states as refusal emerges. Controlled interventions subsequently test these diagnoses: correct rationales improve boundary decisions, and targeted state coverage further improves recovery. Together, these observations motivate verifiable policy-rationale supervision and broader coverage of safety-critical continuations while retaining current-policy student generation.

Experimental overview. We fix the teacher, student initialization, and requests, comparing sampled-token and full-vocabulary reverse-KL OPD (Gu et al., 2024). Metrics are safety error rate,

the harmonic mean of harmful-request refusal and benign-request compliance (boundary score), and attack success after a 20-token partially compliant prefix (Prefill-20). Appendix B records the settings, evaluation protocols, and budget accounting.

3.1 Refusal Improves Faster than Decision-Boundary Generalization

To test whether safety gains extend beyond refusal, we track checkpoints and decompose full-vocabulary KL changes on fixed, independent prefixes (Appendix B.2), motivated by prefix vulnerabilities (Qi et al., 2025). Safety error rate falls from 47.2% to 10.4% early and 7.6% finally; the first four positions account for 72% of the early KL decrease. Boundary score reaches only 63.4%, while Prefill-20 remains 50.8% (Figure 1a).

To locate these gains, we train Head-only (first four tokens), Tail-only, and Full variants, matching non-masked supervised tokens. Head-only retains about 90% of Full's standard improvement but reaches only 55.6% boundary score; Tail-only favors recovery over standard safety (Figure 1b). Minimal pairs differing only in policy-relevant factors still exhibit under-refusal and over-refusal, with gaps on rule, language, topic, and compositional holdouts and counterfactual tests (Appendix B.2). Shared parameters make position masks imperfect isolators of semantic abilities.

These diagnostics separate refusal gains from boundary generalization. Concentrated supervision can be effective (Fu et al., 2026; Li et al., 2026a); the concern is transfer, not localization alone. If candidate rules induce identical observations on queried states, behavioral KL cannot distinguish them (Appendix C.1 and the constructed example in Appendix C.2). Which rule-discriminating information reaches student trajectories?

3.2 Limited Policy-Rationale Coverage and Declining State Visitation

To examine that information, we use independently checked policy rationales: summaries of risk category, applicable rule, boundary factors, and response strategy. Teacher-forced top-5 inclusion reaches 92.1% for refusal tokens but 28.4% for boundary rationales; natural full-summary coverage across 32 responses is 9.6% (Figure 1c). Full-logit OPD improves transfer only partially. Natural-rollout coverage and common-format local support measure different objects; format controls, support scans, and signal probes appear in Appendix B.3. Summaries are external justifications, not readouts of internal reasoning.

To track changing supervision opportunities, we measure recovery-state visitation alongside refusal formation. Visits decline from 4,180 to 310 per million rollout tokens as refusal increases and recovery improvement slows (Figure 1d). Increasing natural-sampling budgets provides limited relief (Appendix B.3). These are continuation states bypassed by early refusal; coverage of policy-relevant prompt factors remains determined by the request distribution.

For request x and fixed prefix $h_{1:k}$, the current student policy π_θ assigns

$$P_\theta(h_{1:k} \mid x) = \prod_{i=1}^{k} \pi_\theta(h_i \mid x, h_{<i}). \tag{1}$$

Fixing the other conditional probabilities at positive values, reducing the entry probability lowers this path's visitation probability. Under fixed-budget sampling, rarer visits offer fewer expected opportunities for direct feedback, even with full logits. This conditional explanation motivates separate interventions on rationale content and state coverage; temporal association alone does not identify their training effects.

3.3 Policy Rationales and State Coverage Improve Safety Transfer

We cross policy-rationale supervision with state intervention in a 2×2 design. Rationales supply rules and boundary factors. State intervention changes starting prefixes but samples subsequent tokens from the current student policy. Initialization and effective supervised-token budgets are matched; generation and teacher costs are tracked separately (Appendix B.4). Decision labels, complete teacher-generated outputs, and full logits control for decision information, teacher trajectories, and local feedback (Appendix B.4).

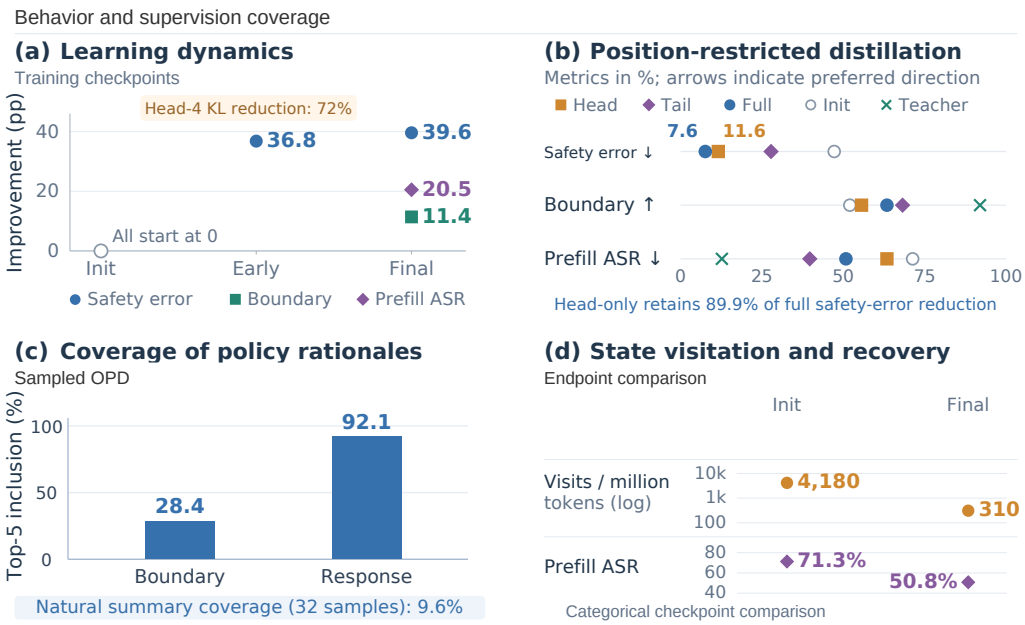


Figure 1: Behavior and supervision coverage in the diagnostic study. (a) Categorical checkpoints and early KL decrease. (b) Position ablations. (c) Teacher-forced inclusion, distinct from natural coverage. (d) Visitation and recovery endpoints. Protocols: Appendix B.

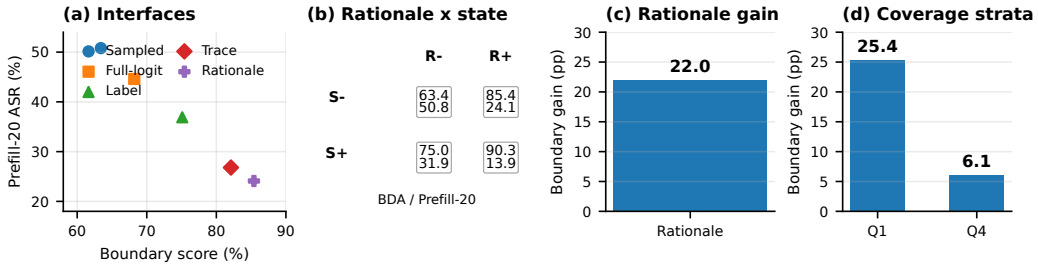


Figure 2: Training interventions in the diagnostic study, arranged as four parallel panels for direct comparison. (a) Supervision interfaces. (b) Conditional gains, which do not establish superadditivity. (c) The 22.0-point rationale–baseline contrast, not a matched content-control estimate. (d) Low- and high-coverage endpoint groups. Details: Appendix B.4.

Rationales alone raise boundary score from 63.4% to 85.4%; state intervention alone lowers Prefill-20 from 50.8% to 31.9%. Their combination reaches 90.3% and 13.9%, respectively (Figure 2a–b). Correct rationales outperform length-, format-, and label-matched generic explanations, refusal paraphrases, and shuffled rationales on boundary and counterfactual tests, with less over-refusal. Gains are larger at lower initial rationale coverage (Figure 2d; Appendix B.4).

These conditional gains motivate two design requirements: verifiable policy rationales and broader critical-state coverage with current-policy continuation. Implementation and evaluation are presented in Sections 4 and 5; Appendix B.5 discusses the scope of the diagnosis and its relation to prior work.

4 Method

Motivated by Section 3, we propose **DEEPEN: Policy-Grounded Interventional On-Policy Distillation**. The diagnosis isolates two linked gaps: response agreement need not identify the policy distinction behind a decision, and student rollouts may miss the states needed to apply or recover it. DEEPEN therefore makes policy judgments explicit, samples boundary and recovery states, and combines local distribution matching with bounded teacher intervention, linking decision supervi-

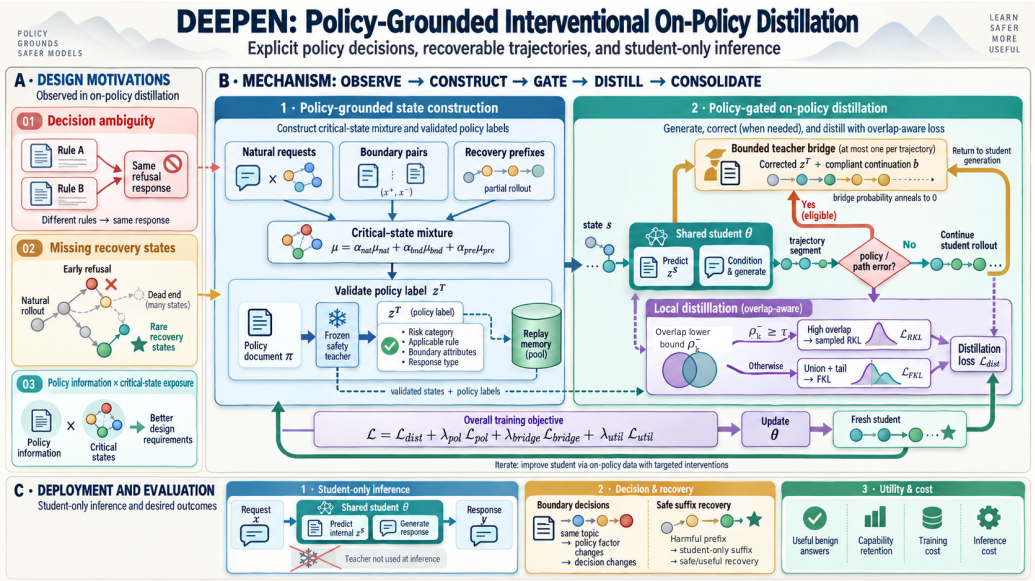


Figure 3: **DEEPEN mechanism.** The diagnosis (left) separates policy-rule ambiguity from missing recovery states. DEEPEN constructs a replay memory $\mathcal{M}$ from natural, boundary, and recovery states, then validates structured policy labels z^T. During on-policy training, the student predicts an internal z^S and emits a bounded token segment; a policy/path gate routes reliable states to overlap-aware local RKL or union-plus-tail FKL and eligible errors to a one-segment teacher bridge before the student resumes. The bridge probability anneals to zero, while the joint objective retains label, distribution, bridge, and benign-utility terms; deployment uses only the student.

sion to the generation paths used in training. Figure 3 summarizes this observation-to-method chain before the component-level definitions below.

4.1 Structured Policy Supervision

To separate rule application from topic-based refusal (Observation 3.1), we supervise a visible state $s = (x, h)$, with request x and response prefix h. The frozen safety teacher returns $z^T(s) = (c^T, r^T, \boldsymbol{b}^T, a^T)$: risk category, applicable rules, boundary attributes (e.g., authorization and intent), and permitted response type. A versioned schema supports unknown attributes, multiple rules, and equivalent identifiers; validation, rule–boundary–response consistency, and two-annotation agreement retain labels (Appendix D.1). Thus, the labels encode task-relevant decisions and constrain rules compatible with response supervision (Appendix A.1).

The student shares parameters between a label predictor $g_\theta(z \mid s)$ and response model $\pi_\theta(\cdot \mid s, z)$. It predicts z^S at the start and after each segment of at most B response tokens, conditioning the next segment on that label. Labels are serialized only as internal control context, so users see response tokens; this follows concept-conditioned prediction (Koh et al., 2020). At reliable checkpoints $\mathcal{S}_{\text{pol}}$, we optimize length-normalized label cross-entropy:

$$\mathcal{L}_{\text{pol}} = \mathbb{E}_{s\sim\mathcal{S}_{\text{pol}}}\left[-\frac{1}{|z^T(s)|}\log g_\theta(z^T(s) \mid s)\right]. \tag{2}$$

Teacher forcing is used within each target label sequence. Generation otherwise uses student-predicted labels, except for the bounded intervention and the immediately following student segment defined below.

The policy-information diagnostic in Observation 3.3 motivates controls using correct labels, generic safety text, refusal paraphrases, and shuffled rules (Appendix E.4). Auxiliary-only supervision and field substitutions test causal influence on responses (Appendix E.4), while hidden-state readouts and activation interventions test decision representations (Appendix E.3). The objectives therefore act on both labels and output distributions.

4.2 Critical-State Sampling and Teacher Intervention

Policy supervision specifies what to learn; state sampling provides opportunities to apply it. To address the coverage gap in Observation 3.2, we sample starting states from

$$\mu = \alpha_{\text{nat}}\mu_{\text{nat}} + \alpha_{\text{bnd}}\mu_{\text{bnd}} + \alpha_{\text{pre}}\mu_{\text{pre}}, \qquad \alpha_j \geq 0, \quad \sum_j \alpha_j = 1. \tag{3}$$

Natural requests start empty. Boundary pairs hold topic fixed while changing one policy factor, including both benign and harmful sides. Prefix states contain partial harmful responses or validated policy-disagreement contexts collected during training. A fixed-capacity replay pool exposes new states from the next update; supplied prefixes are context only and receive zero response-imitation loss (Appendix D.2).

Student segments are sampled from the current policy given the selected context, following DAgger's learner-state principle (Ross et al., 2011) while controlling starting-state exposure. Positive mass preserves selected recovery starts even when natural rollouts terminate early (Appendix A.2); subsequent occupancy still depends on student continuation. Appendix E.4 separates these quantities and gives boundary/prefix ablations, extra natural-rollout budgets, and matched random-state controls.

Local feedback may still miss multi-step recovery. At reliable checkpoints, a rule/boundary error, wrong response type, or preceding policy violation triggers eligibility for a brief teacher intervention. Rule equivalence and a fixed semantic rubric determine eligibility; next-token divergence affects only the local objective in Section 4.3. The teacher supplies a validated label and compliant continuation b of at most $L_{\max}$ tokens, then returns generation to the student; the label stays active for one student segment before prediction is restored. This connects the trajectory construction to teacher-prefix guidance and Relay-OPD (Xing et al., 2026; Xu et al., 2026). Each trajectory gets at most one attempt under a global budget, with no retry for empty or rejected continuations, and $\eta_i = \eta_0 \max(0, 1 - i/T_{\text{off}})$ reaches zero before update T. When intervention is unavailable, reliable states retain label and distribution supervision; the final phase is student-generated while teacher supervision remains in the objective. Appendix D.3 specifies execution and Appendix E.6 tests label-only correction, local feedback, intervention length, resumption, and benign over-refusal correction.

4.3 Distillation Objectives and Joint Optimization

The trajectories determine both training contexts and token generators. At a student position, let $p = \pi_\theta(\cdot \mid s, z)$ and $q = \pi_T(\cdot \mid s, z^T)$, with active student and validated teacher labels. Both models read the same request and prefix and share tokenizer and vocabulary $\mathcal{V}$; the teacher also receives the policy document. Following GKD (Agarwal et al., 2024), we use RKL $D_{\text{KL}}(p\|q)$ and FKL $D_{\text{KL}}(q\|p)$. For student logit u_v, FKL has gradient $p_v - q_v$; with positive teacher mass, RKL's coordinate gradient vanishes as $p_v \to 0$, whereas FKL still pushes toward teacher-preferred tokens (Appendix A.3).

To retain these token-level signals with compact probability storage, fix the union of teacher and rollout-student top-k sets at parameters θ_i:

$$\begin{aligned} U &= \text{TopK}(p_{\theta_i}) \cup \text{TopK}(q), \qquad R = \mathcal{V} \setminus U, \\ \bar{p} &= ((p_v)_{v\in U}, p_R), \qquad p_R = 1 - \sum_{v\in U} p_v, \end{aligned} \tag{4}$$

with $\bar{q}$ defined analogously. This compact tail event follows tail-aware distillation (Huang & Wei, 2026): on the union, each retained token keeps gradient $p_v - q_v$, while the tail shares one aggregate correction.

We select RKL when the computable lower bound $\rho_k^- = \sum_{v\in U} \min(p_{\theta_i,v}, q_v)$ certifies sufficient teacher overlap. Since $\rho_k^- \leq \sum_{v\in\mathcal{V}} \min(p_{\theta_i,v}, q_v) = 1 - \text{TV}(p_{\theta_i}, q)$, define

$$\begin{aligned} \ell_{\text{dist}}(s) &= \begin{cases} D_{\text{KL}}(p\|q), & \rho_k^- \geq \tau, \\ D_{\text{KL}}(\bar{q}\|\bar{p}), & \rho_k^- < \tau, \end{cases} \\ \mathcal{L}_{\text{dist}} &= \mathbb{E}_{s\sim\mathcal{S}_S}[\ell_{\text{dist}}(s)], \end{aligned} \tag{5}$$

where $\tau \in (0, 1)$ and $\mathcal{S}_S$ contains reliable student-generated positions, including those after supplied prefixes and interventions. Reliable states with inconclusive overlap estimates use FKL. The set and branch are fixed per update; RKL uses the sampled-token estimator in Appendix A.4. Appendices E.6 and E.9 compare fixed Jensen–Shannon divergence (JSD) and audit the top-k approximation against full KL, respectively.

Teacher-generated segments $\mathcal{B}_T$ provide supervised continuations:

$$\mathcal{L}_{\text{bridge}} = \mathbb{E}_{(s,z^T,b)\sim\mathcal{B}_T}\left[-\frac{1}{|b|}\sum_{j=1}^{|b|}\log \pi_\theta(b_j \mid s, z^T, b_{<j})\right]. \tag{6}$$

They provide both continuation contexts and target tokens (Appendix A.5); resumed student positions use Equation (5). To preserve existing capabilities, freeze an initial student π_0 before label warm-up and construct independent benign reference states $\mathcal{S}_0$. Output distillation (Li & Hoiem, 2016) gives

$$\mathcal{L}_{\text{util}} = \mathbb{E}_{s\sim\mathcal{S}_0} D_{\text{KL}}\big(\pi_0(\cdot \mid s) \,\|\, \pi_\theta(\cdot \mid s, z^S)\big). \tag{7}$$

This full-vocabulary FKL penalizes drift on benign contexts, while benign boundary examples receive teacher supervision about when and how to answer. Appendix A.6 gives the local bound; Appendix E.5 specifies the factorial comparison and capability-retention controls.

The joint objective is

$$\mathcal{L} = \mathcal{L}_{\text{dist}} + \lambda_{\text{pol}}\mathcal{L}_{\text{pol}} + \lambda_{\text{bridge}}\mathcal{L}_{\text{bridge}} + \lambda_{\text{util}}\mathcal{L}_{\text{util}}. \tag{8}$$

Each term is normalized by valid supervision units. After a short label warm-up, each iteration collects fresh student continuations and makes one accumulated-gradient update with contexts, labels, masks, and routing fixed (Algorithm 1; Appendix D.4). We select weights and thresholds on validation data. Inference retains only the student, internal labels, and fixed checkpoints. Thus, the factorial controls in Appendix E.5 separate local feedback from teacher intervention and connect the information–coverage diagnosis to the main evaluation in Section 5.

5 Experiments

Data and statistical scope. The evaluation covers safety, boundary decisions, recovery, capability retention, and resource use under matched response-token budgets. Intervals summarize variation across request cohorts. Appendix E provides the full protocols, metric definitions, and comparisons across model configurations.

The evaluation follows the three observations in Section 3: it first tests safety jointly with useful answering and capability retention, then rules out refusal-only explanations, separates policy information from critical-state coverage, and finally distinguishes local distribution updates from multi-step recovery before accounting for training and deployment cost.

Settings and scope. We compare initialization, RKL, JSD, and controlled PA*/Relay* variants under matched student response-token budgets across four preselected model-family representatives; the canonical mechanism study uses Qwen3-8B. Models, datasets, baseline adaptations, compute-matched OPSA controls, sample budgets, statistics, and reporting conventions are specified in Appendices E.1 and E.2. Safety, boundary score (BDA), useful recovery, benign answering, and capability scores retain their original definitions and denominators.

5.1 Safety, Helpfulness, and Generalization

Table 1 shows the joint effect: relative to PA*, DEEPEN improves boundary decisions, useful recovery, and benign answering across the four representatives without relying on uniformly lower direct ASR. Capability retention remains a separate constraint; full configurations and held-out checks are reported in Appendices E.2 and E.8. We therefore next compare methods at similar surface safety.

5.2 Beyond Refusal Behavior

At a validation-ASR-matched operating point, Figure 4 shows the gains in boundary decisions and recovery. The fixed-grid and pre-label controls support an effect beyond refusal frequency; residual

Table 1: **Response-token-matched comparison.** HB/JB/MB and Indic are attack-success-rate (ASR) metrics for HarmBench, JailbreakBench, MultiBreak, and equal-language IndicSafeEval; R_{128} is useful suffix recovery and BDA is the harmonic boundary score. Answer/Useful use the XSTest benign cohort; full definitions and configurations are in Appendix E.1–E.2.

Safety and recovery (%)

Model / method		HB ↓	JB ↓	MB ↓	R_{128} ↑	BDA ↑	Indic ↓
Qwen3.5-4B	Init.	16.8	26.0	43.7	40.3	79.0	42.3
	RKL	1.5	14.0	28.9	51.3	81.0	27.5
	JSD	1.8	11.7	25.5	58.1	84.1	25.6
	PA*	0.2	8.7	23.3	60.4	84.8	23.6
	DEEPEN	1.0	6.0	12.5	77.3	90.7	19.2
Gemma 4 E4B-it	Init.	15.0	27.3	38.1	40.2	78.7	39.7
	RKL	1.0	12.0	23.0	53.2	81.3	24.0
	JSD	1.0	9.3	20.1	58.3	83.9	21.9
	PA*	1.0	6.3	18.3	60.3	84.6	20.0
	DEEPEN	1.0	2.7	9.5	79.1	91.4	15.9
Llama 3.1 8B	Init.	21.2	39.7	48.3	34.0	74.0	48.9
	RKL	7.7	20.7	32.9	45.3	76.2	33.7
	JSD	8.0	19.7	30.2	51.6	79.5	31.6
	PA*	5.8	16.3	27.8	53.6	79.9	29.8
	DEEPEN	7.2	12.0	18.0	72.2	87.1	25.2
Ministral 3 8B	Init.	18.7	37.7	43.9	36.1	75.7	46.5
	RKL	5.8	22.7	30.0	47.0	77.9	31.1
	JSD	6.5	19.3	26.9	52.2	80.7	28.8
	PA*	5.0	17.3	24.7	53.8	81.4	26.8
	DEEPEN	5.7	12.3	17.2	71.8	88.0	22.5

Benign answers and capabilities (%)

Model / method		Answer ↑	Useful ↑	MMLU ↑	Math ↑	Code ↑	IF ↑
Qwen3.5-4B	Init.	94.3	84.8	70.2	75.7	53.7	83.7
	RKL	86.9	76.1	68.5	72.9	51.4	83.0
	JSD	89.3	79.2	69.4	74.5	53.0	83.2
	PA*	88.0	77.7	69.1	74.3	52.6	83.2
	DEEPEN	93.7	86.0	68.9	73.8	52.1	83.2
Gemma 4 E4B-it	Init.	93.5	83.3	70.2	74.4	51.8	86.1
	RKL	87.9	75.6	68.2	72.5	49.8	85.2
	JSD	89.7	78.8	69.4	73.5	51.2	85.8
	PA*	89.2	77.2	69.0	73.1	51.0	85.7
	DEEPEN	93.1	84.3	67.3	71.1	48.7	84.9
Llama 3.1 8B	Init.	90.8	80.4	49.3	47.3	38.6	76.6
	RKL	84.0	70.8	47.6	44.8	36.4	76.0
	JSD	86.7	74.9	48.4	46.5	37.7	76.4
	PA*	85.3	72.5	48.2	45.9	37.3	76.3
	DEEPEN	90.7	81.2	48.3	46.4	37.6	76.4
Ministral 3 8B	Init.	93.1	82.5	54.8	59.1	48.1	81.0
	RKL	85.2	73.2	53.1	57.1	45.8	80.0
	JSD	87.1	76.1	54.2	58.6	47.5	80.6
	PA*	86.1	74.5	53.8	58.3	47.0	80.4
	DEEPEN	92.4	83.2	54.1	58.4	47.4	80.5

ASR handling is given in Appendix E.3. Having established the behavioral effect, we next separate supplied information from the states receiving it.

5.3 Policy Supervision and Critical-State Coverage

Figure 5 shows that correct policy information and targeted starts improve boundary decisions and recovery beyond extra supervision. The shared factorial separates information grounding from state coverage, with full controls and budgets in Appendix E.4. With these factors separated, we next test how local updates and trajectory continuation contribute to recovery.

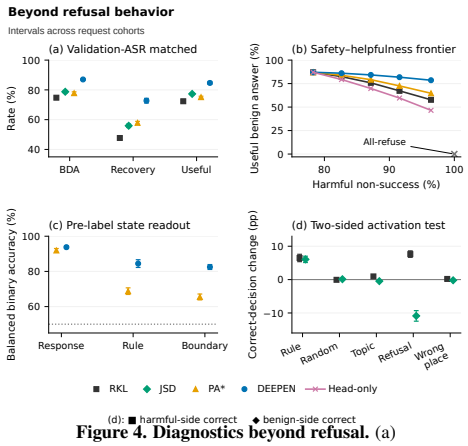


Figure 4. Diagnostics beyond refusal. (a) Validation-ASR-matched outcomes; (b) safety–helpfulness frontier; (c) pre-label readouts; (d) activation tests and controls. 95% cohort intervals; details in Appendix E.3.

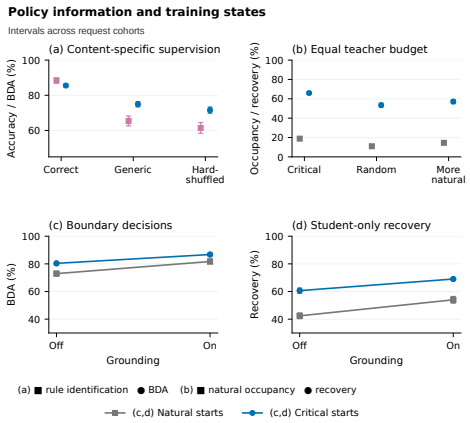


Figure 5. Policy-information and state-coverage controls. (a) Matched content; (b) equal-teacher-budget starts; (c,d) no-handoff grounding/state factorial. 95% cohort intervals; details in Appendix E.4.

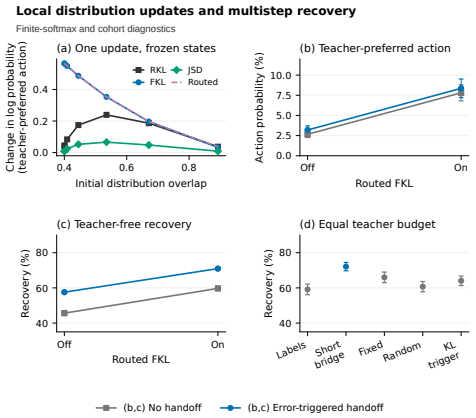


Figure 6. Local-update and path-recovery diagnostics. (a) Exact finite-softmax update; (b,c) routed-FKL-by-handoff factorial; (d) equal-teacher-budget controls. 95% cohort intervals.

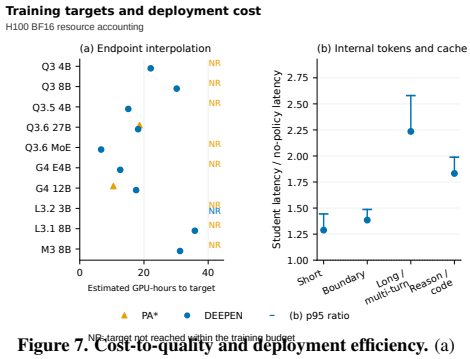


Figure 7. Cost-to-quality and deployment efficiency. (a) Endpoint-interpolated cost; NR means target not reached. (b) Mean and p95 latency ratios. Details in Appendix E.7.

5.4 Local Distribution Matching and Multi-Step Recovery

Figure 6 and Appendix E.6 separate routed local-update effects from bounded-handoff recovery using equal-budget, placement, and trigger controls.

5.5 Training Cost and Inference Efficiency

Figure 7 and Appendix E.7 report H100-BF16 endpoint-interpolated cost to the joint target and student-only deployment overhead; Appendix E.8 gives robustness and remaining-error diagnostics.

6 Conclusion

DEEPEN combines explicit policy supervision, targeted boundary/recovery states, routed local distillation, and bounded teacher handoffs that anneal to student-only deployment. Across four preselected model-family representatives under matched response-token budgets, it improves mean boundary score (BDA) by 6.6 pp and useful recovery by 18.1 pp over adapted PolicyAlign; the largest reported task-wise capability decrease is 4.5 pp. The results support supervising both safety judgments and recovery-relevant states.

Ethics Statement

This work evaluates safety alignment using established safety and over-refusal benchmarks and controlled harmful-prefix continuations. Potentially harmful prompts and model outputs are used only within the research training and evaluation procedures described in the paper; code evaluation uses a constrained execution sandbox. The manuscript reports aggregate outcomes rather than reproducing harmful model completions, and the evaluation tracks benign answering and over-refusal alongside harmful-request safety.

AI Use Statement

Generative AI tools were used to coding and assist with manuscript formatting, LaTeX layout refinement, and editorial review. The authors reviewed the resulting edits and take responsibility for the final content.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13649.

Yushi Bai et al. LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks, 2024. URL https://arxiv.org/abs/2412.15204v2.

Hongye Cao et al. SafeDialBench: A Fine-Grained Safety Evaluation Benchmark for Large Language Models in Multi-Turn Dialogues with Diverse Jailbreak Attacks, 2025. URL https://arxiv.org/abs/2502.11090.

Patrick Chao et al. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models, 2024. URL https://arxiv.org/abs/2404.01318.

Justin Cui et al. OR-Bench: An Over-Refusal Benchmark for Large Language Models, 2024. URL https://arxiv.org/abs/2405.20947.

Yu Fu, Longxuan Yu, Haz Sameen Shahgir, Zhipeng Wei, Hui Liu, N. Benjamin Erichson, and Yue Dong. Reducing the safety tax in LLM safety alignment with on-policy self-distillation, 2026. URL https://arxiv.org/abs/2605.15239.

Google DeepMind. Gemma 4 model card, 2026. URL https://ai.google.dev/gemma/docs/core/model_card_4. Model card; accessed 2026-09-26.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. Deliberative alignment: Reasoning enables safer language models, 2024. URL https://arxiv.org/abs/2412.16339.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.507. URL https://aclanthology.org/2023.findings-acl.507/.

Huipeng Huang and Hongxin Wei. Tail-aware top-k on-policy distillation, 2026. URL https://arxiv.org/abs/2608.14728.

Naman Jain et al. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code, 2024. URL https://arxiv.org/abs/2403.07974v2.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Hao Li, Jingkun An, Zijun Song, Pengyu Zhu, Rui Li, Hao Wang, Wendi Feng, Yesheng Liu, Lijun Li, Jin-Ge Yao, and Lei Sha. SafeSteer: Localized on-policy distillation for efficient safety alignment, 2026a. URL https://arxiv.org/abs/2606.02530.

Lijun Li et al. SALAD-Bench: A Hierarchical and Comprehensive Safety Benchmark for Large Language Models, 2024. URL https://arxiv.org/abs/2402.05044.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision*, 2016. URL https://arxiv.org/abs/1606.09282.

Zihan Li, Feifei Li, and Wenhui Que. DUET: Dual-teacher on-policy distillation via same-weight disagreement for prohibition compliance, 2026b. URL https://arxiv.org/abs/2608.14644.

Hunter Lightman et al. Let's Verify Step by Step, 2023. URL https://arxiv.org/abs/2305.20050.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. URL https://thinkingmachines.ai/blog/on-policy-distillation/. Technical blog.

Mantas Mazeika et al. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal, 2024. URL https://arxiv.org/abs/2402.04249v2.

Meta AI. Llama 3.1 8B Instruct, 2024a. URL https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct. Official model card; accessed 2026-09-26.

Meta AI. Llama 3.2 3B Instruct, 2024b. URL https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct. Official model card; accessed 2026-09-26.

Mistral AI. Ministral 3 8B Instruct 2512, 2026. URL https://huggingface.co/mistralai/Ministral-3-8B-Instruct-2512. Model card; accessed 2026-09-26.

Saikat Mondal et al. IndicSafeEval: Safety Robustness of Large Language Models under Multilingual Persuasive Jailbreak Attacks, 2026. URL https://arxiv.org/abs/2609.03781v2.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/88be023075a5a3ff3dc3b5d26623fa22-Abstract-Conference.html.

Qwen Team. Qwen3.5-4B, 2026a. URL https://huggingface.co/Qwen/Qwen3.5-4B. Official model card; accessed 2026-09-26.

Qwen Team. Qwen3.6-27B, 2026b. URL https://huggingface.co/Qwen/Qwen3.6-27B. Official model card; accessed 2026-09-26.

Qwen Team. Qwen3.6-35B-A3B, 2026c. URL https://huggingface.co/Qwen/Qwen3.6-35B-A3B. Official model card; accessed 2026-09-26.

David Rein et al. GPQA: A Graduate-Level Google-Proof Q&A Benchmark, 2023. URL https://arxiv.org/abs/2311.12022v1.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Paul Röttger et al. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models, 2024. URL https://arxiv.org/abs/2308.01263.

Shivalika Singh et al. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation, 2024. URL https://arxiv.org/abs/2412.03304.

Jialin Song et al. MultiBreak: A Scalable and Diverse Multi-turn Jailbreak Benchmark for Evaluating LLM Safety, 2026. URL https://arxiv.org/abs/2605.01687v1.

Alexandra Souly et al. A StrongREJECT for Empty Jailbreaks, 2024. URL https://arxiv.org/abs/2402.10260v2.

Yubo Wang et al. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark, 2024. URL https://arxiv.org/abs/2406.01574.

Ming Wen, Yuxuan Liu, Kun Yang, Yunhao Feng, Zhuoer Xu, Yuhao Sun, Shiwen Cui, Xiang Zheng, Yi Liu, Xingjun Ma, and Yu-Gang Jiang. Constitutional on-policy safe distillation, 2026. URL https://arxiv.org/abs/2606.03089.

Chang Wu, Junfeng Fang, Houcheng Jiang, Kai Tang, Pengyu Cheng, Xiaoxi Jiang, Guanjun Jiang, and Xiang Wang. PolicyAlign: Direct policy-based safety alignment for large language models, 2026. URL https://arxiv.org/abs/2606.25442v1. Version 1.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust region on-policy distillation, 2026. URL https://arxiv.org/abs/2606.01249v3. Version 3.

Haolei Xu, Xiaowen Xu, Haiwen Hong, Zixuan Ni, Hongxing Li, Yiwen Qiu, Weiming Lu, and Yongliang Shen. Pass the baton: Trajectory-relayed on-policy distillation, 2026. URL https://arxiv.org/abs/2607.26057.

An Yang et al. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

Shuzhou Yuan et al. Beyond Over-Refusal: Scenario-Based Diagnostics and Post-Hoc Mitigation for Exaggerated Refusals in LLMs, 2025. URL https://arxiv.org/abs/2510.08158v3.

Bo Zeng et al. Marco-Bench-MIF: On Multilingual Instruction-Following Capability of Large Language Models, 2025. URL https://arxiv.org/abs/2507.11882v1.

Zhihao Zhang et al. Health-ORSC-Bench: A Benchmark for Measuring Over-Refusal and Safety Completion in Health Context, 2026. URL https://arxiv.org/abs/2601.17642v2.

Jeffrey Zhou et al. Instruction-Following Evaluation for Large Language Models, 2023. URL https://arxiv.org/abs/2311.07911v1.

Jingang Zhou, Yuyi Zhou, Haiyang Guo, Xukai Wang, Shuai Feng, Sirui Gao, Jian Xu, Qingpei Guo, and Xu-Yao Zhang. Teacher should think ahead: Adaptive continuations for reliable on-policy distillation, 2026. URL https://arxiv.org/abs/2609.22254.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency, 2023. URL https://arxiv.org/abs/2310.01405.

A MATHEMATICAL PROPERTIES

We analyze the information supplied by policy labels, exposure to selected starting states, and the local objectives used on collected trajectories. The vocabulary is finite. Unless stated otherwise, student probabilities are a softmax of finite logits, teacher probabilities are fixed and positive, and logarithms are natural. Visible states, active labels, token sets, and supervision masks are held fixed when differentiating. The statements below concern these conditional learning problems; empirical comparisons are specified in Appendix E.

A.1 POLICY INFORMATION AND IDENTIFIABILITY ON THE TRAINING SUPPORT

Let $\mathcal{H}$ be a class of candidate policy rules. A candidate $g \in \mathcal{H}$ assigns a policy judgment $z_g(s)$ and a set of permitted responses to each state. For a finite training support S, let $\mathcal{H}_{\mathrm{resp}}(S)$ contain candidates compatible with the observed response supervision. Canonicalizing equivalent labels under the fixed ontology, additional correct policy annotations restrict this set to

$$\mathcal{H}_{\mathrm{joint}}(S) = \{g \in \mathcal{H}_{\mathrm{resp}}(S) : z_g(s) = z^T(s) \text{ for all } s \in S\} \subseteq \mathcal{H}_{\mathrm{resp}}(S). \tag{9}$$

The inclusion follows directly from imposing additional constraints. It is strict whenever a response-compatible candidate disagrees with a correct label on at least one annotated state. Thus, labels can distinguish candidate rules that permit the same observed refusal. Boundary pairs add states on which permitted responses themselves differ, constraining both judgments and behavior.

This argument describes supervision content. Its informativeness depends on annotation correctness and on the states included in S. In a neural model, accurate label prediction can coexist with a response branch that ignores the label. Conditioning responses on predicted labels supplies a route for their use, following the general idea of supervised intermediate concepts (Koh et al., 2020). The student still reads the request and visible prefix, so labels are supervised decision variables rather than an exclusive information bottleneck. Appendix E.4 tests their functional influence, and Appendix E.3 separately studies hidden-state information.

A.2 EXPOSURE UNDER THE STARTING-STATE MIXTURE

At any training iteration, let K be a set of selected critical starting states. The mixture in Equation (3) satisfies

$$\Pr_{s_0 \sim \mu}(s_0 \in K) = \sum_{j \in \{\mathrm{nat},\mathrm{bnd},\mathrm{pre}\}} \alpha_j \mu_j(K) \geq \alpha_{\mathrm{pre}} \mu_{\mathrm{pre}}(K). \tag{10}$$

Every mixture contribution is nonnegative. Consequently, retaining positive prefix-mixture weight and positive replay-pool mass on K preserves starting-state exposure independently of the probability that the student reaches K from an empty response. The statement applies to the pool available at that iteration; a finite-capacity pool need not retain every previously collected state.

The bound concerns starting states. Occupancy later in a rollout depends on conditional student generation, termination, and any accepted teacher segment. Token-normalized occupancy also depends on trajectory lengths. Accordingly, the evaluation separates externally supplied starting-state exposure from visits during natural student rollouts (Appendix E.4). This use of expert supervision on learner-visited states is related to DAgger (Ross et al., 2011); Equation (10) is the coverage statement used here, without importing its online-learning guarantees.

A.3 KL GRADIENTS, TAIL AGGREGATION, AND OVERLAP BOUNDS

For $p_v = \mathrm{softmax}(u)_v$, the identity $\partial \log p_a / \partial u_v = \mathbf{1}[a = v] - p_v$ gives

$$\begin{aligned} \frac{\partial D_{\mathrm{KL}}(q\|p)}{\partial u_v} &= p_v - q_v, \\ \frac{\partial D_{\mathrm{KL}}(p\|q)}{\partial u_v} &= p_v \left(\log \frac{p_v}{q_v} - D_{\mathrm{KL}}(p\|q) \right). \end{aligned} \tag{11}$$

For FKL, differentiating $-\sum_a q_a \log p_a$ yields the first line. For RKL, differentiating $\sum_a p_a \log(p_a/q_a)$ yields $p_v(\log(p_v/q_v) + 1) - p_v \sum_a p_a(\log(p_a/q_a) + 1)$, which simplifies to the

second line. With fixed positive q, write $q_{\min} = \min_v q_v > 0$. Since $0 \le D_{\mathrm{KL}}(p\|q) \le \log(1/q_{\min})$ and $p_v \log p_v \to 0$, the RKL coordinate gradient tends to zero as $p_v \to 0$. The FKL gradient instead tends to $-q_v$. This is a local logit-gradient comparison, with no corresponding claim that the RKL value tends to zero.

Near a fixed positive teacher distribution, write $p = q + \delta$ with $\sum_v \delta_v = 0$ and small coordinatewise relative deviations. Taylor expansion gives

$$\begin{aligned} D_{\mathrm{KL}}(q\|q+\delta) &= \frac{1}{2}\sum_v \frac{\delta_v^2}{q_v} + O_q(\|\delta\|_2^3), \\ D_{\mathrm{KL}}(q+\delta\|q) &= \frac{1}{2}\sum_v \frac{\delta_v^2}{q_v} + O_q(\|\delta\|_2^3). \end{aligned} \tag{12}$$

The remainder constants depend on q and the neighborhood. High total overlap alone does not ensure small relative deviation on every coordinate. Thus the overlap threshold is an implementable selection rule whose value is evaluated against fixed divergences, as motivated by the flexibility of GKD (Agarwal et al., 2024).

For a fixed set U and its complement R, aggregation gives

$$D_{\mathrm{KL}}(\bar{q}\|\bar{p}) = \sum_{v\in U} q_v \log \frac{q_v}{p_v} + q_R \log \frac{q_R}{p_R}. \tag{13}$$

Expanding each tail probability as $q_v = q_R q(v \mid R)$ and $p_v = p_R p(v \mid R)$ yields the exact decomposition

$$D_{\mathrm{KL}}(q\|p) = D_{\mathrm{KL}}(\bar{q}\|\bar{p}) + q_R D_{\mathrm{KL}}\big(q(\cdot \mid R)\|p(\cdot \mid R)\big). \tag{14}$$

The aggregated KL is a lower bound on the full KL, rather than an unbiased estimate. Its gap is the omitted conditional tail divergence multiplied by teacher tail mass. Small tail mass alone supplies no uniform error bound when the conditional divergence is large. Empty tails are omitted, and zero-mass terms use their continuous limits.

Differentiating Equation (13) gives

$$\frac{\partial D_{\mathrm{KL}}(\bar{q}\|\bar{p})}{\partial u_v} = \begin{cases} p_v - q_v, & v \in U, \\ p_v(1 - q_R/p_R), & v \in R. \end{cases} \tag{15}$$

For $v \in U$, the selected-token contribution is $(1 - q_R)p_v - q_v$, and the tail contribution is $q_R p_v$. For $v \in R$, all tokens share the correction for total tail mass. The teacher–student union explicitly retains teacher-preferred tokens and tokens already assigned high student probability. Preserving full-vocabulary normalization is essential to these identities. Tail-aware OPD studies this aggregation principle for RKL (Huang & Wei, 2026); Equations (14)–(15) specify the FKL construction used here.

Finally, define $\rho = \sum_v \min(p_v, q_v) = 1 - \mathrm{TV}(p, q)$. The unobserved tail overlap obeys $0 \le \sum_{v\in R} \min(p_v, q_v) \le \min(p_R, q_R)$, giving

$$\rho_k^- \le \rho \le \rho_k^- + \min(p_R, q_R), \qquad \rho_k^- = \sum_{v\in U} \min(p_v, q_v). \tag{16}$$

Treating the entire tail as one shared event gives the upper bound: the two distributions can place their tail mass on different tokens. Using the lower bound for selection ensures that $\rho_k^- \ge \tau$ implies $\rho \ge \tau$ at rollout parameters. The other branch includes both genuinely low-overlap states and states whose overlap is unresolved at the chosen k. Appendix E.6 audits this conservatism and the full-KL approximation gap.

A.4 The sampled-token RKL gradient estimator

Let $p_i = \pi_{\theta_i}(\cdot \mid s, z)$ be the rollout distribution and let $v \sim p_i$. At a fixed visible state and active label,

$$\nabla_\theta D_{\mathrm{KL}}(p_\theta\|q)|_{\theta_i} = \mathbb{E}_{v\sim p_i}\left[(\log p_i(v) - \log q(v))\, \nabla_\theta \log p_\theta(v)|_{\theta_i}\right]. \tag{17}$$

Indeed, direct differentiation produces a factor $\log(p_i(v)/q(v)) + 1$. The constant term vanishes because $\mathbb{E}_{v\sim p_i} \nabla_\theta \log p_\theta(v)|_{\theta_i} = 0$. An implementable surrogate is therefore

$$\widehat{\ell}_R(s, v; \theta) = \mathrm{sg}[\log p_i(v) - \log q(v)] \log p_\theta(v), \tag{18}$$

where sg stops gradients. Its expected gradient, evaluated at θ_i, equals Equation (17). Its scalar value is not a KL estimate and can be negative. Differentiating a sampled log-ratio directly, with the sampled token fixed, omits the score-function factor required by Equation (17).

The response distribution used for sampling has temperature one and no top-k or top-p decoding truncation. The top-k set in Equation (4) compresses supervision only. Schema-constrained decoding of the separate label sequence is compatible with conditioning this estimator on the resulting label. We perform one accumulated-gradient update per fresh rollout batch. Reusing the same actions after a parameter update changes the sampling distribution and requires a separately specified estimator.

Branch selection and teacher reliability are determined from the context before sampling the next response token. Multiplication by such a fixed local mask preserves the corresponding masked conditional gradient. Filtering a token after observing whether that token is wrong changes its sampling law. Consequently, later segment and terminal audits affect future interventions and replay collection, while recorded response losses retain their original masks. The unbiasedness statement is conditional on a fixed state and label. Training treats collected states as fixed, following on-policy distillation (Agarwal et al., 2024); it does not differentiate their occupancy distribution or the discrete intervention process. Batch losses use the empirical normalization in Appendix D.4.

A.5 Teacher intervention and continuation contexts

For a particular teacher continuation b of length L and a fixed policy condition z^T, the probability that the student supplies exactly that continuation is

$$\Pr_\theta(b \mid s, z^T) = \prod_{j=1}^{L} \pi_\theta(b_j \mid s, z^T, b_{<j}). \tag{19}$$

If each required conditional probability is at most $\varepsilon < 1$, this probability is at most ε^L. An accepted teacher intervention directly supplies the visible context $(x, h \parallel b)$, where $\parallel$ denotes concatenation. The student can then practice its own continuation from that context. Local FKL changes action probabilities at a given state; inserting a teacher segment supplies a subsequent state immediately. This distinction motivates their factorial comparison in Appendix E.5.

The bound concerns one specified token path. A collection of semantically equivalent paths can have substantially larger probability. Teacher-prefix guidance and trajectory relaying provide related constructions (Xing et al., 2026; Xu et al., 2026); their benefits on other tasks do not establish the safety effects tested here. The cross-entropy in Equation (6) learns accepted segments, while resumed student positions retain conditional on-policy distribution supervision. Because generation can be length-limited, filtered, or deterministic, the accepted-segment distribution can differ from the original teacher distribution. Its cross-entropy is therefore the stated segment-imitation objective, without identifying it with the original teacher's full FKL. Annealing the intervention probability to zero makes final-phase label prediction and response generation fully student-controlled.

A.6 The scope of benign reference distillation

Write $r_s = \pi_0(\cdot \mid s)$ and $p_s = \pi_\theta(\cdot \mid s, z^S)$. Pinsker's inequality and Jensen's inequality give

$$\begin{aligned} \mathbb{E}_{s\sim\mathcal{S}_0} \mathrm{TV}(r_s, p_s) &\le \mathbb{E}_{s\sim\mathcal{S}_0} \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(r_s \| p_s)} \\ &\le \sqrt{\mathcal{L}_{\mathrm{util}}/2}. \end{aligned} \tag{20}$$

The expectation includes the fixed student-label draws used for that update, or their sampling distribution when labels are randomized. For any next-token statistic $f_s : \mathcal{V} \to [0, 1]$, the difference between its expectations under r_s and p_s is at most their TV distance. Thus full FKL controls the average discrepancy of bounded next-token statistics on the fixed benign reference states.

This gives a local interpretation of output-based functional preservation (Li & Hoiem, 2016). Multi-step generation can visit states outside the reference distribution, so capability retention is evaluated on independent tasks in Appendix E.2. If the capability objective were replaced by aggregated top-k FKL, the analogous bound would apply directly only to aggregated distributions. The primary method therefore retains full-vocabulary FKL for Equation (7).

B Observation Study: Setup and Results

B.1 Shared setup and measurement

The observation study fixes the teacher, student initialization, request-family splits, and safety policy across comparisons. The safety policy denotes normative rules; the model policy is a token distribution. Student suffixes are generated on-policy, including after externally supplied starting prefixes. Minimal pairs and their paraphrases stay within one split. The primary budget matches non-masked supervised response and auxiliary-target tokens; compute matching is a separate comparison. Generated tokens, teacher work, and elapsed time need not be equal under token matching.

OPD permits several divergences (Agarwal et al., 2024). The sampled/full-logit comparison fixes RKL (Gu et al., 2024); the practical recipe of Lu & Thinking Machines Lab (2025) is a technical blog. The fixed-state RKL derivative and sampled estimator are given in Equations (11) and (17). They can agree in expectation while differing in variance and cost; unsampled tokens can still change through softmax and shared parameters.

Safety error rate measures harmful responses on ordinary requests; Prefill-20 is attack success after a supplied 20-token partially compliant prefix. Token length does not fix semantic depth, and suffix recovery does not erase harmful content already in the prefix. For harmful-side refusal r_h and benign-side compliance c_b on the same minimal-pair split,

$$H(r_h, c_b) = \begin{cases} 2r_h c_b/(r_h + c_b), & r_h + c_b > 0, \\ 0, & r_h = c_b = 0. \end{cases} \tag{21}$$

Boundary scores report $100H$, not balanced accuracy or F1. Over-refusal is erroneous refusal on benign requests. XSTest motivates separating benign and harmful behavior (Röttger et al., 2024), without identifying the observation-study requests with that dataset. Policy-rule accuracy uses the fixed rule inventory; CPC separates relevant-factor switches from irrelevant-factor invariance. Retained utility is relative to initialization, with a per-task ratio $100u_j/u_{j,0}$ for positive higher-is-better baselines. Independent labels, validation-only checkpoint selection, and paired-request dependence govern evaluation; cohort variation is distinct from training randomness.

B.2 Behavior, position dependence, and held-out boundaries

Fixed diagnostic states separate local teacher matching from changing on-policy visitation. With $\Delta_t(u) = \sum_{s \in \mathcal{S}_t} w_s[D_0(s) - D_u(s)]$, the signed head share is

$$C_K(u) = \frac{\sum_{t \le K} \Delta_t(u)}{\sum_t \Delta_t(u)}, \qquad \sum_t \Delta_t(u) > 0. \tag{22}$$

It can exceed one if the tail worsens and is not a causal attribution. At the early checkpoint, $C_4 = 72\%$ and safety error is 10.4%; the final error is 7.6%, but boundary score and Prefill-20 are 63.4% and 50.8%. The position intervention applies loss to the first four tokens, later tokens, or the full response. Head-only reproduces $(47.2-11.6)/(47.2-7.6) = 89.9\%$ of the full safety improvement, whereas Tail-only better reduces prefix-conditioned error (Table 2). Shared parameters and different rollout lengths prevent treating the masks as separate semantic modules (Qi et al., 2025).

Despite 95.1% in-distribution student–teacher behavioral agreement, held-out language, topic, and policy-factor composition retain boundary gaps (Table 3). Minimal pairs vary one independently labeled relevant factor; surface controls preserve the decision under irrelevant edits. These tests measure generalization rather than establish that concentrated updates are intrinsically ineffective.

Table 2: **Observation-study position controls.** Percentages; Full OPD is the sampled-token baseline.

Condition	Safety error ↓	Prefill-20 ↓	Rule ↑	Boundary ↑	Over-refusal ↓
Initialization	47.2	71.3	51.2	52.0	5.1
Head-only	11.6	63.4	54.8	55.6	16.3
Tail-only	27.8	39.7	66.5	68.2	9.4
Full OPD	7.6	50.8	61.2	63.4	13.8
Teacher	2.3	12.7	94.2	92.0	5.9

Table 3: **Learning and held-out boundary diagnostics.** Checkpoint spacing is categorical, not a cumulative-token axis.

(a) Learning checkpoints: safety error (%)

Checkpoint	Safety error rate ↓
Initialization	47.2
Early	10.4
Final	7.6

(b) Held-out teacher–student boundary gap (pp)

Holdout	Gap (pp)
Language	18.6
Topic	24.1
Composition	31.7

B.3 Coverage, supervision readouts, and natural sampling

Teacher-forced top-k inclusion measures local support for summary tokens given the teacher prefix; semantic Hit@M measures whether any of M natural responses expresses the full summary. Sampled OPD has top-5 inclusion of 28.4% for boundary rationales and 92.1% for response tokens, but semantic Hit@32 is 9.6%. These are different denominators, and absent explanations do not establish absent rule knowledge. The supervision probes in Table 4 likewise characterize the chosen readout, not mutual information; reading an explicit rule field is a reference condition rather than hidden-information extraction.

Recovery-state visits fall from 4,180 to 310 per million generated tokens while Prefill-20 falls from 71.3% to 50.8%. This temporal association does not identify a causal mediator. Learner-induced state distributions motivate the coverage question (Ross et al., 2011); the probability of one token path is not the probability of a semantic class. At the $8\times$ sampling endpoint, safety error, boundary score, and Prefill-20 are 6.3%, 67.1%, and 46.9%, versus 7.6%, 63.4%, and 50.8% at baseline. This limited endpoint improvement is not a universal sampling-efficiency claim.

Table 4: **Coverage and supervision diagnostics.** Natural semantic Hit@32 is 9.6%; it is not the teacher-forced inclusion rate.

(a) Teacher-forced top-5 inclusion (%)

Condition	Boundary	Response
Sampled OPD	28.4	92.1

(b) Supervision-signal readouts (%)

Probe input	Response-mode accuracy ↑	Rule-ID accuracy ↑	Boundary-factor F1 ↑
Sampled log-ratio	92.3	53.7	50.8
Full teacher logits	95.0	64.1	61.9
Teacher structured output	96.2	81.6	80.4
Explicit rationale	97.1	94.5	92.8

Table 4 (continued)
(c) Natural-sampling endpoints (%)

Budget	Safety error rate ↓	Boundary score ↑	Prefill-20 ↓
1× baseline	7.6	63.4	50.8
8×	6.3	67.1	46.9

B.4 Supervision and state interventions

The interface comparison shares initialization, request splits, and effective supervised-token budgets. Decision labels specify answer/refuse; rationales add rule and boundary content; complete teacher traces also change text length and trajectory source. The latter are demonstrations, not access to private reasoning. Table 5 consolidates safety and policy metrics. Relative to sampled OPD, rationales improve boundary score by 22.0 pp and CPC by 25.3 pp while reducing Prefill-20 by 26.7 pp.

Crossing rationale supervision G with targeted starting states I isolates their conditional gains (Table 6). Rationales add 22.0 pp without state intervention and 15.3 pp with it; state intervention adds 11.6 and 4.9 pp, respectively. The interaction is -6.7 pp: the combined cell is best here, but this is neither superadditivity nor an estimated interaction with uncertainty. Extreme-stratum contrasts use pre-intervention coverage, visitation, and independent rule ambiguity; baseline difficulty and headroom constrain their interpretation.

Table 5: **Observation-study supervision interfaces.** All values are percentages; utility is relative to initialization. Teacher reference: safety error 2.3, Prefill-20 12.7, boundary 92.0, over-refusal 5.9, and rule accuracy 94.2. Head/Tail boundary scores are in Table 2; the state-only score is in Table 6.

Condition	Error ↓	Prefill ↓	Boundary ↑	Over-ref. ↓	Rule ↑	CPC ↑	Utility ↑
Initialization	47.2	71.3	52.0	5.1	51.2	47.6	100.0
Sampled OPD	7.6	50.8	63.4	13.8	61.2	57.8	99.3
Full-logit OPD	5.9	44.6	68.2	12.1	66.7	62.9	99.0
Decision label	6.4	36.9	75.1	10.5	73.8	69.4	98.8
Policy rationale	4.8	24.1	85.4	7.2	87.6	83.1	98.7
Teacher trace	5.1	26.8	82.1	8.1	84.2	79.3	98.3
Rationale + state	3.1	13.9	90.3	5.8	92.1	89.1	98.5

Table 6: **Interventions and heterogeneity.** Factorial boundary interaction: -6.7 pp. Strata are defined before intervention.

(a) Rationale G / targeted state I factorial (%)

Rationale G	State I	Boundary score ↑	Prefill-20 ↓
Off	Off	63.4	50.8
On	Off	85.4	24.1
Off	On	75.0	31.9
On	On	90.3	13.9

(b) Extreme-stratum effects (pp)

Pre-training variable	Stratum	Outcome improvement	Gain
Rationale coverage	Low	Boundary-score increase	25.4
Rationale coverage	High	Boundary-score increase	6.1
Recovery-state visitation	Low	Prefill-20 decrease	31.8
Recovery-state visitation	High	Prefill-20 decrease	8.7
Rule ambiguity	High	Counterfactual increase	27.2
Rule ambiguity	Low	Counterfactual increase	5.9

B.5 Interpretation and relation to prior work

The diagnosis concerns the evaluated supervision and coverage conditions, not an inevitable failure of OPD. Shallow alignment and prefix recovery precede this work (Qi et al., 2025). OPSA's main

implementation combines forward and reverse KL (Fu et al., 2026); activation-steered localized RKL also has positive safety results (Li et al., 2026a). Constitutional degradation is specific to the teacher-conditioning and multimodal setting studied by Wen et al. (2026), while same-weight teachers with different prohibition visibility (Li et al., 2026b) and rationale supervision (Hsieh et al., 2023) provide related precedents. Independent teacher audits, matched controls, and generalization tests determine how far our explanation extends. Format effects, generic-text gains, unmatched budgets, or difficulty-adjusted null effects would narrow the attribution. Appendix C supplies a conditional construction, not a proof that real models satisfy its assumptions.

C Conditional Analysis and a Constructed Example

C.1 Identifiability on the Observed Support

To clarify what behavioral matching alone can identify, consider a rule variable $R \in \{r_0, r_1\}$ and a query protocol restricted to a state set $\mathcal{S}$. Let O include every observable teacher signal: sampled outputs, any supplied logits, decision labels, and auxiliary annotations. Assume that, for every $s \in \mathcal{S}$, both the complete teacher conditional distribution and the distribution of all observable signals are the same under the two rules. The request distribution, learner initialization, and protocol's random choices have the same law under both rules; subsequent queries may depend only on previous observations and remain inside $\mathcal{S}$. No extra rule-revealing side information is supplied.

Conditional observational equivalence. Under these assumptions, the full observed dataset $\mathcal{D}$ has the same law under r_0 and r_1. Consequently these data alone cannot distinguish the rules, even when they prescribe different responses at a state $s^* \notin \mathcal{S}$.

Argument. The initial query has the same law in both environments. At any subsequent step, an identical observation history induces the same query distribution, and the assumed equality of conditional observation laws induces the same next observation distribution. Induction gives equality of the joint history distributions. Any estimator applied to that history thus has the same output distribution under either rule. With equal prior probabilities, a rule classifier cannot have average success exceeding one half solely from these observations. Equivalently, wherever defined, the likelihood ratio is one and the posterior odds equal the prior odds. This is an implication of the specified observation model, not an empirical claim about actual teacher distributions.

A rule-revealing label or other valid external information can break this particular equivalence. Querying a disagreement state can also do so: exact full distributions that differ identify the candidate, whereas noisy sampled outputs require enough observations to discriminate their laws. Identical refusal openings are insufficient to establish the assumptions. If full logits or auxiliary signals already distinguish the rules on $\mathcal{S}$, or prior knowledge identifies the rule, the stated ambiguity does not apply.

In finite LLM training, an observed prefix set is not necessarily the population support: an unseen state can still have positive probability. The exact claim above must not be inferred from a finite-sample zero count. Nor does it imply that safe models must generate explicit explanations. The relevant question is whether available observations or prior information distinguish the policy-relevant cases; a revealing decision label may suffice without any rationale text. The empirical diagnostics and interventions test this issue operationally rather than prove that real models meet the equivalence assumptions.

C.2 Two Rules and a Disagreement State

Consider a symbolic two-stage environment to examine the observation model. At the common root s_0, the actions are `stop` and `enter`; the former terminates and the latter reaches s^*. Both candidate teachers use the same root distribution $(1-a, a)$, with $0 < a < 1$. At s^* the actions are `refuse` and `answer`. Rule r_0 assigns probabilities $(1-\epsilon, \epsilon)$, whereas r_1 assigns $(\epsilon, 1-\epsilon)$, with $0 < \epsilon < 1/2$. No other observable at the root reveals the rule. These are symbolic construction parameters, not estimated quantities or proposed language-model hyperparameters.

The restricted-observation condition queries only s_0, even if a generated trajectory enters s^*. Its root distributions and observed signals are identical, so the likelihood-ratio argument in Appendix C.1

applies. A policy-bit condition adds $b = 0$ for r_0 and $b = 1$ for r_1; this reveals the rule without an explanation. An active-query condition requests the full teacher distribution at s^*, which also distinguishes the rules because $\epsilon < 1/2$. A fourth condition supplies both interventions; the construction does not imply that the combination outperforms either revealing intervention alone.

For the natural-access variant, let the current student enter with probability $\rho \in (0, 1)$, and allow a teacher query whenever s^* is visited. For $N \geq 1$ independent episodes with fixed ρ, the probability of observing the state at least once and its expected visit count are

$$\Pr(V_N \geq 1) = 1 - (1 - \rho)^N, \qquad \mathbb{E}[V_N] = N\rho. \tag{23}$$

Both decrease as ρ decreases with N fixed, but non-visitation in one finite run does not imply $\rho = 0$. A predetermined sequence of entry probabilities ρ_n with independent episode randomness instead gives $\Pr(V_N = 0) = \prod_{n=1}^{N}(1-\rho_n)$. This constructed example separates a hard restriction on query support from low-probability natural access; it does not assume every refusal-training procedure reduces ρ.

The four query schemes separate the information available at the root from access to the disagreement state. This construction characterizes conditional ambiguity and its removal; the language-model interventions are evaluated separately.

D Implementation Details

D.1 Policy interface and reliable supervision

The policy document, rule ontology, and serialization schema are fixed within a run. Canonical rule IDs, equivalence classes, multi-label rules, and explicit unknown boundary fields determine the interface in Table 7. Labels are task judgments, not reconstructions of private teacher reasoning. Label mode reads the request and visible prefix to predict z^S; response mode additionally conditions on the active label. Existing tokenizer IDs and constrained label decoding are used. Teacher targets are absent from label-prediction inputs, and response gradients treat sampled label tokens as fixed while updating shared parameters.

A checkpoint occurs at the starting state and after at most B student response tokens. Canonical context $C(x, h, z)$ contains the full visible prefix and exactly one active label; replacing it can require cache reconstruction, which is charged. Teacher and student maintain separate caches. A validated z^T remains active through an accepted bridge and the immediately resumed student segment; the next checkpoint returns to z^S.

Two independent teacher annotations must parse, use valid IDs, agree on decision-relevant fields under ontology equivalence, and permit the response strategy. The preceding student segment is also audited. A failed checkpoint contributes no label loss, next-segment response-distillation loss, or intervention; student generation and independent benign-reference distillation continue. Later audits do not retroactively filter recorded response losses. Annotation agreement is a filter, not proof of accuracy; independent audits and evaluation judges have separate roles, and rejected work remains part of resource accounting.

Table 7: **Structured policy interface.** Fields use the fixed policy ontology.

Field	Content
`risk_category`	One or more categories in the policy taxonomy.
`rule_ids`	Canonical identifiers of the applicable clauses, including permitted multi-rule combinations.
`boundary_attributes`	Typed values for authorization, intent, target, actionability, and other policy-defined conditions, with explicit unknown values.
`response_type`	A permitted response strategy, such as `answer`, `limited_answer`, or `refuse`; alternatives follow the applicable policy.

D.2 STARTING STATES AND DATA SEPARATION

Natural starts use ordinary benign/harmful requests. Boundary starts use independently labeled minimal pairs that change one policy-relevant factor, with balanced benign and harmful sides and no extra pairwise loss. Prefix starts contain controlled partial responses or validated student-disagreement states. A fixed-capacity replay pool stratifies source and length, records semantic depth, and deduplicates request families and similar prefixes. New candidates enter only the next iteration; every replayed start receives a fresh current-policy suffix.

Supplied tokens are context-only; reliable student positions receive distribution supervision and accepted teacher segments receive cross-entropy. An end-of-trajectory audit may save a pre-EOS prefix but never appends a bridge after termination. Family, template, pair, and near-duplicate splits precede variant generation; evaluation prefixes are excluded from replay. Mixture weights are selected on validation data and fixed, while training-pool contents may evolve. Suffix and whole-history safety remain separate (Qi et al., 2025).

D.3 INTERVENTION AND WITHDRAWAL

A reliable checkpoint is eligible when the student's rule/boundary judgment or response type is incompatible with the policy, or its preceding segment violates the permitted strategy. Equivalent compliant phrasings are not errors; both harmful cooperation and erroneous benign refusal qualify. KL and overlap select local supervision, not the primary intervention trigger.

$$\eta_i = \eta_0 \max\{0, 1 - i/T_{\text{off}}\}, \qquad 0 \le \eta_0 \le 1, \quad 0 \le i < T, \quad 0 < T_{\text{off}} < T. \tag{24}$$

An eligible trajectory with remaining budget undergoes this Bernoulli trial. A successful trial consumes its single allowed attempt; a failed trial permits later eligibility. The teacher produces at most $L_{\max}$ tokens, reserving at least one subsequent student position under the trajectory cap. Generation and checking costs count even for rejected segments. A teacher EOS is excluded from the candidate segment; empty or noncompliant candidates are discarded without retry, and the student continues under z^S. Acceptance appends the bridge, rebuilds the student context under z^T, and records the segment-imitation loss. Resumed student probabilities are recomputed on that actual prefix.

Skipped or rejected intervention does not remove otherwise reliable label/KL supervision. After T_{off}, no teacher label or response segment is injected into generation, but the teacher still annotates and scores student states. Withdrawal therefore removes teacher-generated control, not teacher feedback or the starting-state mixture.

D.4 PROBABILITY COMPUTATION AND EXECUTION

Token-level distribution matching assumes the same tokenizer and response vocabulary for teacher and student. At each reliable student position, the current rollout parameters θ_i determine the student top-k set. Teacher scoring supplies its own top-k set and its probabilities on the entire union, including student-selected tokens outside the teacher top-k. Storing only teacher top-k probabilities is insufficient for this operation. At an RKL position, the teacher probability of the sampled student token is also recorded even if that token lies outside the union. The union and branch are determined from pre-action distributions and held fixed during the update.

Selected probabilities are obtained from full-vocabulary log-softmax values. Tail mass is computed using a stable log-complement or log-sum-exp over excluded logits, with at least FP32 probability reductions. The implementation uses $0 \log 0 = 0$ and omits an empty tail. The student's tail probability remains differentiable through its full softmax normalizer. Probability compression reduces storage and transfer; teacher forward passes and full-vocabulary normalization are still generally needed. Annotation, scoring, segment generation, validation, student generation, backpropagation, and checkpoint re-prefilling are measured separately.

The capability reference π_0 is frozen before any policy-label warm-up. Independent benign requests and validated responses generated by that reference define fixed visible states $\mathcal{S}_0$. The current student predicts its own labels on the corresponding fixed-checkpoint schedule. The reference reads the original request and visible prefix; the student additionally reads its active label. Full FKL is computed online or using sufficiently precise cached full reference probabilities. These reference

Algorithm 1 DEEPEN training

Require: Student π_θ, frozen safety teacher π_T, policy/schema, starting-state pools, benign requests, and validated run configuration

Freeze $\pi_0 \leftarrow \pi_\theta$ before warm-up; construct benign reference states $\mathcal{S}_0$
Warm up g_θ using reliable policy labels
for $i = 0, \ldots, T-1$ **do**
    Fix rollout parameters $\theta_i \leftarrow \theta$ and schedule η_i from Equation (24)
    Sample a batch of starting states from the current mixture μ
    Initialize loss records and a next-iteration replay-candidate collection
    for each sampled starting state s_0 **do**
        Collect a conditional rollout using Algorithm 2
        Audit termination; retain validated pre-EOS candidates for future replay
    end for
    Sample benign reference states; predict student labels using the fixed-checkpoint interface
    Record full-vocabulary utility FKL from Equation (7)
    Evaluate differentiable losses with contexts, labels, masks, unions, and branches fixed
    Normalize valid units; accumulate gradients of Equation (8)
    Perform exactly one optimizer update
    Update the bounded replay pool using validated candidates, available from the next iteration
end for
return Student model and its internal label/checkpoint interface

states provide functional distillation, separately from benign policy-boundary examples supervised by the safety teacher. Prefix and reference datasets are audited to preserve this distinction.

We normalize $\mathcal{L}_{\mathrm{dist}}$ over all reliable student response positions, with one shared denominator across the two KL branches. For $\mathcal{L}_{\mathrm{pol}}$, each target is normalized by its label length before averaging over reliable checkpoints. For $\mathcal{L}_{\mathrm{bridge}}$, each accepted teacher segment is normalized by its length before averaging over accepted segments. The utility loss is averaged over reference response positions. An empty collection contributes zero. Starting-source weights are determined by sampling, not by an additional unreported per-source loss. The batch objective uses Equation (18) at RKL positions and the explicit aggregated FKL at other reliable student positions.

A short policy-label warm-up precedes the joint training iterations. Each iteration fixes θ_i, collects fresh continuations, recomputes the required differentiable likelihoods under the same contexts, accumulates gradients, and makes one optimizer update. States, sampled labels, teacher outputs, token sets, branch choices, and masks are detached. Replayed starting prefixes do not imply replayed on-policy actions: every training suffix is newly sampled. The student response sampler uses temperature one without decoding truncation, as required by Appendix A.4. Validation determines mixture weights, loss weights, thresholds, checkpoint spacing, and intervention schedule before confirmatory evaluation.

Algorithm 1 separates the outer optimization loop from the conditional rollout procedure in Algorithm 2. Each record retains the original request, starting prefix, checkpoint labels, actual canonical context, token provenance, teacher reliability, and fixed branch metadata. RKL records additionally retain the rollout log-probability and teacher log-probability of the sampled token. In the algorithms, "record a loss" means storing its supervision and context for differentiable evaluation before the single update; it does not imply a parameter update during generation.

Algorithm 2 Conditional rollout with a single allowed intervention attempt

Require: Starting state (x, h), rollout student π_{θ_i}, teacher π_T, schedule η_i, and remaining budgets
Mark supplied prefix tokens as context-only; set $attempted \leftarrow$ false
while student has not emitted EOS and response capacity remains **do**
  Predict $z^S \sim g_{\theta_i}(\cdot \mid x, h)$ with schema-constrained decoding
  Obtain two teacher annotations and audit the preceding segment
  Set reliability m and, if reliable, the active teacher target z^T
  Set student-side active label $z \leftarrow z^S$
  if $m = 1$ **then**
    Record policy supervision at (x, h)
    Evaluate policy/path eligibility under the fixed semantic rubric
    if eligible, $\neg attempted$, budgets permit resumption, and Bernoulli(η_i) succeeds **then**
      Set $attempted \leftarrow$ true
      Generate one bounded teacher candidate b under z^T; account for its cost
      if b is nonempty and passes segment validation **then**
        Record teacher-segment cross-entropy at the current state
        Append b to h, mark its tokens as teacher-generated, and set $z \leftarrow z^T$
      end if
    end if
  end if
  Rebuild the student context with the current visible prefix and active label z
  for up to B student response positions, subject to EOS and the response cap **do**
    Compute $p_i = \pi_{\theta_i}(\cdot \mid x, h, z)$
    if $m = 1$ **then**
      Score $q = \pi_T(\cdot \mid x, h, z^T)$; fix U, ρ_k^-, and the branch before sampling
    end if
    Sample $v \sim p_i$ at temperature one without decoding truncation
    if $m = 1$ **then**
      Record Equation (18) if $\rho_k^- \geq \tau$, else aggregated FKL
    end if
    Append v to h and mark it as student-generated
    if v is EOS **then**
      Stop the rollout
    end if
  end for
  Retain validated disagreement candidates for the next iteration; preserve recorded masks
end while

Student-only inference alternates two operations: predict a structured label from the visible state, then generate at most B visible response tokens under that label. It terminates on the student's EOS or the response cap. The policy document supplied to the teacher, teacher annotations, reliability checks, KL selection, and intervention logic are absent at inference; the fixed schema and student label interface remain. Any variant that removes this interface is evaluated separately. Deployment decoding settings and checkpoint overhead must be reported alongside safety and utility outcomes.

E Extended Experiments

This section groups each control with its results. Observation-study, response-token-matched, and compute-matched evaluations remain separate protocols. Unless stated otherwise, entries are percentages and intervals describe repeated request cohorts, not independent training-seed uncertainty.

E.1 Measurement, data, and baseline identity

Measurement. Methods are paired within each protocol, metric, and request repetition. The boundary score (BDA) is $100H(r_h, c_b)$, computed within a cohort before averaging; its boundary population is separate from XSTest. CPC equally weights relevant-factor switching and irrelevant-factor invariance. Useful benign answers are a subset of answered requests with the same denominator. Recovery partitions a common prefix cohort into safe/useful, harmful, safe/unhelpful, and invalid outcomes. Limited-answer health compliance and native capability metrics retain their own rubrics. In Table 1, HB/JB/MB denote HarmBench/JailbreakBench/MultiBreak ASR, Indic denotes

the equal-language IndicSafeEval ASR, and MMLU/Math/Code/IF denote MMLU-Pro, MATH500, LiveCodeBench release_v6 pass@1, and IFEval strict, respectively.

Controls. Family splits, decoding, query budgets, and validation-only operating-point selection are fixed before testing. Repeats and dialogue turns remain clustered by request family; native thinking counts toward token budgets. Code evaluation uses a constrained execution sandbox. A frozen same-checkpoint teacher receives privileged policy context; larger-teacher transfer additionally requires exact token-ID compatibility. BF16 includes the documented Ministral FP8 conversion. The model registry follows official Qwen3 technical-report and checkpoint documentation, Meta Llama model cards, and Gemma/Ministral model documentation (Yang et al., 2025; Qwen Team, 2026a;b;c; Meta AI, 2024a;b; Google DeepMind, 2026; Mistral AI, 2026). Effective compute size and total memory-resident parameters are treated separately.

Table 8: **Model registry.** Total / effective-compute billions. Four main representatives are Qwen3.5-4B, Gemma 4 E4B-it, Llama 3.1 8B, and Ministral 3 8B. Gemma E4B includes embeddings in its 8B total; Gemma 12B has 11.95B total. Ministral totals 8.4B language plus 0.4B vision, with text-only execution.

Configuration	Official checkpoint	Total / compute
Qwen3-4B	`Qwen/Qwen3-4B`	4 / 4
Qwen3-8B	`Qwen/Qwen3-8B`	8 / 8
Qwen3.5-4B	`Qwen/Qwen3.5-4B`	4 / 4
Qwen3.6-27B	`Qwen/Qwen3.6-27B`	27 / 27
Qwen3.6-35B-A3B	`Qwen/Qwen3.6-35B-A3B`	35 / 3
Gemma 4 E4B-it	`google/gemma-4-E4B-it`	8 / 4.5
Gemma 4 12B-it	`google/gemma-4-12B-it`	11.95 / 11.95
Llama 3.2 3B	`meta-llama/Llama-3.2-3B-Instruct`	3 / 3
Llama 3.1 8B	`meta-llama/Llama-3.1-8B-Instruct`	8 / 8
Ministral 3 8B	`mistralai/Ministral-3-8B-Instruct-2512`	8.8 / 8.4

Table 9: **Evaluation cohorts and protocols.** n is the per-cohort denominator (per language for IndicSafeEval). Source identifiers refer to papers, not dataset snapshots.

Benchmark	n	Evaluation scope / source
HarmBench (Mazeika et al., 2024)	200	Direct text behaviors; no adaptive search. *Source: 2402.04249v2.*
JailbreakBench (Chao et al., 2024)	100	Held-out single-turn attacks; at most 20 queries/behavior. *Source: 2404.01318.*
MultiBreak (Song et al., 2026)	500	Dialogue-family clusters; at most 8 turns; full-dialogue judge. *Source: 2605.01687v1.*
SALAD-Bench (Li et al., 2024)	1000	Aggregate ASR. *Source: 2402.05044.*
StrongREJECT (Souly et al., 2024)	500	Native continuous 0–1 rubric, not ASR. *Source: 2402.10260v2.*
SafeDialBench (Cao et al., 2025)	500	Turn-aware family clusters; at most 8 turns. *Source: 2502.11090.*
IndicSafeEval (Mondal et al., 2026)	500	500 units/language; equal Hindi, Bengali, Marathi, Punjabi macro. *Source: 2609.03781v2.*
XSTest benign (Röttger et al., 2024)	250	All 250 requests in denominator; refusals/empty outputs fail. *Source: 2308.01263.*
OR-Bench (Cui et al., 2024)	1000	Stratified benign boundary sample. *Source: 2405.20947.*
MS-XSB (Yuan et al., 2025)	500	Scenario holdout independent of XSTest. *Source: 2510.08158v3.*
Health-ORSC-Bench (Zhang et al., 2026)	1000	Limited safe answers allowed; separate from binary BDA. *Source: 2601.17642v2.*
MMLU-Pro (Wang et al., 2024)	1000	Fixed subject-stratified sample. *Source: 2406.01574.*
GPQA Diamond (Rein et al., 2023)	198	Closed-book; answer extraction before judging. *Source: 2311.12022v1.*
MATH500 (Lightman et al., 2023)	500	Normalized exact answer; no tools. *Source: 2305.20050.*
Contest-format proxy	30	30-item contest-format proxy, not official AIME scoring.

Table 9 (continued)

Benchmark	n	Evaluation scope / source
LiveCodeBench release_v6 (Jain et al., 2024)	1055	release_v6, May 2023–April 2025; greedy single sandbox execution. *Source: 2403.07974v2.*
IFEval (Zhou et al., 2023)	541	Original strict instruction checker. *Source: 2311.07911v1.*
Global-MMLU (Singh et al., 2024)	500	Equal macro over five controlled language strata. *Source: 2412.03304.*
Marco-Bench-MIF (Zeng et al., 2025)	500	Equal macro over five controlled language strata. *Source: 2507.11882v1.*
LongBench v2 (Bai et al., 2024)	500	Common 8K-token window; fixed truncation side; not native full length. *Source: 2412.15204v2.*

Baseline identity. PA* retains PolicyAlign's EMA teacher and student-top-k RKL, but uses the common data/budget and an added full-FKL benign anchor (Wu et al., 2026). Relay* adapts the reflection-token trigger, paragraph-delimited legs, and clipped relay objective to safety, with the same added anchor; it is not pure teacher-segment CE (Xu et al., 2026). A frozen-teacher PA* is a separate control. Compute-matched OPSA* retains privileged-context selection and symmetric FKL/RKL rather than JSD (Fu et al., 2026). SFT arms use fixed compliant traces; OPD arms generate fresh suffixes. These controlled adaptations are not strict published-method reproductions; related protocols include trust-region, tail-aware, and adaptive-continuation distillation (Xing et al., 2026; Huang & Wei, 2026; Zhou et al., 2026).

Table 10: **Published PolicyAlign reference, original protocol only.** arXiv:2606.25442v1, Tables 1–2 (pp. 6,16). SR/WJ are judge-ASR, XS is XSTest full compliance; original precision is retained. These values are not comparable to the DEEPEN evaluation rows.

Original model / method	SR	WJ	XS	MMLU-Pro	GPQA	MATH500
LLaMA-3.2-3B / Base	4.15	38.60	92.0	31.07	23.74	19.4
LLaMA-3.2-3B / PolicyAlign	0.32	11.18	90.0	31.43	24.24	18.8
Qwen3-4B / Base	7.35	69.46	94.8	66.29	50.00	72.8
Qwen3-4B / PolicyAlign	0.32	8.51	87.6	68.29	47.98	74.8

Original settings: EMA .99; student top-256 RKL; 50 alignment steps; LR $5e{-6}$; batch 128; training cap 512; evaluation temperature 0/cap 1024. Qwen3Guard-Gen-8B judges harm; GPT-4o judges XS compliance (Wu et al., 2026).

Table 11: **Canonical compute-matched controls and resource multipliers.** GPU-hour matching is separate from response-token matching; student/teacher token budgets are not jointly equalized.

Method	ASR	BDA	Recovery	Useful	GPU h	Student	Teacher
Safety SFT	7.5	70.9	43.3	71.1	1.00	1.70	0.35
Teacher trace SFT	8.2	74.9	49.9	73.7	1.00	1.45	0.55
PA* (EMA teacher)	9.0	79.9	59.3	78.6	1.00	1.20	1.20
PA* (frozen teacher)	9.5	80.9	60.5	79.2	1.00	1.20	1.12
OPSA*	10.0	83.4	64.4	83.6	1.00	1.10	1.16
Fixed JSD	10.5	83.9	64.9	83.6	1.00	1.12	1.05
DEEPEN	11.0	86.8	71.9	85.0	1.00	1.00	1.40

E.2 Full performance and cohort precision

Tables 12–16 expand the same main-protocol checkpoints across ten configurations. All rates are percentages except native StrongREJECT (0–1). Contest-format scores are a proxy, and LongBench v2 uses a common 8K window. The four recovery categories share an exhaustive evaluation population. BDA is the harmonic boundary score, and useful answering shares its denominator with benign answering. The tables place model pairs side by side without pooling their results. Column abbreviations D and Rel. denote DEEPEN and Relay*, respectively.

Table 12: **Full performance: Qwen3-4B and Qwen3-8B.** Common main protocol; native units and metric definitions are given at the start of this subsection.

	Qwen3-4B						Qwen3-8B					
Metric	Init.	RKL	JSD	PA*	Rel.	D	Init.	RKL	JSD	PA*	Rel.	D
HarmBench ASR	21.7	6.5	7.0	5.0	6.5	6.2	17.2	3.2	3.9	2.2	3.2	3.1
JailbreakBench ASR	40.3	25.0	22.7	19.7	19.7	14.7	36.4	19.8	17.8	16.4	16.0	13.0
MultiBreak ASR	48.7	33.9	32.0	29.1	24.9	18.9	45.4	30.0	27.3	24.5	20.8	15.4
SALAD ASR	26.7	12.2	12.8	10.5	12.2	11.5	22.8	7.9	8.5	6.3	7.9	7.4
SafeDialBench ASR	45.9	32.3	29.7	26.3	22.5	17.8	42.4	27.1	24.2	22.1	18.4	12.6
StrongREJECT (0–1)	0.279	0.168	0.153	0.139	0.135	0.106	0.261	0.148	0.132	0.117	0.113	0.083
IndicSafeEval ASR	49.1	34.0	32.0	30.1	29.6	25.7	45.8	31.4	29.2	27.1	26.7	22.4
Boundary harmful	76.8	85.6	84.2	87.4	86.1	86.6	81.0	90.5	89.5	92.0	90.7	91.2
Boundary benign	68.0	64.8	70.9	69.9	71.3	83.1	74.4	71.6	76.7	75.8	77.0	87.7
BDA	72.1	73.7	77.0	77.7	78.0	84.8	77.6	79.9	82.6	83.1	83.3	89.4
Rule identification	55.4	63.4	70.0	73.8	72.8	89.4	56.4	65.2	72.5	76.4	75.6	91.8
Relevant-factor switch	65.6	66.8	69.6	70.1	70.4	76.2	70.7	73.3	75.8	76.1	76.3	82.0
Irrelevant invariance	73.8	75.4	79.3	79.8	79.8	86.1	80.9	83.8	85.7	86.2	86.3	92.8
CPC	69.7	71.1	74.5	75.0	75.1	81.2	75.8	78.5	80.7	81.2	81.3	87.4
Useful recovery	32.2	45.2	52.4	53.9	60.4	71.3	38.1	50.0	56.1	58.2	64.5	74.7
Residual prefix harm	48.1	28.4	24.0	23.3	20.4	14.9	41.4	24.2	20.7	20.1	16.8	11.8
Safe but unhelpful	17.9	24.6	21.8	20.9	17.3	12.0	18.3	23.6	20.9	19.4	16.4	11.3
Invalid recovery	1.8	1.8	1.8	1.8	1.8	1.8	2.3	2.3	2.3	2.3	2.3	2.3
XSTest answered	91.9	83.3	86.3	84.5	85.3	91.2	92.3	85.9	88.5	87.3	88.1	91.9
XSTest useful	81.1	71.9	75.1	72.8	73.7	83.5	80.3	71.6	75.6	73.5	74.7	81.2
OR-Bench answered	82.4	74.7	77.6	76.0	76.7	84.6	83.9	76.4	79.0	77.4	78.1	86.1
MS-XSB answered	83.0	75.5	78.2	76.5	77.3	82.7	84.3	77.9	79.9	78.8	79.4	83.7
Health compliance	65.5	62.3	68.3	67.1	68.8	79.2	69.6	66.2	72.0	70.9	72.4	83.3
MMLU-Pro	60.0	58.0	59.4	59.1	58.9	59.3	64.7	63.0	64.0	63.7	63.6	64.0
GPQA Diamond	35.4	33.7	34.5	34.5	34.5	34.5	44.1	42.0	43.4	42.9	42.7	43.4
MATH500	70.9	68.7	70.3	69.9	69.8	70.1	74.6	72.8	73.8	73.6	73.5	73.8
Contest-format proxy	31.1	30.0	31.1	31.1	31.1	31.1	37.3	36.7	36.7	36.7	36.7	36.7
LCB v6 pass@1	45.4	43.1	44.5	44.2	43.9	44.5	52.3	50.5	51.5	51.2	51.0	51.4
IFEval strict	78.4	77.6	78.1	78.0	77.9	78.0	82.9	82.3	82.8	82.6	82.5	82.7
Global-MMLU	51.9	49.8	51.1	50.8	50.5	51.0	56.6	54.5	55.8	55.4	55.2	55.6
Marco-Bench-MIF	69.2	68.0	68.9	68.7	68.5	68.8	74.4	73.2	74.0	73.8	73.6	73.9
LongBench v2 (8K)	44.6	43.0	44.3	44.0	43.7	44.1	51.8	49.1	50.9	50.2	50.0	50.8
Held-out boundary BDA	64.1	65.9	68.9	69.7	70.0	76.6	66.8	69.0	72.1	72.6	73.0	79.5

Table 13: **Full performance: Qwen3.5-4B and Qwen3.6-27B.** Common main protocol; native units and metric definitions are given at the start of this subsection.

	Qwen3.5-4B						Qwen3.6-27B					
Metric	Init.	RKL	JSD	PA*	Rel.	D	Init.	RKL	JSD	PA*	Rel.	D
HarmBench ASR	16.8	1.5	1.8	0.2	1.5	1.0	9.3	1.0	1.0	1.0	1.0	1.0
JailbreakBench ASR	26.0	14.0	11.7	8.7	8.0	6.0	23.7	10.3	7.7	5.3	4.7	1.7
MultiBreak ASR	43.7	28.9	25.5	23.3	19.0	12.5	33.5	18.4	15.9	13.9	9.8	5.3
SALAD ASR	20.4	5.8	6.4	4.0	5.8	5.2	13.7	4.4	4.4	4.4	4.4	4.4
SafeDialBench ASR	40.1	26.5	23.6	20.7	16.6	10.6	32.0	18.7	15.3	12.3	8.7	2.7
StrongREJECT (0–1)	0.221	0.108	0.093	0.078	0.075	0.045	0.185	0.070	0.055	0.039	0.035	0.010
IndicSafeEval ASR	42.3	27.5	25.6	23.6	23.1	19.2	35.2	21.2	19.1	16.9	16.5	13.1
Boundary harmful	81.7	90.7	89.6	92.8	91.3	91.7	85.9	94.7	94.1	96.9	95.1	95.9
Boundary benign	76.4	73.3	79.2	78.1	79.7	89.6	80.1	77.0	83.3	82.2	84.1	95.4
BDA	79.0	81.0	84.1	84.8	85.1	90.7	82.9	84.9	88.4	88.9	89.2	95.6
Rule identification	59.8	67.9	74.9	77.6	76.8	91.7	60.9	68.4	76.2	80.1	78.9	96.1
Relevant-factor switch	71.5	73.3	76.8	77.7	77.8	83.0	74.5	76.5	80.1	80.6	80.8	87.2
Irrelevant invariance	82.0	84.3	87.3	88.4	88.4	93.6	86.1	88.2	91.8	92.2	92.3	98.2
CPC	76.8	78.8	82.1	83.0	83.1	88.3	80.3	82.3	86.0	86.4	86.6	92.7
Useful recovery	40.3	51.3	58.1	60.4	66.9	77.3	47.8	58.9	65.2	67.4	73.3	83.4
Residual prefix harm	39.8	24.7	20.4	18.9	15.4	11.5	35.3	20.8	17.1	16.2	13.2	7.7
Safe but unhelpful	18.3	22.4	19.9	19.1	16.1	9.7	14.8	18.1	15.6	14.3	11.4	6.8
Invalid recovery	1.6	1.6	1.6	1.6	1.6	1.6	2.1	2.1	2.1	2.1	2.1	2.1
XSTest answered	94.3	86.9	89.3	88.0	88.8	93.7	94.3	88.7	90.0	89.2	90.0	94.0

Table 13 (continued)

	Qwen3.5-4B						Qwen3.6-27B					
Metric	Init.	RKL	JSD	PA*	Rel.	D	Init.	RKL	JSD	PA*	Rel.	D
XSTest useful	84.8	76.1	79.2	77.7	78.5	86.0	84.5	77.1	79.6	77.6	78.7	86.1
OR-Bench answered	83.0	77.3	79.4	78.4	79.1	85.4	85.3	78.0	80.8	79.1	79.8	87.7
MS-XSB answered	84.3	78.3	80.5	79.3	79.9	84.0	85.9	78.9	81.4	80.0	80.8	85.1
Health compliance	70.1	67.1	72.5	71.5	73.0	84.8	77.5	73.8	79.7	78.6	80.0	91.4
MMLU-Pro	70.2	68.5	69.4	69.1	68.9	68.9	77.0	75.1	76.1	75.9	75.7	75.9
GPQA Diamond	49.3	45.8	47.3	46.6	46.1	46.0	65.2	62.8	64.0	63.6	63.3	63.6
MATH500	75.7	72.9	74.5	74.3	73.8	73.8	85.1	82.7	84.6	84.3	83.9	84.3
Contest-format proxy	40.0	37.8	38.9	38.9	37.8	37.8	50.0	47.8	48.9	48.9	47.8	48.9
LCB v6 pass@1	53.7	51.4	53.0	52.6	52.2	52.1	71.9	69.9	71.2	70.7	70.3	70.7
IFEval strict	83.7	83.0	83.2	83.2	83.2	83.2	90.6	89.5	90.1	90.0	89.8	90.0
Global-MMLU	58.8	56.5	57.9	57.5	57.3	57.2	70.4	68.5	69.7	69.5	69.2	69.5
Marco-Bench-MIF	75.3	74.3	74.9	74.7	74.7	74.7	81.7	80.5	81.3	81.0	80.9	81.0
LongBench v2 (8K)	53.5	51.8	52.9	52.5	52.3	52.3	64.4	62.1	63.8	63.5	63.2	63.5
Held-out boundary BDA	71.0	72.2	75.9	76.2	76.7	82.5	75.2	76.7	80.3	81.0	81.2	87.6

Table 14: **Full performance: Gemma 4 E4B-it and Gemma 4 12B-it.** Common main protocol; native units and metric definitions are given at the start of this subsection.

	Gemma 4 E4B-it						Gemma 4 12B-it					
Metric	Init.	RKL	JSD	PA*	Rel.	D	Init.	RKL	JSD	PA*	Rel.	D
HarmBench ASR	15.0	1.0	1.0	1.0	1.0	1.0	10.3	1.2	1.2	1.2	1.2	1.2
JailbreakBench ASR	27.3	12.0	9.3	6.3	6.3	2.7	28.7	12.0	10.3	7.7	6.7	1.0
MultiBreak ASR	38.1	23.0	20.1	18.3	14.8	9.5	35.0	19.3	16.1	13.3	9.7	4.4
SALAD ASR	17.6	4.1	4.6	4.1	4.1	4.1	12.8	4.7	4.7	4.7	4.7	4.7
SafeDialBench ASR	37.5	21.8	19.3	17.2	13.4	8.2	31.8	17.1	13.9	11.1	7.6	1.5
StrongREJECT (0–1)	0.210	0.097	0.082	0.067	0.064	0.034	0.182	0.068	0.053	0.038	0.034	0.010
IndicSafeEval ASR	39.7	24.0	21.9	20.0	19.7	15.9	35.7	20.9	19.0	16.9	16.5	12.5
Boundary harmful	82.6	92.6	91.4	94.6	93.0	93.3	86.4	95.9	95.3	97.7	96.3	96.8
Boundary benign	75.1	72.4	77.5	76.6	78.1	89.5	80.6	77.6	83.6	82.7	84.2	96.0
BDA	78.7	81.3	83.9	84.6	84.9	91.4	83.4	85.7	89.1	89.5	89.8	96.4
Rule identification	58.3	67.2	74.6	78.6	77.3	93.2	62.4	69.8	76.8	80.5	79.6	94.9
Relevant-factor switch	73.1	75.9	78.2	79.1	79.4	85.0	75.9	77.5	81.1	81.7	81.9	87.8
Irrelevant invariance	80.3	83.1	85.8	86.4	86.8	93.4	85.7	88.4	91.8	92.3	92.8	99.2
CPC	76.7	79.5	82.0	82.8	83.1	89.2	80.8	83.0	86.4	87.0	87.3	93.5
Useful recovery	40.2	53.2	58.3	60.3	68.0	79.1	47.1	59.1	66.4	68.6	75.1	85.8
Residual prefix harm	41.6	23.9	21.5	20.8	15.8	10.3	36.9	21.2	17.0	16.1	11.8	6.2
Safe but unhelpful	16.3	20.9	18.2	16.9	14.3	8.6	13.6	17.4	14.2	13.0	10.7	5.7
Invalid recovery	1.9	1.9	1.9	1.9	1.9	1.9	2.4	2.4	2.4	2.4	2.4	2.4
XSTest answered	93.5	87.9	89.7	89.2	89.6	93.1	95.9	88.8	90.9	90.0	90.5	95.1
XSTest useful	83.3	75.6	78.8	77.2	78.0	84.3	85.1	76.7	79.3	78.3	78.9	86.3
OR-Bench answered	84.9	77.7	80.4	78.8	79.9	86.9	87.3	79.6	82.3	80.7	81.5	89.3
MS-XSB answered	85.8	78.9	81.5	80.1	80.8	85.1	86.3	79.3	81.4	80.3	80.7	85.9
Health compliance	71.3	68.2	74.1	73.0	74.7	86.5	77.1	74.7	80.0	78.8	80.4	91.1
MMLU-Pro	70.2	68.2	69.4	69.0	68.9	67.3	75.7	73.6	75.2	74.5	74.2	74.9
GPQA Diamond	47.3	44.6	46.3	46.0	45.6	43.3	59.6	57.7	59.1	58.2	58.2	58.8
MATH500	74.4	72.5	73.5	73.1	72.9	71.1	82.3	80.3	81.8	81.1	80.9	81.2
Contest-format proxy	36.7	33.3	36.7	36.7	36.7	32.2	50.0	47.8	50.0	48.9	48.9	50.0
LCB v6 pass@1	51.8	49.8	51.2	51.0	50.9	48.7	66.2	64.5	65.8	65.0	64.9	65.2
IFEval strict	86.1	85.2	85.8	85.7	85.6	84.9	90.2	89.3	90.0	89.8	89.8	89.8
Global-MMLU	59.8	57.8	59.0	58.7	58.5	56.9	66.9	65.0	66.5	65.8	65.6	66.1
Marco-Bench-MIF	76.5	75.4	76.2	75.9	75.8	74.9	80.1	79.0	79.9	79.6	79.5	79.7
LongBench v2 (8K)	56.5	53.7	55.6	55.0	54.6	52.7	63.0	60.7	62.7	62.0	61.8	62.3
Held-out boundary BDA	70.9	72.3	75.4	76.2	76.7	83.0	76.8	78.3	81.4	81.9	82.1	88.2

Table 15: **Full performance: Llama 3.2 3B and Llama 3.1 8B.** Common main protocol; native units and metric definitions are given at the start of this subsection.

	Llama 3.2 3B						Llama 3.1 8B					
Metric	Init.	RKL	JSD	PA*	Rel.	D	Init.	RKL	JSD	PA*	Rel.	D
HarmBench ASR	28.7	13.3	13.8	10.8	13.3	12.5	21.2	7.7	8.0	5.8	7.7	7.2
JailbreakBench ASR	45.0	31.7	30.7	29.7	29.7	22.3	39.7	20.7	19.7	16.3	15.3	12.0
MultiBreak ASR	56.0	38.9	35.7	33.7	30.1	24.3	48.3	32.9	30.2	27.8	23.6	18.0
SALAD ASR	31.3	16.8	17.6	15.4	16.8	16.4	25.3	11.0	11.6	10.0	11.0	10.7
SafeDialBench ASR	51.1	37.5	34.3	31.5	28.0	22.3	45.5	31.3	27.9	24.9	21.5	16.3
StrongREJECT (0–1)	0.330	0.217	0.202	0.187	0.183	0.153	0.280	0.167	0.152	0.137	0.133	0.103
IndicSafeEval ASR	55.7	40.6	38.6	36.3	35.9	32.4	48.9	33.7	31.6	29.8	29.2	25.2
Boundary harmful	75.7	85.2	84.3	87.6	85.8	86.5	78.3	87.6	86.7	89.6	88.2	88.8
Boundary benign	67.0	64.2	70.1	69.1	70.6	78.0	70.1	67.4	73.4	72.1	74.0	85.5
BDA	71.1	73.2	76.5	77.2	77.4	82.0	74.0	76.2	79.5	79.9	80.5	87.1
Rule identification	46.7	55.8	63.1	67.1	66.2	81.3	52.4	58.9	65.2	69.6	68.4	86.2
Relevant-factor switch	62.6	64.8	67.6	68.1	68.2	73.4	66.2	68.0	70.9	71.4	71.9	79.2
Irrelevant invariance	72.8	75.6	78.8	80.0	80.2	85.4	76.4	79.6	82.6	83.0	83.9	90.8
CPC	67.7	70.2	73.2	74.0	74.2	79.4	71.3	73.8	76.8	77.2	77.9	85.0
Useful recovery	29.2	40.3	46.2	48.1	55.3	65.5	34.0	45.3	51.6	53.6	60.9	72.2
Residual prefix harm	48.1	30.4	27.4	26.7	22.2	17.3	44.6	28.3	24.8	23.4	19.4	12.8
Safe but unhelpful	20.4	27.1	24.2	22.9	20.2	14.9	19.0	24.1	21.2	20.6	17.3	12.6
Invalid recovery	2.3	2.3	2.3	2.3	2.3	2.3	2.4	2.4	2.4	2.4	2.4	2.4
XSTest answered	89.6	82.5	84.5	83.3	84.1	89.1	90.8	84.0	86.7	85.3	86.3	90.7
XSTest useful	77.9	69.6	72.3	70.7	71.5	79.6	80.4	70.8	74.9	72.5	73.7	81.2
OR-Bench answered	80.6	73.4	76.2	74.6	75.5	82.2	83.4	75.4	78.5	76.7	77.9	85.2
MS-XSB answered	82.1	74.9	78.2	75.8	76.9	81.4	82.7	74.7	77.4	75.5	76.7	82.3
Health compliance	63.1	59.9	65.8	64.8	66.6	74.9	68.0	65.2	70.6	69.5	70.8	82.3
MMLU-Pro	34.8	32.6	34.1	33.5	33.3	33.3	49.3	47.6	48.4	48.2	48.2	48.3
GPQA Diamond	23.9	20.7	22.9	22.6	22.2	22.2	33.7	31.6	32.7	32.3	32.2	32.5
MATH500	28.1	26.5	27.2	27.0	27.0	27.0	47.3	44.8	46.5	45.9	45.6	46.4
Contest-format proxy	4.4	2.2	4.4	3.3	3.3	3.3	5.6	4.4	4.4	4.4	4.4	4.4
LCB v6 pass@1	24.7	22.6	24.0	23.6	23.3	23.3	38.6	36.4	37.7	37.3	37.1	37.6
IFEval strict	69.4	69.1	69.3	69.3	69.3	69.3	76.6	76.0	76.4	76.3	76.3	76.4
Global-MMLU	26.5	24.1	25.3	25.0	24.8	24.8	40.9	38.8	40.0	39.7	39.5	39.9
Marco-Bench-MIF	61.5	60.3	61.1	60.9	60.8	60.8	68.2	67.3	67.9	67.6	67.5	67.8
LongBench v2 (8K)	22.9	20.8	22.4	21.8	21.7	21.7	33.8	31.7	32.9	32.6	32.5	32.7
Held-out boundary BDA	62.6	63.5	67.3	67.4	67.9	73.4	65.6	66.7	70.5	71.0	71.2	77.5

Table 16: **Full performance: Qwen3.6-35B-A3B and Ministral 3 8B.** Common main protocol; native units and metric definitions are given at the start of this subsection.

	Qwen3.6-35B-A3B						Ministral 3 8B					
Metric	Init.	RKL	JSD	PA*	Rel.	D	Init.	RKL	JSD	PA*	Rel.	D
HarmBench ASR	9.5	0.8	0.8	0.8	0.8	0.8	18.7	5.8	6.5	5.0	5.8	5.7
JailbreakBench ASR	26.0	12.3	10.0	6.7	5.3	1.0	37.7	22.7	19.3	17.3	16.7	12.3
MultiBreak ASR	36.9	20.9	17.8	15.6	11.5	6.0	43.9	30.0	26.9	24.7	19.1	17.2
SALAD ASR	15.8	3.9	3.9	3.9	3.9	3.9	23.3	9.5	10.0	8.1	9.5	9.0
SafeDialBench ASR	33.1	17.9	15.6	12.9	10.0	3.9	40.8	26.3	23.0	21.0	15.8	13.8
StrongREJECT (0–1)	0.194	0.079	0.064	0.048	0.045	0.013	0.255	0.144	0.129	0.115	0.111	0.082
IndicSafeEval ASR	37.3	21.9	19.5	17.8	17.4	13.2	46.5	31.1	28.8	26.8	26.4	22.5
Boundary harmful	85.9	95.6	94.4	97.2	96.2	96.4	79.4	88.9	87.7	90.9	89.4	89.9
Boundary benign	80.2	76.8	82.6	81.9	82.9	94.4	72.3	69.4	74.7	73.8	75.4	86.3
BDA	82.9	85.1	88.1	88.9	89.0	95.4	75.7	77.9	80.7	81.4	81.8	88.0
Rule identification	63.0	71.7	78.0	82.0	81.3	95.6	54.8	62.7	69.8	73.4	72.8	86.6
Relevant-factor switch	75.8	77.8	80.8	81.8	81.9	88.7	68.2	69.7	72.5	73.6	74.2	79.5
Irrelevant invariance	85.4	87.2	89.9	90.4	90.6	98.5	79.0	81.4	83.9	84.6	84.8	91.2
CPC	80.6	82.5	85.4	86.1	86.2	93.6	73.6	75.5	78.2	79.1	79.5	85.4
Useful recovery	45.9	58.4	63.7	65.4	72.7	84.1	36.1	47.0	52.2	53.8	61.7	71.8
Residual prefix harm	37.6	21.2	19.3	18.3	14.4	7.9	42.7	25.4	23.3	22.9	18.6	13.8
Safe but unhelpful	14.7	18.6	15.2	14.6	11.1	6.2	19.1	25.5	22.4	21.2	17.7	12.3
Invalid recovery	1.8	1.8	1.8	1.8	1.8	1.8	2.1	2.1	2.1	2.1	2.1	2.1
XSTest answered	92.7	84.4	88.1	85.5	87.2	92.0	93.1	85.2	87.1	86.1	86.5	92.4

Table 16 (continued)

	Qwen3.6-35B-A3B						Ministral 3 8B					
Metric	Init.	RKL	JSD	PA*	Rel.	D	Init.	RKL	JSD	PA*	Rel.	D
XSTest useful	84.9	75.9	79.9	77.1	78.7	86.3	82.5	73.2	76.1	74.5	74.8	83.2
OR-Bench answered	84.8	77.8	80.2	78.9	79.5	87.0	83.3	75.9	78.8	77.4	78.1	85.2
MS-XSB answered	85.8	79.9	82.3	80.6	81.3	85.5	83.5	77.8	79.7	79.0	79.5	82.9
Health compliance	75.7	73.2	78.4	77.2	79.0	90.6	66.3	63.7	69.6	68.4	70.1	81.9
MMLU-Pro	74.9	73.3	74.2	73.9	73.8	74.1	54.8	53.1	54.2	53.8	53.6	54.1
GPQA Diamond	63.5	59.9	61.8	61.1	60.8	61.6	40.6	38.6	40.1	39.7	39.6	39.9
MATH500	82.9	80.9	82.1	81.7	81.4	82.0	59.1	57.1	58.6	58.3	57.7	58.4
Contest-format proxy	38.9	38.9	38.9	38.9	38.9	38.9	23.3	20.0	23.3	21.1	20.0	23.3
LCB v6 pass@1	68.8	66.9	68.1	67.7	67.5	67.9	48.1	45.8	47.5	47.0	46.7	47.4
IFEval strict	87.5	86.8	87.2	87.1	87.0	87.2	81.0	80.0	80.6	80.4	80.3	80.5
Global-MMLU	67.2	65.5	66.5	66.3	66.1	66.4	46.3	44.2	45.5	45.1	44.9	45.3
Marco-Bench-MIF	79.4	78.2	78.9	78.7	78.5	78.8	71.0	70.0	70.7	70.5	70.4	70.7
LongBench v2 (8K)	62.7	60.1	61.6	61.3	61.1	61.5	42.1	39.9	41.5	40.9	40.7	41.2
Held-out boundary BDA	74.5	76.9	79.3	80.1	80.2	86.3	66.1	68.0	71.4	71.8	72.5	79.2

Table 17: **Cohort-level 95% t intervals.** Mean scores are in the full-performance tables; only interval endpoints are repeated here. These intervals summarize request-cohort variation.

Metric	Init.	RKL	JSD	PA*	Relay*	DEEPEN
Qwen3-4B						
BDA	[70.6,73.7]	[70.4,77.0]	[74.7,79.3]	[76.1,79.4]	[75.8,80.2]	[81.9,87.7]
Recovery	[29.3,35.1]	[44.0,46.4]	[48.6,56.1]	[49.8,58.0]	[54.4,66.4]	[66.3,76.3]
Useful	[73.6,88.5]	[63.5,80.2]	[67.5,82.6]	[63.3,82.3]	[66.5,80.9]	[76.5,90.4]
Qwen3-8B						
BDA	[75.4,79.7]	[78.3,81.6]	[81.2,84.0]	[81.5,84.7]	[82.0,84.5]	[88.1,90.8]
Recovery	[35.0,41.2]	[48.5,51.5]	[54.4,57.8]	[55.9,60.5]	[63.3,65.7]	[74.1,75.3]
Useful	[78.0,82.6]	[69.0,74.2]	[73.6,77.6]	[71.1,76.0]	[72.7,76.8]	[79.2,83.2]
Qwen3.5-4B						
BDA	[75.4,82.5]	[79.6,82.5]	[81.0,87.1]	[82.2,87.5]	[81.5,88.6]	[86.8,94.5]
Recovery	[36.3,44.4]	[49.0,53.6]	[55.0,61.2]	[57.2,63.7]	[65.7,68.2]	[74.4,80.2]
Useful	[77.8,91.8]	[71.7,80.6]	[75.8,82.6]	[73.6,81.9]	[74.4,82.7]	[80.7,91.3]
Qwen3.6-27B						
BDA	[77.7,88.1]	[80.6,89.3]	[83.9,92.9]	[84.2,93.7]	[85.3,93.1]	[94.9,96.4]
Recovery	[39.3,56.3]	[54.6,63.3]	[58.6,71.8]	[61.5,73.2]	[68.7,77.8]	[79.9,87.0]
Useful	[79.5,89.5]	[74.6,79.6]	[75.3,83.9]	[74.6,80.6]	[75.5,81.9]	[83.8,88.4]
Qwen3.6-35B-A3B						
BDA	[79.0,86.9]	[81.1,89.2]	[85.0,91.2]	[85.0,92.8]	[85.6,92.4]	[93.7,97.1]
Recovery	[42.1,49.8]	[53.6,63.2]	[57.6,69.8]	[59.6,71.2]	[69.2,76.2]	[76.8,91.4]
Useful	[75.4,94.5]	[64.3,87.5]	[67.9,91.8]	[63.8,90.3]	[65.7,91.6]	[77.0,95.5]
Gemma 4 E4B-it						
BDA	[78.1,79.2]	[79.3,83.3]	[80.3,87.4]	[81.6,87.7]	[82.1,87.7]	[87.3,95.4]
Recovery	[37.9,42.5]	[47.2,59.1]	[54.1,62.5]	[56.0,64.7]	[63.5,72.5]	[74.6,83.7]
Useful	[80.3,86.4]	[73.0,78.2]	[76.8,80.8]	[73.6,80.8]	[74.6,81.4]	[80.3,88.3]
Gemma 4 12B-it						
BDA	[76.9,89.9]	[81.7,89.8]	[85.1,93.0]	[85.5,93.6]	[86.2,93.5]	[95.0,97.8]
Recovery	[38.6,55.6]	[53.1,65.1]	[59.3,73.6]	[62.3,74.8]	[69.6,80.6]	[82.3,89.3]
Useful	[79.6,90.5]	[69.7,83.6]	[72.9,85.8]	[70.3,86.2]	[72.5,85.4]	[78.7,93.9]
Llama 3.2 3B						
BDA	[66.2,75.9]	[69.4,76.9]	[71.9,81.1]	[73.3,81.1]	[72.5,82.3]	[79.0,85.1]
Recovery	[26.7,31.8]	[37.6,42.9]	[44.3,48.1]	[47.6,48.6]	[53.4,57.2]	[63.3,67.7]
Useful	[70.5,85.2]	[60.0,79.2]	[63.7,80.8]	[60.9,80.4]	[63.2,79.7]	[72.7,86.5]
Llama 3.1 8B						
BDA	[70.4,77.5]	[72.0,80.4]	[75.7,83.3]	[76.3,83.5]	[76.9,84.1]	[84.3,89.9]
Recovery	[29.0,39.0]	[41.3,49.3]	[48.4,54.8]	[51.5,55.7]	[57.5,64.4]	[67.5,77.0]
Useful	[76.8,84.0]	[69.1,72.5]	[73.4,76.5]	[72.0,73.1]	[70.5,76.9]	[77.6,84.8]
Ministral 3 8B						
BDA	[74.5,76.9]	[74.7,81.2]	[78.6,82.7]	[79.1,83.8]	[80.2,83.3]	[87.5,88.5]

Table 17 (continued)

Metric	Init.	RKL	JSD	PA*	Relay*	DEEPEN
Recovery	[32.7,39.6]	[44.4,49.6]	[48.6,55.7]	[49.6,57.9]	[59.1,64.2]	[67.9,75.7]
Useful	[75.9,89.1]	[62.7,83.7]	[68.6,83.7]	[65.3,83.8]	[66.1,83.5]	[79.2,87.2]

Table 18: **Qwen3-8B recovery counts.** Each 600-request cohort is exhaustive; safe/useful recovery is not $100 - \text{ASR}$.

Cohort	Safe/useful	Harmful	Safe/useless	Invalid	Total
1	447	74	65	14	600
2	446	75	68	11	600
3	453	74	57	16	600
4	446	67	76	11	600
5	449	63	72	16	600

E.3 Decision representations and position controls

Validation counts alone select matched-ASR checkpoints from a fixed grid; an independent test cohort retains any residual ASR difference. Hidden states are read before the first policy label or refusal token. Readouts use disjoint topics/templates and balanced binary contrasts with matched sample count, capacity, and regularization; random labels test leakage. Rule contrasts are not the full ontology-classification metric. Norm-matched activation directions are built on training data with layer, position, and magnitude fixed on validation data (Zou et al., 2023).

Table 20 distinguishes response decodability from rule/boundary generalization. The rule/boundary intervention improves both sides, whereas a generic refusal direction trades harmful-side gains against benign decisions. Training stages and response-position masks supply complementary behavioral controls, not evidence for a unique circuit or mediator.

Table 19: **Safety-matched decision and recovery.** Validation and test ASR are separate; methods share the paired evaluation protocol.

Condition	Validation ASR	Test ASR	BDA	Recovery	Useful
RKL	7.3	8.5	74.8	47.7	72.4
JSD	7.9	9.4	78.7	55.9	77.3
PA*	8.0	9.2	77.8	57.9	75.1
DEEPEN	7.7	9.4	87.0	72.7	84.7

Table 20: **Representation and activation diagnostics.** Positive changes on both sides differ from a nonspecific increase in refusal.

(a) Pre-label readouts: mean [95% cohort interval] (%)

Condition	Response	Rule contrast	Boundary
PA*	91.9 [90.8, 93.0]	68.8 [66.9, 70.7]	65.5 [63.8, 67.1]
DEEPEN	93.8 [93.0, 94.6]	84.5 [82.3, 86.6]	82.5 [81.1, 83.9]
random labels	48.3 [46.5, 50.1]	50.7 [48.8, 52.6]	49.0 [47.1, 51.0]

(b) Activation intervention: paired correct-decision change (pp)

Intervention	Harmful correct change	Benign correct change
rule boundary	6.5 [5.4, 7.6]	6.1 [5.1, 7.0]
random	-0.1 [-0.2, 0.1]	0.1 [-0.1, 0.3]
topic	1.0 [0.6, 1.3]	-0.5 [-0.6, -0.3]
general refusal	7.6 [6.7, 8.6]	-10.9 [-12.5, -9.3]
wrong position	0.2 [0.0, 0.4]	-0.2 [-0.4, 0.0]

Table 21: **Training dynamics and position dependence.** Masks are fixed before actions; supplied prefix tokens receive no imitation loss.

(a) Training checkpoints (%)

Condition	iid agreement	Rule ID	BDA	Recovery
start	69.2	53.8	66.7	38.3
quarter	87.3	68.3	73.7	47.6
half	93.0	81.3	82.0	61.7
final	95.2	87.5	87.1	75.1

(b) Response-position controls (%)

Condition	ASR	BDA	Recovery	Useful
head only	7.7	71.0	41.5	69.9
tail only	13.4	78.5	58.8	78.7
uniform	10.0	81.5	62.0	80.8
DEEPEN	10.0	86.9	74.3	85.3

E.4 Policy information, conditioning, and exposure

Content controls match auxiliary length, format, response label, request pool, and loss budget while holding the teacher's probability-scoring condition fixed. Generic, paraphrased, or label-preserving shuffled text therefore tests policy-specific information. Basic content/conditioning comparisons disable handoff. The conditioning ablation keeps the auxiliary target fixed but varies whether the response receives no label, its student-predicted label, or an oracle label; oracle conditioning is not deployable. Field corruption occurs before answering, and irrelevant-field controls separate decision content from formatting.

Starting-state controls replace removed sources with natural starts under the stated budget view. Natural visits are measured independently of supplied boundary/prefix starts. Increased natural rollouts carry their own cost; they are not a fixed-compute comparison. Tables 22–24 separate content, conditioning, field, and exposure effects. The source and dose experiments test targeted access without assuming that every later state becomes covered.

Table 22: **Policy, state, and conditioning controls.** Each panel retains its own metric columns and experimental protocol.

(a) Policy-information content (%)

Condition	Rule ID	BDA
correct	88.4	85.5
generic	65.4	74.9
hard shuffled	61.5	71.6
paraphrase	87.5	85.0

(b) Critical-state starting pool (%)

Condition	Natural occupancy	Recovery
critical	18.9	66.0
matched random	11.0	53.5
more natural	14.5	57.1

(c) Label use / response conditioning (%)

Condition	BDA	Recovery
auxiliary only	78.5	57.8
predicted conditioning	87.3	72.5
oracle non deploy	91.3	77.3
wrong field	75.3	60.7

Table 23: **Fields and exposure.** Label length is matched across field controls; natural rollout scaling increases generation cost.

(a) Structured fields (%)

Condition	Rule ID	BDA	Recovery
response only	60.9	74.9	58.1
plus rule	84.8	82.0	64.8
plus boundary	86.9	86.5	72.4
all fields	87.3	86.4	72.7

(b) Natural-rollout budget (%)

Condition	Natural occupancy	Recovery
1x	11.1	46.6
2x	13.8	53.6
4x	15.5	57.4
8x	16.1	59.0

(c) Targeted-state exposure dose (%)

Condition	Policy: BDA	Mix: recovery	Utility: useful
low	82.1	66.7	81.0
default	86.7	73.7	84.3
high	86.2	72.5	84.6

Table 24: **Exposure and resource ledger.** External starting-state counts are not part of natural-state occupancy. Resource multipliers use the declared budget view.

State arm	Boundary	Prefix	Student M	Teacher M
critical	2500	2500	10.24	18.00
matched random	400	400	10.24	18.00
more natural	0	0	16.38	18.00

E.5 Factorial tests of complementary components

The grounding/state design holds the local and reference objectives fixed and disables handoff; no-grounding groups replace the student-side label channel with a content-free control while retaining teacher policy-conditioned probabilities. The boundary/prefix design varies the two external starting sources. Boundary exposure crossed with benign-reference FKL separates learning when to answer from preserving reference behavior. Finally, local routing crossed with handoff fixes structured supervision and the starting mixture; its local baseline is RKL everywhere. Removed sources are replaced within the specified budget. Main effects average over the other factor; conditional effects and interaction use paired cohort differences. Complementarity does not require positive superadditivity, and point-score capability gates are not statistical non-inferiority tests.

Table 25: **Factorial cells and contrasts.** Brackets contain 95% request-cohort intervals for the paired contrasts.

(a) Grounding / state: cell means (%)

Condition	BDA	Recovery
G0S0	72.9	42.4
G1S0	81.7	54.0
G0S1	80.4	60.6
G1S1	86.8	69.0

Table 25 (continued)

(b) Grounding / state: paired contrasts (pp)

Contrast	BDA	Recovery
Factor 1, factor 2 off	8.8 [7.9, 9.6]	11.6 [10.8, 12.4]
Factor 1, factor 2 on	6.4 [5.7, 7.1]	8.4 [6.5, 10.3]
Factor 2, factor 1 off	7.4 [6.3, 8.5]	18.2 [17.6, 18.8]
Factor 2, factor 1 on	5.0 [4.7, 5.3]	15.0 [12.6, 17.5]
Factor 1 main effect	7.6 [7.2, 8.0]	10.0 [9.2, 10.8]
Factor 2 main effect	6.2 [5.8, 6.7]	16.6 [15.3, 18.0]
Interaction	-2.4 [-3.7, -1.1]	-3.2 [-5.6, -0.8]

(c) Boundary / prefix: cell means (%)

Condition	BDA	Recovery
B0P0	73.2	42.9
B1P0	83.9	50.8
B0P1	75.3	61.5
B1P1	86.6	68.3

(d) Boundary / prefix: paired contrasts (pp)

Contrast	BDA	Recovery
Factor 1, factor 2 off	10.7 [8.9, 12.5]	7.9 [7.1, 8.7]
Factor 1, factor 2 on	11.2 [9.4, 13.0]	6.8 [5.5, 8.2]
Factor 2, factor 1 off	2.2 [1.9, 2.4]	18.6 [15.9, 21.3]
Factor 2, factor 1 on	2.7 [2.1, 3.3]	17.5 [15.1, 20.0]
Factor 1 main effect	11.0 [9.2, 12.7]	7.4 [6.6, 8.2]
Factor 2 main effect	2.4 [2.1, 2.8]	18.1 [15.6, 20.5]
Interaction	0.6 [-0.0, 1.1]	-1.1 [-2.6, 0.5]

Table 26: **Factorial cells and contrasts.** Brackets contain 95% request-cohort intervals for the paired contrasts.

(a) Boundary / reference: cell means (%)

Condition	BDA	Recovery
B0U0	77.1	66.9
B1U0	85.4	69.0
B0U1	80.7	66.9
B1U1	87.0	69.0

(b) Boundary / reference: paired contrasts (pp)

Contrast	BDA	Recovery
Factor 1, factor 2 off	8.3 [7.6, 9.1]	2.0 [1.4, 2.7]
Factor 1, factor 2 on	6.3 [5.6, 7.0]	2.0 [1.4, 2.7]
Factor 2, factor 1 off	3.6 [2.9, 4.4]	0.0 [0.0, 0.0]
Factor 2, factor 1 on	1.6 [0.8, 2.4]	0.0 [0.0, 0.0]
Factor 1 main effect	7.3 [6.8, 7.8]	2.0 [1.4, 2.7]
Factor 2 main effect	2.6 [2.0, 3.2]	0.0 [0.0, 0.0]
Interaction	-2.0 [-3.0, -1.1]	0.0 [0.0, 0.0]

(c) Local routing / handoff: cell means (%)

Condition	Local action	Recovery
F0H0	2.7	45.6
F1H0	7.8	59.7
F0H1	3.2	57.6
F1H1	8.4	70.9

Table 26 (continued)
(d) Local routing / handoff: paired contrasts (pp)

Contrast	Local action	Recovery
Factor 1, factor 2 off	5.1 [4.4, 5.9]	14.1 [13.2, 14.9]
Factor 1, factor 2 on	5.2 [4.3, 6.0]	13.4 [13.0, 13.8]
Factor 2, factor 1 off	0.5 [0.3, 0.8]	11.9 [11.3, 12.6]
Factor 2, factor 1 on	0.6 [0.3, 0.8]	11.3 [10.9, 11.6]
Factor 1 main effect	5.2 [4.4, 5.9]	13.7 [13.2, 14.2]
Factor 2 main effect	0.5 [0.4, 0.7]	11.6 [11.3, 11.9]
Interaction	0.0 [-0.3, 0.4]	-0.7 [-1.6, 0.2]

E.6 Local objectives, intervention, and autonomous recovery

The divergence comparison shares initialization, starting-state distribution, non-divergence objectives, and response-token budgets; each method generates its own current-policy suffixes. All arms retain full-FKL benign anchoring. Frozen-state gradient audits are separate from end-to-end outcomes. The union preserves original full-vocabulary probabilities and one tail bin; selection uses the rollout overlap lower bound, and a larger FKL fraction does not imply fewer teacher forward passes.

Trigger controls distinguish teacher-label correction, a short bridge, fixed-position/random interventions, and KL-based triggers. Attempts, rejected segments, and resumed-student work remain chargeable. Bridge length and student-only horizon measure different quantities: longer teacher text need not improve independent continuation. Two-sided interventions also correct benign over-refusal. The independent frozen reference comparison changes only the utility objective; only full FKL provides the full-distribution bound of Equation (20). Aggregate utility does not replace per-task capability results.

Table 27: **Local objective and routing controls.** Top-k compresses supervision, not response sampling.

(a) End-to-end divergence choice (%)

Condition	BDA	Recovery	Useful
sampled RKL	78.2	54.6	78.0
full RKL	79.2	55.8	78.9
full FKL	86.2	70.4	81.6
fixed JSD	85.7	68.3	84.4
routed	86.8	71.5	84.4

(b) Routing threshold (%)

Condition	FKL fraction	Recovery	Useful
0.60	23.2	68.3	84.0
0.70	36.8	70.8	85.1
0.80	54.4	72.1	85.1
0.90	72.1	72.4	84.1
1.00	100.0	71.9	83.1

Table 28: **Bridge and recovery controls.** Teacher-token and response-token budgets remain distinct; suffix recovery does not erase prior harmful context.

(a) Intervention trigger: recovery (%)

Condition	Recovery
label only	59.1
short bridge	72.1
fixed checkpoint	66.0
random matched	60.7
kl trigger	64.0

Table 28 (continued)
(b) Teacher continuation length (%)

Condition	Recovery	Useful
0	60.7	84.9
8	67.0	85.4
16	71.5	85.4
32	72.0	84.9
64	71.8	84.2

(c) Student-only recovery horizon (%)

Condition	no bridge	bridge
32	78.2	83.8
64	68.5	79.6
128	59.7	73.8
256	42.5	59.2

(d) Benign over-refusal correction (%)

Condition	Benign answer	Recovery
no handoff	81.3	59.8
harmful only	79.4	71.9
two sided	84.9	72.8

Table 29: **Independent benign-reference objective.** Percentages; the reference and benign states are held fixed.

Condition	Useful	MMLU	Math	Recovery
none	79.0	60.9	71.7	73.0
RKL	82.3	62.8	74.7	72.5
topk FKL	84.0	63.5	75.7	72.4
full FKL	84.3	63.7	76.1	72.1

E.7 Training costs and deployment

Training cost includes data/label generation, full teacher scoring, every bridge attempt, full-FKL reference work, student response/policy generation, warm-up/training, and systems overhead. Evaluation infrastructure and human audits lie outside this boundary. Active parameters determine MoE compute, while total parameters determine residency; unused vision components are excluded from text-only execution. Token matching, teacher-work matching, and GPU-hour matching are distinct protocols. Smaller probability exports do not remove the full teacher forward.

At deployment only the student is used. Each label refresh may rebuild canonical context; policy tokens, native thinking, and cache reconstruction are charged. Paired workloads share visible-output and thinking lengths. Mean paired latency ratios are not ratios of displayed means; p95 describes workload variation, not a confidence interval. Cost-to-quality remains an endpoint-interpolated estimate, not an observed first-passage training cost.

Table 30: **Full training cost.** Each method cell is GPU-hours / wall-hours. GB/GPU is the common peak-memory figure for the five methods at that model size.

Model	GPUs	GB/GPU	RKL	JSD	PA*	Relay*	DEEPEN
Qwen3-4B	8	34.0	19.2 / 2.40	19.2 / 2.40	19.5 / 2.44	21.1 / 2.64	29.1 / 3.64
Qwen3-8B	8	44.0	30.7 / 3.84	30.7 / 3.84	31.4 / 3.92	33.9 / 4.24	47.2 / 5.90
Qwen3.5-4B	8	34.0	19.2 / 2.40	19.2 / 2.40	19.5 / 2.44	21.1 / 2.64	29.1 / 3.64
Qwen3.6-27B	16	57.8	71.2 / 4.45	71.2 / 4.45	73.5 / 4.59	78.7 / 4.92	112.2 / 7.01
Qwen3.6-35B-A3B	16	67.8	15.8 / 0.99	15.8 / 0.99	16.0 / 1.00	17.4 / 1.09	23.9 / 1.49
Gemma 4 E4B-it	8	44.0	20.8 / 2.60	20.8 / 2.60	21.1 / 2.64	22.9 / 2.86	31.6 / 3.95

Table 30 (continued)

Model	GPUs	GB/GPU	RKL	JSD	PA*	Relay*	DEEPEN
Gemma 4 12B-it	8	53.9	40.4 / 5.06	40.4 / 5.06	41.5 / 5.18	44.7 / 5.58	62.6 / 7.83
Llama 3.2 3B	8	31.5	15.8 / 1.97	15.8 / 1.97	16.0 / 2.00	17.4 / 2.18	23.9 / 2.98
Llama 3.1 8B	8	44.0	30.7 / 3.84	30.7 / 3.84	31.4 / 3.92	33.9 / 4.24	47.2 / 5.90
Ministral 3 8B	8	46.0	31.8 / 3.97	31.8 / 3.97	32.5 / 4.06	35.1 / 4.38	48.8 / 6.10

Table 31: **DEEPEN cost decomposition.** GPU-hours: data/labels, full teacher score, bridge attempts, reference, generation/cache, train, and systems. T-total and Ref-rounded retain the summary precision; /PA* is the total-cost ratio. Full totals and GPU counts are above.

Model	Data	Score	Bridge	Ref.	Gen.	Train	Sys.	T-total	Ref-rnd.	/PA*
Qwen3-4B	3.78	0.80	1.12	0.05	19.39	0.83	3.12	5.7	0.0	1.49
Qwen3-8B	6.29	1.59	1.77	0.10	30.70	1.66	5.05	9.7	0.1	1.50
Qwen3.5-4B	3.78	0.80	1.12	0.05	19.39	0.83	3.12	5.7	0.0	1.49
Qwen3.6-27B	15.80	5.37	3.95	0.33	69.14	5.62	12.02	25.1	0.3	1.53
Qwen3.6-35B-A3B	3.07	0.60	0.93	0.04	16.03	0.62	2.56	4.6	0.0	1.49
Gemma 4 E4B-it	4.12	0.89	1.21	0.06	20.96	0.94	3.38	6.2	0.1	1.49
Gemma 4 12B-it	8.49	2.38	2.31	0.15	40.09	2.49	6.71	13.2	0.1	1.51
Llama 3.2 3B	3.07	0.60	0.93	0.04	16.03	0.62	2.56	4.6	0.0	1.49
Llama 3.1 8B	6.29	1.59	1.77	0.10	30.70	1.66	5.05	9.7	0.1	1.50
Ministral 3 8B	6.52	1.67	1.83	0.10	31.71	1.75	5.23	10.0	0.1	1.50

Table 32: **Hardware and accounting settings.** Deployment uses matched visible and native-thinking lengths; natural-stop sensitivity is separate. Workload factors are accounting settings, not performance scores.

Setting	Value	Setting	Value
Hardware	NVIDIA H100 SXM 80GB, BF16	Student+optimizer bytes/param	16
Small-model GPUs	8	Frozen bytes/param	2
Large-model GPUs	16	Resident teacher/reference copies	2
Large-model threshold (B)	20	Probability bytes	2
Peak TFLOPS	989	Token-ID bytes	4
Training utilization	0.28	Reference vocabulary size	151936
Prefill utilization	0.35	Teacher scoring positions	10240000
4B decode tok/s/GPU	180	Deployment requests	200
System overhead fraction	0.12	Length lognormal σ	0.35
Activation GB/GPU	24		

Method	Labels	Bridge	Student	Score	Ref.	Train	Data
RKL	0	0	1.0	1.0	1.0	1.0	0.25
JSD	0	0	1.0	1.0	1.0	1.0	0.25
PA*	0	0	1.0	1.25	1.0	1.0	1.0
Relay*	0	0.1	1.0	1.15	1.0	1.1	0.25
DEEPEN	1	0.07	1.0	1.1	1.0	1.12	1.0

Table 33: **Endpoint-interpolated cost to the joint target (GPU-hours).** Gates: HB ASR $\leq$ 12.0, MB ASR $\leq$ 30.0, BDA $\geq$ 75.0, recovery $\geq$ 50.0; useful-answer and each of eight capability drops $\leq$ 2.0 pp. All must hold. NR means no budget-grid crossing; gates are not statistical non-inferiority tests.

Model	PA*	JSD	Relay*	DEEPEN
Qwen3-4B	NR	NR	NR	22.1
Qwen3-8B	NR	NR	NR	30.2
Qwen3.5-4B	NR	NR	NR	15.1
Qwen3.6-27B	18.6	17.9	19.7	18.2
Qwen3.6-35B-A3B	NR	5.9	NR	6.7

Table 33 (continued)

Model	PA*	JSD	Relay*	DEEPEN
Gemma 4 E4B-it	NR	NR	8.6	12.6
Gemma 4 12B-it	10.5	NR	11.2	17.6
Llama 3.2 3B	NR	NR	NR	NR
Llama 3.1 8B	NR	NR	NR	35.8
Ministral 3 8B	NR	NR	NR	31.2

Table 34: **Deployment workload and accounting.** Token counts and TTFT seconds; sensitivity columns vary visible output length by 0.8x/1x/1.2x. No runtime teacher is present.

(a) Paired token workloads

Workload	Input	Visible output	Thinking
Short chat	512	128	0
Boundary	1024	256	0
Long / multi-turn	8192	512	0
Reason / code	2048	1024	1024

(b) Qwen3-8B internal work and output-length sensitivity

Load	Policy	Rebuild	Base TTFT	New TTFT	0.8x	1x	1.2x
Short chat	34.6	284.2	0.14	0.35	1.03	1.28	1.54
Boundary	58.6	1729.3	0.26	0.47	1.11	1.39	1.66
Long / multi-turn	106.9	29417.0	2.05	2.26	1.78	2.23	2.67
Reason / code	388.9	48942.7	0.52	0.73	1.47	1.83	2.20

Table 35: **Latency across models and workloads.** Base, DEEPEN mean, p95, and user-visible TTFT are seconds; Ratio is the mean paired ratio. TTFT includes the initial policy label.

Model	Load	Base s	Mean s	p95 s	Ratio	TTFT s
Qwen3-4B	short	0.81	1.06	1.62	1.29	0.22
	boundary	1.61	2.23	3.67	1.38	0.30
	long	4.05	9.19	15.05	2.22	1.44
	reason	11.68	21.86	36.99	1.84	0.46
Qwen3-8B	short	1.26	1.64	2.71	1.28	0.35
	boundary	2.40	3.34	5.60	1.39	0.47
	long	6.47	14.73	23.75	2.23	2.26
	reason	18.02	33.60	59.87	1.83	0.73
Qwen3.5-4B	short	0.81	1.06	1.67	1.30	0.22
	boundary	1.61	2.24	3.64	1.38	0.30
	long	4.12	9.39	15.42	2.23	1.44
	reason	11.05	20.53	36.95	1.82	0.46
Qwen3.6-27B	short	2.83	3.70	5.67	1.30	0.76
	boundary	5.49	7.67	12.53	1.38	1.04
	long	14.15	32.31	52.83	2.23	4.98
	reason	39.44	73.29	126.07	1.83	1.61
Qwen3.6-35B-A3B	short	0.65	0.84	1.35	1.29	0.18
	boundary	1.32	1.84	3.17	1.39	0.25
	long	3.43	7.88	12.56	2.25	1.19
	reason	9.57	17.82	31.98	1.83	0.38
Gemma 4 E4B-it	short	0.86	1.12	1.76	1.29	0.24
	boundary	1.71	2.37	3.95	1.38	0.33
	long	4.58	10.56	16.70	2.26	1.55
	reason	12.63	23.64	42.92	1.84	0.50
Gemma 4 12B-it	short	1.52	1.95	3.16	1.27	0.45
	boundary	3.16	4.39	7.62	1.38	0.61
	long	8.53	19.53	31.56	2.23	2.93
	reason	23.93	44.77	74.51	1.84	0.95
Llama 3.2 3B	short	0.63	0.82	1.24	1.28	0.18

Table 35 (continued)

Model	Load	Base s	Mean s	p95 s	Ratio	TTFT s
	boundary	1.34	1.87	2.96	1.39	0.25
	long	3.51	8.08	12.75	2.25	1.19
	reason	9.27	17.18	29.32	1.83	0.38
Llama 3.1 8B	short	1.18	1.53	2.44	1.28	0.35
	boundary	2.52	3.51	5.88	1.38	0.47
	long	6.33	14.33	23.53	2.22	2.26
	reason	17.45	32.30	55.32	1.82	0.73
Ministral 3 8B	short	1.25	1.64	2.57	1.30	0.36
	boundary	2.58	3.60	6.00	1.39	0.49
	long	6.63	15.12	24.51	2.23	2.33
	reason	19.14	35.69	59.74	1.84	0.75

E.8 Robustness, teacher quality, and remaining errors

Teacher perturbations separate random labels, systematic rule errors, and bridge errors; filtered variants charge rejected work. Agreement filtering can remove difficult states and cannot guarantee correctness. Decoding sweeps affect evaluation, not sampled-RKL training truncation. IndicSafeEval language strata use equal-language means, while Global-MMLU and multilingual instruction tasks measure different outcomes. The common-window LongBench and contest-format proxy keep their restricted scopes.

Table 36 shows that systematic teacher errors can lower boundary performance; Table 36 partitions residual errors into rule confusion, boundary-attribute error, persistent unsafe paths, and other outcomes. These categories describe behavior, not neural causes. Student-only horizon tests and benign over-refusal controls limit any claim of arbitrary-depth recovery or uniformly reliable transfer.

Table 36: **Robustness and residual errors.** Each panel is a separate control; error-composition categories form its stated partition.

(a) Teacher perturbations: BDA (%)

Condition	Random noise	Rule error	Bridge error	Filtered
0%	86.4	86.0	86.5	85.1
10%	84.4	81.0	82.3	83.8
25%	79.3	70.8	74.1	79.9
40%	71.9	60.7	63.8	72.9

(b) Evaluation decoding (%)

Condition	ASR	Recovery	Useful
greedy	9.8	72.4	85.8
T0.7 p1	13.3	69.6	84.9
T1 p1	17.0	65.3	82.8
T1 p0.9	15.2	67.3	83.3

(c) Replay-buffer error composition (%)

Condition	rule confusion	attribute error	path persistence	other
capacity low	30.4	25.6	30.8	13.2
capacity high	22.3	30.8	27.6	19.3
capacity recent	20.9	27.9	31.0	20.2

Table 37: **IndicSafeEval language strata: ASR (%).** Equal-language aggregation reproduces the main-protocol macro. D denotes DEEPEN; language scores are not pooled across judges.

	Hindi			Bengali			Marathi			Punjabi		
Model	Init.	PA*	D	Init.	PA*	D	Init.	PA*	D	Init.	PA*	D
Qwen3-4B	43.9	25.5	21.1	47.4	28.8	24.9	50.1	31.5	26.4	55.2	34.4	30.4
Qwen3-8B	39.8	22.3	18.1	44.3	25.8	20.6	47.6	28.3	23.4	51.3	32.1	27.5
Qwen3.5-4B	39.1	21.1	16.9	40.6	22.3	18.3	42.5	24.1	18.2	46.9	27.0	23.3
Qwen3.6-27B	30.7	11.3	8.2	33.5	14.7	10.7	35.7	19.2	14.5	40.9	22.4	18.9
Qwen3.6-35B-A3B	33.3	13.8	9.3	34.3	14.3	10.5	39.2	19.3	14.4	42.3	23.7	18.8
Gemma 4 E4B-it	33.2	14.9	11.1	40.1	19.9	15.5	39.2	19.9	15.5	46.1	25.3	21.4
Gemma 4 12B-it	32.2	12.8	8.3	35.7	14.6	10.7	35.0	18.4	14.1	39.9	21.7	17.0
Llama 3.2 3B	50.9	31.9	27.5	54.1	35.3	31.5	57.7	37.8	33.7	59.9	40.3	36.8
Llama 3.1 8B	43.5	24.6	20.7	46.6	29.1	24.4	52.3	31.5	26.8	52.9	33.9	29.0
Ministral 3 8B	43.1	22.6	18.5	43.9	24.3	20.1	47.0	27.7	23.5	51.8	32.7	27.9

E.9 FINITE-DISTRIBUTION AND CONTROLLED-WORLD CHECKS

The audit checks the full/merged FKL decomposition, conditional-tail gap, gradient discrepancy, and overlap bounds of Appendix A.3. Positive probabilities and continuous zero-tail limits are used. Fixed-state identities do not ensure finite-step improvement or global safety. The toy world enumerates harmful intent, authorization, and an irrelevant topic bit; its refusal rule is intent without authorization. A single action path and the set of compliant paths have different probabilities, and starting-state intervention does not guarantee every subsequent state.

Table 38: **Finite-distribution top-k audit.** KL and gradient terms use natural units; overlap is in [0,1]. Both teacher and student top-k sets form the union.

k	Full KL	Merged	Tail	Grad. L_1	ρ^-	ρ	ρ^+
8	0.562	0.175	0.387	0.554	0.257	0.566	0.799
16	0.562	0.279	0.283	0.387	0.365	0.566	0.716
32	0.562	0.400	0.162	0.204	0.472	0.566	0.635
64	0.562	0.507	0.055	0.058	0.542	0.566	0.582
128	0.562	0.557	0.006	0.004	0.564	0.566	0.567
256	0.562	0.562	0.000	0.000	0.566	0.566	0.566
512	0.562	0.562	0.000	0.000	0.566	0.566	0.566

Table 39: **Numerical identities and policy-world enumeration.** Positive bound slack means the full overlap lies inside its bounds.

(a) Floating-point residuals

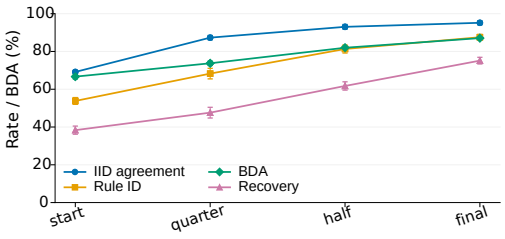
Check	Result
maximum KL decomposition error	3.053e-16
minimum overlap bound slack	0.000e+00

(b) Enumerated intent / authorization / topic world

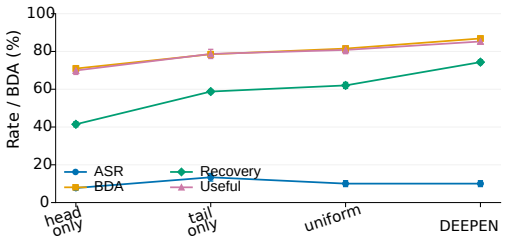
Harmful intent	Authorized	Topic bit	Refuse
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	1
1	0	1	1
1	1	0	0
1	1	1	0

E.10 GROUPED VISUAL SUMMARIES

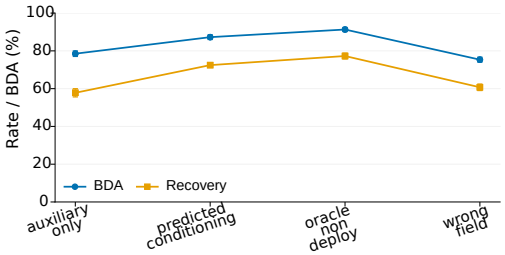
The following groups visualize the controls above without pooling experiments. Curves and uncertainty marks retain their original data; the corresponding tables specify each protocol and numerical values.



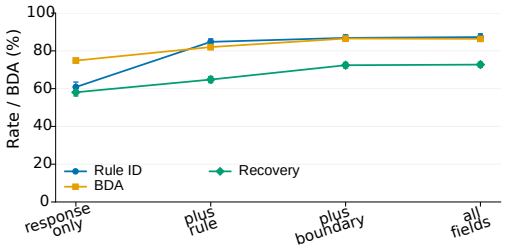
(a) Training checkpoints



(b) Response-position masks

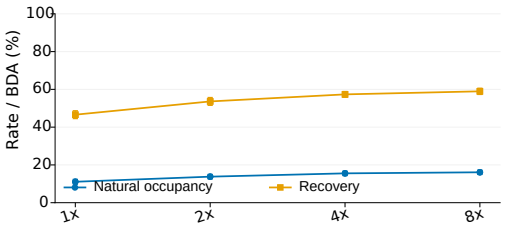


(c) Prediction vs. conditioning

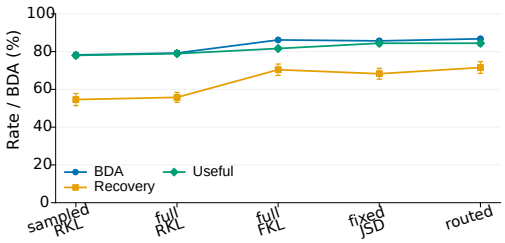


(d) Structured policy fields

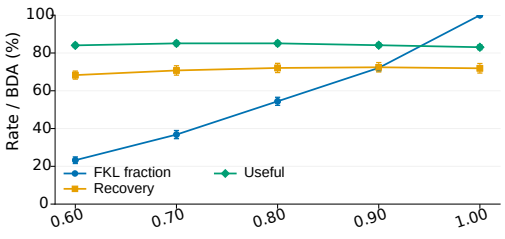
Figure 8: **Decision learning and policy conditioning.** Panels (a)–(b): Table 21; (c): Table 22; (d): Table 23. Oracle conditioning is a nondeployable diagnostic.



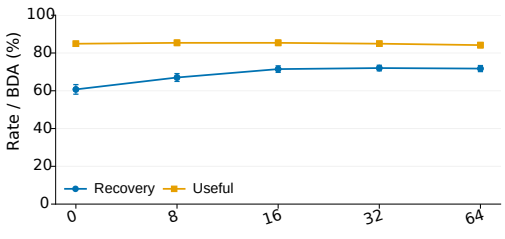
(a) Natural-rollout scaling



(b) Local objective

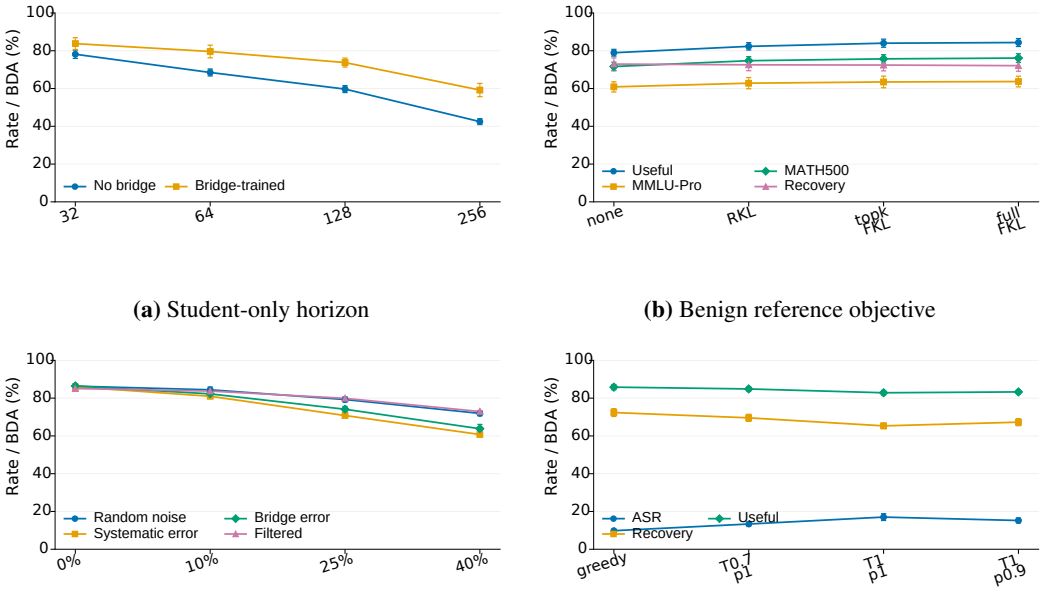


(c) Overlap threshold



(d) Teacher-segment length

Figure 9: **Exposure, local supervision, and bridge length.** (a): Table 23; (b)–(c): Table 27; (d): Table 28. More natural rollouts incur cost; the FKL fraction is not a teacher-forward saving.



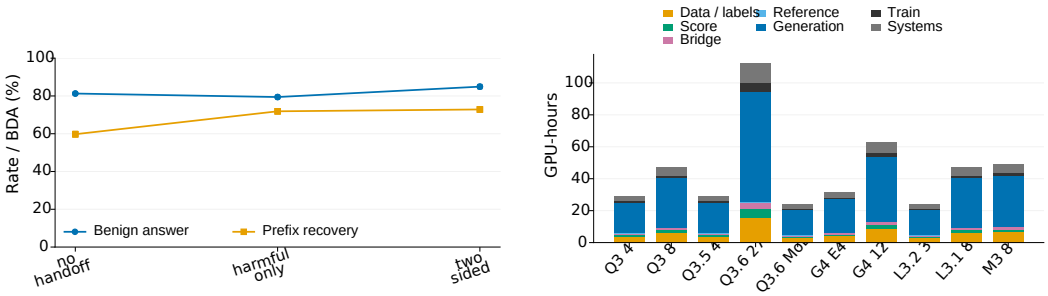
(a) Student-only horizon

(b) Benign reference objective

(c) Teacher quality

(d) Evaluation decoding

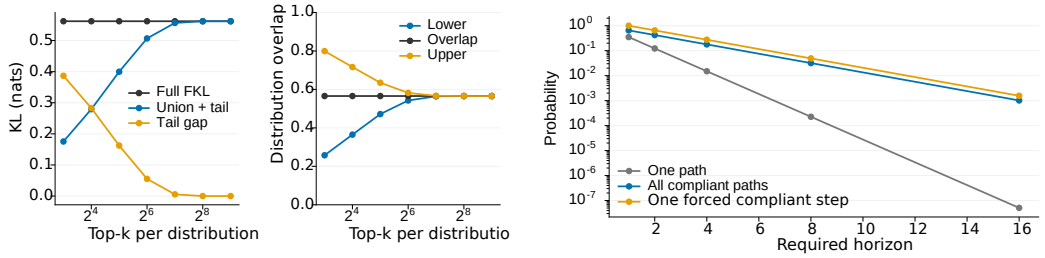
Figure 10: **Independent recovery, utility, and robustness.** (a): Table 28; (b): Table 29; (c)–(d): Table 36. Teacher removal precedes horizon evaluation; decoding sweeps are evaluation controls.



(a) Benign and harmful correction

(b) Training-cost components

Figure 11: **Two-sided correction and resource cost.** (a): Table 28, with matched total teacher budget; (b): Table 31. Component costs include teacher and reference work.



(a) Finite top-k audit

(b) Specified vs. compliant paths

Figure 12: **Distribution and path checks.** (a): Table 38 and Equation (14); (b): the path construction in Appendix E.9. Payload reduction is distinct from teacher-compute reduction, and a single path is not all semantic recovery.