ECHO: ENDOGENOUS CITATION VIA HINDSIGHT-ORACLE SELF-DISTILLATION FOR FACTUAL RECALL

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models often fail to produce facts that are demonstrably encoded in their parameters: in closed-book question answering, access rather than acquisition is often the bottleneck. Reasoning helps by recalling related facts, but one hallucinated premise can derail the answer, and catching such premises in training has required external verifiers that are costly and least reliable on long-tail facts. We exploit two asymmetries: verifying a claim is easier than generating it, and hindsight requires no labels. Conditioned on self-verified claims pooled from its rollouts, the frozen model becomes a *hindsight oracle* that uses no gold answer, retrieval, or external judge, yet is markedly stronger than the student on questions whose answer it can verify but rarely produces. ECHO diagnoses knowledge states without gold answers, builds the oracle from a minimal set of recall-sufficient cues, and distills it into the student on-policy, weighting tokens by the teacher–student entropy gap, with an influence-gated grounding reward. Labels are confined to offline verifier calibration and hyperparameter selection. On Qwen3.5-27B, ECHO is non-inferior to gold-answer RLVR in accuracy (33.1% vs. 33.4% on SimpleQA-Verified) while raising trace factual precision by 6.4 points and FActScore by 4.1. Gains concentrate on tip-of-the-tongue questions, the measured encoding rate is unchanged, and interventions show reliance on recalled cues.

1 INTRODUCTION

Factual hallucinations of language models are commonly attributed to gaps in parametric knowledge (Ji et al., 2023; Huang et al., 2025a). Measured one fact at a time, the picture is different: frontier models encode 95–98% of the facts in the WikiProfile benchmark, yet fail to answer 26–34% of them when asked directly and 11–12% even when allowed to think (Calderon et al., 2026). Scaling removes encoding failures while recall failures persist, above all in the long tail (Kandpal et al., 2023; Mallen et al., 2023). On Qwen3.5-27B, two thirds of the errors on WikiProfile concern facts the model encodes but cannot retrieve, and about half on the smaller 9B (Appendix F). Access, not only acquisition, is a major and addressable bottleneck.

Reasoning is a partial remedy (Wei et al., 2022; DeepSeek-AI, 2025). Generating related facts primes the target, but the mechanism cuts both ways: in the audit of Gekhman et al. (2026) on SimpleQA-Verified, traces whose intermediate facts are all correct reach 41.4% accuracy, whereas those with a hallucinated fact reach only 26.4%. Filtering traces by fact quality helps at test time but needs many samples and a search-based auditor (Min et al., 2023; Wei et al., 2024b), which is costly and least reliable on rare facts; training-time remedies fall short. Reinforcement learning with verifiable rewards (RLVR; Lambert et al., 2024; DeepSeek-AI, 2025) improves closed-book accuracy (Ma & Hewitt, 2026) but rewards answers, not the facts beneath them; in our experiments it leaves the factual precision of traces almost unchanged. Distillation from a larger model (Hinton et al., 2015) is no ready substitute: in mid-training it slows factual recall, because the teacher is itself unsure on knowledge tokens (He et al., 2026). We therefore ask: *can a model learn to reason only over facts it can corroborate itself, without gold answers or an external verifier in the training loop?*

We answer affirmatively by exploiting two asymmetries. *Verification is easier than generation*: a model that cannot produce a fact on demand often recognizes it once the fact is written out and probed

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

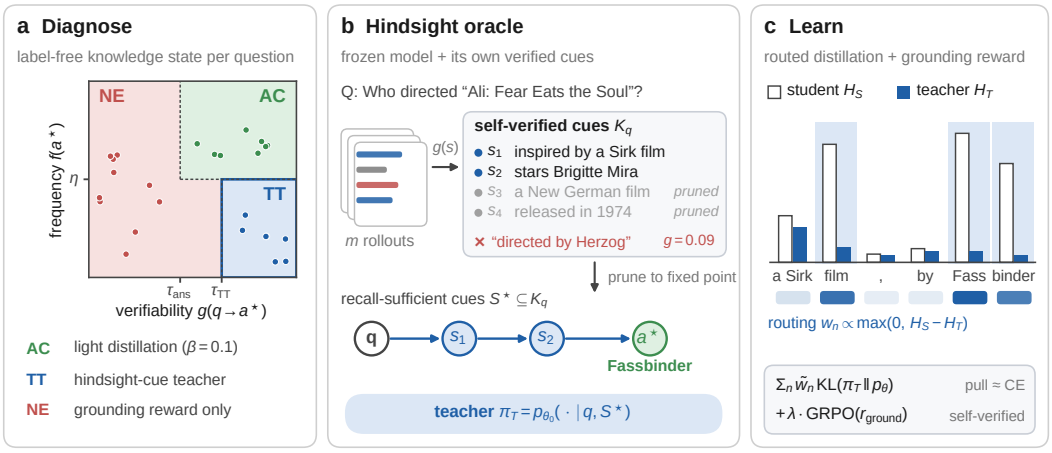

Figure 1: ECHO overview. (a) Without gold answers, each question is placed by the self-verifiability g of its best candidate answer $a^\star$ and by how often the student generates it, $f(a^\star)$. **(b)** Claims from m rollouts are verified by masked-object probes of the frozen model; the survivors K_q are pruned to recall-sufficient cues $S^\star$, and the frozen model conditioned on $S^\star$ is the hindsight oracle (the path is illustrative). **(c)** The student is distilled toward the oracle on its own tokens with weights $\max(0, H_S - H_T)$, plus a grounding reward.

from several directions (Calderon et al., 2026; Li et al., 2024). *Hindsight requires no labels at training time*: the claims in a group of rollouts that survive self-verification can be pooled and handed back to the model as context. Together they yield the *hindsight oracle*, the frozen initial model conditioned on its self-verified cues. It uses no gold answer, retrieval, or external judge, yet is markedly stronger on questions whose answer the model can verify but rarely produces (§4.3): its advantage comes from pooling and filtering its own samples, whereas a larger teacher or a self-distillation teacher that sees gold solutions (Zhao et al., 2026) brings in external information.

ECHO (Figure 1) turns the oracle into a training signal shaped by the structure of recall. It *diagnoses* each question as accessible (AC), tip-of-the-tongue (TT; verifiable yet rarely generated, Brown & McNeill, 1966), or not evocable (NE). It prunes the oracle's context to a 1-minimal set of *recall-sufficient* cues, so that the oracle demonstrates a recall path with few false premises. The student then *learns* from the oracle by on-policy, token-level distillation weighted toward positions where the oracle is more certain than the student, plus a reward for influential, self-verified claims.

On Qwen3.5-9B and 27B, ECHO is non-inferior to gold-answer RLVR in out-of-distribution accuracy while raising trace factual precision over RLVR by 6.2–6.4 points and FActScore by 4.1–4.9. The gains land where the method is designed to act: +19.0 points on TT questions against +8.4 for RLVR; 22.0% of recall failures become answerable while the measured encoding rate stays flat; and on matched cues, deleting a recalled cue costs ECHO 27.4 points of answer retention against 11.2 for RLVR. We contribute (i) the hindsight oracle with a label-free diagnosis of knowledge states; (ii) recall-sufficient cues and entropy-routed on-policy distillation, with an analysis of when routing fails; and (iii) a systematic study across four models, with same-budget baselines, pre-specified ablations, causal interventions, and pretraining-data evidence.

2 Recall, Access, and Endogenous Citation

Let p_θ be the student and θ_0 its frozen initialization. For a question q, the student samples a reasoning trace z followed by an answer a; we write $\hat{y} = (z, a)$ and $\mathrm{ans}(\hat{y}) = a$. An extractor maps $\hat{y}$ to a set $C(\hat{y})$ of decontextualized claims, each a triple $s = (e, r, o)$ of subject, relation, and object.

Definition 1 (Encoding, access, and endogenous citation)**.** Let $\gamma_\theta(x)$ be the model's rate of producing the part of fact f that input x asks for. Following Calderon et al. (2026), f is *encoded*, $f \in \mathcal{E}_\theta$, if $\gamma_\theta(x) > 0.5$ for *some* pretraining-like context x, and *accessible* if $\gamma_\theta(x) > 0.5$ for *every* query x across paraphrases and directions. A trace satisfies *endogenous citation* if $C(\hat{y}) \subseteq \mathcal{E}_\theta$.

The notion is behavioral: it concerns the claims a trace states, not which of them the answer depends on (tested with interventions in §4.5), and possession rather than truth, since a falsehood learned in pretraining satisfies it. ECHO targets fabrications and cannot reliably correct false beliefs the model holds. Since $p_\theta(a \mid q) = \sum_z p_\theta(z \mid q)\, p_\theta(a \mid q, z)$ and the second factor tends to rise with correct related facts in z and fall with fabricated ones, we target the distribution of reasoning traces: we seek a policy that raises the share of endogenously cited traces without losing accuracy, using no gold answers for training questions and no retriever, knowledge base, or external judge during policy optimization.

3 METHOD

3.1 ENDOGENOUS VERIFIER

For a claim $s = (e, r, o)$, the frozen model is probed without the original trace or question (Appendix D): forward probes withhold the object, reverse probes the subject. *Forward QA* asks for the r of e; *forward cloze* completes a neutral, pretraining-like sentence; *reverse QA* asks which entities have o as their r, listing up to five because reverse relations are often one-to-many, and scores a hit if e is listed. Each probe draws 2×4 short samples. With agreement rates u, the semantic entropy H_{sem} of the forward answers (Farquhar et al., 2024), and the log-probability margin Δ_{mg} of o over its strongest competitor,

$$g(s) = \sigma\big(w^\top [\, u_{\text{fwd}},\, u_{\text{cloze}},\, u_{\text{rev}},\, -H_{\text{sem}},\, \Delta_{\text{mg}} \,]\big). \tag{1}$$

The weights w are fit once per model on 1,500 labeled calibration claims and frozen, so g is a calibrated probability of truth, used as a proxy for possession by θ_0. Approximating $\mathcal{E}_\theta$ by $\mathcal{E}_{\theta_0}$ is a working assumption, consistent with the stable encoding rate (§4.5); §4.6 tests g against pretraining exposure. We avoid yes/no probes, which invite acquiescence (AUC 0.63 vs. 0.84).

Thresholds from a target precision. Admitting a fabricated claim plants a false premise the student then learns; rejecting a true claim merely weakens the teacher, so we favor precision. Claim acceptance is set by a target precision $\pi = 0.90$, converted per model to $\tau = \min\{t : \text{Precision}_{\text{calib}}(g \geq t) \geq \pi\}$ (0.80 for 27B, 0.83 for 9B); answer-level thresholds use the same π.

3.2 DIAGNOSING KNOWLEDGE STATES WITHOUT GOLD ANSWERS

ECHO samples $m = 8$ rollouts Y. Because the student rarely says the right answer on tip-of-the-tongue questions, candidates are completed in hindsight by sampling the oracle four times (answer only) from an initial cue set K_q^0 (§3.3), giving $A_q = \text{ans}(Y) \cup \text{ans}(\tilde{Y})$. With $g(q \to a)$ the verifier applied to an answer claim,

$$a^\star = \arg\max_{a \in A_q} g(q \to a), \qquad f(a^\star) = \tfrac{1}{m} \textstyle\sum_j \mathbf{1}[\text{ans}(\hat{y}_j) = a^\star]. \tag{2}$$

A question is AC if $g(q \to a^\star) \geq \tau_{\text{ans}}$ and $f(a^\star) \geq \eta$; TT if $g(q \to a^\star) \geq \tau_{\text{TT}}$ and $f(a^\star) < \eta$; and NE otherwise, with $\eta = 0.5$ as in the labeled protocol of Calderon et al. (2026). Because correct answers are rarer at low f, τ_{TT} converts the same π on the low-frequency part of the calibration set. *Strict* NE questions, with $\max_a g(q \to a) < \tau_{\text{NE}}$ chosen to admit at most $\delta = 5\%$ answerable questions, serve only as abstention targets. Triage is recomputed online at every step.

3.3 THE HINDSIGHT ORACLE AND RECALL-SUFFICIENT CUES

A claim *leaks* if it contains any candidate in A_q, including up to five aliases generated by the frozen model, or if the frozen model judges that it states the answer; the rule uses candidates only, never gold answers. Self-verified claims ($g(s) \geq \tau$) that do not leak the student's candidates form K_q^0; after candidate completion, those that do not leak any candidate in A_q form K_q (Algorithm 1).

Proposition 1 (Information sources)**.** Let the verifier parameters ϕ (the weights w and all thresholds) be fit on the calibration labels before training and then frozen. For every training question q, the oracle's context is a function of $(q, \theta, \theta_0, \phi, \xi)$ alone, where ξ is the sampling randomness; it involves no gold answer for q, no retrieved document, and no external judgment.

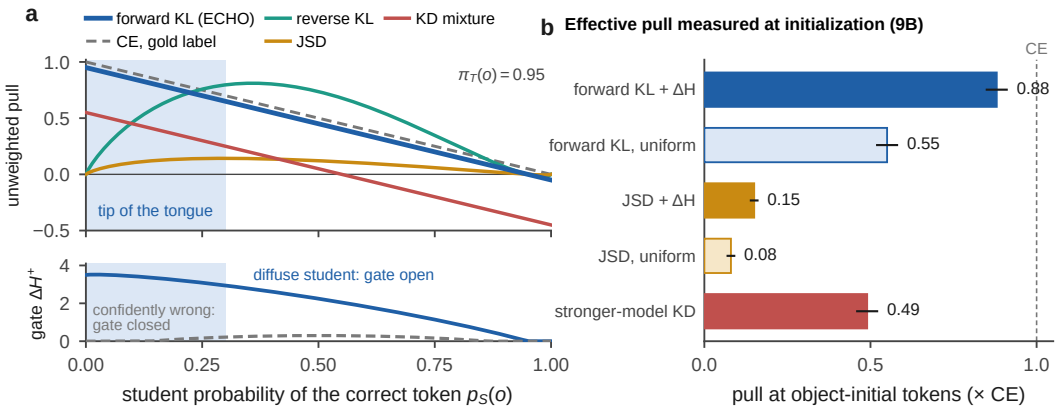

Figure 2: Why forward KL, and where the gate opens. **(a)** Top: unweighted pull on o at $\pi_T(o) = 0.95$: forward KL parallels gold-label cross-entropy (CE), while reverse KL and JSD vanish as $p_S(o) \to 0$ (KD mixture: $p_\Phi(o) = 0.1$, $\alpha = 0.5$). Bottom: the gate $\max(0, H_S - H_T)$ opens when the student spreads its mass over many (here 49) alternatives and can close when the student is confident in one wrong token. **(b)** Measured pull at object-initial tokens relative to CE (9B, initialization, gated-off tokens included; Appendix H).

The oracle differs from the student only in its parameters (θ_0 rather than θ) and in seeing pooled, filtered cues, yet it is stronger: a cue recalled in any rollout becomes available to all, and filtering lowers the share of fabricated premises, which tend to lower the correct answer's probability.

Definition 2 (Recall-sufficient cues)**.** A set $S^\star \subseteq K_q$ is recall-sufficient if $\log p_{\theta_0}(a^\star \mid q, S^\star) \geq \log p_{\theta_0}(a^\star \mid q, K_q) - \epsilon$ and no single cue can be removed without violating this condition.

A greedy search visits the cues in ascending order of influence, deletes each one whose removal preserves the condition, and repeats until no cue can be deleted, yielding a 1-minimal fixed point; $\epsilon = 0.1$ nats allows the oracle to lose at most about 10% of its probability on $a^\star$. The definition preserves relative log-probability only and does not make $a^\star$ correct. The teacher is $\pi_T = p_{\theta_0}(\cdot \mid q, S^\star)$ for AC and TT questions; questions with $S^\star = \varnothing$ are not distilled. Empirically, compression also reduces false premises (the verifier admits about one false claim in ten) and concentrates supervision (§4.3).

3.4 Learning from the Oracle

Entropy-routed distillation. Distillation is on-policy: on the student's own rollouts, already sampled for triage, the student is pulled toward the oracle at every response token, thinking included:

$$\mathcal{L}_{\text{dist}} = \frac{1}{N} \sum_{q \in \mathcal{D}} \omega_{\text{state}(q)} \sum_{\hat{y}} \sum_{n} \tilde{w}_n \, \mathrm{KL}\big(\pi_T(\cdot \mid q, c_q, \hat{y}_{<n}) \,\big\|\, p_\theta(\cdot \mid q, \hat{y}_{<n})\big), \qquad (3)$$
$$w_n = \mathrm{sg}\left[\max(0,\, H_{S,n} - H_{T,n})\right],$$

where $\mathcal{D}$ is the set of questions with a teacher context c_q ($S^\star$, or an abstention note in ECHO-A), $H_{S,n}$ and $H_{T,n}$ are the student and teacher entropies, $\tilde{w}_n = w_n / \bar{w}$ normalizes by the mean over the N tokens of $\mathcal{D}$ in the batch so that traces the student already masters are down-weighted, sg stops gradients, and the KL is taken over the full vocabulary at temperature 1. We set $\omega_{\text{TT}} = 1$ and $\omega_{\text{AC}} = \beta = 0.1$; NE questions are not distilled, except strict NE questions in ECHO-A (below). The divergence is not truncated; its logit gradient $\omega \tilde{w}_n (p_\theta - \pi_T)/N$ has ℓ_1 norm at most $2\omega\tilde{w}_n/N$.

Why forward KL. The unweighted pull on the correct token o is $\pi_T(o) - p_S(o)$. When o appears among the cues, the oracle can copy it, $\pi_T(o) \approx 1$, and the pull approaches the cross-entropy value $1 - p_S(o)$, largest when the student is least likely to produce o (Figure 2a). For JSD and reverse KL the pull vanishes as $p_S(o) \to 0$ (at $p_S(o) = 0.01$: 0.940 for forward KL, 0.075 for reverse KL, 0.022 for JSD). Mixing a larger teacher Φ into cross-entropy with weight α lowers the pull by $\alpha(1 - p_\Phi(o))$, large when Φ is unsure. Equation 3 is the on-policy forward KL of GKD (Agarwal et al., 2024). OPSD and SDPO also distill on-policy, OPSD with a clipped generalized JSD and the current model as teacher (Zhao et al., 2026), SDPO with a reverse KL and a moving-average teacher (Hübotter et al., 2026); ECHO uses forward KL, a frozen teacher θ_0, and ΔH routing, not clipping.

Entropy routing and its failure modes. Forward KL makes the student cover the teacher's whole distribution, including its uncertainty; routing by $\Delta H_n = H_{S,n} - H_{T,n}$ weights only positions where the teacher is more certain.

Hypothesis 1 (Entropy inversion)**.** On TT questions, the oracle is more certain than the student ($\Delta H_n > 0$) at positions that generate a cue or $a^\star$, while $\Delta H_n \approx 0$ at function words.

It can be checked before training (§4.3). It fails when the student is *confidently wrong*, so that H_S is low and the gate closes (Figure 2a, bottom), and when the oracle is *misled* by a false cue or a wrong $a^\star$, which routing amplifies. The entropy gap allocates weight by relative confidence; verification filters the cues that define the teacher.

Influence-gated grounding reward. To reward grounded traces on all questions, the claims of a trace with answer $\hat{a}$ are deduplicated by bidirectional entailment, answer claims are removed by the leak rule, and each remaining claim receives a leave-one-out influence $I(s) = \log p_{\theta_0}(\hat{a} \mid q, C(\hat{y})) - \log p_{\theta_0}(\hat{a} \mid q, C(\hat{y}) \setminus \{s\})$. Then

$$r_{\text{ground}}(\hat{y}) = \frac{\sum_{s \in U} I(s)\, v(s)}{\sum_{s \in U} I(s) + I_0} - \nu \frac{\min(R, 5)}{5}, \qquad U = \{s \in C(\hat{y}) : I(s) \geq \iota_0\}, \tag{4}$$

where $v(s) = +1$ if $g(s) \geq \tau$ and -1 otherwise, and R counts claims that fail self-verification, including the answer claim $q \to \hat{a}$ scored at τ_{ans} ($\iota_0 = 0.05$ nats, $I_0 = 0.1$, $\nu = 0.5$). If adding a claim leaves the other influences unchanged, a passing claim with negligible influence changes nothing, a failing claim lowers the reward by at least $\nu/5$ while $R < 5$, and only influential self-verified claims can raise it; deduplication removes gains from detected repetition (Appendix E). Because the premise holds only approximately, we test padding directly (§4.4).

Full objective. $\mathcal{L} = \mathcal{L}_{\text{dist}} + \lambda \mathcal{L}_{\text{PG}}(A)$ with $\lambda = 0.5$ and GRPO advantages (Shao et al., 2024) computed on the same rollouts. Variant ECHO-A rewards abstaining with $+1$ on strict NE and -1 on AC or TT (weight $\rho = 0.5$), and distills strict NE questions (weight 1) toward the frozen model told that no candidate passed verification. The verifier, cue search, and teacher run only in training; probes are short, deduplicated, and batched, and a run costs about 1.5× an RLVR run (Appendix C).

4 Experiments

4.1 Setup

Models and data. We train Qwen3.5-27B, our main model, and Qwen3.5-9B (Qwen Team, 2026) with thinking enabled, and test generality on Qwen3.5-35B-A3B (MoE) and OLMo-3-7B-Think (Team Olmo, 2025), whose pretraining corpus is public. The training pool holds 90,000 entity questions from 18 EntityQuestions relations (Sciavolino et al., 2021) and Wikidata, used as *question text only*. A 10,000-question development set with gold answers selects learning rates, and a calibration set of 2,000 questions and 1,500 labeled claims per model fits the verifier before training. All pools are split by subject entity and share none with any evaluation set; six EntityQuestions relations are held out entirely. RLVR uses gold answers for all 90,000 questions.

Evaluation and baselines. We measure accuracy (pass@1) on SimpleQA-Verified (Wei et al., 2024a; Haas et al., 2025), the held-out relations, and PopQA (Mallen et al., 2023), whose bottom popularity quintile forms the long tail. The primary trust metric, *trace factual precision* (TFP), is the share of claims in two traces per question judged true by an external search-augmented auditor (Gekhman et al., 2026); evaluation never uses the training verifier. We also report FActScore (Min et al., 2023), WikiProfile knowledge profiles, and reasoning retention on competition math, GPQA-Diamond (Rein et al., 2024), MMLU-Redux (Gema et al., 2025), and IFEval (Zhou et al., 2023). Baselines are label-free self-consistency RL (Zuo et al., 2025) and confidence RL; RFT/STaR (Zelikman et al., 2022), RLVR and a compute-matched RLVR, OPSD with gold solutions (Zhao et al., 2026), and VAKE (Ying et al., 2026), all using gold answers; a test-time filter that returns the least hallucinated of eight traces under external audit; and three baselines with exactly ECHO's labeled budget. All training methods share initialization, questions, rollout budget (256 questions × 8 rollouts for 350 steps), and decoding, except that compute-matched runs train longer and one same-budget

Table 1: Main results (%). SQA: SimpleQA-Verified; EQ: held-out EntityQuestions relations; Tail: bottom PopQA popularity quintile; Avg.: mean of SQA, EQ, PopQA; TFP: trace factual precision (9B: on SQA); FActS.: FActScore. Sup.: supervision (cal.: one-off verifier calibration; gold: gold answers; audit: test-time audit). **Bold**: column best, excluding test-time filter and ECHO + label reward. Differences use unrounded values.

		Qwen3.5-27B								Qwen3.5-9B			
		Accuracy					TFP			Accuracy			
Method	Sup.	SQA	EQ	PopQA	Tail	Avg.	SQA	EQ	FActS.	SQA	Avg.	TFP	FActS.
Base	–	30.0	41.2	36.0	14.2	35.7	81.4	84.3	71.8	12.0	23.5	74.8	63.5
Self-consistency RL	none	30.8	42.0	36.5	14.1	36.4	80.6	83.7	71.3	12.4	23.9	74.1	63.0
Confidence RL	none	30.3	41.5	36.2	13.8	36.0	79.9	83.1	70.6	12.1	23.6	73.4	62.4
Test-time filter	audit	32.5	43.9	37.8	15.3	38.1	90.2	91.3	–	13.6	25.0	85.1	–
RFT / STaR	gold	31.6	43.4	37.4	14.9	37.5	82.0	84.8	72.0	12.9	24.5	75.2	63.7
RLVR	gold	33.4	45.9	38.9	15.8	39.4	82.3	85.0	72.4	14.2	26.2	75.4	64.0
RLVR (compute-matched)	gold	**33.6**	46.2	39.1	15.9	**39.6**	82.1	84.9	72.3	**14.5**	**26.5**	75.3	63.9
OPSD-gold	gold	33.0	45.1	38.4	15.5	38.8	81.0	84.0	71.2	13.9	25.8	74.2	63.1
VAKE	gold	32.4	44.8	38.2	15.6	38.5	84.6	86.9	73.9	13.5	25.4	77.8	65.6
ECHO	cal.	33.1	**46.3**	**39.4**	**17.6**	**39.6**	**88.7**	**90.1**	**76.5**	13.8	26.3	**81.6**	**68.9**
ECHO + label reward	gold	34.6	47.4	40.2	18.1	40.7	88.9	90.3	76.8	14.9	27.2	81.9	69.2
ECHO − RLVR		−0.3	+0.4	+0.5	+1.8	+0.2	+6.4	+5.1	+4.1	−0.4	+0.0	+6.2	+4.9

baseline uses only the calibration questions. Accuracy against RLVR is tested for non-inferiority with margins of 2.0 points on SimpleQA-Verified and 1.0 point on the out-of-distribution macro-average, using 95% paired cluster-bootstrap intervals (hierarchical over three seeds where available); primary tests are pre-specified and Holm-corrected (Holm, 1979). Differences discussed below are significant unless noted; all intervals and tests are in Appendices A and I–N.

4.2 Main results

Same labels, better use. Baselines that spend exactly ECHO's labeled budget on rewarding answers (RLVR on the calibration questions, self-consistency RL with a calibrated gate, and the calibrated verifier as answer reward) reach at most 31.7% on SimpleQA-Verified for 27B (31.8% when compute-matched) and 13.0% for 9B. ECHO exceeds the strongest by 1.4 and 0.8 points, and the compute-matched one by 1.3: the same labels are worth more for filtering cues than for rewarding answers. A fully label-free variant, ECHO-LF (equal-weight verifier, quantile thresholds, label-free learning-rate selection), reaches 32.8% with TFP 87.6.

Accuracy without gold answers. Against RLVR trained with gold answers for every question (Table 1), ECHO is non-inferior: the difference is $-0.3\,[-1.6, +1.0]$ on SimpleQA-Verified and $+0.2\,[-0.3, +0.7]$ on the macro-average for 27B, and $-0.4\,[-1.2, +0.4]$ and $+0.0\,[-0.5, +0.5]$ for 9B, all within the pre-specified margins; ECHO-LF is non-inferior on the macro-average. ECHO beats the strongest label-free baseline by 2.3 and 1.4 points, while majority and confidence rewards barely move accuracy and lower TFP, reinforcing stable errors. ECHO leads RLVR on the long tail (+1.8), whereas RLVR leads by 1.9 points on the 18 training relations it saw with gold answers. Differences in pass@16 are not significant, so we do not claim a wider recall boundary; point estimates on all four reasoning benchmarks move by less than one point.

Trustworthy traces. TFP rises over RLVR by 6.4 points on SimpleQA-Verified and by 5.1 on held-out relations, and the rate of hallucinated claims falls by 39.2% relative to Base. The gain transfers to long-form biographies (FActScore $+4.1$), and it is not bought by saying less: ECHO makes 5.9 claims per trace, Base 6.0. RLVR hardly moves TFP (81.4 to 82.3), and neither compute-matched RLVR (82.1) nor OPSD with gold solutions (81.0) reproduces the gain. Among baselines, only the test-time filter, which needs eight samples and an external audit, reaches a higher TFP.

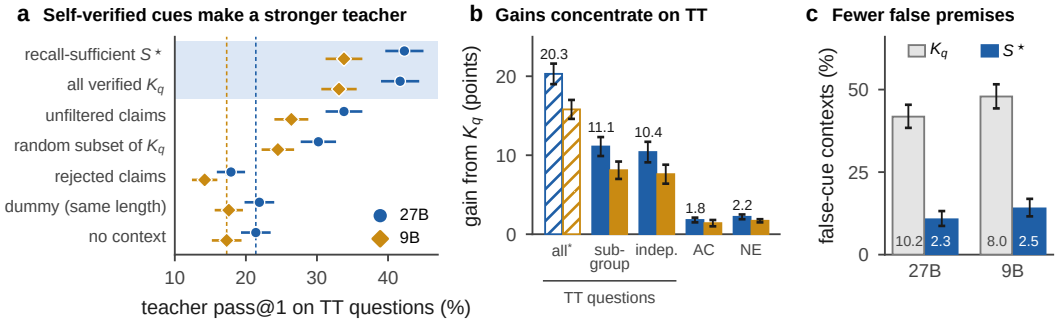

Figure 3: The hindsight oracle at initialization (development TT questions; 16 samples per question and context). **(a)** Oracle pass@1 by context; dashed lines mark no context. **(b)** Gain from K_q. Hatched bars pool all TT questions and carry selection bias; "subgroup" keeps questions whose $a^\star$ came from the student, and "indep." uses a separate pipeline; AC and NE are controls. **(c)** Contexts with at least one false claim; bar labels give mean cue counts.

4.3 The hindsight oracle before training

Self-verification and triage. On held-out calibration claims the verifier reaches AUC 0.84 on 27B (0.78–0.84 across four models) and 89.6% precision at the deployed threshold for a 90% target, while retaining 56% of true claims. Against labeled question-level profiles (Calderon et al., 2026), label-free triage attains Cohen's $\kappa = 0.63$ (27B) and 0.62 (9B), about three quarters of the test–retest ceiling. TT covers only 5.6% of questions but is enriched 4.5× in recall failures, and its target $a^\star$ is correct 92.5% of the time (Appendix G).

A stronger, cleaner teacher. Self-verified cues double the oracle's accuracy on development TT questions (Figure 3a). As the oracle also helped select these questions, we measure the gain where the target came from the student, $+11.1$ points, and in an independent pipeline, $+10.4$ (Figure 3b). The gain is only 2.2 on NE and 1.8 on AC questions, and it is largest on facts that are encoded but not recalled. Verification matters: rejected claims push the oracle below no context (-3.5), and unfiltered claims fall 7.9 points short of K_q. Recall-sufficient cues keep the oracle as strong as K_q (non-inferior; $+0.6$) with 2.3 instead of 10.2 cues, beat a random subset of equal size by 12.1 points, and cut contexts containing a false premise from 41.8% to 10.7% (Figure 3c).

Entropy inversion, with caveats. On verified fact tokens the oracle is more certain than the student by $+0.60$ nats (27B; $+0.57$ for 9B), whereas a standard 27B→9B distillation teacher is not (-0.04). On 9B, routed forward KL therefore exerts 0.88× the cross-entropy pull at object-initial tokens, against 0.49× for stronger-model distillation and 0.15× for JSD (Figure 2b). The inversion holds on only 66.8% of fact tokens, and it also holds on false cues ($+0.40$): certainty is not correctness.

4.4 Ablations and controls

Each ablation removes one component and pre-specifies the metric it should hurt (Figure 4a). All eleven targets deteriorate with intervals excluding zero, and nine of eleven survive Holm correction on 27B (all eleven on 9B). Removing the oracle costs 2.2 points of accuracy; unverified cues cost 4.8 points of TFP. Without state routing, NE questions are distilled as well, and on them hallucinated claims per trace rise from 1.21 to 1.62, above Base (1.49). Uniform token weights and JSD each cost about 1.5 points of accuracy. Part of the JSD gap is a matter of scale: raising the JSD weight until its average pull matches forward KL halves the accuracy gap (-0.8, not significant) but not the tail gap (-1.4), and routing by teacher entropy alone, as in Switch Distillation (He et al., 2026), is likewise weaker on the tail (-1.3; Appendix J). An unweighted reward inflates raw TFP to 89.6 by padding traces with 59% more claims. The influence gate prevents this: injecting 100 true, self-verified, irrelevant claims into fixed traces lifts an ungated influence-weighted ratio reward from 0.31 to 0.63 but leaves ours at 0.24 versus 0.23.

Figure 4b separates cues from answers. An oracle given only $a^\star$ is on par with ECHO in accuracy (-0.5, not significant) but loses 4.9 points of TFP and 1.7 of tail accuracy, and its students state fewer

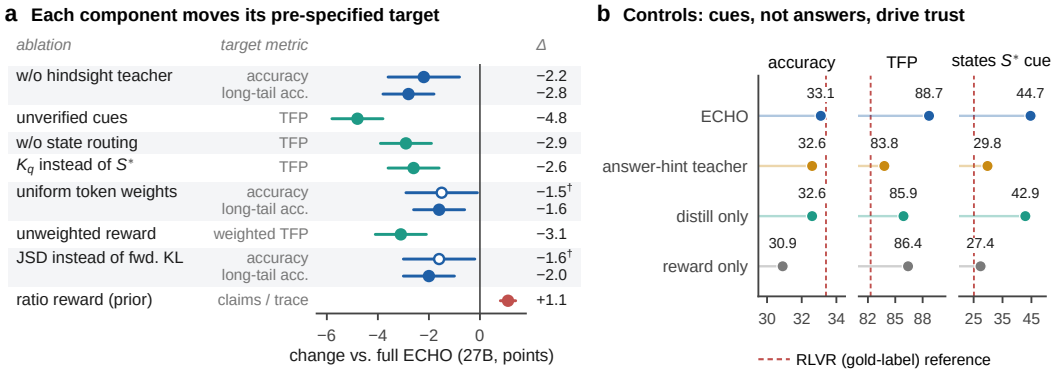

Figure 4: Ablations and controls (27B). **(a)** Change of each ablation's pre-specified target metric relative to full ECHO, with 95% intervals; open markers with † lose significance after Holm correction; for padding, higher is worse. **(b)** Controls (three seeds); the dashed line marks RLVR. "States $S^\star$ cue" is the share of TT traces that state a cue from the $S^\star$ built before training (%).

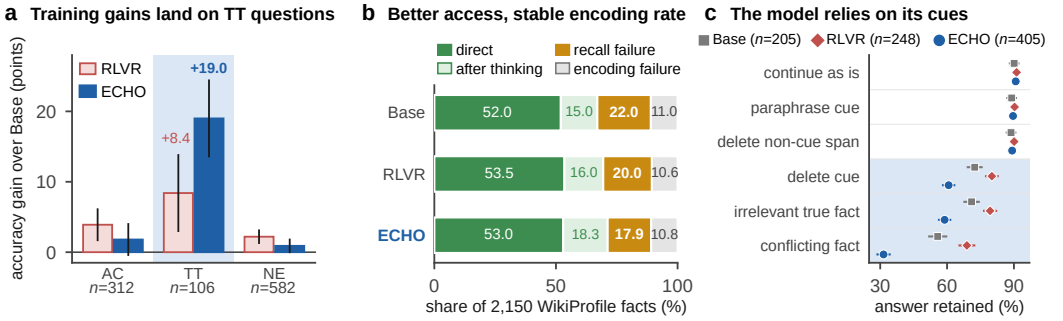

Figure 5: Where the gains come from (27B). **(a)** SimpleQA-Verified gain over Base, grouped by the label-free state before training; 95% intervals. **(b)** Knowledge profiles of 2,150 WikiProfile facts before (Base) and after training. **(c)** Answer retention after an intervention at the first recalled cue of a correct trace, over each model's cue-bearing questions (n in the legend); 95% intervals; shaded rows alter the cue itself.

cues (29.8% vs. 44.7% of traces). Distillation alone keeps most of the accuracy but loses 2.8 points of TFP; the reward alone keeps TFP well above RLVR but loses 2.2 points of accuracy. In these controls, TFP gains come from the cues, not the answer; the two terms are complementary.

4.5 ECHO REPAIRS ACCESS

Gains land where recall fails. Grouping SimpleQA-Verified by the state before training (Figure 5a), ECHO gains 19.0 points on TT questions against 8.4 for RLVR, while RLVR gains more on AC and NE questions. On WikiProfile (Figure 5b), 22.0% of recall failures become answerable after ECHO against 15.0% after RLVR (49.3% among those triaged TT), the measured encoding rate stays at 88.4–88.6%, forgetting is lower (1.4% vs. 2.2%), and ECHO gains more on reverse than forward questions (+4.3 vs. +2.7). The pattern fits improved access rather than new storage.

The model relies on its cues. On the 431 development TT questions that Base, RLVR, and ECHO all answer correctly, a correct trace states a cue of the initial $S^\star$ before the answer for 62.9% of questions under ECHO, 44.6% under RLVR. In each model's cue-bearing correct traces, we intervene at the cue and let the model continue (Figure 5c). Paraphrasing it or deleting an equally long non-cue span barely matters, but deleting it lowers answer retention to 60.7% for ECHO against 80.1% for RLVR, and a conflicting fact to 31.5% against 68.9%. Because each model chooses which cue it states, we also match cues: on the 138 questions where ECHO and RLVR state the same cue, deletion costs ECHO 27.4 points of retention against 11.2 for RLVR, for both recall prompts and deductive premises (Appendix J). Reliance cuts both ways: when the target is wrong (8.2% of development TT

questions), ECHO does not detectably change the correct answer's rate but raises the wrong target's rate (12.8% to 31.6%); such targets make up 8.2–9.1% of TT questions in training.

4.6 Abstention, scale, and pretraining data

ECHO-A reduces wrong answers by 28.6% relative to ECHO at a cost of 0.7 points of accuracy (27B). It abstains 3.4× as often on encoding failures as on recall failures (58.7% vs. 17.3%), whereas RLVR with a gold-label ternary reward abstains on 36.5% of recall failures (1.8×). On Qwen3.5-35B-A3B, ECHO is non-inferior to RLVR (-0.5) while raising TFP by 6.0 points. On OLMo-3-7B-Think, after correcting a co-occurrence proxy with a stratified relation-level audit, 77.2% of self-verified claims are supported by the public pretraining corpus against 30.1% of rejected ones, and ECHO's gains concentrate on facts the corpus contains but rarely ($+4.6$ for 1–9 co-occurrences vs. $+2.8$ for RLVR). At deployment ECHO adds no module; its outputs are 4.7% longer than RLVR's.

5 Related work

Parametric knowledge and recall. Language models store many facts (Petroni et al., 2019) but retrieve them unevenly across phrasings, directions, and popularity (Allen-Zhu & Li, 2024; Berglund et al., 2024; Kandpal et al., 2023; Calderon et al., 2026), and hallucinated facts derail reasoning (Gekhman et al., 2026). RL on QA rewards improves access (Ma & Hewitt, 2026), VAKE inserts knowledge-graph triples (Ying et al., 2026), and CitePretrain cites pretraining documents (Huang et al., 2025b). ECHO uses no graphs or corpora, and no gold answers for training questions.

Distillation, privileged information, and hindsight. Distillation transfers a larger teacher's distribution (Hinton et al., 2015; Kim & Rush, 2016; Gu et al., 2024), on-policy variants compare divergences on the student's own samples (Agarwal et al., 2024), and context distillation makes the same model with extra context the teacher (Snell et al., 2022; Vapnik & Izmailov, 2015). OPSD conditions the teacher on gold solutions (Zhao et al., 2026), SDPO on environment or reviewer feedback (Hübotter et al., 2026). Switch Distillation distills only where the teacher's entropy is low (He et al., 2026); ECHO routes by the teacher–student gap and changes what the teacher sees. Its privileged context is produced and checked by the model itself, extending hindsight relabeling (Andrychowicz et al., 2017; Zhang et al., 2023; Liu et al., 2024) to the teacher's knowledge.

Self-verification, factuality, and abstention. Sampling consistency and semantic entropy detect hallucinations without external sources (Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024); models partly know what they know (Kadavath et al., 2022) and can self-verify at inference (Dhuliawala et al., 2024). Factuality tuning with reference-free confidence (Tian et al., 2024) and label-free RL (Zuo et al., 2025) target answers; ECHO verifies each parsable claim of a trace and turns verified claims into a teacher. The abstention training of Zhang et al. (2024) uses gold answers.

6 Conclusion

In this paper, we introduced ECHO, which teaches language models to reason over facts they can self-verify, without gold answers or an external verifier in the training loop. Because verification is easier than generation and hindsight requires no labels, a frozen model conditioned on its verified cues becomes a hindsight oracle stronger than the student, which ECHO distills on-policy. ECHO is non-inferior to gold-answer RLVR in accuracy while making reasoning traces markedly more factual, and our analyses suggest that the gains come from better access to what the model already knows. More broadly, self-verification is an underused source of supervision when labels are scarce.

Limitations and future work. ECHO still needs labeled data to calibrate its verifier and select hyperparameters, and the verifier is a proxy for possession, not a certificate of truth: confidently held errors can pass self-verification, and a wrong target is reinforced. Triage reaches a quarter of recall failures, gains over Base are modest, and self-verification adds training-time computation. Future work could extend hindsight oracles to multi-hop reasoning and long-form training and pair them with retrieval for facts the model lacks.

AI USE STATEMENT

We used generative AI tools to assist with drafting and editing the text of this manuscript, with writing plotting scripts for the figures, and with checking the consistency of reported numbers across sections. Large language models are also objects of study and components of the method: the models under training verify and extract their own claims, and an external language model extracts and audits claims for evaluation, as described in §4.1 and Appendix A. We have reviewed all AI-assisted content and take full responsibility for the final content of this work, including text, claims, and figures.

ETHICS STATEMENT

This work aims to reduce hallucinated content in model outputs by training models to rest their claims on knowledge they can corroborate. Our method does not introduce new capabilities for harm and uses only publicly available benchmarks, with training questions built from EntityQuestions and Wikidata; no human-subjects research was conducted, and human annotation was limited to small audits of model outputs and labels. Two risks deserve mention. First, self-verification provides evidence that a claim is held by the model, not proof that it is true, so a model trained with ECHO can state encoded misconceptions more fluently; our corpus analysis estimates this share. Second, because the trained model relies more on the cues it states, a false cue misleads it more easily; the verifier is therefore an essential part of the method rather than an optional filter. We report both effects and recommend external fact-checking in high-stakes settings.

REPRODUCIBILITY STATEMENT

Section 3 and Algorithm 1 specify the method, and Appendix B lists every hyperparameter with its source, including all thresholds derived from the target precision. Appendix A documents data construction, subject-entity isolation, the use of every label, baselines, and the training configuration, including the policy-gradient settings; Appendix D gives all probe templates and the full text of every prompt; Appendix N describes the statistical procedures and lists all primary tests with their p-values. Appendices G–M report full tables, per-seed results, and sensitivity analyses. The per-question triage scores behind the recall-versus-encoding AUROC are provided as supplementary material. We will release code, the question-level data splits, and evaluation outputs.

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.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In *International Conference on Machine Learning*, 2024.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, 2017.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". In *International Conference on Learning Representations*, 2024.

Roger Brown and David McNeill. The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 5(4):325–337, 1966.

Nitay Calderon, Eyal Ben-David, Zorik Gekhman, Eran Ofek, and Gal Yona. Empty shelves or lost keys? Recall is the bottleneck for parametric factuality. *arXiv preprint arXiv:2602.14080*, 2026.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

Zorik Gekhman, Roee Aharoni, Eran Ofek, Mor Geva, Roi Reichart, and Jonathan Herzig. Thinking to recall: How reasoning unlocks parametric knowledge in LLMs. *arXiv preprint arXiv:2603.09906*, 2026.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. Are we done with MMLU? In *Proceedings of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2025.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Lukas Haas, Gal Yona, Giovanni D'Antonio, Sasha Goldshtein, and Dipanjan Das. SimpleQA verified: A reliable factuality benchmark to measure parametric knowledge. *arXiv preprint arXiv:2509.07968*, 2025.

Jacqueline He, Howard Yen, Shuyue Stella Li, Margaret Li, Hanqing Zeng, Yinglong Xia, Benyu Zhang, Zhuokai Zhao, Qiang Zhang, Pang Wei Koh, Luke Zettlemoyer, and Wen-tau Yih. Knowledge distillation during mid-training favors reasoning over factual recall, 2026. URL https://arxiv.org/abs/2609.01532.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025a. doi: 10.1145/3703155.

Yukun Huang, Sanxing Chen, Jian Pei, Manzil Zaheer, and Bhuwan Dhingra. Cite pretrain: Retrieval-free knowledge attribution for large language models. *arXiv preprint arXiv:2506.17585*, 2025b.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. doi: 10.1145/3571730.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2017.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, 2023.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2019.

Xiang Lisa Li, Vaishnavi Shrivastava, Siyan Li, Tatsunori Hashimoto, and Percy Liang. Benchmarking and improving generator-validator consistency of language models. In *International Conference on Learning Representations*, 2024.

Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. Chain of hindsight aligns language models with feedback. In *International Conference on Learning Representations*, 2024.

Melody Ma and John Hewitt. Improving parametric knowledge access in reasoning language models. *arXiv preprint arXiv:2602.22193*, 2026.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2023.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2023.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2023.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, 2019.

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

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. In *Conference on Language Modeling*, 2024.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2021.

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

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context. *arXiv preprint arXiv:2209.15189*, 2022.

Team Olmo. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. Fine-tuning language models for factuality. In *International Conference on Learning Representations*, 2024.

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16:2023–2049, 2015.

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

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024a.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models. In *Advances in Neural Information Processing Systems*, 2024b.

Zuocheng Ying, Yang Yang, Yumou Wu, Chuanbo Zhu, Jiarui Wang, Ziqi Wu, Jingming Cai, Junqing Yu, and Zikai Song. From storage to access: Verifiable activation of parametric knowledge in LLMs via explicit priming and implicit reasoning. *arXiv preprint arXiv:2608.18581*, 2026.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, 2022.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-Tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2024.

Tianjun Zhang, Fangchen Liu, Justin Wong, Pieter Abbeel, and Joseph E. Gonzalez. The wisdom of hindsight makes language models better instruction followers. In *International Conference on Machine Learning*, 2023.

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

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. TTRL: Test-time reinforcement learning. In *Advances in Neural Information Processing Systems*, 2025.

APPENDIX

Contents. A Experimental setup in detail B Hyperparameters C Algorithm and training cost D Probe templates and prompts E Structure of the grounding reward F Preliminary measurements G Verifier and triage H The oracle at initialization I Additional main results J Ablations, controls, and mechanism K Abstention L Scale, pretraining data, and long form M Sensitivity and seed variance N Statistical methods and audits

A EXPERIMENTAL SETUP IN DETAIL

Models and decoding. Qwen3.5-27B (dense, main model) and Qwen3.5-9B (dense) are trained and evaluated with thinking enabled; Qwen3.5-35B-A3B (mixture of experts, about 3B active parameters per token) and OLMo-3-7B-Think are used for generalization. Thinking mode uses $T = 0.6$, top-$p = 0.95$, top-$k = 20$; non-thinking mode uses $T = 0.7$, top-$p = 0.8$, top-$k = 20$. The thinking budget is 8,192 tokens for training and factuality evaluation and 32,768 tokens for reasoning retention. Labels under the protocol of Calderon et al. (2026) use that protocol's settings ($T = 1$, eight samples per question, answer only).

Data construction and isolation. The training pool, development set, and calibration set come from the same source: 18 relations of the EntityQuestions training split and entity questions generated automatically from Wikidata. Of 118,420 candidate questions, we remove 9,806 that share a subject entity with any evaluation set, leaving 108,614; removing near-duplicate question texts within the pool (MinHash Jaccard ≥ 0.8) removes a further 4,233, leaving 104,381. These are split by subject entity into 90,000 training, 10,000 development, and 2,000 calibration questions; 2,381 questions straddling group boundaries are discarded. Evaluation sets without subject structure (NQ-Open, TriviaQA) are entity-linked first and deduplicated by the same rule. A separate script confirms zero shared subject entities and zero near-duplicate questions between any two pools and between any pool and any evaluation set. The six remaining EntityQuestions relations were chosen at random before the split and never enter training, development, or calibration. In SimpleQA-Verified, 27 questions (2.7%) have an answer entity that appears as a subject in the training pool; the pool contains no answers, and excluding these questions leaves ECHO $-$ RLVR at -0.3 [-1.6, $+1.0$].

Table 2: Use of every labeled resource (per model). Evaluation sets are never used for fitting or selection.

Use	Data and labels	Enters the training loop?
Verifier weights w, claim threshold τ, evaluation verifier g_{eval}	Fitting half of the calibration claims: 1,500 truth labels	No; fit once and frozen
Answer-level weights; τ_{ans}, τ_{TT}, τ_{NE}	Gold answers of the 2,000 calibration questions	No; converted once and frozen
Learning rate; abstention weight ρ	Gold answers of the 10,000 development questions	No; hyperparameter selection only
Checkpoint	None; the final step is always used (350; 525 or 510 for compute-matched runs)	—
Training questions	90,000 questions, text only	Yes, without answers
External verification during policy optimization	None	—

Calibration claims. For each model, θ_0 generates one thinking trace on each of the 2,000 calibration questions; answer claims and restatements of the question are removed and at most two claims are kept per question. The 1,500 claims from questions 1–1,000 are used for fitting and the 1,500 claims from questions 1,001–2,000 only for evaluating the verifier, so the two halves share no question and no (subject, relation, object) triple. Before sampling, claims whose (subject, relation) matches any evaluation question are removed (0.6%). The bottom popularity quintile is oversampled to 40% so that thresholds are not set by easy, popular facts; deployed precision on held-out claims (Table 12) checks the transfer. Truth labels come from the external search-augmented audit; unverifiable claims (11.0% of those sampled) are dropped, and a human check of 200 claims agrees on 95.0% [91.0, 97.3]. For OLMo, the same claims also receive corpus-support labels.

Table 3: Evaluation sets.

Dataset	Size	Samples	Role
SimpleQA-Verified (Haas et al., 2025)	1,000	32	primary out-of-distribution set; pass@k; three-way grading; TFP
EntityQuestions, 6 held-out relations	5,512	4	out-of-distribution relations; 1,002 audited for TFP
EntityQuestions, 18 training relations	16,563	4	in-distribution relations
PopQA (long tail)	14,267 (2,853)	4	long-tail facts (bottom popularity quintile)
NQ-Open (Kwiatkowski et al., 2019; Lee et al., 2019) / TriviaQA (Joshi et al., 2017)	3,610 / 11,313	4	high-frequency knowledge (closed book)
WikiProfile	2,150 facts	8	knowledge profiles and migration
FActScore biographies	500	1	long-form factuality
AIME 2025 + HMMT Feb. 2025	60	32	reasoning retention
GPQA-Diamond / MMLU-Redux / IFEval	198 / 5,700 / 541	8 / 1 / 1	reasoning, knowledge, instruction following

Metrics. With claims $C(\hat{y})$ extracted from two traces per question by Gemini-3.5-Flash, and $C_{\mathrm{dec}}(\hat{y}) \subseteq C(\hat{y})$ the claims that the audit labels true or false,

$$\mathrm{TFP} = \frac{\sum_{\hat{y}} \sum_{s \in C_{\mathrm{dec}}(\hat{y})} \mathbf{1}[\mathrm{audit}(s) = \mathrm{true}]}{\sum_{\hat{y}} |C_{\mathrm{dec}}(\hat{y})|}, \qquad \mathrm{PP} = \frac{\sum_{\hat{y}} \sum_{s \in C(\hat{y})} \mathbf{1}[g_{\mathrm{eval}}(s) \geq \tau_{\mathrm{eval}}]}{\sum_{\hat{y}} |C(\hat{y})|}. \quad (5)$$

The audit follows the search-augmented hallucination-detection protocol of Gekhman et al. (2026); claims labeled *unknown* or *illegal* are thus excluded from TFP (counting them as false changes the ECHO − RLVR TFP difference from $+6.4$ to $+6.2$), whereas PP uses all extracted claims because g_{eval} scores every claim. Weighted TFP replaces the unit weights of TFP by the gated influence $I(s)\,\mathbf{1}[I(s) \geq \iota_0]$ of the grounding reward, computed with the trace's own answer, and skips traces without such a claim; it guards against padding with correct but idle facts. Provenance precision (PP) uses an evaluation verifier g_{eval} built from the same frozen θ_0 but with two disjoint template sets (Appendix D), fit on the calibration claims and thresholded at the same $\pi = 0.90$ ($\tau_{\mathrm{eval}} = 0.79$ for 27B and 0.82 for 9B). Because it shares the model and calibration labels with the training verifier, PP is auxiliary; TFP is the primary evidence. SimpleQA three-way grading classifies each answer as correct, incorrect, or not attempted; answered precision is correct/(correct+incorrect), and F is the harmonic mean of overall accuracy and answered precision. pass@k uses the unbiased estimator. Reasoning retention reports avg@32 on competition math, avg@8 on GPQA-Diamond, accuracy on MMLU-Redux, and strict prompt-level accuracy on IFEval.

Baselines. All baselines share ECHO's initialization and decoding and, unless stated otherwise below, its question stream and rollout budget. *Self-consistency RL* uses the majority answer of eight samples as a pseudo-label for GRPO, in the style of TTRL (Zuo et al., 2025); *confidence RL* rewards the model's confidence in its answer. The *test-time filter* samples eight traces and returns the one with the fewest hallucinated claims under external audit; it does not train. *RFT/STaR* fine-tunes on correct traces filtered with gold answers. *RLVR* is GRPO with an answer-correctness reward; a compute-matched run trains for 525 steps (1.5× the standard 350) to match ECHO's training cost. *OPSD-gold* is on-policy self-distillation with the gold solution as privileged context (Zhao et al., 2026). *VAKE* follows the two-stage GRPO recipe of Ying et al. (2026), which requires gold answers and a knowledge graph. *ECHO + label reward* adds answer correctness to the reward with weight 0.5. The three *same-budget* baselines use exactly the 2,000 calibration questions and 1,500 calibration claims: RLVR trained on the calibration questions only (350 steps, about 45 epochs); self-consistency RL whose majority answer is used as a pseudo-label only if $g(q \to a) \geq \tau_{\mathrm{ans}}$; and RL with reward $\mathbf{1}[g(q \to \mathrm{ans}(\hat{y})) \geq \tau_{\mathrm{ans}}]$, also run compute-matched (510 steps, since answer-level probes add about 3% per step). *ECHO-LF* removes every label: verifier features are standardized on unlabeled training-pool claims and weighted equally; thresholds are fixed quantile levels on unlabeled training-pool distributions (top 50% of claims for τ, top 50% of candidates for τ_{ans}, top 25% of low-frequency candidates for τ_{TT}, bottom 25% for strict NE); the learning rate is chosen by the mean grounding reward after 50 steps on development-set question text, which selects 1×10^{-6}.

Training. Each step uses 256 questions $\times$ 8 rollouts; runs last 350 steps (89,600 questions, about one epoch), and the final checkpoint is always evaluated. Learning rates are selected on the development set from $\{5 \times 10^{-7}, 10^{-6}, 2 \times 10^{-6}\}$ ($\{5 \times 10^{-6}, 10^{-5}, 2 \times 10^{-5}\}$ for RFT); RL methods use a KL coefficient of 0.001 toward θ_0. The policy-gradient term uses the GRPO surrogate with group-standardized advantages, token-level averaging, and PPO-style ratio clipping at $\epsilon_{\text{clip}} = 0.2$. Each rollout batch is used for a single pass in four mini-batches of 512 rollouts, so the first mini-batch is exactly on-policy and the others are at most three optimizer steps stale; the distillation term is computed on the same mini-batches. All methods use AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, no weight decay), a constant learning rate after 10 warm-up steps, bf16 training, and global gradient-norm clipping at 1.0; ECHO's divergence itself is never truncated. Students, the frozen teacher, and the verifier start from the public checkpoints of the respective models, and the teacher and verifier share the weights θ_0. On 27B, ECHO, RLVR, OPSD-gold, the teacher, token-weight, and divergence ablations, the two additional divergence and routing controls of Appendix J, and the answer-hint and distill-only controls use three seeds; other 27B settings use one. All 9B settings use three seeds.

B HYPERPARAMETERS

Table 4: Hyperparameters and their sources. "Fixed" values were set before any experiment and never changed. Raw thresholds are derived quantities, not hyperparameters.

Hyperparameter	Value	Source
Rollouts per question m	8	Fixed; the eight-sample protocol of Calderon et al. (2026)
Frequency threshold η	0.5 ($f \geq 4/8$ for AC)	Fixed; majority rule of the same protocol
Probe templates $\times$ samples	2×4 per probe	Fixed
Oracle samples for candidate completion	4, answer only	Fixed
Target precision π	0.90	Fixed; shared by all models
Strict-NE miss rate δ	0.05	Fixed
Claim threshold τ	0.80 / 0.83 / 0.81 / 0.82	Derived from π (27B / 9B / 35B-A3B / OLMo)
Answer threshold τ_{ans}	0.78 / 0.81 / 0.79 / 0.80	Derived from π on 2,000 calibration questions
TT threshold τ_{TT}	0.86 / 0.89 / 0.87 / 0.88	Derived from π on the low-frequency subset
Strict-NE threshold τ_{NE}	0.52 / 0.55 / 0.53 / 0.55	Derived from δ
Evaluation threshold τ_{eval}	0.79 / 0.82 / 0.80 / 0.81	Derived from π with evaluation templates
Cue tolerance ϵ	0.1 nats	Fixed; at most $\approx$10% probability loss on $a^\star$
Influence gate ι_0, smoothing I_0, penalty ν	0.05 nats, 0.1, 0.5	Fixed
AC distillation weight β; RL weight λ	0.1; 0.5	Fixed; post-hoc sensitivity in Appendix M
Abstention weight ρ	0 (ECHO), 0.5 (ECHO-A)	Development set, pre-specified rule over $\{0.25, 0.5, 1.0\}$
Divergence	forward KL, full vocabulary, $T = 1$	Method design; no truncation
Routing normalization	batch mean over distilled tokens	Method design
Learning rate	10^{-6} (10^{-5} for RFT)	Development set (label-free rule for ECHO-LF)
Batch; steps	256×8; 350	About one epoch; compute-matched runs: 525 or 510 steps
Policy-gradient clipping; mini-batches	$\epsilon_{\text{clip}} = 0.2$; 4 per rollout batch	Fixed; shared by all RL methods
Optimizer; gradient-norm clipping	AdamW, 10 warm-up steps; 1.0 (global)	Shared by all methods
KL coefficient toward θ_0	0.001	Shared by all RL methods
Maximum thinking length	8,192 tokens	32,768 for reasoning retention

C ALGORITHM AND TRAINING COST

Algorithm 1 One ECHO training step

Require: question batch (text only); student θ; frozen teacher and verifier θ_0; thresholds $\tau, \tau_{\text{ans}}, \tau_{\text{TT}}, \tau_{\text{NE}}$ derived from $\pi, \delta; \eta$

$\mathcal{D} \leftarrow \varnothing$ ▷ questions that are distilled in this step
for each question q **do**
  $Y \leftarrow m$ rollouts from $p_\theta(\cdot \mid q)$ with thinking; $A_q^S \leftarrow \text{ans}(Y)$ ▷ on-policy
  score every distinct claim in $\bigcup_j C(\hat{y}_j)$ by $g(s)$ ▷ deduplicated across rollouts; cached
  $K_q^0 \leftarrow \{s : g(s) \geq \tau \wedge \neg\text{Leak}(s, A_q^S)\}$
  $A_q \leftarrow A_q^S \cup \text{ans}(4 \text{ answer-only samples of } p_{\theta_0}(\cdot \mid q, K_q^0))$ ▷ oracle completes candidates
  $K_q \leftarrow \{s \in K_q^0 : \neg\text{Leak}(s, A_q)\}$
  $a^\star \leftarrow \arg\max_{a \in A_q} g(q \to a)$; $f \leftarrow$ frequency of $a^\star$ in Y
  if $g(q \to a^\star) \geq \tau_{\text{ans}}$ and $f \geq \eta$ **then** state $\leftarrow$ AC; $\omega_q \leftarrow \beta$
  else if $g(q \to a^\star) \geq \tau_{\text{TT}}$ and $f < \eta$ **then** state $\leftarrow$ TT; $\omega_q \leftarrow 1$
  else state $\leftarrow$ NE; strict $\leftarrow [\max_a g(q \to a) < \tau_{\text{NE}}]$
  end if
  if state $\in \{\text{AC}, \text{TT}\}$ **then**
    $S^\star \leftarrow$ RECALLSUFFICIENT$(K_q, a^\star)$
    if $S^\star \neq \varnothing$ **then** $c_q \leftarrow S^\star$; $\mathcal{D} \leftarrow \mathcal{D} \cup \{q\}$
    end if
  else if strict and ECHO-A **then** $c_q \leftarrow c_{\text{abst}}$; $\omega_q \leftarrow 1$; $\mathcal{D} \leftarrow \mathcal{D} \cup \{q\}$
  end if
  if $q \in \mathcal{D}$ **then** ▷ teacher is called only for distilled questions
    for each $\hat{y} \in Y$ and response position n **do**
      $p_S \leftarrow p_\theta(\cdot \mid q, \hat{y}_{<n})$; $p_T \leftarrow p_{\theta_0}(\cdot \mid q, c_q, \hat{y}_{<n})$ ▷ prefill only, full vocabulary
      $w_n \leftarrow \text{sg}[\max(0, H(p_S) - H(p_T))]$
    end for
  end if
  $r(\hat{y}) \leftarrow r_{\text{ground}}(\hat{y}) + \rho\, r_{\text{abst}}(\hat{y})$ for each $\hat{y} \in Y$
end for
$A \leftarrow$ group-standardized rewards ($A = 0$ for zero-variance groups)
$N \leftarrow$ number of response tokens of questions in $\mathcal{D}$; $\bar{w} \leftarrow$ mean of w_n over these tokens; $\tilde{w} \leftarrow w/\bar{w}$
update θ on $\frac{1}{N}\sum_{q \in \mathcal{D}} \omega_q \sum_{\hat{y}} \sum_n \tilde{w}_n \text{KL}(p_T \,\|\, p_S) + \lambda\, \mathcal{L}_{\text{PG}}(A)$ ▷ skip first term if $\mathcal{D} = \varnothing$ or $\bar{w} = 0$

function RECALLSUFFICIENT$(K, a^\star)$
  $S \leftarrow K$
  repeat
    order S by ascending influence $\log p_{\theta_0}(a^\star \mid q, S) - \log p_{\theta_0}(a^\star \mid q, S \setminus \{s\})$
    for $s \in S$ **do**
      if $\log p_{\theta_0}(a^\star \mid q, S \setminus \{s\}) \geq \log p_{\theta_0}(a^\star \mid q, K) - \epsilon$ **then** $S \leftarrow S \setminus \{s\}$
      end if
    end for
  until a full pass removes nothing
  return S in order of first appearance in the rollouts
end function

Answer claim and edge cases. The reward's claim set $C(\hat{y})$ contains the extracted reasoning claims only; the answer claim $q \to \hat{a}$ is constructed separately, scored by $g(q \to \hat{a})$ against τ_{ans}, kept out of U and of the influence terms, and counted in R if it fails. The influence $I(s)$ uses the trace's own answer $\hat{a}$ and context $C(\hat{y})$ in the reward, and $a^\star$ with K_q in the cue search; the frozen model answers without thinking and $\log p$ sums over the normalized answer tokens. A trace without an answer (abstention, or thinking truncated at 8,192 tokens) receives zero for the first reward term while R is still counted; the first term is also zero if $C(\hat{y})$ or $U(\hat{y})$ is empty, and failed extraction is treated as $C(\hat{y}) = \varnothing$. Groups with zero reward variance receive zero advantage. Questions outside $\mathcal{D}$ (ordinary NE questions, strict NE questions in ECHO, and AC or TT questions with $S^\star = \varnothing$) never call the teacher and are excluded from N and $\bar{w}$. A trace whose routing weights are all zero contributes no distillation, and a step whose batch mean $\bar{w}$ is zero skips the distillation term. Over the full 27B run these cases occur for 1.9% of traces (no answer), 0.7% ($C = \varnothing$, including 0.4% extraction failures), 3.2% of groups (zero variance), and 4.1% of TT questions ($S^\star = \varnothing$); no step had $\bar{w} = 0$.

Training cost. Table 5 breaks down the training cost of ECHO relative to one RLVR run on the same model. Three design choices keep the overhead small. Claims are deduplicated by bidirectional entailment across the m rollouts of a question before probing, and verification results are cached, so each distinct claim is probed once. Probe prompts are short, share their prefixes, and are answered without thinking in a few tokens (one name, or up to five for reverse QA), so they batch far more efficiently than long thinking rollouts. The remaining extra passes (influence, cue search, and teacher) are prefill-only.

Table 5: Training cost of ECHO relative to one RLVR run on the same model, measured as wall-clock accelerator time on identical hardware (27B; 1.00 = rollouts and update of RLVR, of which 0.82 is rollout generation and 0.18 the update).

Component	What it does per question	Rel. cost
Student rollouts	$m = 8$ generations, as in RLVR	0.82
Claim extraction	short outputs, one call per rollout	0.06
Verifier probes and semantic clustering	distinct claims only; 3×8 short samples, batched	0.22
Oracle candidate completion	4 answer-only samples conditioned on K_q^0	0.04
Leak filtering and influence	aliases, leak judgments, prefill-only leave-one-out passes	0.03
Recall-sufficient search	21.9 prefill-only passes on average (AC and TT only)	0.04
Teacher forward	one prefill pass per distilled rollout, full vocabulary	0.07
Parameter update	distillation and policy gradient	0.20
Total		1.48

At deployment ECHO has no additional module or sampling; mean output length on SimpleQA-Verified is 1,468 tokens (RLVR 1,402; Base 1,412). The test-time filter and the semantic-entropy threshold of Appendix K each need eight samples per question at inference.

D PROBE TEMPLATES AND PROMPTS

Table 6: Claim-level probe templates. Neither the original trace nor the original question is shown. The evaluation verifier uses disjoint templates.

Probe	Training verifier g	Evaluation verifier g_{eval}

Forward QA	`What is the {r} of {e}?`
	`Answer with the name only.`
	`Name the {r} of {e}. Reply with a single entity.`
		`Who or what is {e}'s {r}?`
		`Give only the name.`
		`Tell me the {r} of {e} in as few words as possible.`

Forward cloze	`The following sentence is taken from an encyclopedia article. Complete it with one entity: "{e}'s {r} is ___"`
	`Complete the statement with a single entity: "According to reference sources, the {r} of {e} was ___"`
		`Fill in the blank with one entity: "{e} --- {r}: ___"`
		`"It is well documented that the {r} of {e} is ___."`
		`Complete with a single entity.`

Reverse QA (set match)	`List up to five entities that have {o} as their {r}. One name per line.` `{o} is the {r} of which entities? Name at most five, one per line.`	`Which entities is {o} the {r} of? Up to five names, one per line.` `Give at most five examples of entities whose {r} is {o}.`
Distributional	Semantic entropy of the eight forward answers clustered by bidirectional entailment; margin between o and the strongest competing object at the cloze position	same
Yes/no (baseline only)	`Is the {r} of {e} {o}? Answer yes or no.`	—

A reverse probe counts as a hit if any listed entity is semantically equivalent to e. Answer-level probes ($q \to a$) replace the forward question by two paraphrases of q generated once per question and cached, never the original wording; cloze and reverse probes use the candidate a as object, with (e, r) from the triple parse of the answer claim. Extracted claims are converted to (e, r, o) triples by the same parsing prompt; a claim without a parse is not probed, cannot enter K_q, and is left out of the reward. Semantic equivalence is judged by the model itself with the entailment prompt of Table 7. Our extraction and restatement prompts adapt those of Gekhman et al. (2026); Table 7 gives the full text of every prompt the method uses. The abstention context c_{abst} reads: *"Note: none of your candidate answers to this question passed self-verification. If you are not sure, say that you do not know."* The answer-hint control uses *"Candidate answer:* $\{a^\star\}$*"*. The dummy context takes claims from K_q of a random other question, truncated to the token length of the current K_q. The external audit uses the search-augmented hallucination-detection prompt of Gekhman et al. (2026) with labels correct, incorrect, illegal, and unknown; it is used only for evaluation and offline labeling. Knowledge profiles use the answer prompt and the two grading prompts of Calderon et al. (2026).

Table 7: Full text of the prompts used by ECHO. Braces mark slots. All auxiliary prompts and the teacher use the frozen model θ_0; teacher distributions are evaluated on prefixes sampled from the student.

Use	Prompt
Claim extraction	`Below is a reasoning trace. List every factual statement it makes about the world, one per line. Rewrite each statement so that it can be understood without the trace, replacing pronouns and references by the entities they denote. Each line must state exactly one fact. Do not list statements about the question, about the reasoning process, or the final answer. If there are none, output NONE.\n\nTrace:\n{trace}`
Triple parsing	`{s}\nRewrite this statement as one line per object in the form subject | relation | object, where the object is the single entity or value the statement asserts. If this is impossible (e.g., a negation or a comparison), output NONE.` For an answer claim, $\{s\}$ is `Question: {q} Answer: {a}`.
Question paraphrase	`Rewrite the question in different words without changing what it asks or adding information. Output only the new question.\n{q}` Two samples at $T = 0.7$, generated once per question and cached.
Restatement removal	`Question: {q}\nStatement: {s}\nDoes the statement only restate or paraphrase the question without adding a fact? Answer YES or NO.`
Alias generation	`List up to five other names, spellings, or short forms by which "{a}" is commonly referred to, one per line. If there are none, output NONE.`
Leak judgment	`Question: {q}\nStatement: {s}\nDoes the statement state the answer to the question, or a fact from which the answer can be read off directly? Answer YES or NO.`
Entailment (dedup., clustering, equivalence)	`Statement A: {a}\nStatement B: {b}\nDoes A imply B? Answer YES or NO.` Applied in both directions; two items are equivalent if both answers are YES.
Answer-only scoring (influence $I(s)$, cue search)	`Facts:\n- {s1}\n- {s2}\n...\nAnswer the question with the name only.\nQuestion: {q}\nAnswer:` Thinking disabled; $\log p$ sums over the tokens of the normalized answer.
Oracle candidates	Same as answer-only scoring with the cues of K_q^0; four samples at $T = 0.7$.
Teacher context	`Here are some facts that may be relevant:\n- {s1}\n- {s2}\n...\n\n{q}` Thinking enabled; the student's response tokens follow unchanged. For strict NE in ECHO-A the fact list is replaced by c_{abst}.

E STRUCTURE OF THE GROUNDING REWARD

Write $N = \sum_{s \in U} I(s)v(s)$ and $D = \sum_{s \in U} I(s)$, so that the first term of Equation 4 is $F = N/(D + I_0)$. Since $I(s) \geq \iota_0 > 0$ on U and $v(s) = \pm 1$, we have $|N| \leq D < D + I_0$, hence $F \in (-1, 1)$ and $r_{\text{ground}} \in (-1.5, 1)$.

Proposition 2 (Structure of the grounding reward)**.** Suppose a non-answer claim s' that is not equivalent to any claim already in $C(\hat{y})$ is added to a trace, the answer is fixed, and the influences of the other claims do not change. (a) If s' passes self-verification and $I(s') < \iota_0$, then U and R are unchanged and so is the reward. (b) If s' fails self-verification, R increases by one, so the reward drops by at least $\nu/5 = 0.1$ while $R < 5$; if also $I(s') \geq \iota_0$, F decreases as well. (c) If s' passes and $I(s') \geq \iota_0$, F strictly increases. Hence the reward can increase only by adding influential, self-verified claims. (d) Independently of the premise, repeating an existing claim verbatim or in paraphrase leaves $C(\hat{y})$, and therefore the reward, unchanged, because $C(\hat{y})$ is deduplicated by bidirectional entailment (for paraphrases, provided the entailment judge detects them).

Proof. (a) and (d) are immediate. For (b), the penalty $\nu \min(R, 5)/5$ grows by $\nu/5$ when $R < 5$; if $s' \in U$ with $v(s') = -1$, $F' = (N - I')/(D + I' + I_0) \leq N/(D + I_0)$ is equivalent to $-(D + I_0) \leq N$, which holds. For (c), $F' = (N + I')/(D + I' + I_0) > N/(D + I_0)$ is equivalent to $D + I_0 > N$, which holds. □

Without deduplication, two copies of a failing claim would each have leave-one-out influence near zero and could drop out of U; deduplication closes this route. The premise of fixed influences is only approximate, so we test padding directly (Table 8).

Table 8: Padding stress test (27B; 500 development questions × 2 traces, answers fixed). M true, self-verified, irrelevant claims (taken from other questions' K_q) are injected.

Injected claims M	0	5	20	100
Influence-weighted ratio reward (no gate, no penalty)	0.31	0.34	0.43	0.63
Grounding reward (Equation 4)	0.24	0.24	0.23	0.23
Claims per trace	6.2	11.2	26.2	106.2
False claims per trace (external audit)	0.52	0.52	0.52	0.52

Of the injected claims, 98.6% have $I(s) < \iota_0$ (95th percentile 0.03 nats) and never enter U; the slight decline of the grounding reward reflects small changes in the influence of existing claims. Injecting five claims that *fail* self-verification barely moves the ratio reward (0.31 to 0.30) but lowers the grounding reward from 0.24 to -0.10 (the penalty saturates at five failing claims for traces that already contain some).

F Preliminary measurements on Qwen3.5

Table 9: Recall boundary on SimpleQA-Verified (100 samples per question with and without thinking; $\Omega(100)$ as defined by Gekhman et al. (2026)).

Model	Thinking pass@1	Thinking pass@100	No-thinking pass@1	No-thinking pass@100	$\Omega(100)$
Qwen3.5-9B	12.0 [10.6, 13.4]	40.8 [37.8, 43.9]	9.0 [7.8, 10.2]	25.8 [23.2, 28.6]	0.58 [0.52, 0.64]
Qwen3.5-27B	30.0 [28.0, 32.0]	63.6 [60.6, 66.5]	24.0 [22.1, 25.9]	46.4 [43.3, 49.5]	0.37 [0.33, 0.41]

Hallucinated premises hurt within a question. On the 412 SimpleQA-Verified questions for which Qwen3.5-27B produces both clean traces (all decidable claims correct) and traces with at least one incorrect claim among 100 samples, clean traces are correct 53.4% of the time and hallucinated ones 44.8%, a gap of 8.7 [7.9, 9.5] points; 345 questions (83.7% [79.9, 87.0]) fall below the diagonal, and the within-question slope is 0.83 [0.80, 0.85], matching the 0.84 reported by Gekhman et al. (2026).

Table 10: Knowledge profiles on WikiProfile following Calderon et al. (2026) (facts, %). Inference without encoding (14 facts for 27B, 19 for 9B) is merged into recall after thinking. Encoding rates are 88.4% (27B) and 69.1% (9B).

Model	Direct recall	After thinking	Recall failure	Encoding failure	Recall failures among errors
Qwen3.5-9B	645 (30.0)	258 (12.0)	602 (28.0)	645 (30.0)	48.3
Qwen3.5-27B	1,118 (52.0)	323 (15.0)	473 (22.0)	236 (11.0)	66.7

Scale mainly removes encoding failures (30.0% to 11.0%), while the share of recall failures among errors rises from 48.3% to 66.7%: the larger model's errors are mostly facts it holds but cannot retrieve.

Table 11: Claim-extraction audit on 100 traces (%).

Extractor	Faithful	Self-contained	Factual	Recall
Gemini-3.5-Flash (evaluation)	98.0 [93.0, 99.4]	94.0 [87.5, 97.2]	92.0 [85.0, 95.9]	85.0 [76.7, 90.7]
Qwen3.5-27B itself (training)	96.0 [90.2, 98.4]	92.0 [85.0, 95.9]	92.0 [85.0, 95.9]	82.0 [73.3, 88.3]

G VERIFIER AND TRIAGE

Table 12: Verifier performance on the 1,500 held-out calibration claims of each model. τ is converted from $\pi = 0.90$ on the fitting half; coverage is the share of true claims that are accepted; "coverage at 0.90" is a property of the held-out precision–coverage curve, and "deployed precision" applies the fitted τ to held-out claims. AUC differences between g and the best single probe or the yes/no probe are tested with a question-level cluster bootstrap, since claims from the same question are dependent (all $p < 0.001$).

Model	AUC (truth)	Yes/no AUC	τ	Coverage at 0.90	Deployed precision
Qwen3.5-9B	0.80 [0.78, 0.82]	0.60 [0.57, 0.63]	0.83	42.0 [38.6, 45.4]	89.2 [85.8, 91.9]
Qwen3.5-27B	0.84 [0.82, 0.86]	0.63 [0.60, 0.66]	0.80	55.0 [51.8, 58.2]	89.6 [86.9, 91.8]
Qwen3.5-35B-A3B	0.83 [0.81, 0.85]	0.62 [0.59, 0.65]	0.81	51.0 [47.8, 54.2]	89.4 [86.5, 91.7]
OLMo-3-7B-Think	0.78 [0.76, 0.80]	0.60 [0.57, 0.63]	0.82	38.0 [34.6, 41.5]	88.9 [85.1, 91.8]

Table 13: Left: probe ablation (27B, AUC against truth; 95% half-widths 0.02–0.03 for all claims and 0.04–0.05 for long-tail claims). Right: training uses the model's own claim extraction (27B).

Setting	All	Long tail
Full $g(s)$	0.84	0.81
− reverse probe	0.81	0.74
− cloze probe	0.82	0.78
− distributional signals	0.79	0.76
Yes/no only	0.63	0.58
Single-answer reverse match	0.83	0.79

	Gemini extract.	Self-extract.
AUC (truth)	0.84	0.83
Precision at frozen w, τ (%)	89.6	89.0
Coverage (%)	56.3	53.3
Claims per trace	6.1	5.4
K_q overlap (%)	—	78.0
Self-contained / faithful (%)	94.0 / 98.0	92.0 / 96.0

Reverse relations are one-to-many for 36.8% of claims; on these, set matching reaches AUC 0.82 against 0.78 for single-answer matching. Restatement filtering removes 96.0% of the 7.0% of claims that merely restate the question (AUC 0.85 before and 0.84 after). On OLMo, g predicts corpus support (co-occurrence of subject and object within 50 tokens at least three times) with AUC 0.86 [0.84, 0.88], higher than its AUC against truth (0.78). A relation-level audit of the proxy (Table 45) estimates its precision on these claims (84.3%) and the share of proxy-negative claims that are nonetheless supported (14.0%); with the estimator of Appendix L, 5.8% [3.4, 8.3] of corpus-supported calibration claims are false, a share that self-verification cannot reliably rule out. Separately, the proxy misses 18.4% [14.4, 23.2] of true claims that Wikipedia documents (55 of 299 sampled); this is a miss rate among documented true claims, not $P(\text{supported} \mid \text{proxy-negative})$.

Table 14: Answer-level thresholds (2,000 calibration questions; thresholds are converted on all questions, diagnostics are out-of-fold under five-fold cross-fitting).

	Qwen3.5-9B	Qwen3.5-27B
Answer-level AUC	0.76 [0.74, 0.78]	0.80 [0.78, 0.82]
τ_{ans} ($\pi = 0.90$)	0.81	0.78
Out-of-fold precision at τ_{ans} (%)	89.3 [86.1, 91.8]	89.8 [87.1, 92.0]
Candidates accepted (%)	36.0	48.0
Spearman $\rho(g, f)$	0.52	0.56
NE questions whose majority answer is correct (%)	9.0	7.0
NE questions never answered correctly in 100 samples (%)	84.0	87.0
τ_{TT} ($\pi = 0.90$ on $f < \eta$)	0.89	0.86
τ_{NE} ($\leq \delta = 5\%$ answerable in strict NE)	0.55	0.52

Triage against labeled profiles. We triage the 2,150 forward natural questions of WikiProfile (reverse questions as a slice) and compare with question-level classes obtained from the labeled protocol: *answerable* (majority correct in eight samples), *recall failure* (fact encoded, question not answered), and *encoding failure*. For 27B the class distribution is 1,677 / 260 / 213; for 9B, 1,195 / 361 / 594. Rerunning the labeled protocol with a new seed gives the test–retest ceiling. A grader from a different model family agrees with the protocol's grader on 97.3% [94.8, 98.6] of 300 answers. The per-question scores for 27B ($g(q \to a^\star)$, $f(a^\star)$, labeled class, and triage state) are provided as supplementary material, from which Table 16 and the recall-versus-encoding AUROC of 0.68 can be recomputed.

Table 15: Agreement between label-free triage and labeled question-level profiles on WikiProfile. "Target correct" is the share of questions whose $a^\star$ matches the gold answer, i.e., whether the training target is right. The recall-vs.-encoding AUROC uses $\max_a g(q \to a)$; the student's majority-answer frequency alone gives 0.58 and semantic entropy 0.57 (27B).

	Qwen3.5-9B	Qwen3.5-27B
Three-class agreement (%)	77.5 [75.7, 79.2]	85.3 [83.8, 86.8]
Cohen's κ (test–retest ceiling)	0.62 [0.59, 0.64] (0.86)	0.63 [0.60, 0.66] (0.84)
AC target correct (%)	95.8 [94.5, 96.9]	98.4 [97.7, 98.9]
TT share (%)	6.8 [5.8, 7.9]	5.6 [4.7, 6.6]
TT target correct (%)	87.7 [81.4, 92.1]	92.5 [86.4, 96.0]
Recall failures within TT (%); enrichment	45.2 [37.4, 53.3]; 2.7×	54.2 [45.3, 62.8]; 4.5×
Recall failures triaged TT (%)	18.3 [14.6, 22.6]	25.0 [20.1, 30.6]
NE questions indeed unanswered (%)	90.5 [88.4, 92.3]	86.6 [83.0, 89.5]
Encoding failures within NE (%)	60.4 [57.1, 63.5]	45.5 [40.8, 50.3]
Encoding failures triaged NE (%)	91.2 [88.7, 93.3]	90.6 [85.9, 93.8]
Recall vs. encoding failure (AUROC)	0.66 [0.62, 0.70]	0.68 [0.63, 0.73]

Table 16: Confusion matrices (questions; rows: labeled class, columns: label-free triage).

	Qwen3.5-27B				Qwen3.5-9B			
Labeled class	AC	TT	NE	Total	AC	TT	NE	Total
Answerable	1,577	43	57	1,677	1,058	52	85	1,195
Recall failure	21	65	174	260	24	66	271	361
Encoding failure	8	12	193	213	24	28	542	594
Total	1,606	120	424	2,150	1,106	146	898	2,150

Harm arises when the distillation target $a^\star$ is wrong, which is not determined by the confusion matrix alone. Among the 120 TT questions of 27B, $a^\star$ is correct for all 43 answerable questions, for 63 of 65 recall failures, and for 5 of 12 encoding failures (the cues sometimes allow the answer to be inferred), so 9 targets (7.5%) are wrong. AC targets are correct for 98.4% of AC questions and receive only light distillation ($\beta = 0.1$). Recall failures triaged as NE forgo a rescue but receive no distillation target. NE is not the same as unencoded: in 27B, only 45.5% of NE questions are encoding failures.

Table 17: Triage components (27B). Oracle candidate completion raises the share of recall failures whose gold answer enters A_q; the dedicated τ_{TT} restores target correctness. With the global threshold, 9B yields 228 TT questions at 70.6% [64.4, 76.1] target correctness, against 146 at 87.7% with τ_{TT}.

Setting	Gold in A_q (%)	Recall fail.→TT (%)	TT	Target correct (%)	Enrich.	κ
Student samples + τ_{TT}	31.2	16.5	95	92.6 [85.6, 96.4]	3.7×	0.60
+ oracle candidates, global τ_{ans}	44.6	31.9	157	82.2 [75.4, 87.4]	4.4×	0.63
+ oracle candidates + τ_{TT} (full)	44.6	25.0	120	92.5 [86.4, 96.0]	4.5×	0.63

Slices and stability. TT is more frequent for tail than for head questions (27B: 8.6% vs. 3.4% for the bottom and top popularity quintiles) and for reverse than for forward questions (8.1% vs.

5.6%). Rerunning triage with a new seed agrees on 91.8% of questions overall, but only 64.2% of TT questions are TT in both runs, which is why triage is recomputed online at every step. A fact-level variant (a fact is AC only if all four knowledge questions are) gives $\kappa = 0.55$. In the training pool, TT covers 7.5% (27B) and 7.0% (9B) of questions, about 6,750 and 6,300 questions.

H THE ORACLE AT INITIALIZATION

Development-set triage yields, for 27B and 9B respectively, 6,110 and 4,430 AC, 748 and 701 TT, and 3,142 and 4,869 NE questions; TT target correctness is 91.8% [89.7, 93.6] and 87.0% [84.3, 89.3], and $a^\star$ comes from student samples for 591 and 547 TT questions. Each context is evaluated with 16 fresh samples per question, independent of the triage samples; all contexts except the gold-answer ceiling pass through the leak filter, and all share one template.

Table 18: Oracle strength on development-set TT questions (%). pass@16 half-widths are 2.5–3.8 points.

Context	27B pass@1	27B pass@16	9B pass@1	9B pass@16
None	21.4 [19.5, 23.3]	66.2	17.3 [15.4, 19.2]	59.8
Dummy (same length)	21.9 [20.0, 23.8]	66.9	17.6 [15.8, 19.4]	60.3
Rejected claims	17.9 [16.1, 19.7]	60.8	14.2 [12.6, 15.8]	54.1
Unfiltered claims	33.8 [31.4, 36.2]	80.1	26.4 [24.2, 28.6]	72.6
Random subset of K_q ($\lvert S^\star \rvert$ cues)	30.2 [27.9, 32.5]	76.7	24.5 [22.4, 26.6]	70.4
K_q	41.7 [39.2, 44.2]	86.0	33.1 [30.8, 35.4]	79.4
$S^\star$	42.3 [39.8, 44.8]	86.4	33.8 [31.4, 36.2]	80.1
$S^\star$, shuffled	41.6 [39.1, 44.1]	86.0	33.3 [30.9, 35.7]	79.6
Gold answer (ceiling)	98.6 [98.2, 99.0]	100.0	97.9 [97.4, 98.4]	100.0

Table 19: Key oracle comparisons (pass@1 differences, points).

Comparison	27B	9B
K_q- none, subgroup with $a^\star$ from student samples	+11.1 [+9.9, +12.3]	+8.1 [+7.0, +9.2]
K_q- none, independent pipeline (603 / 556 questions)	+10.4 [+9.1, +11.7]	+7.6 [+6.4, +8.8]
K_q- none, all TT (selection-biased; reference only)	+20.3 [+19.0, +21.6]	+15.8 [+14.6, +17.0]
K_q- none, NE control	+2.2 [+1.9, +2.5]	+1.7 [+1.5, +1.9]
K_q- none, AC control	+1.8 [+1.5, +2.1]	+1.4 [+1.0, +1.8]
Specificity: subgroup gain − NE gain	+8.9 [+7.7, +10.1]	+6.4 [+5.3, +7.5]
$S^\star - K_q$ (non-inferiority margin −3.0)	+0.6 [−0.4, +1.6]	+0.7 [−0.2, +1.6]
$S^\star-$ random subset of equal size	+12.1 [+11.0, +13.2]	+9.3 [+8.2, +10.4]
K_q- unfiltered claims	+7.9 [+6.8, +9.0]	+6.7 [+5.8, +7.6]
Rejected claims − none	−3.5 [−4.4, −2.6]	−3.1 [−3.9, −2.3]
Dummy − none	+0.5 [−0.4, +1.4]	+0.3 [−0.5, +1.1]
$S^\star - S^\star$ shuffled	+0.7 [−0.3, +1.7]	+0.5 [−0.5, +1.5]

Selection bias and the independent pipeline. Oracle samples partly determine which questions are TT, so comparisons over all TT questions overstate the oracle. On the subgroup whose $a^\star$ came from student samples, the target is not supplied by the oracle, although the fact that no oracle candidate won still enters the selection; the estimate is therefore a subgroup effect. The independent pipeline fixes $a^\star$ and the state from eight student rollouts only, builds cues from eight further rollouts, and evaluates the oracle on a third set of samples; its target population is the set of questions triaged TT from student candidates alone.

By labeled class. On WikiProfile TT questions, cues raise pass@1 from 14.6% to 39.8% for recall failures (+25.2 [+21.1, +29.3]; 9B +20.2), from 38.2% to 52.9% for answerable questions (+14.7), and from 6.3% to 13.5% for the 12 encoding failures (too few for conclusions): the oracle evokes knowledge that is encoded but inaccessible.

Table 20: Recall-sufficient cues: fewer cues, fewer false premises. Relative reduction of contexts with a false premise: 74.4% (27B) and 70.8% (9B). Within K_q on 27B, the oracle is correct 34.2% of the time with a false premise in context and 47.1% without.

	27B: K_q	27B: $S^\star$	9B: K_q	9B: $S^\star$
Audited claims	7,630	1,720	5,608	1,753
Mean cues per context	10.2	2.3	8.0	2.5
Claim precision (%)	93.1 [92.3, 93.9]	95.4 [94.1, 96.5]	90.4 [89.4, 91.4]	93.6 [92.2, 94.8]
Contexts with ≥ 1 false claim (%)	41.8 [38.4, 45.4]	10.7 [8.7, 13.2]	47.9 [44.3, 51.6]	14.0 [11.6, 16.9]
Oracle pass@1 (%)	41.7	42.3	33.1	33.8

Leak filter and cue statistics (27B). A human audit of 300 deleted and 300 retained claims finds 7.3% [4.9, 10.9] unnecessary deletions and 1.3% [0.5, 3.4] missed leaks. K_q still contains the gold answer string for 15 TT questions (2.0%; 9B: 17, 2.4%); excluding them changes the all-TT gain only from 20.3 to 19.4. The median $|K_q|$ is 9 (IQR 6–14); $S^\star$ is empty for 31 questions (4.1%), which are not distilled; the median $|S^\star|$ is 2 (IQR 1–3), $|S^\star| \leq 3$ for 80.9% of questions, and tail questions need more cues (2.8 vs. 1.9 for the top quintile). The search uses 21.9 short forward passes per question on average (19.4 with a single pass; 17.7 for 9B), and 92.1% of questions converge after the first pass.

Table 21: Entropy gap $\Delta H = H_{\text{student}} - H_{\text{teacher}}$ at initialization (nats) on the rollouts of TT questions (5,984 for 27B, 5,608 for 9B). Token types come from aligning extracted claims to trace spans. The distillation teacher is the 27B model without context, with 9B as student.

Token type	Share (%)	Oracle, 27B	Oracle, 9B	KD, 27B→9B
Verified fact	7.9 / 7.6	+0.60 [0.56, 0.64]	+0.57 [0.53, 0.61]	−0.04 [−0.08, 0.00]
audited true	7.4 / 6.9	+0.62 [0.58, 0.66]	+0.59 [0.55, 0.63]	—
audited false	0.5 / 0.7	+0.40 [0.31, 0.49]	+0.38 [0.30, 0.46]	—
Rejected fact	3.5 / 4.5	−0.07 [−0.11, −0.03]	−0.08 [−0.12, −0.04]	−0.21 [−0.26, −0.16]
Answer	0.9 / 0.9	+0.67 [0.61, 0.73]	+0.62 [0.56, 0.68]	−0.02 [−0.08, +0.04]
$a^\star$ correct	0.8 / 0.8	+0.69 [0.63, 0.75]	+0.64 [0.58, 0.70]	—
$a^\star$ wrong	0.1 / 0.1	+0.49 [0.37, 0.61]	+0.49 [0.39, 0.59]	—
Other	87.7 / 87.0	+0.03 [0.02, 0.04]	+0.02 [0.01, 0.03]	−0.11 [−0.12, −0.10]
Fact tokens where the teacher is more certain (%)	—	66.8	64.9	41.7
Routing weight on fact and answer tokens (%)	—	54.8	52.3	—

Effective pull. Following He et al. (2026), we compare the gradient at object tokens with that of cross-entropy. Let J be the set of object-initial positions, i.e., the first token of the original object span of each evaluable claim in the TT rollouts at initialization (9B), including positions where the gate is closed, and let o_n be the first token of the correct object at $n \in J$. We compute $\sum_{n \in J} \tilde{w}_n \left(\pi_T(o_n) - p_S(o_n)\right) / \sum_{n \in J}(1 - p_S(o_n))$, where both distributions condition on the trace prefix before n and the denominator is the cross-entropy pull with unit weights; unrouted variants use $\tilde{w}_n = 1$, and divergences other than forward KL use their own logit gradient on o_n. The effective pull is 0.88 [0.85, 0.91] for forward KL with ΔH routing, 0.55 [0.52, 0.58] with uniform weights, 0.15 [0.14, 0.16] and 0.08 [0.07, 0.09] for JSD with and without routing, and 0.49 [0.46, 0.52] for a standard 27B distillation teacher, the same order as the roughly 0.5 reported for mid-training. Split by audit, routed forward KL exerts 0.95 [0.92, 0.98] at object-initial positions of audited-true claims, but only 0.19 [0.08, 0.30] toward the correct object at those of verified claims that are audited false, where the oracle favors the false object. Stylistic tokens ("wait", "hmm", "alternatively", and 39 other word forms) make up 1.9% of 27B response tokens but receive only 2.8% of the routing weight; their per-token forward KL is 0.21 nats, 0.4× that of verified fact tokens (0.52). Clipping the per-token KL at 5.0 would affect only 0.3% of fact tokens, which carry less than 0.1% of the routed pull, so ECHO does not truncate the divergence. The correct object is the claim's own object for audited-true claims and, for audited-false claims, the value in the auditor's cited evidence, which usually does not occur in the trace: it is evaluated at the start of the original (false) object span, located by string match. Claims whose error lies in the subject or relation rather than the object are excluded. These positions are a subset of the full claim spans over which Table 21 reports entropy gaps. The analytic values

underlying Figure 2a are given in Table 22; in its gate strip, the teacher keeps $\pi_T(o) = 0.95$ and spreads its remaining mass evenly over the same 49 alternatives, the diffuse student spreads $1 - p_S(o)$ evenly over them, and the binary student puts it on one wrong token.

Table 22: Analytic pull on the correct token (binary vocabulary, $\pi_T(o) = 0.95$, $p_\Phi(o) = 0.1$, $\alpha = 0.5$).

$p_S(o)$	0.01	0.05	0.1	0.3	0.5	0.9
Cross-entropy (gold)	0.990	0.950	0.900	0.700	0.500	0.100
Forward KL (ECHO)	0.940	0.900	0.850	0.650	0.450	0.050
JSD	0.022	0.070	0.103	0.143	0.121	0.014
Reverse KL	0.075	0.280	0.463	0.796	0.736	0.067
KD mixture	0.540	0.500	0.450	0.250	0.050	−0.350

I ADDITIONAL MAIN RESULTS

Table 23: Additional metrics on SimpleQA-Verified and in-distribution and high-frequency sets. Half-widths (27B): pass@16 ±2.5, weighted TFP ±1.0–1.5, PP ±1.4–1.7, EQ training relations ±0.6, NQ ±1.3, TriviaQA ±0.7. Compute is relative to one RLVR run.

	Qwen3.5-27B									Qwen3.5-9B		
Method	pass@16	W-TFP	PP	Claims	Halluc.	EQ-train	NQ	TQA	Compute	pass@16	W-TFP	PP
Base	56.1	79.2	71.3	6.0	1.12	44.6	44.1	74.6	—	34.4	72.1	63.9
Self-consistency RL	54.0	78.3	70.9	6.1	1.18	45.9	44.5	75.1	1.0×	32.6	71.2	63.5
Confidence RL	53.1	77.4	70.2	6.2	1.25	45.0	44.2	74.8	1.0×	31.9	70.3	62.8
Test-time filter	—	87.5	75.0	6.0	0.59	47.3	45.0	75.9	8× inf.	—	82.9	68.2
RFT / STaR	55.0	79.9	71.8	6.0	1.08	49.1	45.0	75.5	0.6×	33.5	72.6	64.3
RLVR	54.3	80.1	72.0	6.0	1.06	52.8	45.9	76.0	1.0×	32.9	72.8	64.6
RLVR, compute-matched	53.6	79.9	71.9	6.0	1.07	53.5	46.0	76.2	1.5×	32.4	72.7	64.5
OPSD-gold	55.8	78.6	69.0	6.1	1.16	52.0	45.5	75.7	1.2×	34.0	71.4	61.7
VAKE	55.2	82.6	72.9	6.0	0.92	50.2	45.4	75.6	1.8×	33.6	75.3	65.4
ECHO	57.4	86.9	79.6	5.9	0.67	50.9	45.6	75.7	1.5×	35.6	79.4	71.4
ECHO + label reward	56.8	87.1	79.8	6.0	0.67	53.6	46.4	76.6	1.5×	35.1	79.7	71.6

Table 24: Qwen3.5-9B: remaining columns of Table 1 and in-distribution and high-frequency sets. Half-widths: SQA-V ±1.5, EQ ±1.0, PopQA ±0.6, tail ±0.8, TFP ±1.3–1.5.

Method	EQ	PopQA	Tail	TFP (EQ)	EQ-train	NQ / TriviaQA
Base	30.9	27.5	8.9	78.6	33.8	33.0 / 61.3
Self-consistency RL	31.4	27.9	8.8	78.0	34.9	33.3 / 61.7
Confidence RL	31.0	27.6	8.6	77.4	34.2	33.1 / 61.4
Test-time filter	32.7	28.7	9.5	86.8	36.1	33.9 / 62.3
RFT / STaR	32.2	28.5	9.2	78.9	37.4	33.8 / 62.1
RLVR	34.6	29.9	9.8	79.1	41.3	34.6 / 63.0
RLVR, compute-matched	34.9	30.0	9.9	79.0	41.9	34.8 / 63.2
OPSD-gold	34.0	29.5	9.6	78.3	40.6	34.1 / 62.5
VAKE	33.5	29.3	9.7	81.0	38.8	34.0 / 62.4
ECHO	34.8	30.2	11.0	84.0	39.0	34.2 / 62.6
ECHO + label reward	35.7	30.9	11.3	84.3	42.0	35.0 / 63.4

Table 25: Labeled budget: baselines using exactly ECHO's calibration data, and the fully label-free ECHO-LF (SimpleQA-Verified, %). ECHO-LF has one seed on 27B and three on 9B.

Method	Sup.	27B acc.	27B TFP	9B acc.	9B TFP
RLVR on the 2,000 calibration questions	cal.	31.4	81.9	12.6	75.0
Self-consistency RL + calibrated gate	cal.	31.5	82.4	12.8	75.6
Calibrated-verifier answer reward	cal.	31.7	83.0	13.0	76.3
Calibrated-verifier answer reward, compute-matched	cal.	31.8	83.1	13.0	76.4
ECHO	cal.	33.1	88.7	13.8	81.6
ECHO-LF	none	32.8	87.6	13.5	80.5

Against the strongest same-budget baseline, ECHO's TFP advantage is $+5.7\,[+4.7, +6.7]$ (27B) and $+5.3\,[+4.0, +6.6]$ (9B). ECHO-LF's verifier reaches AUC 0.81 and its triage $\kappa = 0.59$ with TT target correctness 88.3%; on held-out relations and PopQA it scores 45.8 and 39.1 (27B) and 34.4 and 29.9 (9B), for out-of-distribution averages of 39.2 and 25.9; its macro-average difference to RLVR is $-0.2\,[-0.7, +0.3]$ and $-0.3\,[-0.8, +0.2]$, its 27B FActScore is 75.4, and all reasoning benchmarks change by less than one point.

Table 26: Reasoning retention. The first row gives Base; other rows give changes relative to Base (27B / 9B). Paired half-widths: math ± 1.9, GPQA-Diamond ± 1.6, MMLU-Redux ± 0.3, IFEval ± 1.5. Only the math criterion is a statistical (non-inferiority) test; that no point estimate drops by more than 1.5 points on the other benchmarks is descriptive, and the intervals allow larger drops (e.g., 9B IFEval $-0.8\,[-2.3, +0.7]$). Officially reported Qwen3.5 scores are 85.5 / 81.7 (GPQA-Diamond), 93.2 / 91.1 (MMLU-Redux), and 95.0 / 91.5 (IFEval); our reproduction is slightly lower because of sampling settings.

Method	Competition math	GPQA-Diamond	MMLU-Redux	IFEval
Base (absolute)	86.4 / 77.9	84.3 / 80.2	92.9 / 90.6	94.6 / 91.0
Self-consistency RL	-0.4 / -0.6	-0.2 / -0.3	$+0.1$ / $+0.1$	-0.4 / -0.5
Confidence RL	-0.9 / -1.2	-0.5 / -0.6	0.0 / 0.0	-0.7 / -0.9
RFT / STaR	-0.3 / -0.5	-0.1 / -0.2	$+0.1$ / 0.0	-0.3 / -0.4
RLVR	-0.8 / -1.1	-0.2 / -0.4	$+0.2$ / $+0.1$	-0.6 / -0.7
RLVR, compute-matched	-1.1 / -1.4	-0.4 / -0.6	$+0.2$ / $+0.1$	-0.8 / -0.9
OPSD-gold	-1.3 / -1.7	-0.6 / -0.8	-0.1 / -0.1	-0.9 / -1.1
VAKE	-0.9 / -1.2	-0.3 / -0.5	$+0.1$ / 0.0	-0.7 / -0.9
ECHO	-0.6 / -0.9	-0.3 / -0.5	$+0.3$ / $+0.1$	-0.5 / -0.8
ECHO + label reward	-0.7 / -1.0	-0.3 / -0.5	$+0.3$ / $+0.1$	-0.6 / -0.8

Table 27: Key comparisons. Rows marked $*$ are pre-specified primary tests and pass after Holm correction within the experiment group; the others are exploratory. LB: lower bound of the 95% interval.

Comparison	27B	9B	Criterion
$*$ ECHO − self-consistency RL, SQA-V	$+2.3\,[+1.1, +3.5]$	$+1.4\,[+0.5, +2.3]$	LB > 0
$*$ ECHO − best same-budget baseline	$+1.4\,[+0.2, +2.6]$	$+0.8\,[+0.1, +1.5]$	LB > 0
$*$ ECHO − RLVR, SQA-V	$-0.3\,[-1.6, +1.0]$	$-0.4\,[-1.2, +0.4]$	LB > -2.0
$*$ ECHO − RLVR, OOD macro-average	$+0.2\,[-0.3, +0.7]$	$+0.0\,[-0.5, +0.5]$	LB > -1.0
$*$ TFP (SQA-V), ECHO − RLVR	$+6.4\,[+5.4, +7.4]$	$+6.2\,[+4.9, +7.5]$	$\geq +3.0$
$*$ TFP (EQ held-out), ECHO − RLVR	$+5.1\,[+4.0, +6.2]$	$+4.9\,[+3.6, +6.2]$	$\geq +3.0$
$*$ PP (auxiliary), ECHO − OPSD-gold	$+10.6\,[+9.3, +11.9]$	$+9.7\,[+8.2, +11.2]$	$\geq +5.0$
$*$ Retention: lower bound on math	-2.5	-2.8	LB > -3.0
ECHO − compute-matched RLVR, accuracy	$-0.5\,[-1.8, +0.8]$	$-0.7\,[-1.5, +0.1]$	—
ECHO − compute-matched RLVR, TFP	$+6.6\,[+5.6, +7.6]$	$+6.3\,[+5.0, +7.6]$	—
FActScore, ECHO − RLVR	$+4.1\,[+3.1, +5.1]$	$+4.9\,[+3.9, +5.9]$	—
Weighted TFP, ECHO − RLVR	$+6.8\,[+5.8, +7.8]$	$+6.6\,[+5.3, +7.9]$	—
Hallucinated-claim rate, drop vs. Base	39.2%	27.0%	—
Long tail, ECHO − RLVR	$+1.8\,[+1.0, +2.6]$	$+1.2\,[+0.5, +1.9]$	—
In-distribution relations, ECHO − RLVR	$-1.9\,[-2.4, -1.4]$	$-2.3\,[-2.9, -1.7]$	—
NQ and TriviaQA (mean), ECHO − RLVR	$-0.3\,[-0.9, +0.3]$	$-0.4\,[-1.0, +0.2]$	—
pass@16, ECHO − Base	$+1.3\,[-0.6, +3.2]$	$+1.2\,[-1.0, +3.4]$	—
pass@16, RLVR − Base	$-1.8\,[-3.7, +0.1]$	$-1.5\,[-3.7, +0.7]$	—
ECHO − test-time filter, accuracy	$+0.6\,[-0.7, +1.9]$	$+0.2\,[-0.7, +1.1]$	—
Retention: largest point-estimate drop elsewhere (descriptive)	-0.5	-0.8	≥ -1.5

Training dynamics. During training, ECHO's accuracy curve closely tracks RLVR's (final 33.1 vs. 33.4), while its TFP rises steadily to 88.7 and RLVR's stays flat (82.3). Online triage of the training stream shows TT questions shrinking from 7.5% to 4.6%, AC growing from 61.1% to 65.9%, and NE falling from 31.4% to 29.5%: tip-of-the-tongue questions become accessible. External audit agrees with human judgment on 94.3% [91.1, 96.4] of 300 claims; a blind per-method audit of 200 claims each gives 94.0 (ECHO), 95.0 (RLVR), 93.0 (Base), and 95.0 (OPSD-gold), with no detectable difference in agreement (χ^2 test on the 4×2 table of agreements, $p = 0.80$). Equal agreement does not exclude method-dependent errors, so Table 54 reports per-method confusion matrices and a human-corrected TFP difference.

J Ablations, controls, and mechanism

Table 28: Component ablations (Qwen3.5-27B). Pre-specified target metrics are in bold. Paired half-widths: accuracy ± 1.4 and tail ± 1.0 for three-seed ablations (teacher, token weights, divergence), ± 1.3 and ± 0.9 otherwise; TFP and weighted TFP ± 1.0; PP ± 1.3; claims ± 0.3. The unweighted reward is (passing − failing)/all claims; the ratio reward is influence-weighted but has neither the influence gate nor the failure penalty.

Setting	Component	Acc.	Tail	TFP	W-TFP	PP	Claims	Halluc.
ECHO (full)	—	33.1	17.6	88.7	86.9	79.6	5.9	0.67
No oracle (reward only)	hindsight oracle	**30.9**	**14.8**	86.4	84.3	77.1	5.9	0.80
Unverified cues	endogenous verifier	31.9	16.3	**83.9**	81.6	74.6	6.0	0.97
No state routing	state-dependent objectives	32.2	16.6	**85.8**	83.9	76.8	6.0	0.85
K_q instead of $S^\star$	recall-sufficient cues	33.0	17.3	**86.1**	84.0	77.4	6.3	0.88
Uniform token weights	ΔH routing	**31.6**	**16.0**	87.9	86.0	78.9	5.9	0.71
Unweighted reward	influence weighting	32.0	17.0	89.6	**83.8**	80.3	9.4	0.98
JSD instead of forward KL	forward KL	**31.5**	**15.6**	88.2	86.3	79.1	5.9	0.70
Ratio reward	influence gate and penalty	32.9	17.4	89.3	86.4	79.9	**7.0**	0.75

Table 29: Change in each ablation's target metric relative to full ECHO (points; claims per trace for the ratio reward, where an increase is worse). 9B uses three seeds throughout. After Holm correction over the eleven targets, all remain significant on 9B; on 27B the accuracy targets of uniform token weights and JSD do not, while their tail targets do.

Ablation	Target	27B	9B
No oracle	accuracy	$-2.2\,[-3.6, -0.8]$	$-1.3\,[-2.2, -0.4]$
	long-tail accuracy	$-2.8\,[-3.8, -1.8]$	$-1.5\,[-2.3, -0.7]$
Unverified cues	TFP	$-4.8\,[-5.8, -3.8]$	$-5.6\,[-6.9, -4.3]$
No state routing	TFP	$-2.9\,[-3.9, -1.9]$	$-3.4\,[-4.7, -2.1]$
K_q instead of $S^\star$	TFP	$-2.6\,[-3.6, -1.6]$	$-2.9\,[-4.2, -1.6]$
Uniform token weights	accuracy	$-1.5\,[-2.9, -0.1]$	$-1.1\,[-2.0, -0.2]$
	long-tail accuracy	$-1.6\,[-2.6, -0.6]$	$-1.1\,[-1.9, -0.3]$
Unweighted reward	weighted TFP	$-3.1\,[-4.1, -2.1]$	$-3.6\,[-4.9, -2.3]$
JSD instead of forward KL	accuracy	$-1.6\,[-3.0, -0.2]$	$-0.9\,[-1.7, -0.1]$
	long-tail accuracy	$-2.0\,[-3.0, -1.0]$	$-1.2\,[-2.0, -0.4]$
Ratio reward	claims per trace	$+1.1\,[+0.8, +1.4]$	$+1.3\,[+0.9, +1.7]$

Further observations. Without state routing, hallucinated claims per trace on SimpleQA-Verified questions triaged NE rise from 1.21 to 1.62 ($+0.41\,[+0.28, +0.54]$), above Base (1.49). Without the oracle, pass@16 drops by $2.6\,[-4.5, -0.7]$ points. With uniform token weights, competition math drops by $2.3\,[-4.2, -0.4]$ and IFEval by $1.4\,[-2.9, +0.1]$ relative to ECHO: without routing, distillation also changes the model on many non-factual tokens. Unweighted rewards increase claims per trace by 59.3% (9B: 64.1%). An online teacher (not frozen) gives accuracy 32.5 and TFP 87.2; shuffling $S^\star$ gives 33.0 and 88.6, neither significantly different from ECHO.

Divergence scale and routing controls. Two exploratory controls, not part of the pre-specified ablations, probe the divergence and routing results further (Table 30; 27B, three seeds each). *Scale-matched JSD* multiplies the JSD term by $c = 5.2$, chosen so that its mean per-token logit-gradient norm on routed tokens equals that of forward KL at initialization; on the development set, $c \in \{2, 4, 8\}$ gives accuracy within 0.3 points of $c = 5.2$. *Teacher-entropy routing* follows Switch Distillation (He et al., 2026): a token is distilled with unit weight if the teacher's entropy is below its batch median and not at all otherwise (there is no gold-label fallback in our setting), and the weights are renormalized to mean 1 over distilled tokens as in ECHO, so that total distillation strength is matched. Scale matching halves the accuracy gap of JSD but not its long-tail gap, consistent with the vanishing pull at low $p_S(o)$ rather than with a mere difference in loss strength. Teacher-entropy routing is better than uniform weights on reasoning retention but places only 21.3% of its weight on fact and answer tokens (54.8% for ΔH routing), because the teacher is also confident on tokens the student already predicts well; it is weaker on the tail.

Table 30: Divergence scale and routing controls (27B, three seeds; differences to full ECHO with 95% hierarchical-bootstrap intervals). Math is the change on competition math relative to ECHO.

Setting	Acc.	Tail	TFP	ΔAcc.	ΔTail	ΔMath
ECHO (full)	33.1	17.6	88.7	—	—	—
JSD instead of forward KL	31.5	15.6	88.2	$-1.6\,[-3.0, -0.2]$	$-2.0\,[-3.0, -1.0]$	$-0.5\,[-2.4, +1.4]$
JSD, scale-matched ($c = 5.2$)	32.3	16.2	88.3	$-0.8\,[-2.2, +0.6]$	$-1.4\,[-2.4, -0.4]$	$-0.4\,[-2.3, +1.5]$
Uniform token weights	31.6	16.0	87.9	$-1.5\,[-2.9, -0.1]$	$-1.6\,[-2.6, -0.6]$	$-2.3\,[-4.2, -0.4]$
Teacher-entropy routing	32.4	16.3	88.0	$-0.7\,[-2.1, +0.7]$	$-1.3\,[-2.3, -0.3]$	$-0.9\,[-2.8, +1.0]$

Table 31: Controls (27B, three seeds). "States $S^\star$ cue" is the share of all development-set TT traces, correct or not, that state a cue from the $S^\star$ built before training (Base 23.9, RLVR 25.1). ECHO − answer-hint: accuracy $+0.5\,[-0.6, +1.6]$, tail $+1.7\,[+0.8, +2.6]$, TFP $+4.9\,[+3.9, +5.9]$; ECHO − distill-only: accuracy $+0.5\,[-0.6, +1.6]$, TFP $+2.8\,[+1.8, +3.8]$. On 9B, answer-hint minus ECHO is $-0.4\,[-1.1, +0.3]$ in accuracy and $-4.6\,[-5.8, -3.4]$ in TFP; distill-only minus ECHO is $-0.3\,[-1.0, +0.4]$ and $-2.5\,[-3.7, -1.3]$.

Setting	Teacher context	Reward	Acc.	Tail	TFP	W-TFP	States $S^\star$ cue
ECHO	$S^\star$	yes	33.1	17.6	88.7	86.9	44.7
Answer-hint teacher	$a^\star$ only	yes	32.6	15.9	83.8	81.7	29.8
Distill only	$S^\star$	no	32.6	17.0	85.9	84.0	42.9
Reward only	—	yes	30.9	14.8	86.4	84.3	27.4

The answer-hint teacher leads students to state the answer first and justify it afterwards, with 6.4 claims per trace of which 1.04 are hallucinated.

Table 32: Gain by knowledge state before training (SimpleQA-Verified, 27B, %), using the model's own label-free triage. On TT, ECHO exceeds RLVR by $+10.6\,[+5.0, +16.2]$.

State before training	Questions	Base	ECHO	ECHO gain	RLVR	RLVR gain
Accessible (AC)	312	82.0	83.8	$+1.8\,[-0.4, +4.0]$	85.9	$+3.9\,[+1.7, +6.1]$
Tip-of-the-tongue (TT)	106	19.0	38.0	$+19.0\,[+13.6, +24.4]$	27.4	$+8.4\,[+3.0, +13.8]$
Not evocable (NE)	582	4.1	5.0	$+0.9\,[0.0, +1.8]$	6.3	$+2.2\,[+1.3, +3.1]$

Table 33: Knowledge-profile migration on WikiProfile (27B, labeled protocol). ECHO recovers $7.0\,[+3.8, +10.1]$ points more recall failures than RLVR. Recall failures whose forward question was triaged TT before training (71 facts) are recovered at 49.3% [38.0, 60.7], those triaged AC (221) at 21.7% [16.8, 27.6], and those triaged NE (181) at 11.6% [7.7, 17.1].

Class	Base	ECHO	RLVR
Answerable	1,441 (67.0%)	1,532 (71.3%)	1,493 (69.4%)
direct / after thinking	1,118 / 323	1,139 / 393	1,150 / 343
Recall failure	473 (22.0%)	385 (17.9%)	430 (20.0%)
Encoding failure	236 (11.0%)	233 (10.8%)	227 (10.6%)
Encoding rate	88.4%	88.6%	88.5%
Recall failure → answerable	—	104 (22.0% [18.5, 25.9])	71 (15.0% [12.1, 18.5])
Encoding failure → answerable	—	7 (3.0% [1.4, 6.0])	12 (5.1% [2.9, 8.7])
Answerable → not answerable	—	20 (1.4% [0.9, 2.1])	31 (2.2% [1.5, 3.0])

Table 34: Stratified outcomes on development-set TT questions (27B): pass@1 of the *correct* answer before and after training (16 fresh samples). "Oracle confidently wrong": $a^\star$ wrong and the oracle's pass@1 on $a^\star$ given $S^\star$ is at least 50%. "Student confidently wrong": the modal answer of eight samples is wrong and appears at least five times.

Stratum	Questions	Before	ECHO after	ECHO change	RLVR change
$a^\star$ correct	687	22.6	43.2	$+20.6\,[+18.8, +22.4]$	$+9.3$
$a^\star$ wrong	61	8.0	7.2	$-0.8\,[-3.3, +1.7]$	$+5.1$
oracle confidently wrong	35	5.4	4.1	$-1.3\,[-4.3, +1.7]$	—
All cues in $S^\star$ true	640	22.4	43.0	$+20.6\,[+18.7, +22.5]$	$+9.2$
$S^\star$ contains a false cue	77	15.6	24.7	$+9.1\,[+4.2, +14.0]$	$+7.6$
$S^\star$ empty (not distilled)	31	15.2	24.2	$+9.0\,[+1.6, +16.4]$	$+7.4$
Student uncertain	570	25.0	46.2	$+21.2\,[+19.1, +23.3]$	$+9.8$
Student confidently wrong	178	9.9	21.3	$+11.4\,[+8.2, +14.6]$	$+6.0$
All	748	21.4	40.3	$+18.9\,[+17.2, +20.6]$	$+8.9$

On the 61 questions with a wrong $a^\star$, the correct answer's rate shows no detectable change ($-0.8\,[-3.3, +1.7]$), while the rate of the specific wrong target rises from 12.8% to 31.6% (13.9% to

41.6% on the 35 confidently wrong ones). Averaged over all 748 TT questions, this is a 1.5-point increase in the rate of the selected wrong targets, not in the total error rate: probability also moves away from other wrong answers, and these questions change the correct-answer rate over all TT questions by only -0.07 points. Across the development-set TT pools at steps 0, 175, and 350, wrong-target rates are 8.2%, 8.7%, and 9.1%, respectively; these pools change over training, so the rates are not transitions of fixed questions. On the step-0 pool, wrong-target questions receive 7.6% of the TT distillation weight. Of these 35 questions, 19 have a false cue in $S^\star$ that entails $a^\star$, 9 have true cues from which the oracle reasons wrongly (confusing relation, time, or direction), and 7 differ from the gold answer only in granularity or alias. With $\pi = 0.95$, the 35 fall to 13 while TT shrinks by 31%. On 9B, $a^\star$-correct questions gain $+15.9$ (610 questions) and $a^\star$-wrong ones change by $-1.2\,[-3.4, +1.0]$ (91 questions). All 152 development TT questions with a wrong $a^\star$ across both models were inspected manually; 72 (47.4% [39.6, 55.3]) of the wrong candidates came from the oracle and 80 from student samples.

Table 35: Gradient at the answer position at initialization (27B, development-set TT). $o^\star$ is the first token of the gold answer; $\pi_T(o^\star) - p_S(o^\star)$ is the forward-KL pull on its logit before weighting (positive means toward the correct answer); routing weight is normalized so that the mean over all response tokens is 1.

Stratum (questions)	ΔH (nats)	$\pi_T(o^\star) - p_S(o^\star)$	Pull toward correct (%)	Routing weight
$a^\star$ correct, all cues true (603)	$+0.73$	$+0.39$	88.1	7.1
$a^\star$ correct, a cue false (58)	$+0.55$	$+0.18$	69.0	5.6
$a^\star$ wrong, $S^\star$ non-empty (56)	$+0.53$	-0.02 (toward $a^\star$: $+0.36$)	7.9	5.4
Student uncertain (570)	$+0.86$	$+0.34$	77.8	8.4
Student confidently wrong (178)	$+0.08$	$+0.29$	74.2	0.9 (6.0 on cue tokens)

Both failure modes of entropy routing appear. With a wrong $a^\star$, routing still assigns high weight but the pull points to the wrong answer. With a confidently wrong student, the pull points the right way but the answer position receives almost no weight; the weight distribution suggests that cue positions may contribute to the $+11.4$ gain of this stratum (not tested separately).

Table 36: Recall paths (27B). Top: WikiProfile accuracy (2,150 questions each, eight samples). Bottom: development-set TT questions. Restricted to the questions each model answers correctly, the shares are 41.4%, 44.2%, and 63.9% (495, 561, 634 questions), but question composition differs; conclusions rest on the common-correct pool.

	Base	RLVR	ECHO
WikiProfile forward questions	78.0	81.2	80.7
change vs. Base	—	$+3.2\,[+2.0, +4.4]$	$+2.7\,[+1.5, +3.9]$
WikiProfile reverse questions	73.1	75.5	77.4
change vs. Base	—	$+2.4\,[+1.2, +3.6]$	$+4.3\,[+3.1, +5.5]$
Forward–reverse gap	4.9	5.7	3.3
Common-correct pool (431): correct traces stating an $S^\star$ cue first (%)	41.8 [37.2, 46.5]	44.6 [39.9, 49.3]	62.9 [58.3, 67.3]
All traces stating an $S^\star$ cue (%)	23.9	25.1	44.7
Newly correct under ECHO (152): stating an $S^\star$ cue first (%)	—	—	71.7 [64.1, 78.2]

Intervention protocol. The selection rules were fixed before any intervention was run. For each question, we take the first correct trace in sampling order that states a cue of the initial $S^\star$ before its answer and intervene at the first such cue, located by aligning the extracted claim to its span in the trace. Every condition edits only that span (deleting it, or replacing it with a paraphrase, an irrelevant true fact, or a false fact of the same relation), and the non-cue control deletes an equally long span without an $S^\star$ cue elsewhere in the prefix. All conditions share one truncation point, the end of the sentence that contains the cue in the original trace, and 16 continuations are sampled from there with a cleared cache. Questions are excluded if the prefix up to the truncation point already contains the answer or one of its aliases, or if the cue's sentence states a second $S^\star$ cue, so that each edit changes only the targeted cue. In the matched analysis, the shared cue is the first $S^\star$ cue that both traces state; cue position and confidence strata are defined on the ECHO trace.

Table 37: Interventions at the first recalled $S^\star$ cue in a correct trace (development-set TT; answer retention over 16 regenerations with a cleared cache, %). Difference-in-differences (drop relative to the unmodified continuation, ECHO − RLVR): deletion $+19.0$ [$+14.6$, $+23.4$], conflicting fact $+37.0$ [$+31.2$, $+42.8$]; deletion within the cue-bearing subsets of the common-correct pool $+17.8$ [$+12.9$, $+22.7$]. Each model's questions are those on which it states a cue, so the columns are different subsets; Table 38 compares matched cues. 95% half-widths: ± 1.6–2.4 for rows near 90%, ± 3.0–4.3 otherwise.

Intervention	Base	RLVR	ECHO
Questions	205	248	405
Continue unmodified	90.1	91.2	90.8
Paraphrase the cue	89.0	90.3	89.6
Delete an equally long non-cue span	88.7	90.1	89.2
Delete the cue	72.4	80.1	60.7
Replace with an equally long irrelevant true fact	71.0	79.3	58.9
Replace with a false fact of the same relation	55.8	68.9	31.5
Cue-bearing subsets of the common-correct pool: questions	180	192	271
unmodified / cue deleted	90.4 / 72.9	91.0 / 80.6	90.6 / 62.4
Newly correct under ECHO (109 of 152 state a cue): unmodified / deleted	—	—	88.1 / 49.5

Restricting to the common-correct pool shows that the difference is not an artifact of which questions each method answers, but which of those questions state a cue, and which cue, is still decided by each model. We therefore match cues: among the 431 questions, 138 have a correct ECHO trace and a correct RLVR trace that state the same $S^\star$ cue (bidirectional entailment), and we delete that cue in both (Table 38). The paired effect is smaller than the unmatched one but remains large, for both cue types, early and late cue positions, and high and low confidence of the unmodified continuation. Cues fall into two types by human annotation (200 cues, 91.5% agreement): recall prompts about related entities, such as a film's lead actress (68.4%), and deductive premises from which the answer follows (31.6%). Over ECHO's own cue-bearing traces (Table 37), deletion leaves 64.2% and 53.1% of answers respectively: both types are used. Table 39 shows one success and one failure end to end.

Table 38: Matched-cue interventions (27B; 138 development TT questions from the common-correct pool on which ECHO and RLVR state the same $S^\star$ cue; that cue is deleted; 16 regenerations). Retention drop is relative to the unmodified continuation; DiD is ECHO − RLVR, paired over questions. Confidence is the answer probability in the unmodified continuation, split at the median. Retention on all matched questions is 90.9 (ECHO) and 91.4 (RLVR) unmodified, 63.5 and 80.2 after deletion.

Subset	Questions	ECHO drop	RLVR drop	DiD
All matched	138	27.4	11.2	$+16.2$ [$+10.9$, $+21.5$]
recall prompts	94	24.8	10.4	$+14.4$ [$+8.0$, $+20.8$]
deductive premises	44	33.0	12.9	$+20.1$ [$+10.9$, $+29.3$]
cue in first third of the trace	71	—	—	$+15.8$ [$+8.4$, $+23.2$]
cue later in the trace	67	—	—	$+16.6$ [$+8.9$, $+24.3$]
high unmodified confidence	69	—	—	$+15.1$ [$+7.9$, $+22.3$]
low unmodified confidence	69	—	—	$+17.3$ [$+9.6$, $+25.0$]

Table 39: Two cases end to end (27B, development TT questions; counts over 16 regenerations or samples).

Success: a recall prompt the student relies on	
Question	Who directed the 1974 film *Ali: Fear Eats the Soul*?
$S^\star$	(1) *Ali: Fear Eats the Soul* stars Brigitte Mira. (2) The film was inspired by Douglas Sirk's *All That Heaven Allows*.
ECHO trace	"… a West German film from 1974. It stars Brigitte Mira. Mira was a regular in the films of Rainer Werner Fassbinder, and the story was inspired by Sirk's *All That Heaven Allows* … The director is Rainer Werner Fassbinder."
Intervention at cue (1)	All conditions truncate after "It stars Brigitte Mira.", so the prefix never names the director. Unmodified: Fassbinder 15/16. Paraphrased: 15/16. Cue deleted: 9/16 (Wim Wenders 4, Werner Herzog 3). Replaced by "It stars Liv Ullmann": 4/16 (Ingmar Bergman 10, other 2). RLVR, same cue deleted: 13/16.
Failure: a true cue, a confused relation, a wrong target	
Question	In which town was the composer of the symphonic poem *Vltava* (*The Moldau*) born? (Gold: Litomyšl)
$S^\star$	(1) *Vltava* is part of the cycle *Má vlast* by Bedřich Smetana. (2) Smetana is buried at Vyšehrad Cemetery.
What happens	Both cues are true, but the oracle conflates place of burial with place of birth and, since Vyšehrad lies in Prague, assigns high probability to Prague, which passes τ_{TT} and becomes $a^\star$. Prague itself never appears in a cue: once candidate completion adds it to A_q, the leak rule removes student claims that name it, such as "Smetana died in Prague". After training, ECHO answers Prague in 7/16 samples (before: 2/16) and Litomyšl in 3/16 (before: 3/16), the pattern of the wrong-target stratum in Table 34.

K ABSTENTION

ECHO-A uses only strict NE questions as abstention targets ($\tau_{\mathrm{NE}} = 0.52$ for 27B and 0.55 for 9B), which make up 63.8% (27B) and 71.2% (9B) of NE; other NE questions receive only the grounding reward. The weight ρ is selected before training as the value in $\{0.25, 0.5, 1.0\}$ with the highest development-set F whose accuracy loss is at most one point. Baselines are Base with an abstention prompt; a semantic-entropy threshold applied to ECHO at test time (eight samples), tuned to ECHO-A's abstention rate; and RLVR with a gold-label ternary reward (+1 correct, 0 abstain, -1 wrong).

Table 40: SimpleQA-Verified three-way grading (%). 27B half-widths: correct ± 2.0, wrong ± 2.4, abstain ± 1.8, answered precision ± 2.2, F ± 1.8. For 9B, the semantic-entropy threshold matches ECHO-A's abstention rate of 27.8%.

		Qwen3.5-27B					Qwen3.5-9B				
Method	Sup.	Corr.	Wrong	Abst.	Prec.	F	Corr.	Wrong	Abst.	Prec.	F
Base	—	30.0	69.2	0.8	30.2	30.1	12.0	87.3	0.7	12.1	12.0
Base + abstention prompt	—	29.1	58.3	12.6	33.3	31.1	11.5	76.1	12.4	13.1	12.2
ECHO	cal.	33.1	66.1	0.8	33.4	33.2	13.8	85.5	0.7	13.9	13.8
Semantic-entropy threshold (8×)	none	31.2	48.4	20.4	39.2	34.7	12.4	59.8	27.8	17.2	14.4
RLVR + ternary reward	gold	30.6	34.9	34.5	46.7	37.0	11.9	51.2	36.9	18.9	14.6
ECHO-A ($\rho = 0.5$)	cal.	32.4	47.2	20.4	40.7	36.1	13.3	58.9	27.8	18.4	15.4

Table 41: Abstention comparisons (SimpleQA-Verified).

Comparison	27B	9B
Relative reduction of wrong answers vs. ECHO	28.6%	31.1%
Accuracy change vs. ECHO	$-0.7\,[-1.6, +0.2]$	$-0.5\,[-1.1, +0.1]$
AC questions abstained, ECHO-A	1.9% [0.9, 4.1]	2.6%
AC questions abstained, RLVR + ternary reward	7.4% [5.0, 10.8]	9.1%
Accuracy vs. semantic-entropy threshold (same abstention rate)	$+1.2\,[+0.4, +2.0]$	$+0.9\,[+0.2, +1.6]$
Accuracy vs. RLVR + ternary reward	$+1.8\,[+0.6, +3.0]$	$+1.4\,[+0.7, +2.1]$
F vs. RLVR + ternary reward	$-0.9\,[-2.1, +0.3]$	$+0.8\,[-0.2, +1.8]$

Table 42: Where abstentions fall (27B, WikiProfile forward questions by labeled class before training; Wilson 95% intervals).

Class before training	Questions	ECHO-A	RLVR + ternary reward
Answerable	1,677	1.8% [1.3, 2.5]	5.9% [4.9, 7.1]
Recall failure	260	17.3% [13.2, 22.4]	36.5% [30.9, 42.5]
Encoding failure	213	58.7% [52.0, 65.1]	67.1% [60.6, 73.1]
Ratio encoding / recall failure	—	3.4×	1.8×

Sweeping ρ on 27B gives (correct, wrong, abstain, precision, F) of (32.8, 53.6, 13.6, 38.0, 35.2) at $\rho = 0.25$ and (31.3, 39.9, 28.8, 44.0, 36.6) at $\rho = 1.0$; the latter has a higher F on the test set but loses 1.8 points of accuracy and therefore violates the pre-specified rule. Teaching abstention on all NE questions instead of strict NE gives (31.2, 45.1, 23.7, 40.9, 35.3) and raises abstention on answerable questions from 1.8% to 4.9%. ECHO-A retains TFP 89.1 and tail accuracy 17.3, and reasoning benchmarks change by less than one point. In our comparison, the gold-label ternary-reward baseline abstains heavily on recall failures, whereas ECHO-A uses the TT/NE triage to concentrate abstention on encoding failures.

L SCALE, PRETRAINING DATA, AND LONG FORM

Table 43: Qwen3.5-35B-A3B (MoE; one seed per method, same configuration as 27B).

Method	SQA-V	TFP	W-TFP	PP	EQ held-out	PopQA tail
Base	24.6	79.2	77.0	69.5	38.7	12.0
RLVR	27.6	80.1	77.8	70.2	42.9	13.4
OPSD-gold	27.0	79.0	76.9	67.4	42.1	13.1
ECHO	27.1	86.1	84.0	76.8	43.2	14.9

Relative to RLVR, ECHO's accuracy differs by $-0.5\,[-1.8, +0.8]$ and its TFP by $+6.0\,[+5.0, +7.0]$; relative to OPSD-gold, PP differs by $+9.4\,[+8.1, +10.7]$. ECHO-A yields 26.5 correct, 52.4 wrong, and 21.1 abstained, reducing errors by 27.3% relative to ECHO. Reasoning changes relative to Base are -0.5 (math), -0.2 (GPQA-Diamond), $+0.2$ (MMLU-Redux), and -0.7 (IFEval).

Table 44: OLMo-3-7B-Think: testing "endogenous" against the public pretraining corpus. Corpus support is the raw co-occurrence proxy of Appendix G; audit-corrected rates are given in the text. SimpleQA-Verified accuracy is too low to be informative (4.8 / 5.3 / 5.2), so PopQA is the main accuracy set.

	Base	RLVR	ECHO
PopQA	21.6	23.7	24.0
EQ held-out	26.3	29.9	29.4
PopQA tail	6.1	6.8	7.9
TFP (PopQA traces)	68.4	69.3	75.8
Corpus-supported claims (raw proxy)	66.9	67.8	75.2

Co-occurrences	Questions	ECHO	RLVR
0 (none detected)	2,210	+0.3	+0.9
1–9	4,105	+4.6	+2.8
10–99	4,520	+2.9	+2.7
≥100	3,432	+0.8	+1.6

Of self-verified claims, 81.3% [80.2, 82.3] (4,066 of 5,002) are corpus-supported by the raw proxy, against 34.7% [32.5, 37.0] (590 of 1,701) of rejected claims. The proxy has false positives (co-occurrence without the relation) and false negatives (paraphrases, aliases, and windows it misses), so we correct it with a stratified relation-level audit (Table 45). For a pool G of claims and the proxy decision $x \in \{+, -\}$,

$$\hat{P}(\text{supp} \mid G) = \sum_x \hat{P}(x \mid G)\, \hat{P}(\text{supp} \mid x, G),$$

$$\hat{P}(\text{false} \mid \text{supp}, G) = \frac{\sum_x \hat{P}(x \mid G)\, \hat{P}(\text{supp}, \text{false} \mid x, G)}{\hat{P}(\text{supp} \mid G)},$$

where $\hat{P}(x \mid G)$ comes from the full pool, the conditional terms from the audit sample of each cell, and truth from the external audit; intervals resample both the pool and the audit samples. The corrected support rates are 77.2% [73.0, 81.4] for self-verified and 30.1% [27.2, 33.3] for rejected claims. Among supported claims, 5.8% [3.4, 8.3] of calibration claims and 6.9% [3.5, 10.8] of claims in ECHO traces are false (raw proxy for ECHO traces: 7.6%). Relative to RLVR, ECHO raises the raw corpus-supported share by 7.4 points, TFP by $+6.5$ [$+5.3$, $+7.7$], and PopQA accuracy by $+0.3$ [-0.3, $+0.9$]. PopQA gain half-widths are about 0.8–0.9 points. Gains concentrate on facts that the corpus contains but rarely and are negligible for facts with no detected co-occurrence; because the proxy misses some occurrences, this bin does not prove absence from the corpus. Co-occurrence does not verify the relation itself, so "support" is a corrected estimate rather than a per-claim proof.

Table 45: Stratified relation-level audit of the co-occurrence proxy (OLMo). Each cell is a random sample; two annotators judge whether a retrieved window expresses the claimed relation (supported), and truth comes from the external audit. For proxy-positive claims the windows are the co-occurrence hits; for proxy-negative claims annotators query a full-text index of the corpus with up to five aliases each for subject and object and inspect windows of 200 tokens. Claims from the same question are resampled together. "Share" is the proportion of the pool with that proxy decision.

Pool	Proxy	Share (%)	Audited	Supported	Supported & false
Calibration claims	+	60.0	300	253 (84.3%)	13
	−	40.0	300	42 (14.0%)	5
Trace claims, self-verified	+	81.3	150	134 (89.3%)	—
	−	18.7	150	37 (24.7%)	—
Trace claims, rejected	+	34.7	150	119 (79.3%)	—
	−	65.3	150	6 (4.0%)	—
Claims in ECHO traces	+	75.2	150	133 (88.7%)	9
	−	24.8	150	35 (23.3%)	3

Table 46: Long-form factuality on 500 FActScore biographies (half-width ≈ 1.2). Training used short questions only.

Method	27B FActScore	Facts per bio	Unsupported per bio	9B FActScore
Base	71.8	34.2	9.6	63.5
RLVR	72.4	34.6	9.5	64.0
OPSD-gold	71.2	35.1	10.1	63.1
ECHO	76.5	33.8	7.9	68.9
ECHO-A	78.2	31.2	6.8	70.6

ECHO improves FActScore over Base by $+4.7$ [$+3.7$, $+5.7$] with 1.2% fewer facts per biography (not significant); ECHO-A produces 8.8% fewer facts because it expresses uncertainty.

Consolidation into direct recall. Qwen3.5 shares parameters between thinking and non-thinking modes. As a preliminary, single-seed observation, a short label-free self-distillation after ECHO, with cross-entropy of the non-thinking mode on the thinking mode's majority answer for training-pool AC and TT questions, teaches the non-thinking mode to give that answer directly. Non-thinking SimpleQA-Verified accuracy rises from 24.9 (ECHO) to 26.8 ($+1.9$ [$+0.9$, $+2.9$]; Base 24.0, RLVR 25.3) while thinking accuracy stays at 33.0; WikiProfile direct recall grows from 1,139 to 1,214 facts, and 71 of the 75 new facts previously required thinking.

M SENSITIVITY AND SEED VARIANCE

Table 47: Threshold settings. Claim precision and coverage (share of true claims accepted) apply each setting's thresholds to the held-out calibration claims; TT counts and target correctness are on WikiProfile forward questions; parentheses give paired differences to the model's default. For 27B, "shared raw thresholds" coincides with the default because the shared values are 27B's.

Setting	Precision	Coverage	TT	Target correct	SQA-V accuracy	TFP
27B, $\pi = 0.85$	85.3	66.0	152	88.2	32.8 (-0.3 [-1.2, $+0.6$])	87.9
27B, $\pi = 0.90$ (default)	89.6	56.3	120	92.5	33.1	88.7
27B, $\pi = 0.95$	94.7	38.0	83	96.4	32.7 (-0.4 [-1.3, $+0.5$])	89.0
27B, quantile thresholds (no labels)	88.4	58.0	131	89.9	32.9 (-0.2 [-1.1, $+0.7$])	88.2
9B, $\pi = 0.90$ (default)	89.2	44.2	146	87.7	13.8	81.6
9B, 27B's raw thresholds	87.1	47.5	181	83.4	13.5 (-0.3 [-0.9, $+0.3$])	80.4

Higher π gives cleaner cues and higher TFP but fewer TT questions and slightly lower accuracy; $\pi = 0.90$ lies in the middle of a flat region, consistent with a sweep of $\tau_{\mathrm{TT}} \pm 0.03$ (141 questions at 89.4% and 97 at 94.8% target correctness). Sharing raw thresholds across models would lower 9B's claim precision to 87.1% and its TT target correctness to 83.4%, which is why the shared hyperparameter is π.

Table 48: Triage sensitivity to η and m (27B, post hoc; both were fixed in advance).

η	0.250	0.375	0.500	0.625	0.750
κ	0.59	0.61	0.63	0.62	0.59
TT share (%)	3.2	4.3	5.6	7.0	8.6
Enrichment of recall failures in TT	5.6×	5.0×	4.5×	3.7×	3.0×
m	4	8	16		
Gold answer in A_q for recall failures (%)	38.1	44.6	50.4		
κ	0.60	0.63	0.64		
TT target correct (%)	91.0	92.5	92.9		
Relative inference cost	0.5×	1.0×	2.0×		

Table 49: Post-hoc sensitivity to β and λ (9B, SimpleQA-Verified; one seed each, paired with ECHO's first seed, which is the default column; accuracy half-width $\approx \pm 0.7$, no difference is significant). Larger λ weights the policy gradient more, raising TFP and slightly lowering accuracy.

	Default ($\beta = 0.1, \lambda = 0.5$)	$\beta = 0.05$	$\beta = 0.2$	$\lambda = 0.25$	$\lambda = 1.0$
Accuracy	13.7	13.7	13.6	13.9	13.4
TFP	81.8	81.4	81.8	80.3	82.0

Table 50: Seed standard deviations on SimpleQA-Verified (three seeds). Comparisons with several seeds use hierarchical bootstraps, which include seed variance; single-seed comparisons cover question sampling only.

Model	Method	Accuracy	TFP	PP
27B	ECHO	0.4	0.3	0.4
	RLVR	0.5	0.3	0.4
	OPSD-gold	0.5	0.4	0.5
	No oracle	0.5	0.4	0.5
	Uniform token weights	0.4	0.3	0.4
	JSD instead of forward KL	0.5	0.3	0.4
9B	ECHO	0.3	0.4	0.5
	RLVR	0.4	0.3	0.4

Table 51: Per-seed results on SimpleQA-Verified (accuracy / TFP, %). Seed 1 is the seed against which single-seed 27B ablations are paired; their intervals cover question sampling only.

Model	Setting	Seed 1	Seed 2	Seed 3
27B	ECHO	33.0 / 88.6	32.7 / 88.4	33.5 / 89.0
	RLVR	33.3 / 82.2	33.9 / 82.6	32.9 / 82.0
	OPSD-gold	32.9 / 81.2	32.5 / 80.5	33.5 / 81.2
	No oracle	30.8 / 86.6	31.4 / 86.6	30.4 / 85.9
	Uniform token weights	31.5 / 87.8	31.2 / 87.6	32.0 / 88.2
	JSD instead of forward KL	31.4 / 88.1	32.0 / 88.5	31.0 / 87.9
	JSD, scale-matched	32.2 / 88.2	31.9 / 88.0	32.7 / 88.6
	Teacher-entropy routing	32.3 / 87.9	32.8 / 88.3	32.0 / 87.7
	Answer-hint teacher	32.5 / 84.0	32.2 / 83.3	33.0 / 84.0
	Distill only	32.5 / 85.8	33.1 / 86.2	32.1 / 85.6
9B	ECHO	13.7 / 81.8	13.5 / 81.1	14.1 / 81.8
	RLVR	14.4 / 75.3	13.7 / 75.1	14.4 / 75.7

All methods except RFT select 10^{-6} from $\{5 \times 10^{-7}, 10^{-6}, 2 \times 10^{-6}\}$ on the development set; RFT selects 10^{-5} from $\{5 \times 10^{-6}, 10^{-5}, 2 \times 10^{-5}\}$; ECHO-LF selects 10^{-6} with its label-free rule.

N Statistical methods and audits

Intervals. Proportions are reported in percent with one decimal; κ, AUC, and correlations with two decimals; entropy gaps in nats with two decimals. Brackets give 95% intervals. Independent proportions use Wilson intervals. When a question contributes several samples or claims, we resample questions (cluster bootstrap, 10,000 replicates) and report the 2.5% and 97.5% quantiles; comparisons between methods resample the same questions for both (paired). With several training seeds, a hierarchical bootstrap first resamples seeds and then questions, so intervals cover training randomness as well as question sampling. κ intervals use a multinomial bootstrap of the confusion matrix, and AUC differences are tested with the same question-level cluster bootstrap rather than DeLong's test (DeLong et al., 1988), which assumes independent observations. pass@k uses $1 - \binom{n-c}{k} / \binom{n}{k}$ per question.

Testing hierarchy. Primary tests and their criteria were fixed before each group of experiments; within a group, primary tests are Holm-corrected (Holm, 1979). All other comparisons are exploratory and reported with intervals only. Non-inferiority compares the lower 95% bound with a margin: 2.0 points on SimpleQA-Verified and 1.0 point on the out-of-distribution macro-average for accuracy against RLVR, and 3.0 points for the oracle with $S^\star$ against K_q. Non-inferiority does not imply equivalence. Families are defined per experiment and model: the main-results family has eight tests and the ablation family eleven targets (Table 52). p-values are bootstrap p-values: one-sided for superiority and non-inferiority tests (the share of replicates at or beyond the null boundary) and two-sided for ablation targets; with 10,000 replicates, values below 0.001 are reported as $<$0.001. Holm-adjusted values are $\max_{j \le i}(m - j + 1)\, p_{(j)}$ for the i-th smallest p-value in a family of m tests.

Table 52: All primary tests with raw and Holm-adjusted p-values (27B / 9B). NI: non-inferiority with the stated margin; for ablations, the change in the pre-specified target relative to full ECHO. An adjusted value below 0.05 passes.

Family	Test	Raw p	Holm-adjusted p
Main results	ECHO > self-consistency RL (SQA-V)	<0.001 / 0.001	<0.01 / <0.01
	ECHO > best same-budget baseline (SQA-V)	0.012 / 0.013	0.018 / 0.026
	NI vs. RLVR, SQA-V (margin 2.0)	0.006 / <0.001	0.018 / <0.01
	NI vs. RLVR, OOD macro-average (margin 1.0)	<0.001 / <0.001	<0.01 / <0.01
	TFP gain over RLVR ≥ 3.0 (SQA-V)	<0.001 / <0.001	<0.01 / <0.01
	TFP gain over RLVR ≥ 3.0 (EQ held-out)	<0.001 / 0.002	<0.01 / <0.01
	PP gain over OPSD-gold ≥ 5.0	<0.001 / <0.001	<0.01 / <0.01
	NI on competition math (margin 3.0)	0.008 / 0.015	0.018 / 0.026
Ablations	No oracle: accuracy	0.004 / 0.006	0.012 / 0.024
	No oracle: long-tail accuracy	<0.001 / <0.001	<0.02 / <0.02
	Unverified cues: TFP	<0.001 / <0.001	<0.02 / <0.02
	No state routing: TFP	<0.001 / <0.001	<0.02 / <0.02
	K_q instead of $S^\star$: TFP	<0.001 / <0.001	<0.02 / <0.02
	Uniform token weights: accuracy	0.042 / 0.019	**0.068** / 0.038
	Uniform token weights: long-tail accuracy	0.003 / 0.008	0.012 / 0.024
	Unweighted reward: weighted TFP	<0.001 / <0.001	<0.02 / <0.02
	JSD instead of forward KL: accuracy	0.034 / 0.031	**0.068** / 0.038
	JSD instead of forward KL: long-tail accuracy	<0.001 / 0.004	<0.02 / 0.020
	Ratio reward: claims per trace	<0.001 / <0.001	<0.02 / <0.02

Table 53: Human audits, data checks, and calibration.

Audit	Sample	Result	Used for
Claim extraction (Gemini / 27B)	100 traces	faithful 98.0 / 96.0	self-extraction in training
Claim truth labels	200 claims	agreement 95.0 [91.0, 97.3]	verifier evaluation
OLMo co-occurrence proxy (relation level, stratified)	1,500 claims in 8 cells	support rates per cell (Table 45); misses 18.4 [14.4, 23.2] of documented true claims	corpus-support correction
Cross-family regrading of profiles	300 answers	agreement 97.3 [94.8, 98.6]	labeled classes
Leak filter (deleted / kept)	300 / 300 claims	unnecessary 7.3; missed 1.3	oracle contexts
External audit vs. humans (overall)	300 claims	agreement 94.3 [91.1, 96.4]	TFP
External audit vs. humans (blind, per method)	200 claims each	94.0 / 95.0 / 93.0 / 95.0 (ECHO / RLVR / Base / OPSD-gold), χ^2 test $p = 0.80$	per-method error rates (Table 54)
Unknown and illegal labels	all audited claims	4.1 / 4.3 / 4.6 / 4.2%	counting as false: +6.2 TFP gap
Cue precision in training pool (step 0)	500 + 500 claims	27B 91.6 [88.8, 93.7]; 9B 90.8 [87.9, 93.0]	thresholds transfer to training data
Data isolation	all data	0 shared subject entities; 0 near-duplicates	Appendix A
Cue type (recall prompt / premise)	200 cues	agreement 91.5	interventions
Wrong $a^\star$ on development TT	152 questions	72 from the oracle (47.4 [39.6, 55.3])	failure analysis

Table 54: Blind per-method audit: external auditor versus human judgment on 200 claims per method, drawn uniformly at random from the audited traces (at most one claim per question). Human-corrected TFP is $r \cdot \mathrm{PPV} + (1 - r) \cdot \mathrm{FOR}$, where r is the method's raw TFP and PPV and FOR are the shares of auditor-true and auditor-false claims that humans judge true. After correction, ECHO $-$ RLVR is $+6.3$ [$+2.0$, $+10.6$], against $+6.4$ raw; the interval resamples each method's audit sample and the evaluation questions and is dominated by the few auditor-false claims.

Method	Both true	Auditor true, human false	Auditor false, human true	Both false	Agreement	Corrected TFP
ECHO	170	6	6	18	94.0	88.5
RLVR	158	5	5	32	95.0	82.2
Base	156	8	6	30	93.0	80.5
OPSD-gold	157	5	5	33	95.0	81.0