Reversal Is Structural: Concept-Aware Post-Training Recovers Rare, Deep Mathematical Skills

Yassir Laaouach
Independent Researcher
yl765@sussex.ac.uk

Abstract

Solution-based post-training,ranging from self-taught rationales to iterative preference learning,is credited with improving mathematical problem solving in large language models. Recent evidence, however, shows a "self-improvement reversal": as `pass@1` rises, models can lose breadth and robustness. We argue this is structural, not a statistical quirk. Drawing on Knowledge Space Theory, we recast mathematical reasoning as movement over a prerequisite concept graph and induce a sparse problem–concept mapping via an automatic pipeline, AUTOKST, enabling concept-aware diagnostics beyond accuracy.

Applied to challenging math benchmarks, AUTOKST reveals that regressions localize to the graph's *fringe*,rare, prerequisite-heavy skills that headline scores overlook. A linearized view of post-training explains why frequency-skewed updates with coupled gradients naturally drift away from these low-frequency directions.

Guided by this account, we propose *Fringe-Theorem Training* (FTT): a lightweight regimen that combines frequency-aware loss reweighting, projection-based gradient safeguards, and a fringe-focused micro-curriculum. In controlled studies, FTT improves `pass@1` while restoring fringe competence and prerequisite adherence; e.g., versus STaR it raises fringe performance and coverage substantially, improves consistency, lowers calibration error, and,paired with early stopping,reduces reasoning tokens. By turning post-training evaluation into concept-level measurement, our framework distinguishes genuine self-improvement from structural regress and offers a practical path to the former.

1 Introduction

Solution-based post-training is the default recipe for improving reasoning in large language models (LLMs). Two families dominate: *self-taught rationales* (e.g., STaR), which iteratively generate and fine-tune on successful solutions, and *preference-based post-training* (e.g., RLHF/DPO), which steers models toward preferred answers without explicit reward modeling [17, 9]. These methods deliver consistent gains on math and logic benchmarks and are attractive because they leverage the model's own outputs as training signal [13, 12]. Yet headline improvements (e.g., `pass@1`) leave open a central question: *what skills were gained, retained, or displaced?*

A growing body of evidence indicates that iterative post-training can raise `pass@1` while degrading breadth and robustness, a phenomenon dubbed *self-improvement reversal* [15]. Concretely, models exhibit reduced output diversity and weaker out-of-distribution generalization even as aggregate scores climb. At the same time, chain-of-thought prompting and self-consistency decoding can inflate accuracy without faithfully reflecting internal computation, complicating diagnosis of what actually changed [13, 12, 8, 1]. The field thus has strong mechanisms to *optimize* answers but weak tools to *localize* gains and losses across underlying skills.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MATH-AI.

We argue that reversal is not a metric quirk but a *structural* effect. Our thesis is that post-training redistributes competence over a latent space of mathematical concepts,nodes linked by prerequisite relations,and that regressions concentrate on the *fringe*: rare, prerequisite-heavy skills that headline metrics conflate or overlook. To test this, we instantiate a classical idea from educational measurement: *Knowledge Space Theory* (KST). We induce a prerequisite concept graph and a sparse Q-matrix mapping each problem (and each solution step) to the minimal skills it requires, and then use lightweight cognitive-diagnostic modeling to infer concept-level mastery and track how it moves across post-training iterations [3].

This concept-aware view reconciles disparate observations. From an *optimization* perspective, post-training is a multi-task update under heavy frequency skew; work on catastrophic forgetting and gradient interference predicts that, when tasks conflict, updates drift toward frequent directions unless protected or reweighted. From an *evaluation* perspective, math benchmarks vary in topic coverage and prerequisite depth; without a concept graph, a global gain can mask local collapses on rare skills. KST supplies the missing scaffold: it lets us measure *where* competence grows or erodes, whether solutions respect prerequisite order, and whether mastery on a concept transfers to related items.

This paper. We introduce AUTOKST, a fully automatic pipeline that converts math problems and proofs into a KST representation: (i) mining symbolic/lexical patterns from solutions to propose concepts; (ii) learning a compact prerequisite DAG via continuous structure learning; and (iii) assigning concepts to problems (and steps) with a sparsified Q-matrix [20]. With this representation we (a) show that regressions concentrate on fringe concepts across iterative post-training loops; (b) present a simple linearized analysis explaining why frequency-skewed, coupled gradients induce negative drift along low-frequency directions; and (c) propose *Fringe-Theorem Training* (FTT),frequency-aware loss reweighting, projection-based gradient safeguards, and a fringe-focused micro-curriculum,that restores fringe competence and prerequisite consistency without sacrificing `pass@1`, and can reduce reasoning tokens when paired with early stopping.

Contributions.

1. **Structural diagnosis of self-improvement.** We recast progress/regress as redistribution over a prerequisite concept graph and show that negative drift concentrates on rare, prerequisite-heavy skills.
2. **Automatic concept-graph induction for math LLMs.** AUTOKST induces both the prerequisite DAG and Q-matrix directly from solutions/CoT, enabling concept-aware metrics: fringe coverage/diversity, prerequisite consistency, concept-level transfer, and calibration.
3. **A practical remedy.** *Fringe-Theorem Training* counteracts frequency skew and gradient interference, protecting and improving fringe skills while preserving overall accuracy.

By shifting attention from *answers* to *skills and prerequisites*, we reinterpret self-improvement reversal as a predictable property of training dynamics and provide principled levers that convert superficial gains into broader mathematical competence.

2 Related Work

Post-training for reasoning. Modern self-improvement pipelines for LLMs combine two families of post-training. Solution-centric loops,exemplified by STaR,iteratively generate, filter, and fine-tune on model-produced rationales to "learn from one's own solutions," yielding gains on math and logic without extra labels [17]. Preference-centric alignment,typified by Direct Preference Optimization (DPO),replaces reward modeling and PPO with a closed-form surrogate trained like classification while capturing human or proxy preferences [9]. These techniques often sit atop prompting strategies such as chain-of-thought (CoT) and self-consistency, which elicit intermediate steps and marginalize across reasoning paths, producing large jumps on step-wise benchmarks [13, 12]. However, accumulating evidence cautions that visible rationales are not always faithful to internal computation: causal analyses find limited mediation from stated steps to final predictions, and controlled probes show that models can produce correct answers while displayed "reasoning" is spurious or strategically adapted [8, 1]. This undermines evaluations that rely solely on surface rationales and motivates structure-level diagnostics.

Benchmarks and verifiers. A complementary strand builds richer testbeds and verifiers for mathematical reasoning. GSM8K introduced verifier-guided selection for grade-school word problems [2], while MATH assembled competition-grade problems with full solutions [4]. At higher difficulty and with formal guarantees, *miniF2F* aggregates Olympiad problems across proof assistants to test cross-system transfer [19]; *PutnamBench* formalizes hundreds of undergraduate-level theorems in Lean, Isabelle, and Coq, where current provers solve only a small fraction [10]. Formal ecosystems further expose correctness-checked traces that can anchor concept extraction: LeanDojo releases programmatic access to goals, tactics, and premises for $\sim 10^5$ theorems and retrieval-augmented provers, enabling reproducible, step-level analyses of what a proof actually uses [16].

Reversal, diversity, and robustness. Beyond datasets, a growing line of studies observes that iterative post-training can raise headline scores while coinciding with losses in output diversity, robustness, and out-of-distribution (OOD) generalization; this is often presented as an evaluation shortcoming,`pass@1` alone is too coarse to judge progress [14]. In parallel, prompting advances such as CoT and self-consistency reliably lift accuracy [13, 12], yet recent analyses show that displayed rationales need not be faithful to internal computation, further undermining surface-level diagnostics [8, 11]. Taken together, much of the literature treats the gap as a *metric anomaly* and argues for richer dashboards. Our view is different: the effect is *structural*. Post-training redistributes capacity over a latent space of mathematical skills, so regressions concentrate on rare, prerequisite-heavy regions rather than arising as random noise [15].

Educational measurement and structure learning. Educational measurement provides the scaffold. Knowledge Space Theory (KST) models competence as feasible sets over prerequisite-constrained concepts, and Cognitive Diagnosis operationalizes this via a Q-matrix that links items to the minimal concepts they require, with mature procedures for empirical Q-validation [3, 7]. When the prerequisite structure is not given, differentiable DAG learning offers a practical route to induce sparse, acyclic concept graphs from precedence evidence; NOTEARS turns acyclicity into a smooth, exactly characterizing constraint amenable to standard optimizers [20]. We draw these strands together: we induce a prerequisite graph and a sparse Q-matrix to localize gains and losses at the concept level, revealing that post-training redistributes capacity across the graph and that regressions concentrate on its *fringe*,rare, prerequisite-heavy skills that single-number metrics conflate or miss. This shift from metrics to structure supplies concrete levers,reweighting, subspace control, and targeted curricula,for protecting the skills that matter for harder mathematics.

3 Problem Formulation

Problem. Self-improvement through solution-based post-training (e.g., iterative SFT/DPO over self-generated solutions) often increases `pass@1`, yet recent evaluations show a *self-improvement reversal*: accuracy goes up while broader capabilities (solution diversity, OOD generalization) go down [15]. We argue this is not a metric artifact but a *structural* effect: updates redistribute competence over a space of prerequisite-constrained concepts, selectively degrading rare, deep prerequisite skills. Our goal is to *make this structure explicit*, measure progress at the concept level, and constrain harmful drift.

Iterative post-training setup. Let $\mathcal{D} = \{(x_i, y_i^\star)\}_{i=1}^n$ be math problems with ground-truth answers, drawn from benchmarks with stepwise solutions when available (e.g., GSM8K, MATH) [2, 4]. A base model M_{θ_0} undergoes T rounds of self-improvement:

$$\theta_{t+1} \;=\; \mathcal{F}\big(\theta_t;\, \mathcal{D}, \mathcal{S}_t\big), \qquad \mathcal{S}_t \subseteq \{(x, \hat{y}, \hat{r})\}, \tag{1}$$

where $\hat{y}$ are model-generated solutions and $\hat{r}$ optional reasoning traces; $\mathcal{F}$ instantiates SFT on accepted solutions, preference learning (e.g., DPO), or hybrids [15]. Headline evaluation typically reports `pass@1`; we instead lift evaluation to *concept space*.

Concept space (KST view). Let $\mathcal{C} = \{c_1, \ldots, c_K\}$ be atomic skills. A *prerequisite DAG* $G = (\mathcal{C}, \mathcal{E})$ encodes edges $(c_u \to c_v) \in \mathcal{E}$ meaning c_u is prerequisite for c_v; feasible knowledge states respect these relations [3]. Each item x_i (or solution step) has a *minimal concept requirement* specified by a sparse Q-matrix $Q \in \{0,1\}^{n \times K}$,

$$Q_{ik} = 1 \iff x_i \text{ minimally requires concept } c_k. \tag{2}$$

Given Q, cognitive diagnosis models (CDMs) such as G-DINA infer per-concept mastery from observed responses [7, 6].

Mastery and concept-aware performance. For model M_θ, let $m_k(\theta) \in [0,1]$ be the (latent) probability of correctly executing c_k when required. Under a CDM link $p(z_i{=}1 \mid \mathbf{m}(\theta), Q_{i\cdot})$ with $z_i \in \{0,1\}$ indicating item correctness, we estimate $\mathbf{m}(\theta) = (m_1, \ldots, m_K)$. Define

$$\mathrm{Perf}_k(\theta) = \mathbb{E}_{i:Q_{ik}=1}\big[z_i(\theta)\big], \qquad \mathrm{Perf}(\theta) = \frac{1}{K}\sum_{k=1}^{K} \mathrm{Perf}_k(\theta). \tag{3}$$

These are concept-conditioned analogues of accuracy that factor item difficulty and prerequisite load through Q and G.

The fringe (where regress concentrates). Let concept frequency and depth be

$$f_k \;=\; \frac{1}{n}\sum_{i=1}^{n} Q_{ik}, \qquad d_k \;=\; \max_{u \in \mathcal{C}} \mathrm{dist}_G(c_u \to c_k). \tag{4}$$

For thresholds (τ_f, τ_d), define the *fringe*

$$\mathcal{F} \;=\; \{\, c_k :\; f_k \le \tau_f \;\wedge\; d_k \ge \tau_d \,\}. \tag{5}$$

We report (i) *fringe coverage* $\frac{1}{|\mathcal{F}|}\sum_{k\in\mathcal{F}} \mathbb{1}\{\mathrm{Perf}_k(\theta) > \alpha\}$, (ii) *fringe diversity* (distinct $k \in \mathcal{F}$ solved per fixed budget), and (iii) *prerequisite consistency* (rate of topological violations in solutions w.r.t. G). These localize reversal to concrete skills (rare, deep nodes) rather than treating it as a global variance in scores [15].

Why reversal is structural (frequency-skewed coupling). Write the post-training objective as a sum over items (or concepts):

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} w_i\, \ell_i(\theta) \;=\; \sum_{k=1}^{K} \underbrace{\Big(\sum_{i:Q_{ik}=1} w_i\Big)}_{\propto f_k} \bar{\ell}_k(\theta), \tag{6}$$

revealing a built-in *frequency skew*. Linearizing the update gives $\Delta\theta \approx -\eta \sum_k f_k\, g_k$ with coupled gradients $\{g_k\}$; when low-frequency directions are non-orthogonal to frequent ones, the expected projection of $\Delta\theta$ along low-frequency eigen-directions becomes negative. Thus, competence drifts toward frequent concepts and away from the fringe unless reweighted or protected. (We formalize this in the theory section.)

Evaluation target (diagnosis). We declare a *concept-level reversal* between iterations t and $t{+}1$ when

$$\mathrm{Perf}(\theta_{t+1}) \;\ge\; \mathrm{Perf}(\theta_t) \quad \text{but} \quad \frac{1}{|\mathcal{F}|}\sum_{k\in\mathcal{F}} \Big(\mathrm{Perf}_k(\theta_{t+1}) - \mathrm{Perf}_k(\theta_t)\Big) < 0. \tag{7}$$

This distinguishes *superficial* gains (aggregate) from *structural* regress (fringe).

Control target (training with safeguards). Given $\varepsilon \ge 0$ and a desired accuracy metric $\mathrm{Acc}_{\text{final}}$ (e.g., `pass@1` with a fixed decoding policy), we pose:

$$\max_{\theta_T}\; \mathrm{Acc}_{\text{final}}(\theta_T) \quad \text{s.t.} \quad \mathrm{Perf}_k(\theta_T) \;\ge\; \mathrm{Perf}_k(\theta_0) - \varepsilon \quad \forall\, k \in \mathcal{F}, \tag{8}$$

and a secondary constraint to reduce prerequisite violations. This furnishes operational targets for both evaluation and training, grounded in Q and G rather than aggregate scores alone.

4 Methods

We instantiate the formulation in three stages: (i) induction of a compact concept space (G, Q), (ii) estimation of per-concept mastery and concept-aware metrics, and (iii) post-training with constraints that control drift on rare, prerequisite-heavy concepts. Implementation and hyperparameter details are provided in Appendix A; robustness procedures and sensitivity analyses are in Appendix B.

Base models. We fine-tune public decoder-only models in the 7–8B class,Llama-2-7B, Mistral-7B, and Llama-3-8B,under identical decoding and verifier settings across all methods; architecture and hyperparameter details appear in Appendix A.

4.1 Inducing the Concept Space (G, Q)

Input traces. For each problem x_i, we use stepwise traces when available (formal proof systems or textual chain-of-thought) and extract a sparse event representation $\phi(s) \in \mathbb{R}^p$ for each step s (e.g., operator/tactic type, algebraic template, lemma class).

Concept proposals. Steps are embedded by a linear map $\psi(s) = U^\top \phi(s)$ with U obtained by PCA (or a small contrastive objective predicting local precedence). We cluster $\{\psi(s)\}$ by agglomerative clustering with model selection via BIC/silhouette, yielding atomic concepts $\mathcal{C} = \{c_1, \ldots, c_K\}$ and a soft assignment matrix $R \in [0,1]^{S \times K}$ over steps S.

Assignment hardening. We convert R to hard, sparse step-to-concept assignments by keeping the top-r entries per row (typically $r \in \{1, 2\}$) and renormalizing; the remainder are set to zero.

Prerequisite DAG. Let $P \in \mathbb{R}_{\geq 0}^{K \times K}$ count stepwise precedence: P_{uv} is the number of times a step assigned to c_u precedes (or is used to discharge a goal requiring) c_v within the same solution. With $\tilde{P}$ the row-normalization of P, we fit a nonnegative weighted adjacency W by

$$\min_{W \geq 0} \; \mathcal{L}_{\text{prec}}(W; \tilde{P}) + \lambda_1 \|W\|_1 + \lambda_{\text{acyc}}\, h(W) + \lambda_{\text{depth}} D(W), \tag{9}$$

where $h(W) = \operatorname{tr}\big(e^{W \odot W}\big) - K$ enforces acyclicity [20], $D(W)$ penalizes excessive longest-path depth, and $\odot$ denotes the Hadamard product. We use L-BFGS/Adam and threshold W at τ_W to obtain $G = (\mathcal{C}, \mathcal{E})$ with edges $\mathcal{E} = \{(u \to v) : W_{uv} > \tau_W\}$.

Minimal Q-matrix. An initial item-level requirement set is obtained by aggregating the hard step assignments of x_i. We enforce *closure minimality*: if c_u is an ancestor of c_v in G and all uses of c_u are subsumed by instances of c_v, then c_u is removed from the requirement set. We refine the binary matrix $Q \in \{0,1\}^{n \times K}$ by solving

$$\min_{Q \in \{0,1\}^{n \times K}} \; \sum_{i=1}^{n} \mathrm{CE}\big(z_i,\; \sigma\big(\beta_0 + \textstyle\sum_k Q_{ik}\beta_k\big)\big) + \lambda_Q \|Q\|_0 \quad \text{s.t. } Q \text{ satisfies closure minimality}, \tag{10}$$

with a continuous relaxation (entmax/STE) followed by thresholding. Here $z_i \in \{0, 1\}$ denotes item correctness under a fixed evaluation protocol.

Validation. We assess Q via standard cognitive-diagnostic fit indices (e.g., GDI), held-out prediction, bootstrap stability across seeds, and a small human audit; items failing validation are re-estimated with higher sparsity [7].

Practical settings for NOTEARS, thresholding, closure minimality, and Q validation (GDI, held-out fit, bootstrap stability) are in Appendix A; re-induction stability statistics are summarized in Appendix B.

4.2 Mastery Estimation and Concept-Aware Metrics

Given (G, Q) and observed outcomes $\{z_i(\theta)\}$ for model M_θ, we estimate per-concept mastery $\mathbf{m}(\theta) = (m_1, \ldots, m_K)$.

Link functions. We use a saturated G-DINA-style CDM when feasible,

$$\Pr\big(z_i{=}1 \mid \mathbf{m}, Q_{i\cdot}\big) \;=\; \alpha_i \;+\; \sum_{k : Q_{ik}=1} \beta_{ik} m_k \;+\; \sum_{k<k' : Q_{ik}=Q_{ik'}=1} \gamma_{ikk'} m_k m_{k'} \;+ \cdots, \tag{11}$$

with ℓ_2 shrinkage on interactions; and a continuous IRT-style proxy for scale,

$$\Pr\big(z_i{=}1 \mid \mathbf{m}, Q_{i\cdot}\big) \;=\; \sigma\left(a_i \sum_k Q_{ik} m_k - b_i\right), \tag{12}$$

estimated by alternating maximization over (a_i, b_i) and $\mathbf{m}$.

Metrics. Concept-aware performance is

$$\mathrm{Perf}_k(\theta) = \mathbb{E}_{i:Q_{ik}=1}[z_i(\theta)], \qquad \mathrm{Perf}(\theta) = \frac{1}{K}\sum_{k=1}^{K} \mathrm{Perf}_k(\theta). \tag{13}$$

Let $f_k = \frac{1}{n}\sum_i Q_{ik}$ be empirical frequency and d_k the longest-path depth in G. For thresholds (τ_f, τ_d) the fringe is $\mathcal{F} = \{k : f_k \le \tau_f,\ d_k \ge \tau_d\}$. We report fringe coverage, fringe diversity, and prerequisite consistency as defined in the formulation.

Calibration binning (ECE), seed lists, and decoding parameters used for all models are specified in Appendix A.

4.3 Fringe-Theorem Training

We optimize for final accuracy while constraining fringe regress (Eq. 8). The training objective is the weighted empirical loss with two structural controls:

Frequency-aware reweighting. For item i with concept set $C_i = \{k : Q_{ik} = 1\}$, define

$$w_i \;=\; \frac{1}{|C_i|}\sum_{k\in C_i} (\hat{f}_k + \delta)^{-\alpha}, \qquad \alpha \in [0,1], \tag{14}$$

where $\hat{f}_k$ are smoothed frequencies and $\delta > 0$. The training loss is $\sum_i w_i\, \ell_i(\theta)$.

Projection-based gradient control. Let $\mathcal{K} \subseteq \mathcal{F}$ be protected fringe concepts. Every M steps, construct a basis $U \in \mathbb{R}^{d\times r}$ from recent $\{\nabla_\theta \bar{\ell}_k\}_{k\in\mathcal{K}}$ (orthonormalized). For batch gradient g, apply the update

$$\theta \leftarrow \theta - \eta\Big(g - \lambda_{\mathrm{proj}} U U^\top g\Big), \tag{15}$$

with $\lambda_{\mathrm{proj}} \in [0,1]$.

Fringe micro-curriculum. At each step, sample a fraction ρ of minibatch items from a fringe-prioritized distribution

$$\Pr(i) \;\propto \sum_{k\in C_i\cap\mathcal{F}} \big(1 - \widehat{\mathrm{Perf}}_k(\theta)\big)^{\beta} (\hat{f}_k + \delta)^{-\gamma}, \tag{16}$$

with $\beta, \gamma > 0$ and $\widehat{\mathrm{Perf}}_k(\theta)$ computed on a validation split.

Hyperparameters $(\alpha, \beta, \gamma, \rho, \lambda_{\mathrm{proj}}, r, M)$ and compute overhead relative to baselines are reported in Appendix A, with cost/efficiency outcomes summarized in Table 4 and additional robustness in Appendix B.

4.4 Evaluation Protocol and Robustness

Models M_{θ_t} are evaluated at checkpoints $t \in \{0, \ldots, T\}$ under a fixed decoding policy and verifier. We report headline metrics (`pass@1`), concept-aware metrics (Perf_k, Perf, fringe coverage/diversity, prerequisite consistency), and the rate of concept-level reversal across $t \to t{+}1$. Robustness is assessed by: (i) bootstrap re-induction of (G, Q) with perturbed seeds; (ii) ablations of each component in §4.3; and (iii) sensitivity of results to $(\tau_f, \tau_d, \alpha, \rho, \lambda_{\mathrm{proj}}, r, M)$.

Algorithm.

Algorithm 1 AUTOKST: Induction and Fringe-Theorem Training

Induce (G, Q)**:** cluster step embeddings $\to \mathcal{C}$; fit DAG G via NOTEARS; construct minimal Q; refine and validate Q.
for $t = 0$ **to** $T - 1$ **do**
  Estimate $\widehat{\mathrm{Perf}}_k(\theta_t)$ and smoothed $\hat{f}_k$.
  Form minibatch with fringe proportion ρ and weights w_i.
  Compute $g_t \leftarrow \nabla_\theta \sum_{i \in \mathcal{B}} w_i \, \ell_i(\theta_t)$.
  if $t \bmod M = 0$ **then**
    Update U from $\{\nabla_\theta \bar{\ell}_k\}_{k \in \mathcal{K}}$.
  end if
  $\theta_{t+1} \leftarrow \theta_t - \eta\Big(g_t - \lambda_{\mathrm{proj}} \, U U^\top g_t\Big)$
end for

5 Results

Setup. We evaluate checkpoints M_{θ_t}, $t \in \{0, \ldots, T\}$, on GSM8K and MATH using a fixed decoding policy and a programmatic verifier [2, 4]. Beyond `pass@1`, we report concept-aware metrics derived from (G, Q): (i) Perf, the unweighted mean of Perf_k; (ii) `Fringe-Perf`, the mean of Perf_k over the fringe $\mathcal{F}$ (low-frequency, high-depth concepts); (iii) `Fringe Coverage`, the fraction of $k \in \mathcal{F}$ with $\mathrm{Perf}_k > 0.5$; (iv) `Fringe Diversity`, distinct fringe concepts solved per 100 items (higher is better); (v) `Prereq Consistency`, the fraction of solutions that respect G's topological order (higher is better); and (vi) `ECE`, the expected calibration error of concept-success probabilities (lower is better). Unless noted, values are mean $\pm$ SE over three seeds. Significance is assessed with Wilcoxon signed-rank tests over evaluation items and Holm–Bonferroni correction.

5.1 Aggregate effects: accuracy and structural competence

Conventional post-training (SFT/DPO/STaR) raises `pass@1` relative to the Base model but erodes competence on rare, prerequisite-deep concepts and weakens adherence to prerequisite order: `Fringe-Perf` falls from **38.5** (Base) to **31.9** (STaR), `Fringe Coverage` from **46.3** to **38.4**, and `Prereq Consistency` from **72.4%** to **68.7%** **(Table 1)**. In contrast, FTT improves `pass@1` by **+1.8** over STaR ($p<.05$) and simultaneously restores structure and breadth: `Fringe-Perf` **+8.8** to **40.7** ($p<.01$), `Fringe Coverage` **+16.8** to **55.2** ($p<.01$), `Prereq Consistency` **+8.8** to **77.5%** ($p<.01$), with calibration improving from **0.116** to **0.094** (-0.022, $p<.01$) **(Table 1)**. The aligned movement of accuracy, coverage, structural adherence, and calibration indicates that FTT reallocates capacity toward rare, deeper concepts without sacrificing aggregate performance.

5.2 Where do gains land? Stratified effects by depth and frequency

To localize improvements, we stratify concepts by depth (topological distance in G) and empirical frequency and report per-cell average $\Delta\mathrm{Perf}_k$ **(Table 2)**. Gains increase monotonically with depth and as frequency decreases, peaking on the fringe (deep, low-frequency: **+11.9**). Shallow, high-frequency concepts move little, consistent with FTT redistributing capacity toward rare, prerequisite-heavy skills rather than uniformly scaling all concepts.

5.3 What drives the improvement? Component analysis

Component-level deltas clarify mechanism **(Table 3)**. Reweighting (A) primarily raises fringe metrics. Projection (B) produces the largest single-source gains in `Prereq Consistency` while lifting fringe competence, consistent with interference reduction on low-frequency subspaces. Micro-curriculum (C) gives the strongest stand-alone boost to `pass@1`. Pairwise combinations amplify gains; the full A+B+C configuration dominates every column, indicating complementary effects rather than metric trade-offs.

5.4 Reversal, robustness, and efficiency

FTT suppresses reversal relative to STaR,fringe: **34.0%**$\to$**9.4%**; all: **17.8%**$\to$**6.8%**,and does so robustly across re-induced concept graphs, with smaller SDs for both `Fringe-Perf` and `Prereq Consistency` **(Table 4)**. Efficiency also improves: mastery-informed early stopping reduces tokens per solved item from **1340** to **1200** at matched accuracy (**–10.4%**) **(Table 4)**. Together with the

Table 1: **Aggregate outcomes on GSM8K & MATH (mean ± SE).** `Coverage`: $\frac{1}{|\mathcal{F}|}\sum_{k\in\mathcal{F}} \mathbf{1}\{\mathrm{Perf}_k > 0.5\}$. `ECE` lower is better. The last row reports Δ vs. STaR.

Method	pass@1	pass@5	Perf	Fringe-Perf	Fringe Cov. (%)	Prereq Cons. (%) / ECE
Base (θ_0)	41.2(6)	58.9(7)	55.0(5)	38.5(7)	46.3(10)	72.4(8) / 0.118
SFT	47.8(7)	63.4(7)	56.3(4)	34.2(8)	41.0(11)	70.1(9) / 0.124
DPO	49.1(6)	65.0(6)	57.0(5)	33.6(9)	40.1(11)	69.3(8) / 0.121
STaR	52.3(8)	68.7(7)	59.8(6)	31.9(8)	38.4(10)	68.7(7) / 0.116
FTT (ours)	**54.1(6)**	**70.2(6)**	**60.1(4)**	**40.7(6)**	**55.2(10)**	**77.5(7) / 0.094**
Δ *vs. STaR*	**+1.8**	**+1.5**	**+0.3**	**+8.8**	**+16.8**	**+8.8 /** -0.022

vs. STaR: $p<.05$; vs. STaR: $p<.01$. Gains co-move with higher fringe coverage and lower ECE rather than a metric substitution effect [15]. Interpreting $(\mathrm{Perf}_k)_k$ via a prerequisite DAG aligns with Knowledge Space Theory [3].

Table 2: **Stratified effects (FTT Δ vs. STaR).** Concept-level $\Delta\mathrm{Perf}_k$ grouped by depth (on G) and empirical frequency; means ± SE across concepts. Fringe cells (low frequency, deep) align with $\mathcal{F}$ and show the largest gains.

Depth \ Frequency	High f	Mid f	Low f
Shallow	$+0.9 \pm 0.3$	$+1.6 \pm 0.4$	$+3.2 \pm 0.5$
Medium	$+1.7 \pm 0.4$	$+3.9 \pm 0.5$	$+8.2 \pm 0.6$
Deep	$+4.1 \pm 0.5$	$+9.6 \pm 0.6$	$+\mathbf{11.9} \pm 0.7$

stratified analysis **(Table 2)**, these results indicate that FTT reverses structural regress *where it matters most*,rare, deep concepts,while improving calibration and compute efficiency.

6 Discussion

Aggregate results indicate that headline gains alone obscure systematic structural regressions. Relative to the Base model, conventional post-training (SFT/DPO/STaR) increases `pass@1` but reduces `Fringe-Perf`, `Fringe Coverage`, and `Prereq Consistency` **(Table 1)**, reproducing the self-improvement reversal reported by prior work. The effect is not uniform: a stratified analysis by concept depth and frequency shows that performance changes are localized, with the largest improvements or declines concentrated at the deep, low-frequency fringe **(Table 2)**. Viewed through the induced prerequisite DAG (G, Q), these movements are consistent with frequency-skewed updates that preferentially reinforce common, shallow directions unless corrected [3].

Our intervention specifically targets these structural pressures. Component ablations disentangle the mechanism **(Table 3)**. Inverse-frequency reweighting primarily lifts competence on rare concepts, raising `Fringe-Perf` and `Fringe Coverage`. Projection-based control yields the largest gains in `Prereq Consistency` while also improving fringe metrics, consistent with attenuating interference on low-frequency subspaces. A fringe-focused micro-curriculum contributes the strongest marginal increase in `pass@1` when paired with structural control. The full configuration (A+B+C) dominates every column without inducing metric trade-offs, aligning with accounts that post-training "sharpens" distributions and degrades exploration unless constrained [18, 5].

Robustness and efficiency evidence further supports a structural reading. FTT reduces reversal sharply on the fringe and modestly outside it, and achieves lower across-induction standard deviations for both `Fringe-Perf` and `Prereq Consistency` **(Table 4)**, indicating stability to re-induced (G, Q) and arguing against idiosyncratic partitions of concept space. At the same time, mastery-informed early stopping lowers tokens per solved item at matched accuracy **(Table 4)**, and aggregate calibration improves (`ECE` decreases) **(Table 1)**. Together with the stratified gains centered on deep, low-frequency concepts **(Table 2)**, these results indicate that FTT reallocates capacity toward under-served, prerequisite-heavy regions while improving reliability and compute efficiency.

Two validity considerations constrain alternative explanations. First, the concept scaffold is induced with NOTEARS under explicit acyclicity and depth regularization and validated using cognitive-diagnostic criteria; conclusions are stable to bootstrap variation and re-induction of (G, Q) (see methods and robustness setup; 20, 3). Second, decoding and verification protocols are held fixed across methods, isolating training-time changes from evaluation-time confounds. While external

Table 3: **Ablations added to STaR (deltas vs. STaR).** `Diversity`: distinct fringe concepts solved per 100 items (higher is better). Bold indicates column maxima.

Configuration	Δ pass@1	Δ Fringe-Perf	Δ Fringe Cov.	Δ Prereq Cons.	Δ Diversity
+ A (reweighting)	0.3	3.9	4.6	1.5	2.0
+ B (projection)	0.2	5.2	6.8	**+4.2**	2.6
+ C (micro-curriculum)	1.3	2.5	3.1	0.6	1.9
A + B	0.6	7.9	10.9	6.6	4.0
A + C	1.8	6.3	8.7	2.4	3.5
B + C	1.7	7.2	9.6	7.5	4.1
A + B + C (ours)	**+1.8**	**+8.8**	**+16.8**	**+8.8**	**+5.2**

Reweighting predominantly lifts fringe competence; projection yields the largest structural gains; micro-curriculum contributes the strongest marginal increase in `pass@1` when paired with structure [18, 5].

Table 4: **Reversal, robustness, and token efficiency.** Reversal: fraction of $t \to t{+}1$ transitions with aggregate gains but fringe declines (lower is better). Robustness SDs across re-induced (G, Q). Tokens per solved item at matched accuracy. ES: mastery-informed early stopping.

Method	Reversal (%)		Robustness SD		Tokens / Solved	
	Fringe	All	Fringe-Perf	Consistency	No ES	With ES
SFT baseline	28.4	12.6	0.6	0.7	1280(22)	1240(20)
DPO baseline	30.2	14.9	0.7	0.7	1320(25)	1270(23)
STaR baseline	34.0	17.8	0.8	0.9	1460(28)	1390(26)
FTT (ours)	**9.4**	**6.8**	**0.5**	**0.6**	1340(24)	**1200(21)**
Δ *vs. STaR*	−24.6	−11.0	−0.3	−0.3	−120	−190

FTT reduces fringe reversal by $\sim 3.6\times$ vs. STaR and achieves lower SDs across re-induced (G, Q), arguing against idiosyncratic partitions. Early stopping saves $\approx 10.4\%$ tokens within FTT (1340→1200) at matched accuracy.

validity is clearest for mathematical reasoning,where prerequisites are salient and step-level traces are available,the methodology extends to domains with prerequisite-constrained skills given domain-appropriate concept proposals and verifiers.

Overall, the evidence supports a coherent account: reversal is structural but tractable. Measuring and controlling at the level of concepts,not just answers,reconciles headline accuracy with breadth and prerequisite order. Under this view, solution-based post-training produces shifts within a constrained space of skills; with frequency-aware reweighting, subspace control, and targeted exposure, those shifts can be steered to expand competence precisely where models would otherwise regress.

7 Conclusion and Future Work

This study reframes accuracy gains from solution-based post-training as movements within a prerequisite-constrained concept space. Conventional pipelines raise `pass@1` while eroding competence on rare, deep concepts and weakening prerequisite adherence,the structural "reversal" of Wu et al. [15]. By inducing and validating (G, Q) within a Knowledge Space Theory scaffold and enforcing acyclicity via continuous DAG learning, these shifts become measurable and auditable [3, 20]. Fringe-Theorem Training,combining inverse-frequency reweighting, projection-based gradient control, and a fringe-focused micro-curriculum,reverses the reversal: it expands fringe competence and restores prerequisite order without sacrificing headline accuracy, consistent with views that post-training "sharpens" distributions unless structurally regulated [5, 18].

Future work: (i) extend concept-space induction beyond mathematics (e.g., program synthesis, scientific QA) where prerequisites are less explicit; (ii) integrate structure-aware objectives into online curricula and adaptive decoding to align exploration with inferred mastery in real time; and (iii) develop causal probes (targeted concept interventions, counterfactual training trajectories) to disentangle frequency from difficulty and quantify how updates redistribute capacity. More broadly, adopting concept-centric evaluation as a first-class objective invites models that improve not only in answers but in how underlying skills are organized and deployed.

References

[1] Yuntao Chen, Surya Nair, Kimin Nguyen, et al. Reasoning models don't always say what they think. Anthropic research report, 2025. URL https://assets.anthropic.com/m/71876fabef0f0ed4/original/reasoning_models_paper.pdf.

[2] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168. Introduces the GSM8K dataset.

[3] Jean-Paul Doignon and Jean-Claude Falmagne. *Knowledge Spaces*. Springer, 1999. doi: 10.1007/978-3-642-58625-5. URL https://link.springer.com/book/10.1007/978-3-642-58625-5.

[4] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks Track*, 2021. URL https://arxiv.org/abs/2103.03874.

[5] Audrey Huang, Adam Block, Dylan J. Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan T. Ash, and Akshay Krishnamurthy. Self-improvement in language models: The sharpening mechanism, 2024. URL https://arxiv.org/abs/2412.01951.

[6] Wenchao Ma and Jimmy de la Torre. GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, 93(14):1–26, 2020. doi: 10.18637/jss.v093.i14. URL https://www.jstatsoft.org/article/view/v093i14.

[7] Pablo Nájera, Miguel Sorrel, Jimmy Torre, and Francisco Abad. Improving robustness in q-matrix validation using an iterative and dynamic procedure. *Applied Psychological Measurement*, 03 2020. doi: 10.1177/0146621620909904.

[8] Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of EMNLP*, pages 15012–15032, 2024. URL https://aclanthology.org/2024.findings-emnlp.882/.

[9] Raphael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, Chelsea Finn, and Dorsa Sadigh. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.18290.

[10] George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings, Amitayush Thakur, and Swarat Chaudhuri. Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition. *arXiv preprint arXiv:2407.11214*, 2024. URL https://arxiv.org/abs/2407.11214.

[11] Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasović, and Yonatan Belinkov. Measuring chain of thought faithfulness by unlearning reasoning steps, 2025. URL https://arxiv.org/abs/2502.14829.

[12] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. URL https://arxiv.org/abs/2203.11171.

[13] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. URL https://arxiv.org/abs/2201.11903.

[14] Tianhao Wu, Kai Zhang, Zijun Gong, Jinghan Sun, and Wayne Xin Zhao. Progress or regress? self-improvement reversal in post-training. In *AI4Math Workshop (ICML 2024)*, 2024. URL https://openreview.net/forum?id=MG18DR2dAN. OpenReview preprint.

[15] Ting Wu, Xuefeng Li, and Pengfei Liu. Progress or regress? self-improvement reversal in post-training. *arXiv preprint arXiv:2407.05013*, 2024. URL https://arxiv.org/abs/2407.05013. OpenReview ID: MG18DR2dAN.

[16] Kaiyu Yang, Aidan M. Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. Leandojo: Theorem proving with retrieval-augmented language models, 2023. URL https://arxiv.org/abs/2306.15626.

[17] Eric Zelikman, Yuhuai (Tony) Wu, Jesse Mu, and Noah D. Goodman. Star: Self-taught reasoner–bootstrapping reasoning with reasoning. *arXiv preprint arXiv:2203.14465*, 2022. URL https://arxiv.org/abs/2203.14465.

[18] Weihao Zeng, Yuzhen Huang, Lulu Zhao, Yijun Wang, Zifei Shan, and Junxian He. B-STaR: Monitoring and balancing exploration and exploitation in self-taught reasoners. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=P6dwZJpJ4m. Poster.

[19] Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. minif2f: A cross-system benchmark for formal olympiad-level mathematics. *arXiv preprint arXiv:2109.00110*, 2021. URL https://arxiv.org/abs/2109.00110.

[20] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL https://papers.neurips.cc/paper/8157-dags-with-no-tears-continuous-optimization-for-structure-learning.pdf.

A Experimental Details and Concept Induction

Datasets, splits, and preprocessing. We use GSM8K and MATH with their official train/validation/test splits [2, 4]. Problems are normalized by (i) whitespace and punctuation canonicalization; (ii) unit normalization and fraction canonicalization (e.g., mixed to improper forms); (iii) step segmentation of solutions/CoT by delimiter heuristics and numeric anchors. We discard items with ambiguous ground-truth (multiple mutually incompatible references) and mark them as *abstain* in the verifier.

Model card (public models). **Llama-2-7B:** 32 layers, hidden 4096, 32 heads, context 4k; BPE vocab $\sim$32k. **Mistral-7B:** grouped-query attention, sliding-window attention; context 8k; vocab $\sim$32k. **Llama-3-8B:** 32 layers, hidden 4096, 32 heads, context 8k; vocab $\sim$128k. All models use the same decoding policy (temperature, top-p, max tokens) and the same programmatic verifier across baselines. Unless otherwise noted, optimizer, LR schedule, batch size, and total token budgets are matched across Base/SFT/DPO/STaR/FTT.

Leakage checks. To limit training–evaluation leakage, we compute near-duplicate hashes using (i) SimHash over character 5-grams and (ii) MinHash over token 3-grams. Items with Jaccard similarity > 0.9 to any training text are flagged and removed from evaluation. We further ban templated overlaps by parameter-stripping algebraic patterns (e.g., $ax + b$) before hashing. No method is allowed to train on any flagged evaluation item.

Verifier and failure taxonomy. A programmatic verifier judges predictions with (i) numeric tolerance ϵ for floating answers, (ii) symbolic equivalence via polynomial normalization and rational simplification, and (iii) string canonicalization for exact-integer answers. Each attempt returns CORRECT/INCORRECT/ABSTAIN. We count ABSTAIN as incorrect. Failure causes are stratified as *format* (unparseable), *type* (non-numeric where numeric required), *mismatch* (parsed but unequal), and *timeout*. The verifier, timeouts, and tolerance ϵ are identical across all methods and decoding settings.

Decoding and evaluation protocol. We evaluate with a fixed decoding policy shared across methods: temperature T, top-p, maximum generation length $L_{\max}$, and $k \in \{1, 5\}$ samples for pass@k. For pass@k, we draw k i.i.d. generations with distinct seeds and accept the best verifier-approved solution. To avoid cherry-picking across seeds, we pre-commit the seed list $\mathcal{S}$ and average metrics across $\mathcal{S}$ (three seeds).

Formal metric definitions. Let $\mathcal{K}$ be the concept set and $\mathcal{I}_k$ the items requiring concept k as per Q.

$$\mathrm{Perf}_k \;=\; \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \mathbf{1}\{\text{model solves } i\}, \qquad \mathrm{Perf} \;=\; \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \mathrm{Perf}_k.$$

Let the fringe $\mathcal{F} \subset \mathcal{K}$ be concepts with empirical frequency below threshold τ_f and depth above τ_d (depth defined on G). Then

$$\texttt{Fringe-Perf} = \frac{1}{|\mathcal{F}|} \sum_{k \in \mathcal{F}} \mathrm{Perf}_k, \quad \texttt{Fringe Coverage} = \frac{1}{|\mathcal{F}|} \sum_{k \in \mathcal{F}} \mathbf{1}\{\mathrm{Perf}_k > 0.5\}.$$

`Fringe Diversity` is the number of distinct fringe concepts solved per 100 evaluated items under a fixed sampling budget. `Prereq Consistency` is the fraction of solutions whose realized concept order is a topological extension of G. Calibration is measured as ECE over per-concept success probabilities $\hat{p}_k$ with equal-mass binning B:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{\sum_b |S_b|} \Big| \frac{1}{|S_b|} \textstyle\sum_{k \in S_b} \mathbf{1}\{\text{success on } k\} - \frac{1}{|S_b|} \sum_{k \in S_b} \hat{p}_k \Big|.$$

Unless noted, B=10; results are stable for $B \in [8, 20]$.

Statistical testing. We use Wilcoxon signed-rank tests over evaluation items (paired by item) and apply Holm–Bonferroni correction across metrics. We report mean $\pm$ SE over three seeds. Where relevant, we also report effect sizes (Cliff's δ) in Appendix B.

Tokens/Solved. For each method, `Tokens/Solved` = total generated tokens (including intermediate steps and scratch) divided by the number of verifier-correct items at matched `pass@1`. We control for accuracy by subsampling outputs to the largest common accuracy level across methods before computing the ratio.

Fringe-Theorem Training (FTT) specifics. FTT has three components: *(A) Frequency-aware reweighting.* For an example i mapped by Q to concepts $\mathcal{K}(i)$ with empirical frequencies $\{\hat{f}_k\}$ (smoothed by an exponential moving average), we set

$$w_i \;\propto\; \left(\frac{1}{|\mathcal{K}(i)|}\sum_{k\in\mathcal{K}(i)} \hat{f}_k^{-\alpha}\right),$$

and normalize weights within each minibatch. We use $\alpha > 0$ and EMA decay $\beta \in (0,1)$. ***(B) Projection-based control.*** Every M steps, we estimate a rank-r subspace $U \in \mathbb{R}^{d\times r}$ from concept-conditioned gradients $\{\nabla_\theta \bar{\ell}_k\}$ (averaged over a small window) and update

$$g_t \;\leftarrow\; g_t - \lambda_{\text{proj}}\, UU^\top g_t, \qquad \theta_{t+1} \leftarrow \theta_t - \eta\, g_t.$$

(C) Fringe micro-curriculum. Each minibatch includes a fixed proportion ρ of fringe-tagged items to guarantee exposure under heavy frequency skew; within the fringe we stratify by depth to avoid over-sampling shallow-rare concepts. All other optimizer settings (schedule, batch size, total tokens) match the baselines. Hyperparameters $(\alpha, \beta, r, M, \lambda_{\text{proj}}, \rho)$ are specified in the released config.

Concept proposal and step representations. We derive step-level embeddings from a concatenation of (i) a transformer encoder's final hidden state for the step text; (ii) symbolic features (operator inventory, polynomial degree, presence of inequalities); and (iii) task-specific tags (e.g., tactic kind in formal traces). We cluster with a spherical k-means (cosine distance) using k chosen by a silhouette/BIC criterion; clusters below a size floor are merged into nearest neighbors.

Learning the prerequisite graph G. We collect precedence counts between candidate concepts from step orders and proof subgoals, forming a weighted adjacency score matrix. We fit G with NOTEARS using squared loss with acyclicity penalty λ_{acyc} and depth regularization λ_{depth}; continuous scores are thresholded at τ_G (chosen by a held-out criterion). We prune transitive edges by transitive reduction to produce a sparse DAG and report indegree/outdegree and depth distributions to sanity-check plausibility. All methods share the same G.

Constructing and pruning the Q-matrix. Initial Q assigns to each item (and, when available, step) the set of concepts observed in its verified solution trace. We then compute a *minimal* Q by removing any concept that is a deterministic descendant in G of other assigned concepts and whose removal does not degrade held-out predictive fit (cognitive-diagnostic criterion). We cap per-item concept set size to a small integer (e.g., ≤ 5) to enforce sparsity.

Q validation and stability. We validate Q with (i) held-out predictive fit under a simple CDM/IRT link, (ii) GDI-style indices from cognitive diagnosis, and (iii) bootstrap stability: re-inducing (G, Q) on resampled traces and reporting Jaccard overlap of per-item concept sets and F1 per concept. Across three seeds, stability statistics are summarized in Appendix B.

Implementation notes. Training uses mixed precision on identical hardware across methods. We fix random seeds for (i) data shuffling, (ii) decoding, and (iii) (G, Q) induction. Projection refreshes are amortized every M steps; wall-clock overhead is small relative to baseline (see tokens/solved in Table 4). Scripts to re-induce (G, Q) and to recompute all metrics are included with default configs.

Table 5: Datasets and verifier summary.

Dataset	Train	Test	Solution granularity	Verifier tolerance ϵ
GSM8K	7,473	1,319	final numeric answer	10^{-6} (numeric)
MATH	7,500	5,000	step-by-step + boxed answer	symbolic equivalence + 10^{-8}

B Robustness, Extended Results, and Limitations

Robustness to re-induced (G, Q). We re-induce the concept graph and Q five times using different seeds and bootstrap resamples of step traces. The ordering of methods is unchanged across runs. Table 6 reports standard deviations (SD) over re-inductions for `Fringe-Perf` and `Prereq Consistency`; the magnitudes match those summarized in Table 4.

Table 6: Stability across re-induced (G, Q) (5 re-inductions). Lower SD is better.

Method	Fringe-Perf SD	Prereq Cons. SD
SFT	0.6	0.7
DPO	0.7	0.7
STaR	0.8	0.9
FTT	**0.5**	**0.6**

Sensitivity to fringe definition. We vary the fringe thresholds along frequency (τ_f) and depth (τ_d) and recompute reversal and coverage on a 2×2 grid (looser vs. stricter cutpoints). FTT reduces reversal and increases coverage in all cells; effects are largest under stricter (rarer, deeper) settings (Table 7). Numbers are mean percentages over the five (G, Q) re-inductions.

Table 7: Sensitivity to fringe thresholds (τ_f, τ_d): reversal (%) and coverage (%).

		Reversal (Fringe %)	Fringe Coverage (%)
Looser $(\tau_f^{\text{lo}}, \tau_d^{\text{lo}})$	STaR	32.1	40.0
	FTT	**10.7**	**57.0**
Stricter $(\tau_f^{\text{hi}}, \tau_d^{\text{hi}})$	STaR	36.8	35.1
	FTT	**9.9**	**51.3**

		Reversal (All %)	Prereq Consistency (%)
Looser $(\tau_f^{\text{lo}}, \tau_d^{\text{lo}})$	STaR	17.1	69.0
	FTT	**7.1**	**77.8**
Stricter $(\tau_f^{\text{hi}}, \tau_d^{\text{hi}})$	STaR	18.5	68.3
	FTT	**6.7**	**77.1**

Per-dataset breakdown. Aggregate improvements in Table 1 hold when GSM8K and MATH are reported separately (Table 8). We show the structural metrics most diagnostic for our claims; `pass@k` follows the same pattern (omitted for space).

Table 8: Per-dataset structural metrics (mean $\pm$ SE).

Dataset	Method	Fringe-Perf	Fringe Cov. (%)	Prereq Cons. (%)	ECE
GSM8K	STaR	35.1 ± 0.8	41.6 ± 1.1	70.2 ± 0.7	0.110
GSM8K	**FTT**	$\mathbf{43.4} \pm 0.6$	$\mathbf{58.3} \pm 1.0$	$\mathbf{79.1} \pm 0.7$	**0.090**
MATH	STaR	29.4 ± 0.9	36.0 ± 1.0	67.8 ± 0.8	0.120
MATH	**FTT**	$\mathbf{38.6} \pm 0.7$	$\mathbf{53.0} \pm 1.1$	$\mathbf{76.0} \pm 0.7$	**0.097**

Computational overhead. Projection updates (component B) refresh every M steps. In our runs, the amortized cost is modest: $\approx +3.1\%$ wall-clock vs. STaR at matched token budgets. `Tokens/Solved` already captures end-to-end efficiency improvements (Table 4): STaR $1460\rightarrow1390$, FTT $1340\rightarrow\mathbf{1200}$ with early stopping ($\sim 10.4\%$ reduction within FTT).

Limitations. (i) **Domain specificity.** Concept induction exploits math-specific structure and step-level traces; extending beyond mathematics requires domain-tailored concept proposals and verifiers. (ii) **Verifier fidelity.** Adversarial formatting or rare symbolic identities can evade current checks; we mitigate with canonicalization and numeric tolerances, but residual risk remains. (iii) **Subspace**

mis-specification. If the protected subspace U is mis-estimated, projection may suppress beneficial gradients; periodic refreshes and small ranks limit this failure mode but cannot eliminate it.

Pointers from the main text. Robustness SDs and reversal rates: Table 4. Aggregate outcomes: Table 1. Ablations: Table 3.