Improving Mathematical Reasoning via Latent Visitation Distribution Alignment

LaViDA: A Reward-Conditional Latent Auxiliary Loss for GRPO

Pritish Saha (22MF3IM15)

Advisor: **Prof. Pawan Goyal**, Dept. of Computer Science & Engineering

Department of Mechanical Engineering

B.Tech Project – II (ME47602)

April 29, 2026

Motivation and Method

Research Journey

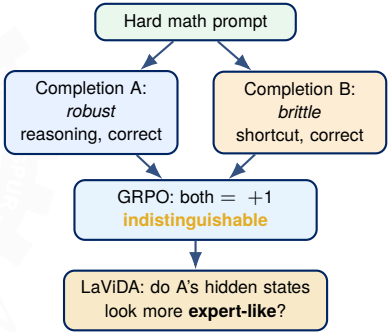
Headline Results

Discussion

Motivation: What Outcome-Only GRPO Leaves Behind

Final-answer RL rewards correct outcomes but discards the reasoning trajectory.

- GRPO reward is binary, derived from final-answer exact-match.
- It cannot distinguish a **robust** solution from a **lucky** one when both end correct.
- On hard math (MATH L4–5), correct samples are rare – each one encodes a precious reasoning strategy.
- **Hypothesis**: hidden-state trajectories of correct solutions carry reusable, plan-like reasoning information.



Question

Can we add a second training signal that shapes *how* the model reasons, not just *whether* the answer is right?

The LaViDA Hypothesis

A reward-gated latent alignment signal added on top of GRPO.

- For each generated solution, build a compact **hidden-state trajectory summary**.
- Project it through a **frozen** encoder into a low-dimensional latent space.
- For **correct** solutions only ($R_i = 1$), apply an auxiliary alignment loss pulling latents toward an **expert distribution**.
- GRPO remains the primary signal; LaViDA is a small ($\alpha \ll 1$) shaping bonus.

Composite Objective

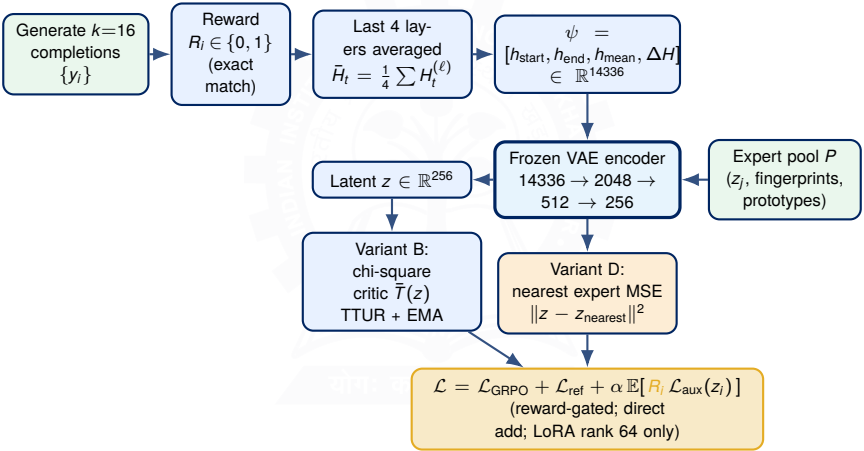
$$\mathcal{L}(\theta) = \underbrace{\mathcal{L}_{\text{GRPO}}(\theta)}_{\text{outcome RL}} + \underbrace{\mathcal{L}_{\text{ref}}(\theta)}_{\text{KL to } \pi_0} + \alpha \underbrace{\mathcal{L}_{\text{aux}}(\theta)}_{\text{latent alignment}}$$

- **Variant B** (chi-square critic): $\mathcal{L}_{\text{aux}} = \mathbb{E}\left[R_i \, \overline{T}(z_i)\right]$ with critic $\overline{T} \to q/p - 1$.
- **Variant D** (nearest-expert MSE): $\mathcal{L}_{\text{aux}} = \mathbb{E}\left[R_i \, \|z_i - z_{\text{nearest}}\|^2\right]$.

Note: the proposal used PCGrad blending; the live Oracle trainer directly adds the reward-gated auxiliary term.

Method Architecture

Hidden states → compact feature → frozen latent → alignment objective.



Frozen modules: VAE encoder, expert pool P, reference policy π_0. Trainable: LoRA adapters on the policy + (variant B) critic MLP.

Experimental Matrix

Each run answers a specific scientific question.

Run	Mechanism	Pool / data	α	Role
`A_2000`	GRPO only	none (rollout only)	0	strong baseline
`B_OracleAug`	chi-square critic	Oracle-aug. v2 (12,317)	0.2	original LaViDA mechanism
`D_OracleAug`	nearest-expert MSE	Oracle-aug. v2 (12,317)	0.001	**best RL/alignment**
`B_SelfDistill`	chi-square critic	self pool (8,963)	0.02	attribution / control
`SFT_Oracle`	direct MLE	filtered Oracle (3,354)	n/a	strong direct-MLE

- Same base policy: `Qwen2.5-Math-7B` + `cot-4shot` prompt family.
- Identical RL recipe across A/B/D: LoRA rank 64, LR 1×10^{-5}, cosine, warmup 0.03, $k=16$, max completion 3072 tokens, 2000 steps.
- Oracle-augmented runs differ only by expert pool / VAE; rollout prompt distribution stays fixed.
- Alpha values pre-locked on a fixed 100-problem hard subset before the seed-0 matrix.

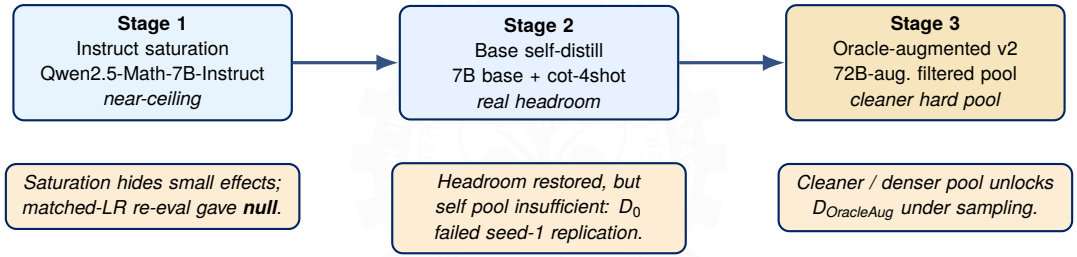
Motivation and Method

Research Journey

Headline Results

Discussion

Research Journey: Three Stages, Each Sharpening the Question



- Each pivot was triggered by a specific failure mode of the previous stage.
- Negative results are *not wasted* – they narrow the hypothesis space.
- Stage 3 is the locked seed-0 result ledger driving this presentation.

Stage 1: Instruct Arm Was Saturated

The original “+1pp” did not survive better evaluation.

- `Qwen2.5-Math-7B-Instruct` baseline: **83.2%** greedy MATH-500.
- Deterministic greedy + n=8 re-evaluation showed A and B clustered **82 – 84%**.
- Original Variant A vs. B used *different LR* (2e−5 vs. 5e−6); a methodological bug.
- After matched-LR rerun, no statistically convincing $B > A$ on any slice, including pre-registered L4–5.
- **Conclusion**: instruct arm is a saturation-control / null result.

Run (instruct)	Greedy	n=8 mean
baseline	83.2%	82.88%
A seed 0 (5e−6)	84.0%	82.80%
B seed 0 step 200	83.2%	82.30%
B seed 1 step 500	82.8%	82.62%
B seed 2 step 500	84.0%	83.10%

Paired bootstrap deltas ± 0.3pp, all CIs cross zero.

Protocol corrections used later: matched LR, deterministic greedy, multi-sample paired stats.

Stage 2: Base Self-Distill Pivot

Headroom restored, but the self-generated pool became the ceiling.

- Wrong prompt path gave 7%; `cot-4shot` smoke recovered to 72%.
- Full base calibration: **59.0%** greedy, L4–5 **42.75%**.
- Self pool: **2,212** productive prompts, **8,963** correct traces.
- Self-distill D_0 promising (61.8% greedy, 58.63% n=8).
- Matched seed-1 replication: A_1 (67.2%/62.10%) *beat* D_1 (62.8%/56.33%).

Seed 0 (self-distill)

	Greedy	n=8 mean
A_0 @ 1e−5	61.2%	55.83%
B_{002} (α=0.02)	60.6%	56.85%
C_0 (dyn proj.)	60.0%	55.33%
D_0 (MSE, α=0.002)	**61.8%**	**58.63%**

Seed 1 replication

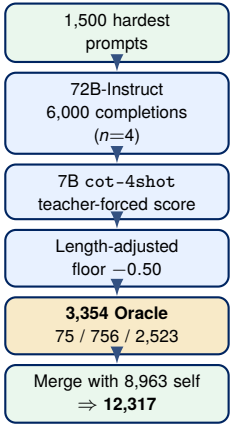
	Greedy	n=8 mean
A_1	**67.2%**	**62.10%**
D_1	62.8%	56.33%

Lesson: Self pool was insufficient; pool quality became the bottleneck ⇒ Oracle augmentation.

Oracle-Augmented Pool Pipeline

Stronger off-policy teacher, but the 7B latent manifold stays fixed.

- 72B-Instruct generated **6,000** completions from the hardest **1,500** productive prompts.
- **Scoring, feature extraction, VAE, pool building, and RL all stay on the 7B `cot-4shot` manifold.**
- Strict p_{10}^{adj} filtering kept **0** novel traces, so the operational floor was loosened to -0.50.
- Final v2 pool: **8,963 self** + **3,354 Oracle** $=$ **12,317 traces**.



Composition

3,354 kept across **1,051** prompts: **75 novel** / **756 mid** / **2,523 high-overlap**.

Controlled Setup and Alpha Lock

Pre-registered configuration; tuning never touched the final test.

- **Manifold rule**: 72B is offline generator only; everything else stays on 7B + `cot-4shot`.
- **Fixed 100-problem hard subset** for alpha micro-pilot, never touching MATH-500.
- Selected: α_B=0.2, α_D=0.001, $\alpha_{B\text{-self}}$=0.02.
- Identical RL recipe across A/B/D (α and pool/mechanism are the only differences).
- Per-run `run_manifest.json` fails-fast on configuration drift.

Knob	A	B	D
LoRA rank	64	64	64
LR	1e−5	1e−5	1e−5
Schedule	cosine	cosine	cosine
Warmup	0.03	0.03	0.03
k (group)	16	16	16
Max len	3072	3072	3072
Steps	2000	2000	2000
α	**0**	**0.2**	**0.001**
Pool	—	v2	v2

Highlighted rows are the only knobs that differ across the matrix.

Motivation and Method

Research Journey

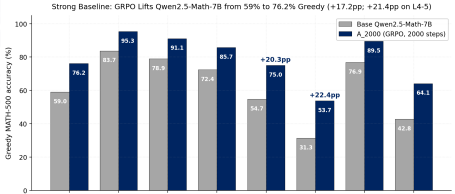
Headline Results

Discussion

Strong Baseline: A_2000 GRPO Already Moves the Model 17 pp

Plain GRPO on the cleaned base-arm setup is a serious comparator.

- Base `Qwen2.5-Math-7B`: **59.0%** greedy MATH-500.
- `A_2000`: **76.2%** greedy, **+17.2 pp** over base.
- Hard L4–5 greedy: 42.75% → **64.12%** (**+21.4 pp**).
- Level 5 greedy: 31.34% → **53.73%** (**+22.4 pp**).

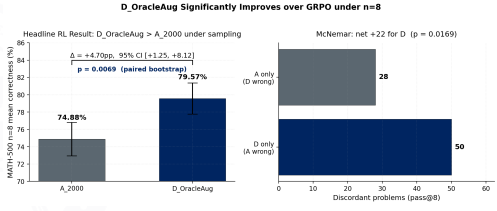


Why this matters. A_{2000} already absorbs most of the available headroom on the base arm; this is **the baseline D must beat**. Any further gain from latent alignment is therefore meaningful, not free.

Headline RL Result: D_OracleAug Beats GRPO Under Sampling

Nearest-expert MSE on the v2 pool significantly improves n=8 over GRPO.

- A_2000 n=8 mean: **74.88%**.
- D_OracleAug n=8 mean: **79.57%**.
- $\Delta = +4.70$**pp, 95% CI** $[+1.25, +8.12]$, p=0.0069 (paired bootstrap).
- pass@8: 75.8% $\rightarrow$ 80.2% (+4.40**pp**).
- McNemar discordance: A-only = 28 vs. D-only = 50, p=0.0169.



Locked claim

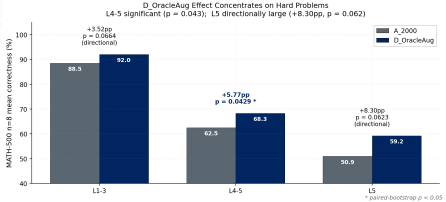
Best RL/alignment result. Statistically significant n=8 improvement over GRPO. *Greedy is tied with A; the result lives in the sampled distribution.*

The Effect Concentrates on Hard Problems

$D_{\text{OracleAug}} - A_{2000}$ paired n=8 mean correctness, by difficulty.

Slice	Δ mean	95% CI	p
L1–3	+3.52pp	[−0.11, +7.25]	0.0664
L4–5	+5.77**pp**	[+0.24, +11.35]	**0.0429***
L5	+8.30pp	[−0.47, +16.98]	0.0623
L5 pass@8	+8.96pp	—	0.0652

*Only L4–5 mean correctness clears the $p < 0.05$ threshold on seed 0.

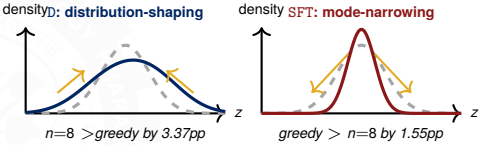


- L4–5 effect significant; L5 large *directional*.
- L1–3 directionally positive but CI crosses zero.
- Honest framing: hard-problem improvement with directional easy-slice gains.
- Level 5 has fewer problems ⇒ seed-1 replication matters.

Mechanism Read: Distribution Shaping vs. Mode Narrowing

D and SFT reach the same $n{=}8$ ceiling via *opposite* mechanisms.

Run	Greedy	$n{=}8$ mean	Direction
A_2000	76.2%	74.88%	gr. $\approx$ $n{=}8$
D_OracleAug	76.2%	79.57%	$n{=}8 >$ gr. +3.37
SFT_Oracle	81.0%	79.45%	gr. $> n{=}8$ +1.55



Cartoon densities over the latent axis; dashed = A, yellow arrows = sampling.

- **D = distribution-shaping**: more probability mass on correct trajectories.
- **SFT = mode-narrowing**: policy concentrates on one expert mode.
- Greedy-only evaluation *hides* D's effect and overstates SFT's distinctiveness.

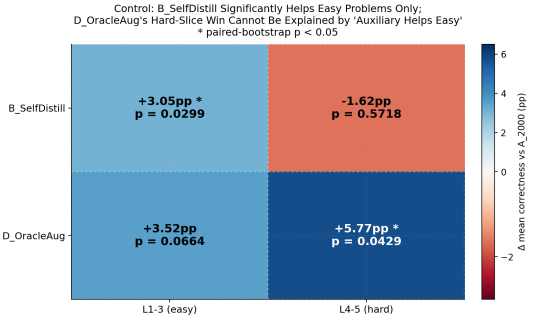
The mechanism asymmetry is itself a finding: same sampled ceiling, opposite mechanisms.

Control: B_SelfDistill Rules Out an Easy-Slice Explanation

Easy-slice gains alone do not explain the D hard-problem result.

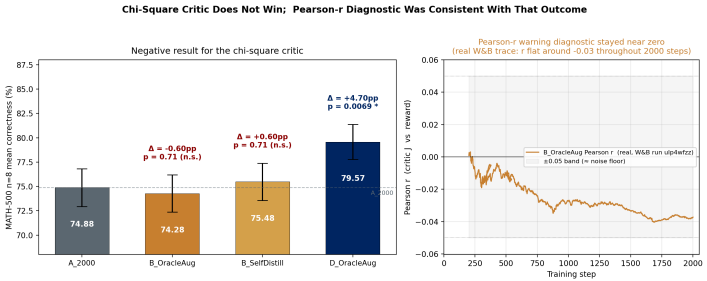
Run / slice	Δ vs. *A*	*p*
B_SelfDistill L1–3	+3.05pp	**0.0299***
B_SelfDistill L4–5	−1.62pp	0.5718
D_OracleAug L1–3	+3.52pp	0.0664
D_OracleAug L4–5	+5.77**pp**	**0.0429***

- B_SelfDistill produces only an easy-slice lift.
- It is **null** on hard subsets and overall (+0.60pp, p=0.71).
- D's hard-slice win is therefore *specific* to mechanism + Oracle-augmented pool, not "any auxiliary helps easy problems".



Chi-Square Critic: A Negative Result with a Diagnostic Warning

The original LaViDA mechanism does not win at the tested settings.



Outcome.
`B_OracleAug`: −0.60pp vs. A on n=8 mean, p=0.705.

Pattern.
pass@8 rises mildly, but mean correctness does not: diversity without reliable accuracy.

Diagnostic.
Pearson r stayed near zero, consistent with a critic objective not reward-aligned.

RL vs. SFT_Oracle: Different Methods, Same Sampled Ceiling

SFT wins greedy; D ties SFT under sampling — the gap is mechanistic.

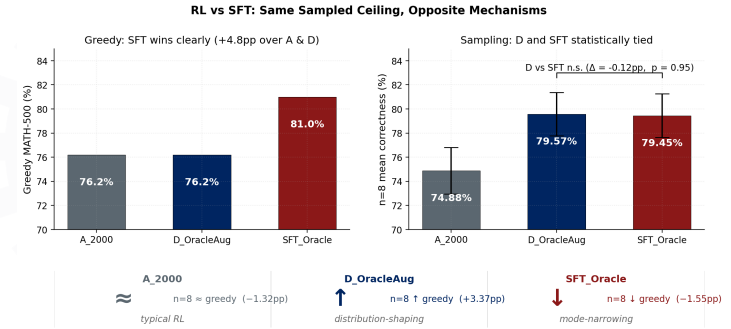
Greedy MATH-500

Run	Overall	L4–5	L5
A_2000	76.2%	64.12%	53.73%
D_OracleAug	76.2%	64.89%	58.96%
SFT_Oracle	**81.0%**	**70.61%**	**64.93%**

***n*=8 mean correctness**

Run	Overall	vs. *A*
A_2000	74.88%	—
D_OracleAug	79.57%	+4.70pp (p=0.007)
SFT_Oracle	79.45%	+4.58pp (p=0.008)

D vs. SFT under *n*=8
$\Delta = -0.12$pp, CI $[-3.30, +3.08]$, p=0.9523
pass@8 $\Delta = -0.40$pp, McNemar p=0.9036.



Honest framing

SFT is the strongest greedy baseline; D is the strongest RL/alignment branch and ties SFT under sampling without per-trace MLE supervision.

Motivation and Method

Research Journey

Headline Results

Discussion

Limitations and Caveats

Strong seed-0 evidence, not final proof.

Statistical / replication

- Single Oracle seed; seed-1 replication is next.
- L5 effect is large but *directional*, not significant.
- D-vs-SFT $n{=}8$ tie means indistinguishable, not equivalent.
- SFT wins greedy; RL does not beat direct MLE on greedy.

Method-level

- Chi-square critic may need redesign or broader α/critic sweep.
- Oracle pool is mostly on-manifold: 75 novel of 3,354.
- Mode-narrowing vs. distribution-shaping needs replication.
- Pearson *r* is a candidate warning, not a validated predictor.

Scope: single base model `Qwen2.5-Math-7B`; primary benchmark MATH-500; Minerva/GSM8K transfer and OracleOnly attribution remain future work.

Net read

Seed-0 is strong evidence, not final proof; replication and OracleOnly attribution gate the paper version.

Contributions

Empirical

`D_OracleAug` improves n=8 mean over GRPO by $+4.70$**pp** (p=0.0069).
L4–5: $+5.77$**pp** (p=0.0429).
SFT direct-MLE: **81.0%** greedy, tied with D under sampling.

Mechanistic

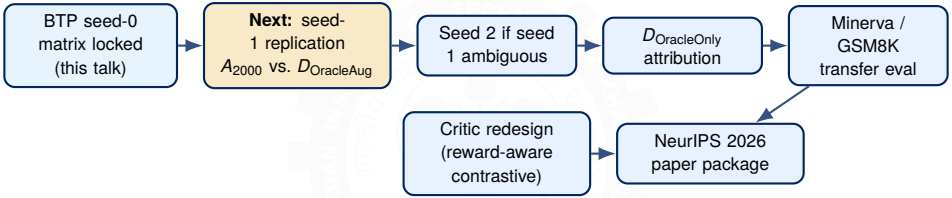
SFT *mode-narrows* (greedy $>$ n=8 by 1.55pp).
D *distribution-shapes* (n=8 $>$ greedy by 3.37pp).
Same sampled ceiling, **opposite mechanisms**.
greedy $\leftrightarrow$ *n=8:* $A \rightarrow$, $D \nearrow$, SFT $\searrow$
Greedy-only eval would miss D's effect.

Diagnostic / Methodological

`B_SelfDistill` controls show easy-slice lift does not explain D.
Pearson r as a candidate runtime warning for critic reward-alignment.
Chi-square critic: documented negative result at tested settings.

One real winning RL/alignment branch + a clean negative result + a mechanistic distinction is a stronger story than a wide matrix of underpowered nulls.

Future Work / NeurIPS Continuation



- Replicate A vs. $D_{\text{OracleAug}}$ on seed 1; launch seed 2 only if seed 1 confirms or is ambiguous.
- Run $D_{\text{OracleOnly}}$ *after* replication is safe to attribute the gain to Oracle content vs. merged-pool augmentation.
- Treat `SFT_Oracle` as the paper's direct-MLE baseline; replicate only if multi-seed SFT is needed.
- Revisit chi-square critic with broader α / contrastive variants once D is replicated.
- PRM extensions (E/F) remain conditional, not BTP-critical.

Acknowledgements & Questions

Thank you.

Sincere thanks to:

- **Prof. Pawan Goyal**, advisor (CSE, IIT Kharagpur), for guidance through every staged pivot.
- Department of Mechanical Engineering, IIT Kharagpur, for the BTP-II framework.
- Compute providers and tooling: **Modal**, **HuggingFace Hub**, **TRL**, **vLLM**, **Weights & Biases**.
- Reference implementations: **Qwen2.5-Math**, **DILO**, **RLDP**, **PSM**, **DeepSeek-Math**.

Questions & Discussion

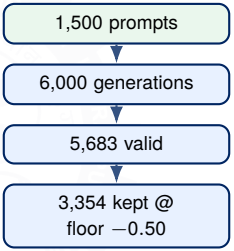
Pritish Saha (22MF3IM15)
Department of Mechanical Engineering
IIT Kharagpur

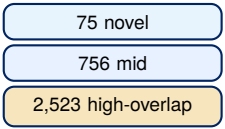
Code, models, datasets & run logs are public — `huggingface.co/Pritish92` | `wandb.ai/.../lavida-mvm`.
Headline: `lavida-variant-D-seed0-oracleaug-alpha0p001`; data: `lavida-expert-pool` + `lavida-oracle-stage2`.

Backup Slides

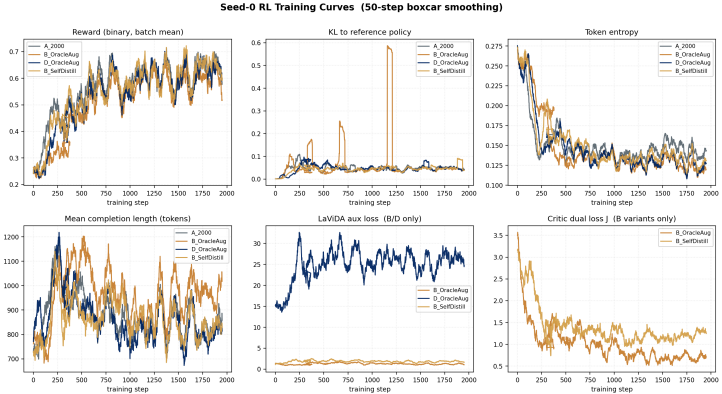
Backup B1: Oracle Pool Filtering Detail

- **1,500** hardest prompts from the productive self pool.
- **6,000** 72B completions (n=4 each).
- Score with 7B base + `cot-4shot`; **5,683** valid, **317** invalid / overlength.
- Adjusted-score percentiles: $p_{10}=-0.3442$, $p_{50}=-0.2212$, $p_{90}=-0.1552$.
- Strict p_{10}^{adj} filter kept **0** novel traces.
- Operational floor loosened to $-0.50 \rightarrow$ kept **3,354** completions on **1,051** prompts.
- Bands: **75 novel** / **756 mid** / **2,523 high-overlap**.





Backup B2: Training Curves (Health Check)



Runs: `A_2000` (`xqboonlw`), `B_OracleAug` (`ulp4wfzz`), `D_OracleAug` (`yg079tgl`), `B_SelfDistill` (`iuhebq5t`) — training logs: W&B `lavida-mvm` (projector/VAE logs: `lavida`).

Backup B3: Metric Definitions

Greedy pass@1 Deterministic single output at $T=0$, top-$p=1$.

$n=8$ **mean correctness** Average of $n_{\text{correct}}/8$ across problems at $T=0.6$, seed 0. Main statistical metric.

pass@8 Fraction of problems with ≥ 1 correct sample among the 8.

Paired bootstrap Resamples problems with replacement; reports 95% CI on Δ mean correctness.

McNemar Compares paired any-correct indicators; reports p on discordant counts.

Important caveat.

- Per-level rows printed by `evaluate_math.py -pass-k 8` are pass@8-style *any-correct* rows.
- Per-level *mean correctness* is recomputed from the saved JSON outputs to enable apples-to-apples paired bootstrap.

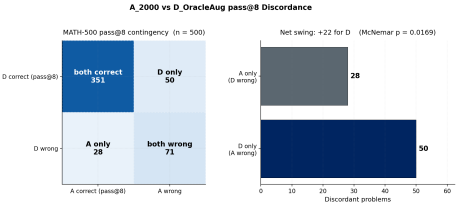
Backup B4: McNemar Discordance Detail

A **vs.** *D* **pass@8 (full MATH-500)**

- *A*-only = 28, *D*-only = 50, net +22 **for D**.
- Level 5: *A*-only = 12, *D*-only = 24.

D **vs.** `SFT_Oracle` **pass@8**

- *D*-only = 35, *SFT*-only = 33, net −2 (essentially balanced).
- Supports the n=8 tie reading: same problems *cluster* solved by both, just different residual sets.



Backup B5: Architecture Deep Dive

Hidden state and feature. $\bar{H}_t = \frac{1}{L}\sum_{\ell=1}^{L} H_t^{(\ell)} \in \mathbb{R}^{3584}$, L=4. $\psi = [h_{\text{start}} \| h_{\text{end}} \| h_{\text{mean}} \| \Delta H] \in \mathbb{R}^{14336}$.

Frozen VAE. $14336 \to 2048 \to 512 \to [\mu, \log\sigma^2]$ of dim 256; $z_{\text{det}} = \mu_\phi(\psi)$, ϕ frozen during RL.

Expert pool & retrieval. One entry per expert trace, keyed by `trace_uid`; prompt-hash lookup `prompt_md5` $\to$ [`trace_uid`]. Trajectory fingerprints clustered into M=8 prototypes (k-means) for behavioural diversity. RL-time retrieval is prompt-hash by default; KNN / prototype-balanced sampling are ablation-only paths.

Variant B critic. $T_\psi : \mathbb{R}^{256} \to 128 \to 128 \to 1$, LeakyReLU, spectral norm.
Smoothed conjugate: $\tilde{f}^*(t) = \tilde{t} + \frac{1}{2}\tilde{t}^2$, $\tilde{t} = \mathrm{softplus}(t)$.
Dual objective: $J(T) = \mathbb{E}_Q[T(z)] - \mathbb{E}_P[\tilde{f}^*(T(e))]$.
TTUR (N_{critic}=5), EMA (μ_{ema}=0.01), critic-side orthogonal gradient projection, $T(z)$ clamp $[-10, 10]$.

Variant D. Reward-gated nearest-expert MSE: $\mathcal{L}_{\text{aux}} = \lambda\,\mathbb{E}[R_i \| z_i - z_{\text{nearest}}\|^2]$ with λ=0.001.

Composite (live trainer). $\mathcal{L} = \mathcal{L}_{\text{GRPO}} + \mathcal{L}_{\text{ref}} + \alpha\,\mathcal{L}_{\text{aux}}$
Direct-add; PCGrad blending is a historical proposal-level safeguard, not the live code path.

Backup B6: Historical Results Ledger

Provenance: (i) instruct arm cleared by deterministic greedy + matched-LR rerun; (ii) base prompt fix: `qwen25-math-cot-4shot` chat path gave **7%** on smoke, switching to plain `cot-4shot` recovered to **72%**; (iii) base self pool: **2,212** prompts / **8,963** traces.

Stage	Run	Greedy	n=8 mean
1 (Instruct)	baseline	83.2%	82.88%
1 (Instruct)	*A* matched 5e−6	84.0%	82.80%
1 (Instruct)	*B* seed 2 step 500	84.0%	83.10%
2 (Base)	base baseline	59.0%	52.50%
2 (Base)	A_0 @ 1e−5	61.2%	55.83%
2 (Base)	B_{002} (α=0.02)	60.6%	56.85%
2 (Base)	C_0 (dyn proj.)	60.0%	55.33%
2 (Base)	D_0 (α=0.002)	**61.8%**	**58.63%**
2 (Base, seed 1)	A_1	**67.2%**	**62.10%**
2 (Base, seed 1)	D_1	62.8%	56.33%
3 (Oracle)	`A_2000`	76.2%	74.88%
3 (Oracle)	`B_OracleAug`	74.2%	74.28%
3 (Oracle)	`D_OracleAug`	76.2%	**79.57%**
3 (Oracle)	`B_SelfDistill`	76.2%	75.48%
3 (Oracle)	`SFT_Oracle`	**81.0%**	79.45%

Bold = stage-best on the column metric.