ML Lab Research Proposal

Discrimination Premise in Self-Teacher RLVR for Mathematical Reasoning

Team: Ethan Wang (solo)

Date: July 22, 2026

Project: REFLEX-RLVR: Falsification and Measurement of LLM Self-Verification

Target Compute: 4x AMD Instinct MI300X GPU, approx. 2 weeks

---

1. PROBLEM STATEMENT

Self-teacher Reinforcement Learning with Verifiable Rewards (RLVR) trains a base language model with only a correctness checker and no external teacher. Methods like STaR (Zelikman et al., NeurIPS 2022) and self-rewarding language models (Yuan et al., ICML 2024) assume the model discriminates valid reasoning chains better than it generates them. Whether this holds for multi-step math is untested. Our pre-registered behavioral test on Qwen2.5-1.5B found the answer-rate difference was exactly 0.000, but this traced to a measurement confound. Follow-up probes show the discrimination signal is real: verdict accuracy ranges from 67% (1.5B base) to 92% (7B Instruct), and answer-relabeled corruption produces 2.0–3.3 logprob units of discrimination on 100% of problems. The open question is whether this signal bootstraps a self-teacher training loop.

2. PROPOSED APPROACH

We measure discrimination across 11 model configurations spanning five families (Qwen2.5 1.5B–72B, Llama-3.1-8B, Mistral-7B, Gemma-2-9B, DeepSeek-R1-Distill) using three complementary probes: consistent-corruption logprob comparison, YES/NO verdict prompting, and per-token surprise at corruption sites. All probes use 51 AIME 2018–2023 problems where the models have pass@1024 = 0 (confirmed inability to solve).

For AAAI, we extend in five directions: (1) complete multi-seed error bars on all 11 configurations (4 seeds each), (2) run a simplified STaR-style self-teacher loop for 3 iterations on Qwen2.5-7B to test whether 80% verdict accuracy bootstraps improvement, (3) measure discrimination on frontier API models (GPT-4o, Claude Sonnet) as a ceiling, (4) extend per-token surprise to step-level verification and compare against trained process reward models, and (5) fit discrimination scaling laws across 7 model sizes (1.5B to 72B) to predict the scale threshold for effective self-teaching.

3. DATASET

Dataset	Source	Size	Use
AIME 2018–2023 (H_K hard set)	Art of Problem Solving	51 problems with oracle solutions	Primary test set. All models have pass@1024 = 0. Measures discrimination where generation fails.
AIME expanded substrate	AoPS + curated	600 problems spanning difficulty range	STaR training substrate (Experiment 2). Includes easier problems where models can generate correct chains.
MATH-500	Hendrycks et al., NeurIPS 2021	500 problems	Forgetting control during STaR training. Monitors catastrophic forgetting.
Synthetic control set	Generated	50 problems with known-correct solutions	Calibration: verifies discrimination probes work on solvable problems.

Preprocessing

- Oracle solutions manually verified against AoPS answer keys. Corrupted solutions generated by changing one intermediate calculation step.
- Answer-relabeled corruption: both the intermediate step AND the final boxed answer are modified, removing the measurement confound.
- vLLM inference with temperature=0.6, top-p=0.95. Gemma-2-9B requires gpu_memory_utilization=0.85, max_model_len=2048 (softcapping + sliding window workaround).

4. EVALUATION METRICS

Experiment	Metric	Target / Interpretation
Multi-seed consolidation (11 models x 4 seeds)	sigma_disc CV, verdict delta CV, per-problem ICC	CV < 3% for sigma_disc, < 10% for verdict delta. Confirms headline numbers are stable.
STaR self-teacher loop (3 iterations on Qwen2.5-7B)	pass@1 and pass@32 on H_K, verdict accuracy, MATH-500 (forgetting)	pass@32 increases from 0 to 1–5 problems solved. Verdict accuracy improves from 80% to 85%.
Frontier API probe (GPT-4o, Claude Sonnet)	delta_verdict on 51 AIME problems	delta_verdict >> 7.34 (our best open model). Establishes discrimination ceiling at > 95% accuracy.
Step-level verification (per-token surprise + verdict-at-step)	Step-level AUROC for corruption localization, top-1 accuracy	AUROC > 0.7 for step identification. Verdict-at-step degrades at and after corrupted step.
Scaling law (7 model sizes, 1.5B to 72B)	Power-law exponent for sigma_disc and delta_verdict vs. parameters	Smooth scaling (no mirage). Extrapolate threshold for self-teacher viability.

5. COMPUTE REQUIREMENTS

Inference-only experiments dominate the compute budget. The STaR training loop is the single largest expense, requiring generation, filtering, and LoRA fine-tuning over 3 iterations.

Phase	Hardware	Est. Wall-Clock	Notes
Multi-seed consolidation (21 model-seed combos)	4x MI300X	~4 hours	$27 equivalent. Parallelizable across models.
STaR self-teacher (3 iter x generate + filter + train)	4x MI300X	~3 days	$220 per loop ($440 with ablation). LoRA rank-32, effective batch 256.
Frontier API probe	API only	~30 minutes	$3–5 (OpenAI + Anthropic API). No GPU needed.
Step-level PRM comparison (3 models x 51 problems)	4x MI300X	~3 hours	$17 equivalent. Per-token surprise + verdict-at-step.
Scaling law sweep (6 new model sizes + probes)	4x MI300X	~6 hours	$24 equivalent. Qwen2.5 3B/14B/32B base + Instruct.
Total	4x MI300X	~5 days within 2-week window	~$513 total (~$740 with STaR ablation). Well within $2,000 ceiling.

Software Dependencies

- Python 3.10+, PyTorch 2.x with ROCm backend.
- vLLM for batched LLM inference (supports Qwen, Llama, Mistral, Gemma architectures).

• PEFT/LoRA for STaR fine-tuning (rank-32, all linear layers).

• SymPy for mathematical answer verification (used alongside self-verdict as a two-stage filter).

6. EXPECTED DELIVERABLES

• Complete 11-model discrimination panel with 4-seed error bars on all headline numbers (sigma_disc, delta_verdict, per-token LMG).

• STaR self-teacher results: pass@32 improvement curve over 3 iterations, with verdict-only vs. SymPy-only filtering ablation isolating the discrimination signal's contribution.

• Frontier discrimination ceiling: delta_verdict for GPT-4o and Claude Sonnet, anchoring the 67–92% open-model range.

• Discrimination scaling curves: power-law fits across 7 Qwen sizes (1.5B to 72B) for sigma_disc and delta_verdict, with extrapolated threshold for self-teacher viability.

• An AAAI-27 paper draft with all figures generated from per-seed JSON result files.

7. RISK & MITIGATION

Risk	Likelihood	Mitigation
STaR loop shows no improvement on H_K at any iteration (null result)	Medium-High	Still informative: pins the practical threshold above measured discrimination levels. Report as "discrimination is necessary but not sufficient." Reduces to Experiments 1,3,4,5 which are self-contained.
Catastrophic forgetting during STaR (MATH-500 drops > 5pp)	Medium	KL anchor at 0.002 against frozen base. Reduce SFT learning rate to 1e-5. Increase replay data mix ratio.
Verdict filter too aggressive at 67% accuracy (1.5B base rejects too many correct chains)	Medium	Use softer threshold (verdict > -0.5 instead of > 0). Combine verdict with SymPy for two-stage filter. Run STaR on 7B-Instruct (92% accuracy) as the primary model.
Gemma-2-9B seed runs trigger vLLM OOM (softcapping + sliding window fragility)	Low	Pin exact vLLM version from successful Round 8 runs. Use --enforce-eager if flash attention fails.
72B seed runs exceed memory on 4x MI300X	Low	Lower max_model_len to 1536. Use fp8 quantization if needed. GPU memory utilization at 0.85.