Bridging the Judgment Gap: Cross-Domain Reasoning Transfer for Dual-Control Agents

Jarrod Barnes
Jarrod@arc.computer

Aman Jaglan
Aman@arc.computer

August 20, 2025

Abstract

This paper addresses the Judgment Gap: the difference between an agent's reasoning and its collaborative implementation in dual-control environments. Here, we propose a Teacher–Student framework for transferring abstract reasoning across domains. We train a Teacher model with reinforcement learning on logical problems in mathematics, a data-rich domain, to produce "thinking traces" that guide a Student agent's execution in data-scarce enterprise settings. Our experiments support this method, showing a statistically significant ($p < .01$) 6× performance gain on τ^2-bench and generalization to CRMArena-Pro with high recall but low precision. These results lead to three main conclusions: (1) Abstract reasoning transfers as a procedural skill across domains, substituting data-efficiently for in-domain fine-tuning. (2) Separating strategic reasoning in the Teacher from tactical execution in the Student closes the Judgment Gap in complex, collaborative settings. (3) This framework lays the groundwork for durable "organizational memory," through generating high-quality interaction data for continual learning.

Keywords: Reinforcement Learning, Agentic Systems, Cross-Domain Learning, Dual-Control Environments, Judgment Gap

1 Introduction

Modern AI agents reason well in static environments, but their reliability drops sharply in real-world workflows requiring user collaboration. We call this failure the *Judgment Gap*: the disconnect between an agent's ability to plan a solution and its ability to apply collaborative judgment in implementing it. The gap shows up clearly in dual-control environments — formalized as Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs) — where both an agent and a user can modify a shared state. The authors of τ^2-bench report a 28-point performance drop in baseline models when shifting from a reasoning-only to a dual-control mode, identifying this "coordination failure" as a primary obstacle [1].

We address this gap with a Teacher–Student framework that transfers abstract reasoning across domains. We hypothesize that agents learn logical problem-solving as a procedural skill in math, where logic is clear and data plentiful, and then use it to guide actions in unrelated, data-poor enterprise domains. This method avoids the "cold start" problem, where missing in-domain preference data usually blocks RL application.

We make four primary contributions: (1) a framework for transferring abstract reasoning skills; (2) a statistically significant demonstration of improved agent performance on the coordination-heavy `mms_issue` tasks in τ^2-bench; (3) a demonstration of the generalizability of a single, math-trained

System	Success@1	95% CI
Student-only	0.041	$[0.011, 0.137]$
Ours (Teacher→Student)	0.240	$[0.146, 0.381]$

(a) τ^2-bench ablation (n=49). Two-proportion z-test: $p = 0.0039$ (12/49 vs 2/49).

	Predicted	
Actual	Violation	No Violation
Violation	9	4
No Violation	33	7

High recall (0.692), low precision (0.214). $N = 53$.

(b) CRMArena-Pro policy compliance.

Figure 1: Experimental results. Single attempt; 95% Wilson CIs.

Teacher on the CRMArena-Pro benchmark; and (4) an outline of how this framework provides the foundation for a full Reinforced Continual Learning (RCL) pipeline, which we frame as future work.

2 Methodology

Our approach relies on a Teacher–Student architecture separating strategic planning from tactical execution.

Teacher training. The Teacher model (Qwen3-8B) is trained on mathematics via two stages: (i) **Supervised fine-tuning (SFT)** on $\sim$ 7,000 problems to teach step-by-step reasoning; and (ii) **Reinforcement learning (RL)** with Group Relative Policy Optimization (GRPO) to improve the Teacher's generation of instructional traces. The reward is formulated as

$$R \;=\; \alpha\,\mathrm{Usefulness} - \beta\,\mathrm{KL}(\pi_{\text{teacher}} \,\|\, \pi_{\text{ref}})\,, \tag{1}$$

where *Usefulness* measures a frozen student-predictor's accuracy conditioned on the trace, and the KL term regularizes against a reference policy.

Teacher–Student inference pipeline. At inference time, the Teacher generates a strategic trace for a task, which is passed into the Student's system prompt to guide execution. This pipeline logs successful trajectories ("golden paths") for offline RCL and the creation of an organizational memory. The math SFT/RL data has no content overlap with our evaluation domains.

3 Experiments and Results

All results are reported with 95% Wilson confidence intervals. All reported success rates are single-attempt (Success@1).

Experiment 1: Evaluating collaborative judgment (τ^2-bench). We evaluate on the 49 `mms_issue` tasks in τ^2-bench, using a Qwen3-8B Student. A main ablation isolates the Teacher's contribution: a Student-only baseline achieves a success rate of 4.1% (2/49), while adding the Teacher's thinking traces gives 24.0% (12/49). This improvement is statistically significant (two-proportion z-test, $p = 0.0039$). Performance varied by user persona, with the lowest success rate on 'Hard' persona tasks (17.6%).

Experiment 2: Assessing deep reasoning (CRMArena-Pro). To check generalization, we ran Policy Compliance tasks with a Qwen3-32B Student. Our system correctly identifies 9 of 13 true policy violations. The confusion matrix (TP=9, FN=4, FP=33, TN=7; $N = 53$) implies Precision 0.214, Recall 0.692, Specificity 0.175, F1 0.327, and Balanced Accuracy 0.434, showing a bias toward recall, with room to improve calibration.

4 Discussion and Future Work

Our framework shows that transferring reasoning processes provides a data-efficient boost to agent reliability in dual-control settings. Limitations include using a single source domain for the Teacher and not providing Teacher traces to closed-model baselines. Comparing with trace-equipped baselines is future work. In future work, we plan to develop a production RCL loop that converts successful interactions and expert corrections into durable organizational memory using continual SFT and RL.

References

[1] V. Barres, H. Dong, X. Si, S. Ray, and K. Narasimhan. τ^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment. *arXiv preprint arXiv:2506.07982*, 2025.

[2] E. Cetin, T. Zhao, and Y. Tang. Reinforcement learning teachers of test time scaling. *arXiv preprint arXiv:2506.08388*, 2025.

[3] K.-H. Huang et al. CRMArena-Pro: Holistic assessment of LLM agents across diverse business scenarios and interactions. *arXiv preprint arXiv:2505.18878*, 2025.

[4] W. Ping et al. AceReason-Nemotron: Advancing math and code reasoning through reinforcement learning. *arXiv preprint arXiv:2505.16400*, 2025.