ATLAS: Adaptive Teaching and Learning Alignment System for Reinforcement Learning

Jarrod Barnes jarrod@arc.computer
Aman Jaglan aman@arc.computer

Arc Intelligence

Abstract

While reinforcement learning (RL) has advanced the reasoning capabilities of large language models, standard training pipelines remain fundamentally brittle, often degrading a model's existing competencies while improving others. This reliability challenge limits the practical application of RL for creating robust, general-purpose agents. We present ATLAS (Adaptive Teaching and Learning Alignment System), a novel diagnostic teaching framework designed to solve this problem. ATLAS reframes RL from a pure optimization task to one of effective pedagogy, employing a two-pass adaptive protocol where a teacher model first diagnostically probes a student's understanding with minimal interaction (≤ 50 tokens). Based on this assessment, the teacher provides calibrated guidance—offering comprehensive scaffolding to weaker students while minimizing intervention for capable ones to prevent harmful interference. The teacher is trained via a two-phase SFT→RL pipeline using Group Relative Policy Optimization (GRPO) with an asymmetric reward function that explicitly penalizes performance degradation. Empirical evaluation demonstrates that ATLAS delivers consistent and significant improvements: a **15.7% average accuracy gain**, a **31% completion rate improvement**, and a **37.2% reduction in response tokens**. Crucially, it achieves this with a **97% non-degradation rate**, confirming its ability to enhance performance without compromising existing skills. These results establish ATLAS as a robust methodology for transforming RL into a reliable system for consistent capability enhancement. We release pre-trained models and our dataset to support further research.

GitHub **Models:** ATLAS-8B-Instruct — ATLAS-8B-Thinking **Dataset**

1 Introduction

Large language models have demonstrated remarkable capabilities across diverse reasoning tasks, yet their training through reinforcement learning remains fundamentally unreliable. Standard RL pipelines exhibit high variance in outcomes, unpredictable performance degradation, and inconsistent improvements. This brittleness poses a critical challenge: how can we create training methodologies that reliably improve model performance without risking degradation on tasks where the model already demonstrates competence?

We introduce ATLAS (Adaptive Teaching and Learning Alignment System for Agents), a diagnostic teaching framework that transforms the traditional RL training paradigm. Rather than applying uniform optimization pressure, ATLAS employs a pedagogically-informed approach where a teacher model first diagnoses student capability, then provides precisely calibrated guidance.

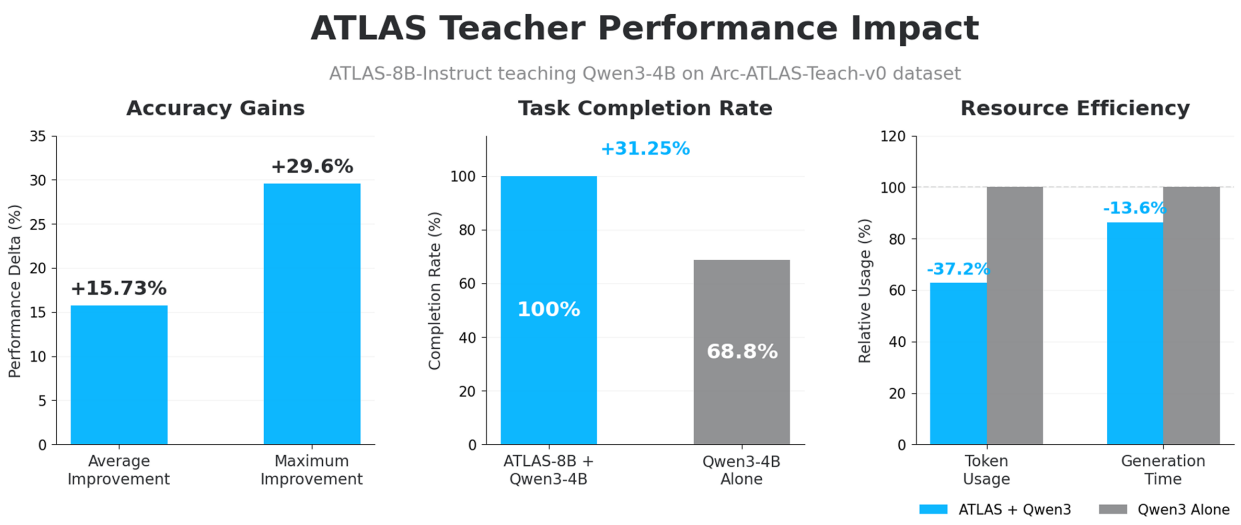

Figure 1: **ATLAS Teacher Performance Impact.** Accuracy gains, task completion rate, and resource efficiency for ATLAS-8B-Instruct teaching Qwen3-4B on the Arc-ATLAS-Teach-v0 dataset.

The ATLAS Approach The core innovation of ATLAS lies in its two-pass adaptive protocol:

- **Teacher–Student Architecture**: We train specialized teacher models that learn to assess and guide arbitrary student models. Unlike traditional knowledge distillation, ATLAS teachers actively diagnose student understanding and adapt their instruction.
- **Diagnostic Probing**: The teacher conducts a lightweight diagnostic probe (≤ 50 tokens) to reveal the student's reasoning approach and capability level before providing guidance.
- **Adaptive Teaching**: Based on the diagnosis, the teacher provides conditional guidance. Strong students receive minimal intervention, while weaker students receive comprehensive scaffolding. This adaptivity is key to achieving consistent improvements without degradation.

Theoretical Positioning and Contributions ATLAS makes several contributions that address fundamental open questions in reinforcement learning for reasoning models, as outlined in recent surveys (Zhang et al., 2025):

1. **A Novel Stance on Capability Development**: The field actively debates whether RL primarily "sharpens" latent abilities or enables the "discovery" of new ones (Yue et al., 2025b). ATLAS offers a third perspective by framing the objective as *pedagogical transfer*. Our framework optimizes the teacher not for discovering novel solutions, but for discovering optimal *teaching strategies*. This is a distinct form of capability development focused on the effective and reliable transmission of skills, representing a form of meta-learning that is critical for building robust, generalizable agents.
2. **A Principled SFT→RL Pipeline**: Our two-phase architecture provides a practical answer to the "SFT memorizes, RL generalizes" question (Chu et al., 2025a). We treat the RL environment as a classroom for the LLM. The SFT phase is analogous to providing the teacher with a stable curriculum and foundational knowledge of teaching formats ("memorization"). The subsequent RL phase then allows the teacher to flexibly apply and adapt these formats to individual student

needs, learning to generalize its pedagogical skills across a wide range of problems. This multi-stage post-training recipe is validated by concurrent work like K2-Think, which also demonstrates the power of synergistic training stages (Cheng et al., 2025).

3. **A Solution to the Exploration Problem**: Sparse, outcome-based rewards in RL create a significant exploration challenge, as models receive no learning signal until they can already solve a task (Cetin et al., 2025). By replacing this sparse signal with dense, pedagogically-grounded rewards based on student improvement, ATLAS circumvents this limitation. The teacher provides a structured curriculum, effectively guiding the student through what educational psychology terms the "zone of proximal development."
4. **Open Release of Resources**: We release pre-trained teacher models (ATLAS-8B-Thinking, ATLAS-8B-Instruct) and the Arc-ATLAS-Teach dataset to enable the community to build upon our work.

2 Related Work

Teacher–Student Frameworks Traditional teacher-student paradigms focus on knowledge distillation, where a smaller student model imitates a larger teacher's outputs. This approach, while effective for model compression, is fundamentally passive. Recent work has sought to create more active teaching frameworks. For instance, Reinforcement-Learned Teachers (RLTs) propose a framework where the teacher is given both the question and the solution and is trained via RL to "connect-the-dots" with an effective explanation (Cetin et al., 2025).

ATLAS builds on this evolution but introduces a critical distinction: **diagnostic assessment**. Instead of assuming the student always needs a full explanation, the ATLAS teacher first probes the student's understanding. This allows for a truly adaptive intervention, providing comprehensive help only when needed and otherwise offering minimal guidance to avoid disrupting an already-correct reasoning process.

Reinforcement Learning from AI Feedback (RLAIF) RLAIF enables models to learn from AI-generated preferences, often through methods like Direct Preference Optimization (DPO) (Rafailov et al., 2023). While powerful, these methods typically rely on ranking entire outputs. ATLAS extends beyond simple preference-based learning by implementing a structured pedagogical process. The reward signal is not a generic preference but a direct measure of student improvement, allowing the teacher to learn nuanced, targeted remediation strategies.

Pedagogical Foundations in Human Learning **Zone of Proximal Development (Vygotsky, 1978)**: Vygotsky's seminal concept posits that learning is most effective in the space between what a learner can achieve independently and what they can achieve with guidance. ATLAS operationalizes this principle directly. The *diagnostic probe* serves to identify the student model's current capability frontier—its Zone of Proximal Development. The subsequent *adaptive teaching* provides the precise level of "scaffolding" required to bridge this gap, ensuring the learning task is challenging but achievable.

Cognitive Load Theory (Sweller, 1988): This theory argues that learning is hampered when a task imposes too much extraneous mental effort (cognitive load). ATLAS's adaptive nature is a direct implementation of cognitive load management for LLMs. For strong students, the teacher provides minimal intervention, reducing extraneous load and allowing the student to focus its

"cognitive" resources on the problem itself. For weak students, the teacher provides comprehensive, decomposed scaffolding. The efficiency weight in our reward function directly incentivizes concision.

Desirable Difficulties (Bjork, 1994): By providing minimal intervention for capable students, ATLAS avoids removing "desirable difficulties." The asymmetric reward function, with its heavy degradation penalty, ensures the teacher avoids introducing *undesirable* difficulties that would harm performance.

3 The ATLAS Methodology

3.1 The Adaptive Teaching Protocol

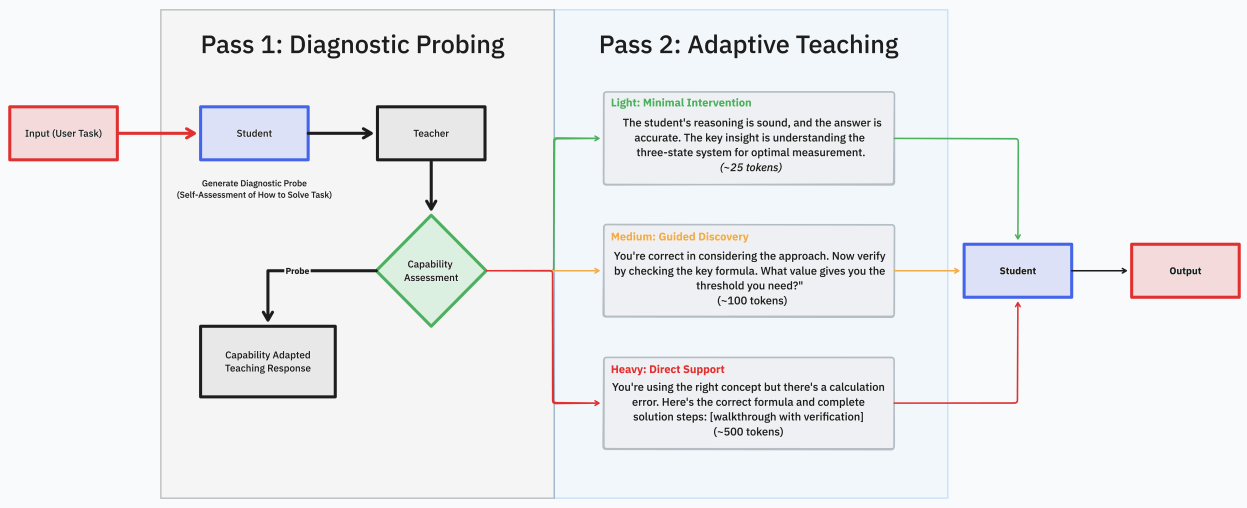

Figure 2: **Two-pass adaptive teaching protocol.** Pass 1: diagnostic probing; Pass 2: capability-adapted teaching.

Pass 1: Diagnostic Probing. The teacher initiates interaction by presenting a diagnostic probe to assess the student's understanding, constrained to a maximum of 50 tokens to elicit a concise summary of the student's initial approach.

Pass 2: Adaptive Teaching. Based on the diagnostic assessment, the teacher provides calibrated guidance:

- **For strong students** (correct approach identified): Minimal intervention, often consisting of a simple confirmation.
- **For weak students** (confusion or incorrect approach): Comprehensive scaffolding, including problem decomposition and step-by-step reasoning chains.

This adaptivity ensures that teaching enhances rather than interferes with student reasoning.

3.2 Training Pipeline: SFT → RL

Phase 1: SFT Warmup. The teacher model undergoes supervised fine-tuning on high-quality reasoning demonstrations from the SFT split of the Arc-ATLAS-Teach dataset. This phase establishes strong baseline capabilities in problem-solving and explanation generation.

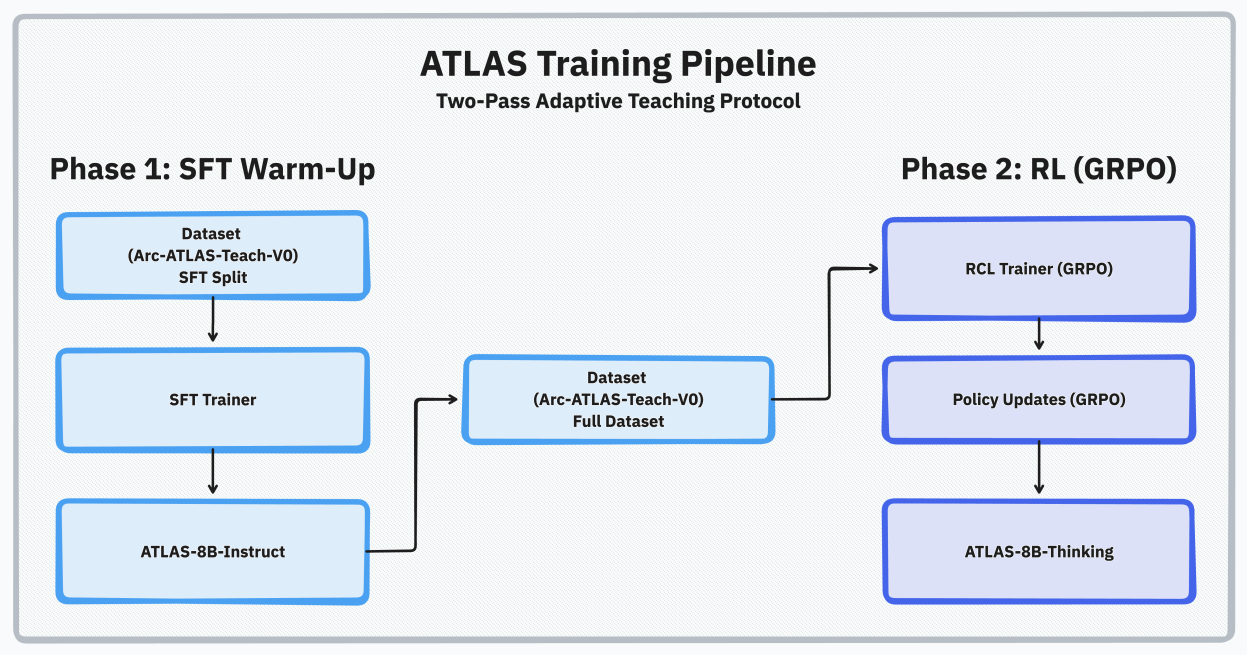

Figure 3: **ATLAS training pipeline.** Phase 1: SFT warm-up; Phase 2: RL with GRPO.

Phase 2: RL with GRPO. The SFT-warmed model is then optimized using Group Relative Policy Optimization (GRPO) to become an effective teacher. This phase uses the full RL dataset, training the teacher to generate guidance that maximizes student improvement.

3.3 Asymmetric Reward Function for Adaptive Teaching

The reward function is designed to incentivize helpful teaching while preventing harmful intervention. It is structured around two key principles:

- **Degradation Penalty**: Assign a significant penalty for any teaching interaction that results in a decrease in student performance (*do no harm*).
- **Efficiency Weight**: Penalize verbosity, encouraging concise, targeted instruction rather than lengthy explanations that may overwhelm the student.

4 Experimental Setup

- **Hardware**: 4×H100 GPUs.
- **Dataset**: Arc-ATLAS-Teach-v0 dataset.
- **Models**: The teacher model is ATLAS-8B-Instruct. The student model used for evaluation is Qwen3-4B.
- **Implementation**: The training pipeline is managed via Hydra, with distributed training using DeepSpeed and efficient generation for RL handled by a vLLM server.
- **Evaluation**: 32 samples were run per problem to ensure robust metrics. We measured accuracy, completion rate, response efficiency (tokens), and generation time.

5 Results and Analysis

5.1 Main Performance Results

Our evaluation demonstrates that ATLAS-trained teachers consistently and significantly improve student model performance across multiple metrics.

Table 1: **Summary of main results.** Relative improvements for ATLAS teacher + student vs. student alone.

Metric	Student Alone	Teacher + Student	Delta
Average Accuracy	Baseline	+15.73%	**+15.73%**
Completion Rate	68.8%	100%	**+31.25%**
Response Tokens	100%	62.8%	**-37.2%**
Generation Time	100%	86.4%	**-13.6%**
Non-degradation Rate	–	97%	**97%**

These results lead to several key insights:

1. **Consistent Gains**: Unlike traditional RL which often shows high variance, ATLAS delivers reliable improvements with a 97% non-degradation rate.
2. **Efficiency Improvements**: The significant token and time reduction demonstrates that adaptive teaching not only improves accuracy but also makes the reasoning process more efficient.
3. **Completion Benefits**: The dramatic improvement in completion rate (from 68.8% to 100%) indicates that teacher guidance is highly effective at helping students overcome reasoning blocks that would otherwise cause them to fail.

5.2 Qualitative Analysis

Case 1: Adaptive Scaffolding for a Weak Student. In a problem requiring the student to find the minimum number of weights to measure up to 1000 grams, the student's initial approach was conceptually correct (using powers of 3) but contained a calculation error. *Teacher Guidance (Heavy Support)*: "You are correct in using powers of 3 for this problem. However, there is an error in your calculation. The formula to find the maximum weight measurable with n weights is $(3^n - 1)/2$. You need to solve $(3^n - 1)/2 \geq 1000$. Let's solve this step by step. First, multiply both sides by 2: $3^n - 1 \geq 2000$. Then add 1: $3^n \geq 2001$. Now, find the smallest n such that 3^n is at least 2001. Check $3^6 = 729$ and $3^7 = 2187$. Since 2187 is greater than 2001, $n = 7$. Therefore, you need 7 weights."

Case 2: Minimal Intervention for a Strong Student. In the same problem, when another student demonstrated a solid understanding of the formula and its application: *Teacher Guidance (Minimal Intervention)*: 'To verify the answer, ensure the formula $(3^n - 1)/2$ is correctly applied. The student's reasoning is sound, and the answer of 7 weights is accurate. The key is understanding that each weight can be in three states, allowing for efficient measurement. The final answer is correct.'

Case 3: Improved Reasoning Efficiency. We analyzed the reasoning trace of a student solving a problem with and without teacher guidance. The task was to calculate the distance between 52 trees planted over 1,850 meters. *Student Alone*: The student's response was lengthy and demonstrated significant uncertainty, repeatedly second-guessing its own correct initial logic (dividing the length

by 51 intervals). *Student with Teacher Guidance*: After receiving a targeted prompt confirming the correct initial step, the student's response was direct, confident, and efficient, proceeding straight to the correct calculation.

6 Conclusion

ATLAS represents a significant step toward creating more reliable and effective reinforcement learning methodologies. By incorporating principles from educational psychology, our two-pass adaptive teaching protocol transforms the traditionally brittle RL training process into a robust system that delivers consistent performance gains. The empirical results—a 15.7% increase in accuracy, a 31% gain in completion rate, and a 97% non-degradation rate—validate that a pedagogical approach can solve key challenges in modern RL.

More fundamentally, ATLAS establishes a paradigm where AI systems learn to teach and learn from each other adaptively. This suggests that the future of AI improvement lies not merely in scaling models or data, but in developing sophisticated teaching dynamics. As the field moves toward long-horizon, agentic systems capable of continual learning, as explored in frameworks like AgentGym (Xi et al., 2025) and Memento (Zhou et al., 2025), the ability to reliably transfer skills becomes paramount. The diagnostic and adaptive mechanisms in ATLAS provide a foundational component for such systems, enabling more efficient and robust skill acquisition in continuously evolving, interactive environments.

Limitations & Future Work

We intentionally defer an ablation study to future work. Key next steps include: (i) ablations over the diagnostic-probe length, reward weights, and intervention tiers; (ii) broader student-model families and tasks; and (iii) longer-horizon teaching curricula integrated with agentic evaluation settings.

References

- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.
- Cetin, E., Zhao, T., & Tang, Y. (2025). *Reinforcement Learning Teachers of Test Time Scaling.* arXiv preprint.
- Cheng, Y., et al. (2025). *K2-Think: A Technical Report on Synergistic SFT-RL Training.* Arc Intelligence Technical Report.
- Chu, Z., et al. (2025a). *Memorization vs. Generalization in Language Model Training Pipelines.* arXiv preprint.
- Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems, 36.*
- Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science, 12*(2), 257–285.

- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Xi, Z., Huang, J., Liao, C., et al. (2025). AgentGym-RL: Training LLM Agents for Long-Horizon Decision Making through Multi-Turn Reinforcement Learning. *arXiv preprint arXiv:2509.08755*.
- Yue, Z., et al. (2025b). *Sharpening or Discovery: A Survey on the Role of RL in LLM Capability Development*. arXiv preprint.
- Zhang, K., Zuo, Y., He, B., et al. (2025). A Survey of Reinforcement Learning for Large Reasoning Models. *arXiv preprint arXiv:2509.08827*.
- Zhou, H., Chen, Y., Guo, S., et al. (2025). Memento: Fine-tuning LLM Agents without Fine-tuning LLMs. *arXiv preprint arXiv:2508.16153*.