ICLR
International Conference On Learning Representations

Georgia Tech

CISCO

Enhancing Language Model Reasoning with Structured Multi-Level Modeling

Siheng Xiong[1], Ali Payani[2], Faramarz Fekri[1]

[1]Georgia Institute of Technology, [2]Cisco Research

Paper

Code

Motivation: What is missing in current long-horizon reasoning

- Single-policy long CoT mixes planning and execution.
- Without guidance or structure, errors can accumulate and cause the implicit plan to drift.

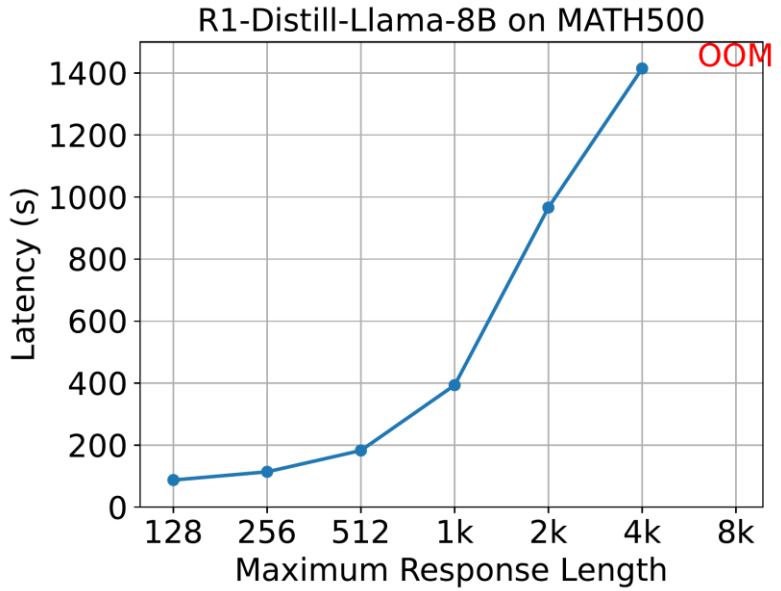
- Outcome-only RL uses only the final answer correctness as the reward signal.
- This supervision is sparse, delayed, and increasingly costly for long trajectories.

Example: Flat CoT mixes planning and execution

Task: Find the smallest multiple of 30 using only digits 0 and 2.

What flat CoT does:
- finds some local constraints correctly
- repeatedly checks candidates such as 2220
- revisits the same partial conclusions
- fails to maintain a clear global search plan

Takeaway:
Flat CoT can satisfy local checks, but still struggles to sustain structured long-horizon reasoning.

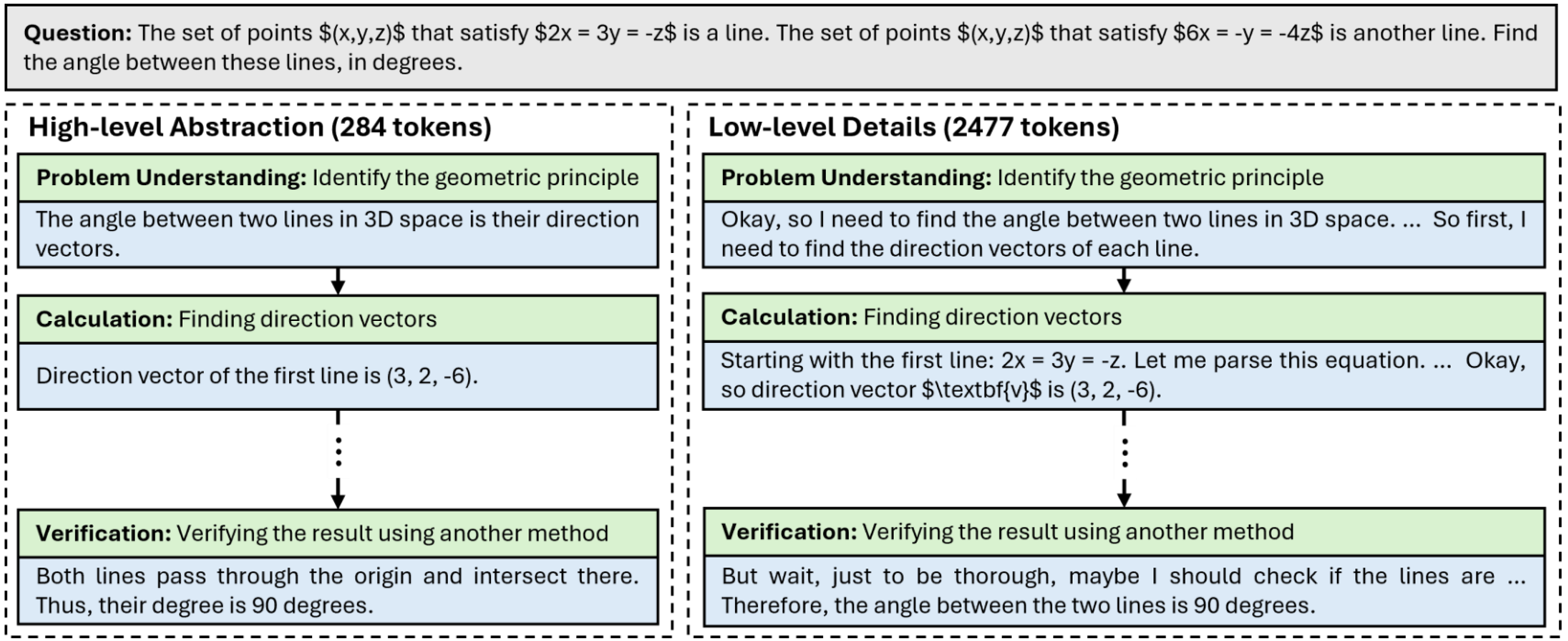


Outcome-only RL becomes increasingly costly as trajectories grow longer.

Methodology: What is MLR

MLR explicitly separates reasoning across multipel levels of abstraction.

- high level organizes coarse plans and key intermediate states
- low level handles detailed local inference

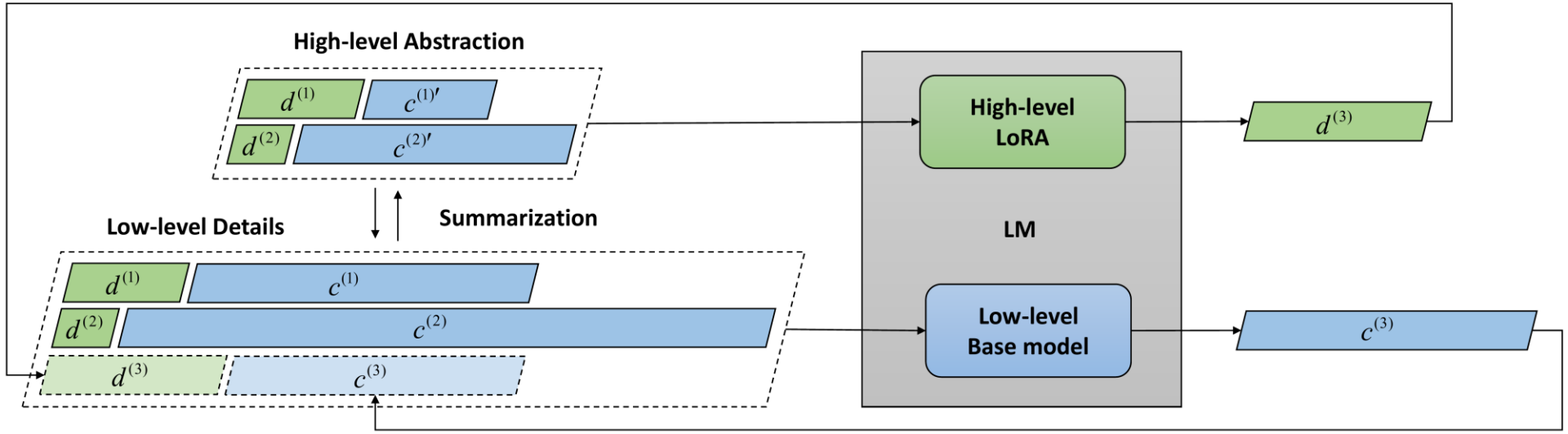


MLR represents the same reasoning process at different granularities.

Methodology: How MLR is implemented

MLR uses two coupled reasoning streams: high-level abstraction for global structure, and low-level generation for detailed execution.

- low-level reasoning is incrementally summarized into high-level states
- the next-step prompt is built from both levels, so global plans remain aligned with local derivations



Cross-level context construction: abstract subgoals guide generation, while detailed reasoning history supports execution.

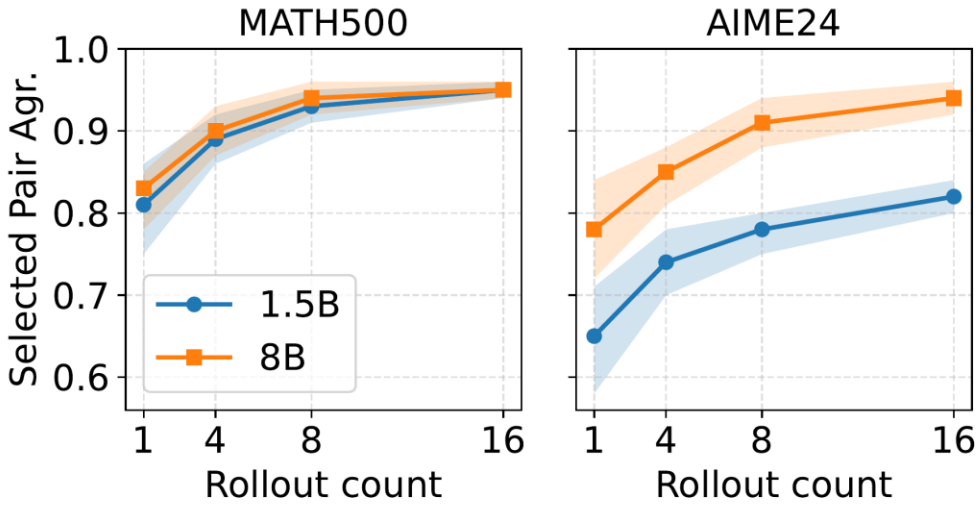
Training: Iterative Multi-level Step-DPO with Scalable Preference Annotation

1. SFT initialization
- long CoT → **steps, descriptors, summaries**
- train **low-level base model** + **high-level LoRA**

2. Process-level preferences
- score steps by **future solvability**
- construct preferences with **TSMC**
- **reliable with only a few rollouts**

3. Iterative optimization
- alternate **high-/low-level Step-DPO**
- **refresh preference data** online

Takeaway:
Hierarchical initialization + online process-level optimization

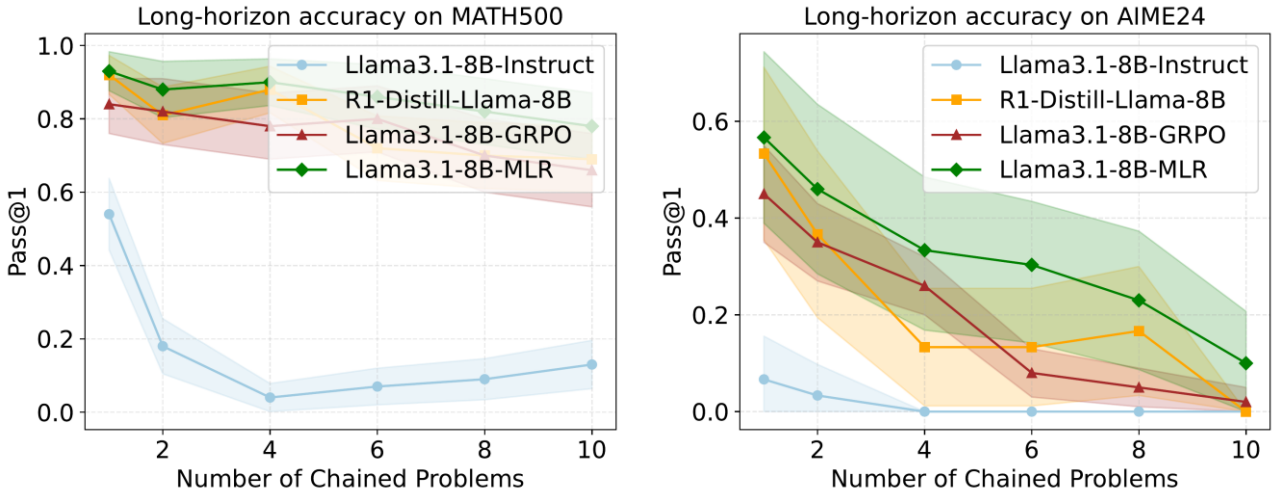


Preference pairs quickly become reliable with a few rollouts.

Main results

Method (Llama-3.1-8B)	MATH500	AIME24	GPQA	Avg.
DeepSeek-R1-Distill	89.1	50.4	49.0	62.8
SFT + GRPO	86.5	42.0	47.0	58.5
MLR	**91.5**	**53.2**	**52.8**	**65.8**

- MLR outperforms strong baselines
- Largest gains appear on hard multi-step settings
- Structure matters beyond prompting



MLR remains more robust as reasoning horizon increases.

Takeaways

- **Problem:** Flat reasoning entangles long-horizon reasoning
- **Method:** Explicit multi-level reasoning structure
- **Result:** Stronger accuracy and better robustness

Stronger reasoning comes from a better structure, not merely more tokens.

Thanks for your listening!

Paper

Code