



上海交通大学
SHANGHAI JIAO TONG UNIVERSITY



Learning to Reason in Structured In-context Environments with Reinforcement Learning

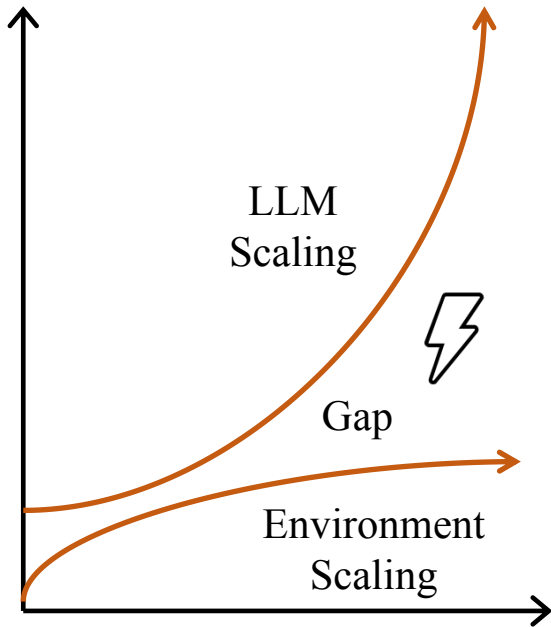
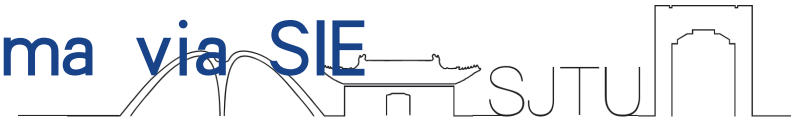
Peng Yu

Shanghai Jiao Tong University

pursuit_yp@sjtu.edu.cn

Mar, 2026

Motivation: Solving the Environment Scaling Dilemma via SIE



The Scaling Dilemma

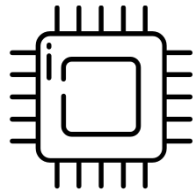
Math & Code: Hard to Scale
Game Engine: Hard to Transfer



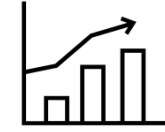
Massive Structured Data



Automated Construction Pipeline



Structured In-context Environment



Scalability

Automated extraction from real-world data sources.



Generalizability

Learning how to think and transfers to math & logic.



Verifiability

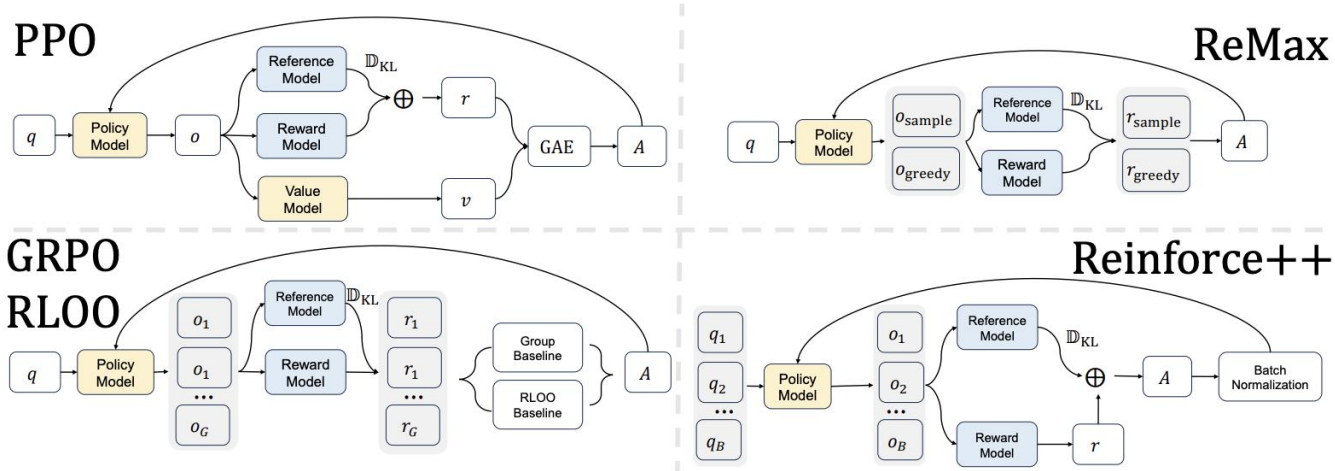
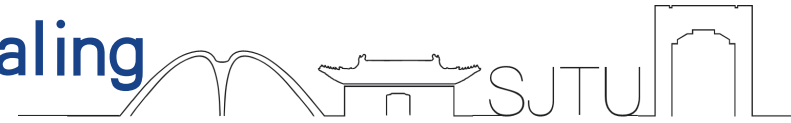
Schemas support objective and rule-based verification.

The SIE Framework

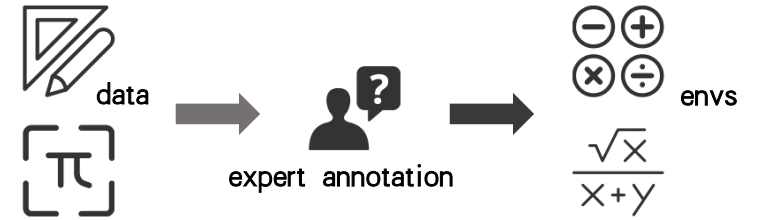
Static Data → RL Playground
Zero Human Intervention

The SIE framework addresses reasoning bottlenecks by automatically constructing scalable, generalizable, and verifiable RL environments from massive structured data.

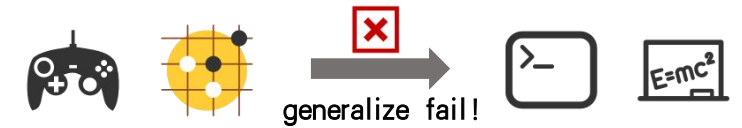
Background: from RL Algorithm to Environment Scaling




internalized-rule envs are difficult to scale.



externalized-rule envs are hard to generalize.

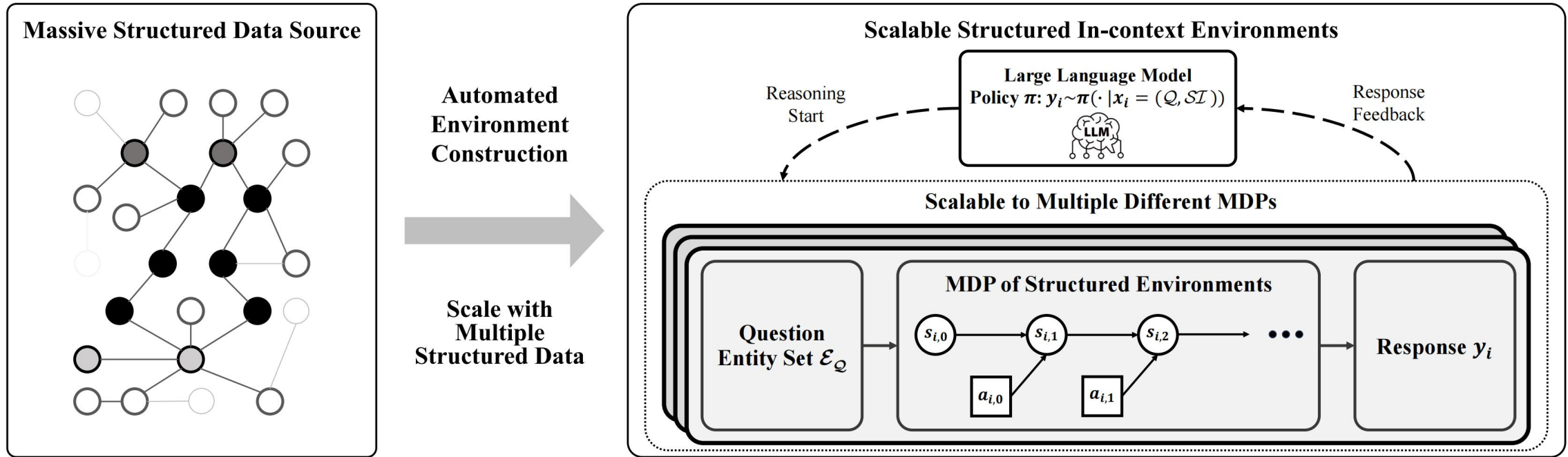
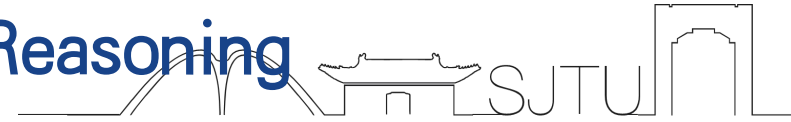


RL fine-tuning **algorithm** optimizations: 
 from PPO to GRPO, Reinforce++, DAPO, GSPO...

The characteristics of an ideal LLM reasoning environment: 
Scalability,
Generalizable Reasoning, **Verifiability**.

Research on RL for LLMs has focused primarily on algorithmic optimizations, while the environment directly determine which capabilities can be incentivized for LLMs.

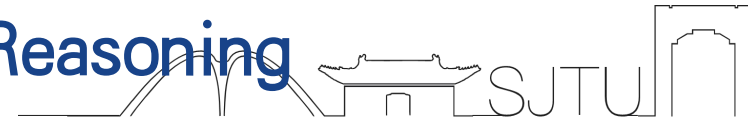
Structured In-context Environments (SIE) for LLM Reasoning



- SIE achieves **scalability** by automatically constructing reasoning environments from large-scale structured data,
- where the rich multi-hop compositional patterns naturally support **generalizable reasoning**.
- The explicit schemas and reasoning chains in structured data provide a foundation for rule-based **verifiability**.

Constructing high-quality LLM reasoning environments from massive structured data is not only feasible but also promising for balancing scalability and generalizability.

Structured In-context Environments (SIE) for LLM Reasoning



Automatically construct SIEs from large-scale KGs:

$$\mathcal{G}_{seed} = \text{MultiHopSearch}(\mathcal{G}, \mathcal{E}_Q, q_{hop}) \cup \text{MultiHopSearch}(\mathcal{G}, \mathcal{E}_A, a_{hop}),$$

$$\mathcal{G}_{support} = \text{ShortestPathSearch}(\mathcal{G}_{seed}, \mathcal{E}_Q, \mathcal{E}_A, n_{hop}),$$

$$rel_{retain} = \text{FilterRelation}(\mathcal{G}_{seed} \setminus \mathcal{G}_{support}, \mathcal{Q}),$$

$$\mathcal{G}_{distract} = \text{FilterTriple}(\mathcal{G}_{seed} \setminus \mathcal{G}_{support}, \mathcal{Q}, rel_{retain}),$$

$$\text{SIE-ratio} = \text{Shuffle}(\text{Retain}(\mathcal{G}_{support}, ratio) \cup \mathcal{G}_{distract})$$

for $ratio \in \{100\%, 75\%, 50\%, 25\%, 0\%\}$,

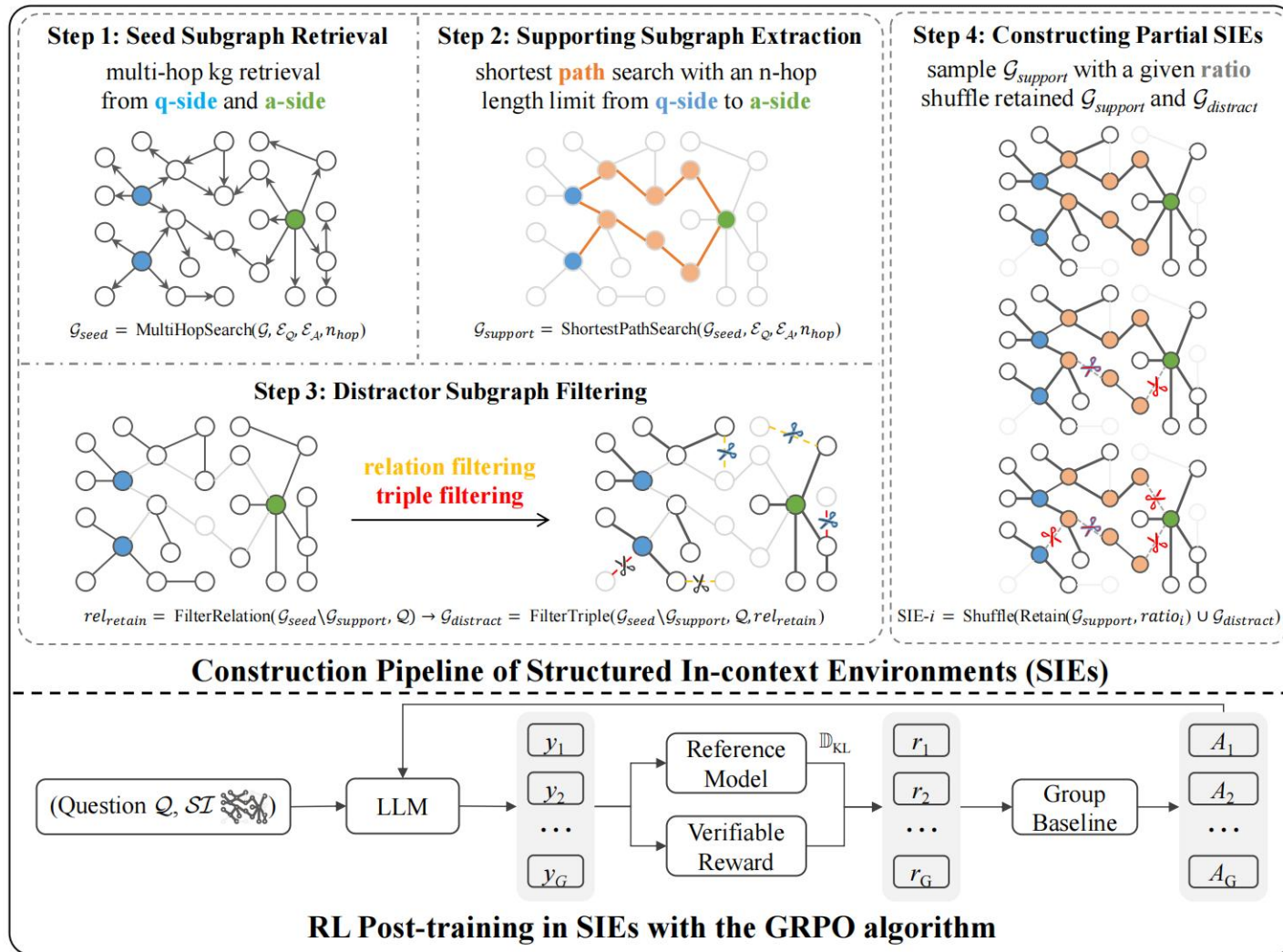
Learning to reason in SIEs with the GRPO alg.:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{\substack{(x, y^*) \sim SIE \\ \{y_i\}_{i=1}^G \sim \pi_{\theta_{old}}(\cdot | x)}} \left[\frac{1}{G} \sum_{i=1}^G \left(\min \left(\frac{\pi_{\theta}(y_i | x)}{\pi_{\theta_{old}}(y_i | x)} A_i, \right. \right. \right. \\ \left. \left. \left. \text{clip} \left(\frac{\pi_{\theta}(y_i | x)}{\pi_{\theta_{old}}(y_i | x)}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) - \beta \mathbb{D}_{KL}(\pi_{\theta} \parallel \pi_{ref}) \right],$$

$$\mathbb{D}_{KL}(\pi_{\theta} \parallel \pi_{ref}) = \frac{\pi_{ref}(y_i | x)}{\pi_{\theta}(y_i | x)} - \log \frac{\pi_{ref}(y_i | x)}{\pi_{\theta}(y_i | x)} - 1,$$

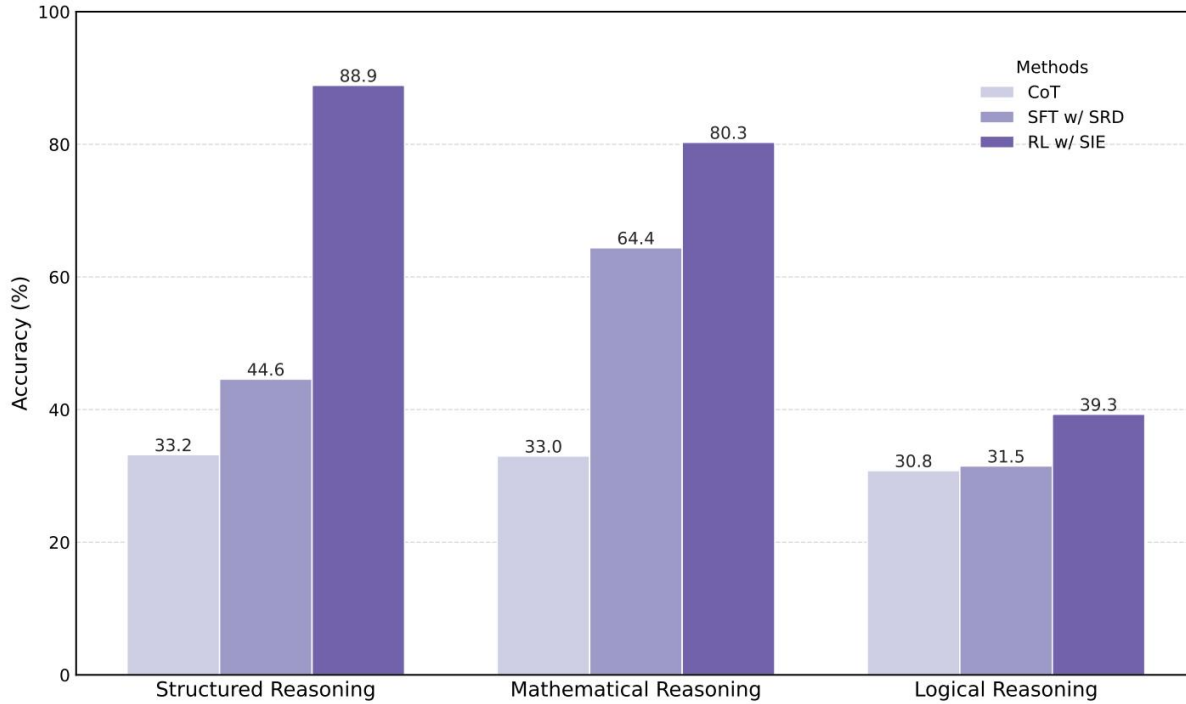
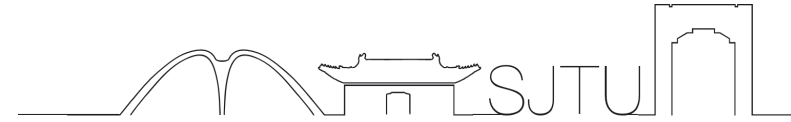
answer reward
format reward

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})},$$



SIEs as the in-context soft environment to fine-tune LLMs with Reinforcement Learning.

Experiments: Overall Performance of RL in SIEs



Setting:

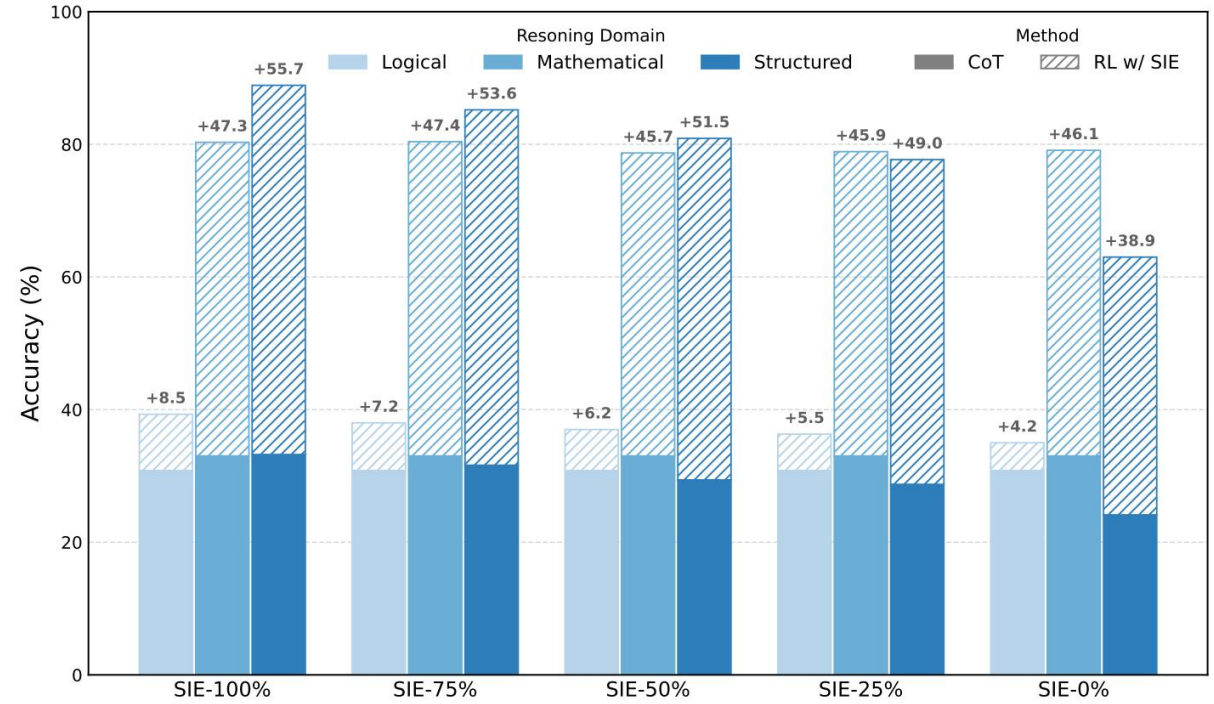
Model: Qwen2.5-7B-Instruct (CoT/SFT on SRD/RL in SIE)

Training: **SIE-100%** constructed from WebQSP and CWQ

Test: **Structured** (WebQSP/CWQ/GraIQa),
Math (GSM8K/MATH500), **Logic** (K&K)

Results:

RL in SIEs can **effectively** improve structured reasoning abilities and is more **efficient** than SFT. Furthermore, the reasoning ability can be **generalized** to the fields of mathematics and logic.



Setting:

Model: Qwen2.5-7B-Instruct (CoT/RL in SIE)

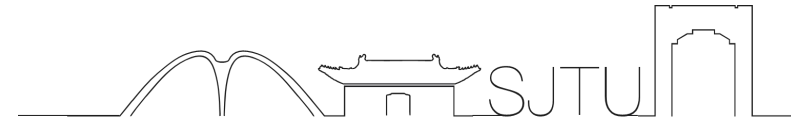
Training: **Partial SIEs** constructed from WebQSP and CWQ

Test: **Structured** (WebQSP/CWQ/GraIQa),
Math (GSM8K/MATH500), **Logic** (K&K)

Results:

RL fine-tuning in partial SIEs can lead to deep **compositional** reasoning, and **robustly** improves the structured reasoning and **generalization** capabilities of LLMs.

Experiments: Performance of RL in SIE–100%



Datasets	Qwen2.5-7B-Instruct		Llama3.1-8B-Instruct		Qwen2.5-7B		Qwen3-8B	
	w/o Context	w/ SIE	w/o Context	w/ SIE	w/o Context	w/ SIE	w/o Context	w/ SIE
WebQSP	59.7	93.4 (+33.7)	61.3	93.2 (+31.9)	62.8	93.2 (+30.4)	48.6	90.2 (+41.6)
CWQ	36.7	87.7 (+51.0)	39.7	89.7 (+50.0)	38.4	89.3 (+50.9)	29.7	78.6 (+48.9)
GrailQA	20.8	85.8 (+65.0)	24.9	85.0 (+60.1)	19.5	81.5 (+62.0)	21.8	85.1 (+63.3)

- **Setting:** RL without Context / RL in SIE–100%
- **Results:** The SIE Framework **Effectively** Enhances LLM Structured Reasoning.

Datasets	Qwen2.5-7B-Instruct			Llama3.1-8B-Instruct		
	CoT	SFT w/ SRD	RL w/ SIE	CoT	SFT w/ SRD	RL w/ SIE
WebQSP	26.3	40.5 (+14.2)	93.4 (+67.1)	36.5	43.4 (+6.9)	93.2 (+56.7)
CWQ	34.4	43.3 (+8.9)	87.7 (+53.3)	37.2	49.5 (+12.3)	89.7 (+52.5)
GrailQA	40.5	55.7 (+15.2)	85.8 (+45.3)	43.6	60.0 (+16.4)	85.0 (+41.4)

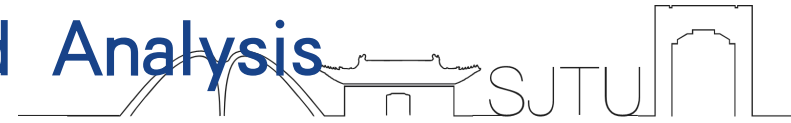
- **Setting:** CoT / SFT on SRD / RL in SIE–100%
- **Results:** RL Fine–tuning in SIE is More **Efficient** than SFT on SRD for LLM Structured Reasoning.

Datasets	Qwen2.5-7B-Instruct		Llama3.1-8B-Instruct		Qwen2.5-7B		Qwen3-8B	
	CoT	RL w/ SIE	CoT	RL w/ SIE	CoT	RL w/ SIE	CoT	RL w/ SIE
GSM8K	29.2	87.4 (+58.2)	67.4	82.6 (+15.2)	27.0	86.2 (+59.2)	71.0	91.9 (+20.8)
MATH500	43.0	61.6 (+18.6)	38.4	47.0 (+8.6)	30.2	59.2 (+29.0)	20.4	36.6 (+16.2)
KK-easy	42.0	49.5 (+7.5)	20.5	37.0 (+16.5)	37.5	52.0 (+14.5)	79.5	90.0 (+10.5)
KK-hard	19.5	29.0 (+9.5)	6.0	15.5 (+9.5)	15.5	27.5 (+12.0)	59.5	73.5 (+14.0)

- **Setting:** CoT / RL in SIE–100%, Generalization
- **Results:** Structured Reasoning Skills Learned in SIEs **Generalizes** to Out–of–Domain Reasoning Domains.

Experimental results not only validate the **effectiveness and efficiency** of RL fine–tuning in SIEs but also reveal that the learned cognitive pattern and compositional strategies can be **generalized** to boarder mathematical and logical reasoning domains.

Experiments: Performance of RL in Partial SIEs and Analysis



Setting	Qwen2.5-7B-Instruct		Llama3.1-8B-Instruct		Qwen2.5-7B		Qwen3-8B	
	CoT	RL w/ SIE	CoT	RL w/ SIE	CoT	RL w/ SIE	CoT	RL w/ SIE
SIE-100%	26.3	93.4 (+67.1)	36.5	93.2 (+56.7)	2.6	93.2 (+90.6)	47.8	90.2 (+42.4)
SIE-75%	23.6	89.2 (+65.6)	33.8	90.4 (+56.6)	2.0	90.2 (+88.2)	47.3	88.0 (+40.7)
SIE-50%	22.3	86.4 (+64.1)	31.1	89.4 (+58.3)	2.5	87.2 (+84.7)	44.9	84.0 (+39.1)
SIE-25%	22.0	85.4 (+63.4)	31.5	86.8 (+55.3)	1.6	85.8 (+84.2)	44.8	82.6 (+37.8)
SIE-0%	17.8	72.8 (+55.0)	26.1	75.1 (+49.0)	1.7	73.4 (+71.7)	36.7	70.8 (+34.1)

Setting	GSM8K	MATH500	KK-easy	KK-hard
CoT	29.2	43.0	42.0	19.5
SIE-100%	87.4 (+58.2)	61.6 (+18.6)	49.5 (+7.5)	29.0 (+9.5)
SIE-75%	87.7 (+58.5)	61.0 (+18.0)	50.0 (+8.0)	26.0 (+6.5)
SIE-50%	86.2 (+57.0)	59.0 (+16.0)	48.5 (+6.5)	25.5 (+6.0)
SIE-25%	86.0 (+56.8)	60.2 (+17.2)	48.0 (+6.0)	24.5 (+5.0)
SIE-0%	87.1 (+57.9)	58.0 (+15.0)	47.0 (+5.0)	23.0 (+3.5)

- **Setting:** CoT / RL in Partial SIEs (SIE-100%, SIE-75%, SIE-50%, SIE-25%, SIE-0%)
- **Results:** RL in Partial SIEs Achieves **Robust** Reasoning and Generalization. Information-constrained environments can effectively shift the model's reasoning paradigm from shallow memory retrieval to **deeper compositional reasoning**.

Methods	WebQSP	CWQ	GrailQA	GSM8K	MATH500	KK-easy	KK-hard
CoT	26.3	34.4	40.5	29.2	43.0	42.0	19.5
GRPO	93.4	87.7	85.8	87.4	61.6	49.5	29.0
REINFORCE++	93.1	88.4	83.2	86.7	62.2	49.0	24.5
PPO	85.4	73.4	81.4	78.4	59.6	49.0	25.0

- **Setting:** CoT / GRPO / Reinforce++ / PPO
- **Results:** The SIE Framework is **Applicable** to Mainstream RL Fine-tuning Algorithms.

Methods	WebQSP	CWQ	GrailQA	GSM8K	MATH500	KK-easy	KK-hard
SFT w/ SRD	40.5	43.3	55.7	68.1	54.8	41.5	21.5
RL w/ SIE	93.4	87.7	85.8	87.4	61.6	49.5	29.0
RL w/ SIE f/ SFT	88.5	79.6	81.7	88.7	62.0	52.0	33.5

- **Setting:** SFT on SRD / RL in SIE-100% / RL in SIE-100% from SFT Checkpoint
- **Results:** Starting RL Fine-tuning from an SFT Checkpoint **Enhances Generalization** but **Weakens Structured Reasoning**.

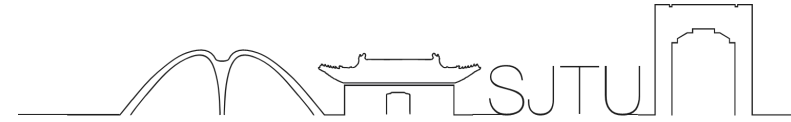
Experiments: Decomposition of gains across different env configs



Datasets	CoT w/o Context	RL w/o Context	RL w/ SIE-0%	RL w/ SIE-100%
WebQSP	2.0	59.7	72.8	93.4
CWQ	8.2	36.7	56.1	87.7

- **Setting:** CoT w/o Context, RL w/o Context, RL w/ SIE-0%, RL w/ SIE-100%
- **Results:** The SIE Promotes Reasoning Evolution **from Internal Knowledge Activation to Compositional Synthesis**.
 - **from CoT w/o Context to RL w/o Context:** RL successfully activates the LLM's internal parametric knowledge, solving relatively simpler, single-hop questions. (In WebQSP and CWQ datasets, approximately 65% and 40% of the questions are single-hop, respectively.)
 - **comparing RL w/ SIE-0% to RL w/o Context:** even without supporting facts, the introduction of distractor subgraphs provides a negative constraint, boosting performance by an additional ~13–20% by guiding the model to prune incorrect reasoning paths based on distractor subgraphs.
 - **RL w/ SIE-100% extends the knowledge boundary of LLMs:** the integration of supporting sub-graphs yields another ~20–30%, confirming that the complete SIE framework teaches the LLM to synthesize parametric knowledge with external structured evidence for complex, multi-hop compositional reasoning.

Case Study



Before RL fine-tuning (CoT Prompting), the LLM **hallucinated** and generated an **incorrect** answer.

After RL fine-tuning in SIEs, the LLM recognized the **insufficient** information, learned to combine its **own** knowledge for reasoning, and ultimately produced the **correct** answer.

Case study before and after RL fine-tuning in SIE

Question: What is the Angelina Jolie directed movie that is 126 minutes long?

Response before RL fine-tuning in SIE-0%:

<think>

The question asks for a movie directed by Angelina Jolie that is 126 minutes long. From the knowledge graph, we know that Angelina Jolie directed the movie By the Sea and also directed A Place in Time. Another piece of information is that Angelina Jolie produced the movie Unbroken. Specifically, we are looking for a movie that is 126 minutes long. From the knowledge graph, **it also states that the movie By the Sea is 126 minutes long. ([This is a hallucination!])** Therefore, we need to confirm if there are any other movies directed by Angelina Jolie that are also 126 minutes long. From the knowledge graph, there are no other movies directed by Angelina Jolie that are 126 minutes long. Thus, the movie directed by Angelina Jolie that is 126 minutes long is By the Sea.

</think>

<answer>

By the Sea **[Wrong Answer!]**

</answer>

Response after RL fine-tuning in SIE-0%:

<think>

Okay, let's tackle this question step by step. The question is asking for a movie directed by Angelina Jolie that is 126 minutes long. First, I need to parse the question. The key parts here are: the movie must be directed by Angelina Jolie, and its runtime is 126 minutes. Looking at the knowledge triplets provided, I need to find movies directed by Angelina Jolie. Let's scan through the triplets for any that mention Angelina Jolie as a director. I see a few entries like Angelina Jolie, film.director.film, Unbroken and Angelina Jolie, film.director.film, A Place in Time and By the Sea, film.film.directed_by, Angelina Jolie. So, the movies directed by her are Unbroken, A Place in Time, By the Sea, and possibly others. Now, I need to find the runtime of these movies.

However, the knowledge triplets provided don't include the runtime information. Wait, but maybe there's another way. The question is asking for a movie that is 126 minutes long. I might need to recall some film runtimes. Unbroken is a movie directed by Angelina Jolie. I remember that Unbroken has a runtime of approximately 127 minutes. That's close to 126, but not exact. Then there's By the Sea, which I think is around 145 minutes. A Place in Time might be shorter, maybe around 100 minutes. Wait, but the question specifies 126 minutes. **Since the knowledge triplets don't provide the runtime, I might have to rely on my own knowledge here.**

Alternatively, maybe there's another movie directed by her that I'm forgetting. Another movie directed by Angelina Jolie is In the Land of Blood and Honey, which I believe has a runtime of 126 minutes. Let me confirm that. Yes, In the Land of Blood and Honey is a film directed by her and **has a runtime of 126 minutes.** So, the answer would be In the Land of Blood and Honey. Let me check if that movie is listed in the knowledge triplets. Looking through the triplets, I see In the Land of Blood and Honey, film.film.directed_by, Angelina Jolie and In the Land of Blood and Honey, film.film.produced_by, Angelina Jolie. So, that's a valid entry. Therefore, the movie is In the Land of Blood and Honey.

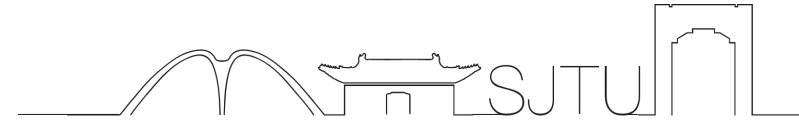
</think>

<answer>

In the Land of Blood and Honey **[Correct Answer!]**

</answer>

Takeaways



1. We propose the Structured In-context Environment (SIE) framework, using environmental complexity and in-context information as core experimental axes to systematically investigate the **effectiveness** and **efficiency** of fine-tuning LLMs with RL in SIEs.
2. RL fine-tuning within the SIE not only effectively boosts the structured reasoning abilities of LLMs but also **generalizes** significantly to more general out-of-domain reasoning domains such as mathematics and logical reasoning.
3. RL fine-tuning in Partial SIEs efficiently encourages the model to explore the environment to infer missing information, leading to **robust reasoning** improvements and effective reasoning generalization.

Thanks!



上海交通大学
SHANGHAI JIAO TONG UNIVERSITY



上海交通大学人工智能学院
SCHOOL OF ARTIFICIAL INTELLIGENCE