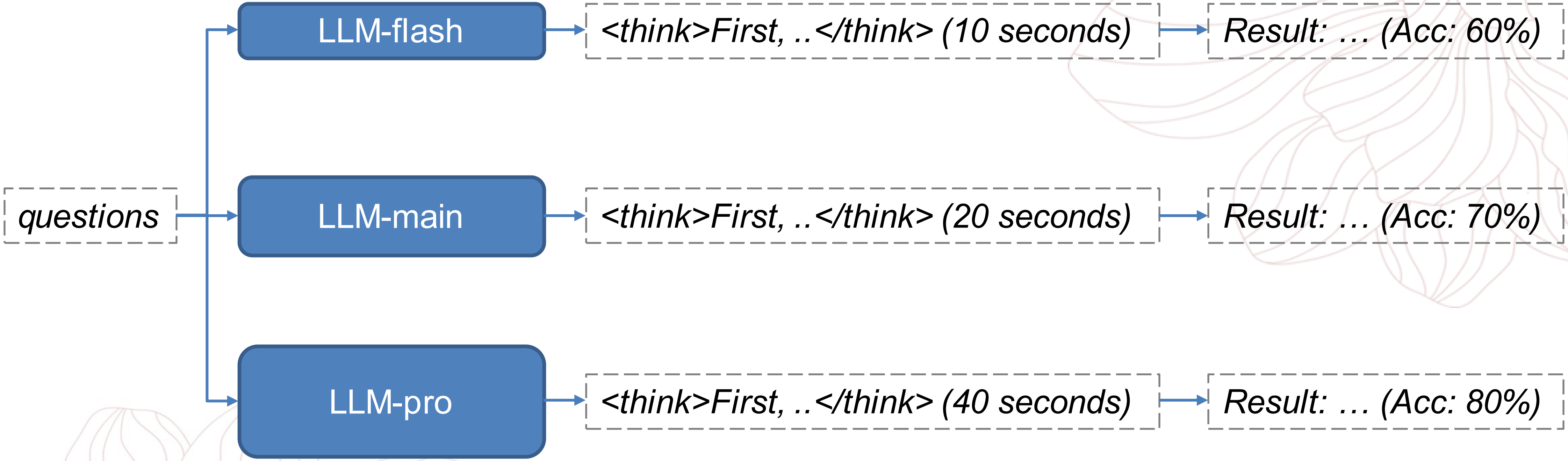
Think Silently, Think Fast: Dynamic Latent Compression of LLM Reasoning Chains

NeurIPS 2025

CoLaR-latent-reasoning.github.io

Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Ruihua Song, Jian Luan

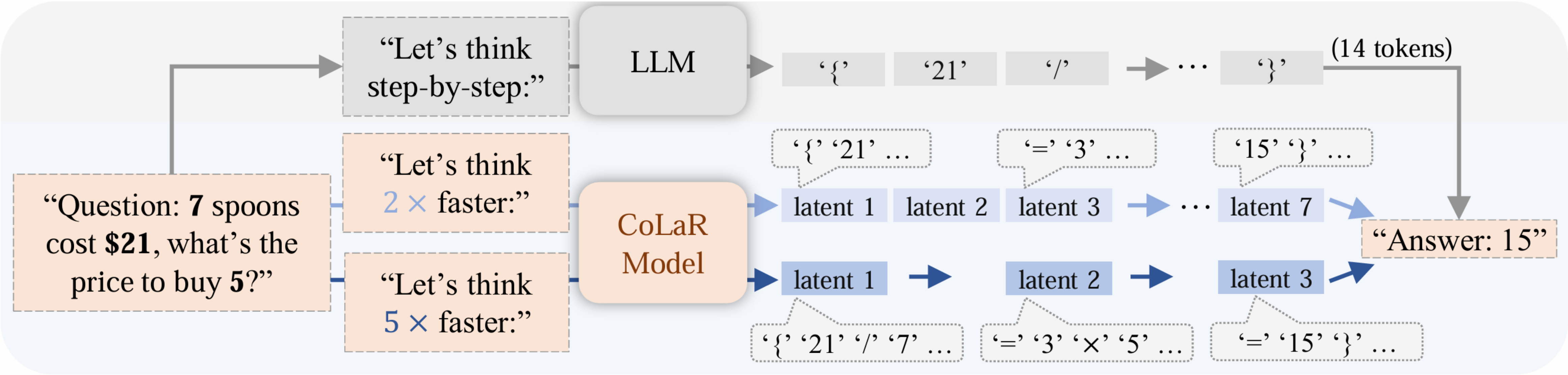
Background: LLM reasoning



Can we use **one omnivorous model** rather than separated models?

Word tokens are too expensive! Must LLMs think in **textual space?**

Overview: Let's think *silently* and *fast*!



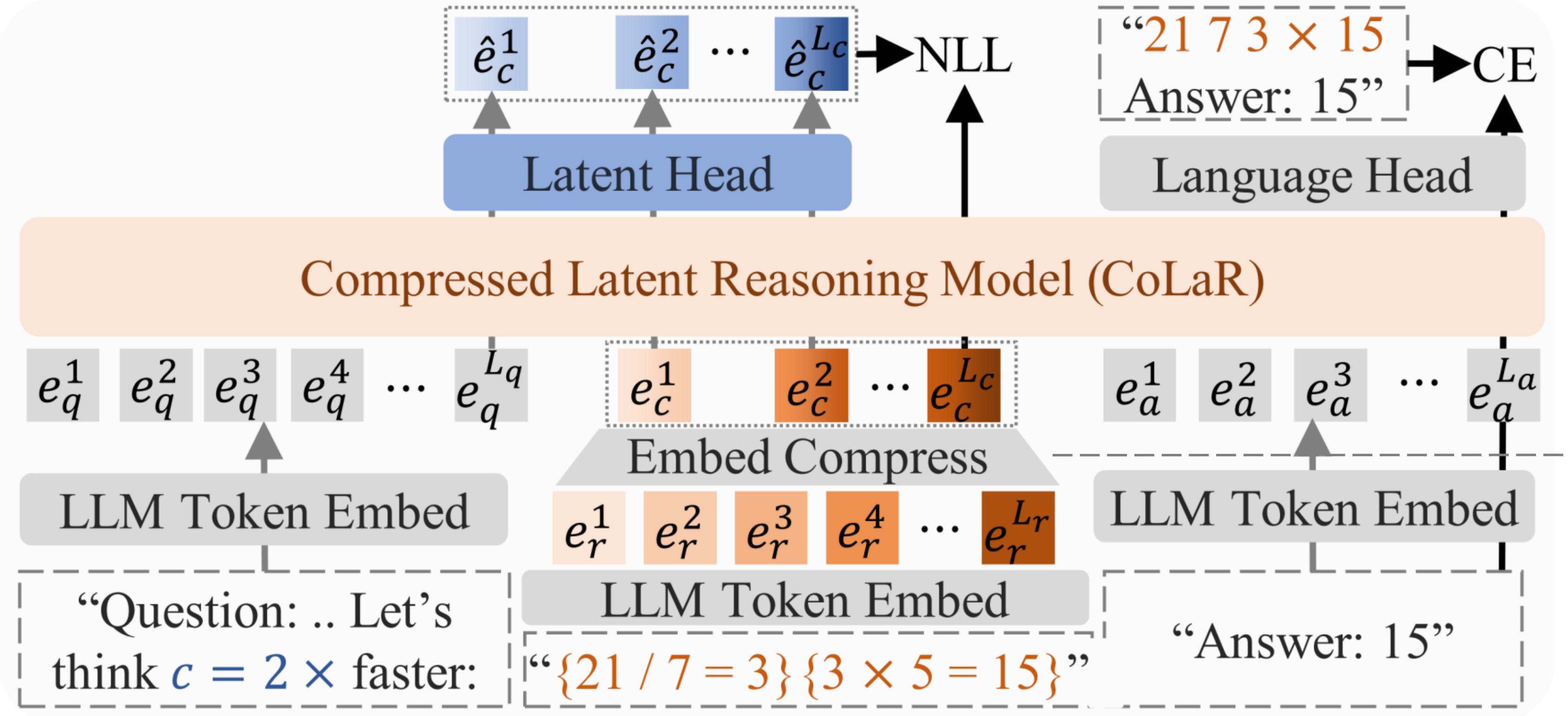
- **Think latent-by-latent**, where **one latent** compresses semantics from **multiple word tokens**
- **Dynamic** and controllable **compression factor** by prompting the thinking speed

Method: SFT training

➢ The model should **compress & predict & understand** latents.

$$\mathcal{L}_{\text{latent}}(i) = -\log p(e_c^i \mid \hat{\mu}_c^i, \hat{\sigma}_c^i) = \frac{(e_c^i - \hat{\mu}_c^i)^2}{2\hat{\sigma}_c^i} + \log \hat{\sigma}_c^i$$

$$\mathcal{L}_{\text{latent}}(i) = \underbrace{\mathbb{E}_\epsilon\left[(\hat{\mu}_c^i + \hat{\sigma}_c^i\epsilon - e_c^i)^2\right]}_{\text{MSE term}} - \alpha \underbrace{\left(\frac{1}{2}\log\left(2\pi e \left(\hat{\sigma}_c^i\right)^2\right)\right)}_{\text{entropy term}}$$

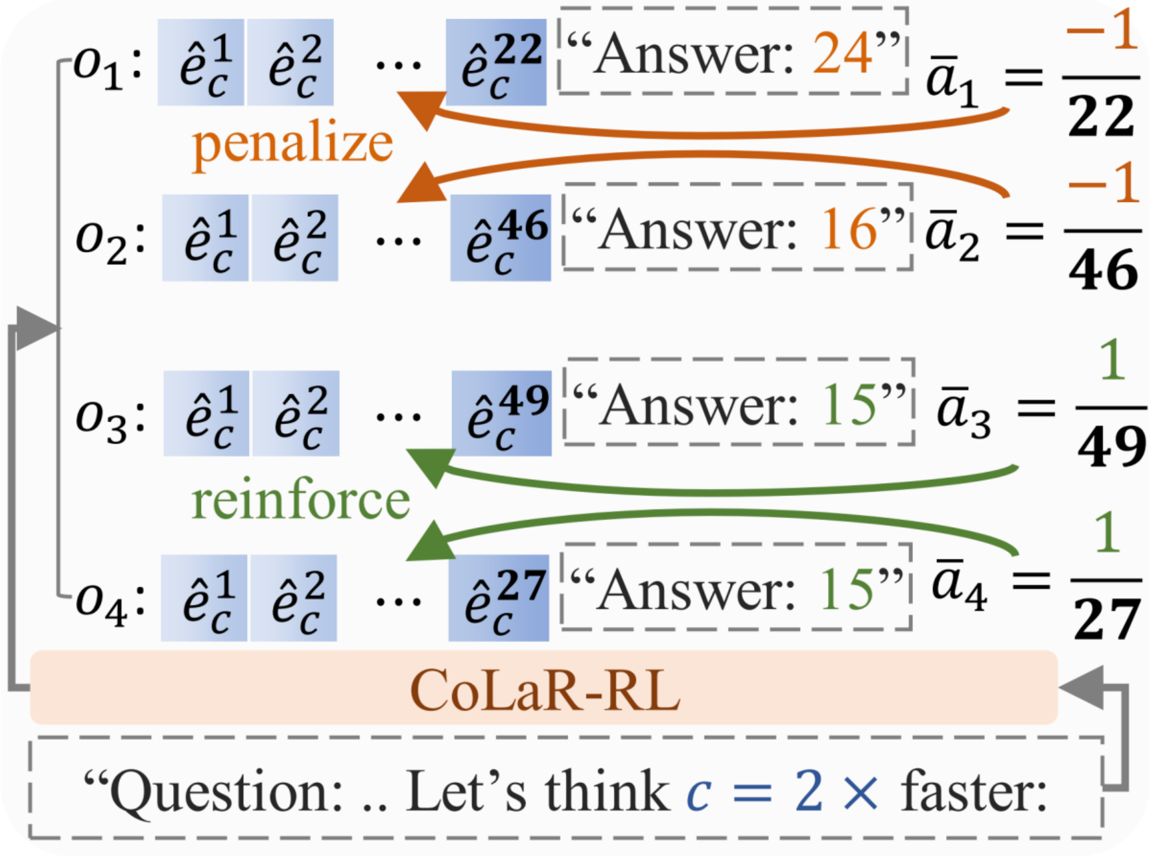


$$\mathcal{L}_{\text{comp}} = -\frac{1}{L_a + L_c} \sum_{i=1}^{L_a + L_c} \log p([\mathbf{t}_c, \mathbf{t}_a]^i \mid [\mathbf{e}_c, \mathbf{e}_a]^{1:i-1}, \mathbf{e}_q),$$

- $e \sim \mathcal{N}(0, \sigma_e)$
- Random two embeddings could be highly uncorrelated (high dimensionality)
- Mean Pooling -> Add & divide by $\sqrt{c}$

Method: RL training

➢ Encourage model to **explore correct** reasoning pathways, and **exploit shorter** ones.



$$\mathcal{L}_{\text{GRPO}} = -\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1-\epsilon, 1+\epsilon\right)A_i\right)\right)$$

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

Experiments: CoLaR vs. SOTA

Table 1: Experiment results of baseline methods and CoLaR on four grade-school math reasoning datasets. We test the methods for five times with different random seeds to report the averaged number and 95% confidence interval ($\pm$) on accuracy (Acc. %) and reasoning chain length (# L). CoLaR-c denotes a same CoLaR model tested with different compression factors c. For ablation methods (marked in gray), suffixes DL, OC, MP and NLL denote CoLaR with a Deterministic Latent head, training withOut Compressed reasoning chain in cross entropy labels, using Mean Pooling to compress embeddings, and training with NLL loss, respectively.

	GSM8k-Aug		GSM-Hard		SVAMP		MultiArith		Average	
	Acc.	# L	Acc.	# L	Acc.	# L	Acc.	# L	Acc.	# L
CoT	$49.4_{\pm.72}$	$25.6_{\pm.11}$	$11.9_{\pm.16}$	$34.2_{\pm.11}$	$59.8_{\pm.29}$	$12.1_{\pm.03}$	$93.2_{\pm.49}$	$13.7_{\pm.09}$	53.6	21.4
iCoT	$19.8_{\pm.23}$	$0.00_{\pm.00}$	$3.87_{\pm.16}$	$0.00_{\pm.00}$	$36.4_{\pm.51}$	$0.00_{\pm.00}$	$38.2_{\pm.66}$	$0.00_{\pm.00}$	24.6	0.00
Coconut	$23.1_{\pm.28}$	$6.00_{\pm.00}$	$5.49_{\pm.33}$	$6.00_{\pm.00}$	$40.7_{\pm.65}$	$6.00_{\pm.00}$	$41.1_{\pm.24}$	$6.00_{\pm.00}$	27.6	6.00
Distill	$13.3_{\pm.62}$	$6.00_{\pm.00}$	$2.97_{\pm.24}$	$6.00_{\pm.00}$	$21.7_{\pm.73}$	$6.00_{\pm.00}$	$19.2_{\pm.83}$	$6.00_{\pm.00}$	14.3	6.00
CoLaR-5	$26.8_{\pm.17}$	$5.57_{\pm.02}$	$5.87_{\pm.10}$	$6.53_{\pm.01}$	$48.4_{\pm.45}$	$2.95_{\pm.02}$	$86.4_{\pm.35}$	$3.21_{\pm.01}$	41.7	4.57
- DL	$26.7_{\pm.11}$	$5.74_{\pm.01}$	$5.53_{\pm.11}$	$8.20_{\pm.04}$	$48.3_{\pm.05}$	$2.90_{\pm.01}$	$84.5_{\pm.19}$	$3.22_{\pm.01}$	41.3	5.02
- OC	$24.8_{\pm.27}$	$5.14_{\pm.12}$	$6.46_{\pm.11}$	$5.49_{\pm.06}$	$46.5_{\pm.18}$	$2.85_{\pm.01}$	$85.9_{\pm.22}$	$3.13_{\pm.01}$	40.1	4.15
- MP	$20.6_{\pm.22}$	$5.61_{\pm.02}$	$4.20_{\pm.07}$	$6.18_{\pm.02}$	$47.7_{\pm.41}$	$2.96_{\pm.01}$	$80.7_{\pm.59}$	$3.20_{\pm.01}$	38.3	4.49
- NLL	$20.3_{\pm.64}$	$5.99_{\pm.06}$	$4.52_{\pm.39}$	$16.6_{\pm.25}$	$43.9_{\pm.43}$	$3.06_{\pm.03}$	$81.6_{\pm.23}$	$3.20_{\pm.02}$	37.6	8.01
CoLaR-2	$40.1_{\pm.20}$	$12.7_{\pm.02}$	$9.08_{\pm.03}$	$14.0_{\pm.07}$	$54.9_{\pm.20}$	$6.11_{\pm.01}$	$91.3_{\pm.12}$	$7.35_{\pm.01}$	48.8	10.0
- DL	$39.7_{\pm.18}$	$12.8_{\pm.01}$	$8.84_{\pm.06}$	$17.2_{\pm.09}$	$54.3_{\pm.23}$	$6.10_{\pm.01}$	$90.1_{\pm.17}$	$7.46_{\pm.01}$	48.2	10.9
- OC	$39.1_{\pm.33}$	$12.3_{\pm.04}$	$8.96_{\pm.01}$	$16.9_{\pm.13}$	$54.7_{\pm.18}$	$6.08_{\pm.02}$	$90.1_{\pm.25}$	$7.36_{\pm.01}$	48.2	10.6
- MP	$36.9_{\pm.30}$	$12.4_{\pm.02}$	$8.46_{\pm.19}$	$12.0_{\pm.05}$	$54.1_{\pm.42}$	$6.14_{\pm.01}$	$86.8_{\pm.20}$	$7.43_{\pm.01}$	46.6	9.49
- NLL	$32.3_{\pm.51}$	$12.2_{\pm.04}$	$7.57_{\pm.16}$	$16.6_{\pm.25}$	$51.0_{\pm.24}$	$5.50_{\pm.03}$	$88.3_{\pm.41}$	$7.09_{\pm.02}$	44.8	10.3

- 14.1% ↑ performance compared to Coconut with shorter reasoning chains
- Reduces reasoning chain length by 53.3% ↓ with only a 4.8% ↓ performance degradation compared to CoT

Table 2: Experimental results on the challenging MATH dataset. We evaluate our proposed method CoLaR on two base models and three settings: -DL denotes using a Deterministic Latent head, -NLL denotes CoLaR trained with NLL Loss as $\mathcal{L}_{\text{latent}}$, which is our main method, and - /w GRPO denotes the post-trained CoLaR-NLL with GRPO reinforcement learning process. We calculate the performance gain between CoLaR-NLL and CoLaR-NLL-RL to highlight the effectiveness of reinforcement learning. Compression factor c and # L_{max} are set to 2 and 128, respectively.

	DeepSeek-R1-Distill-Qwen-1.5B		Llama-3.2-1B-Instruct	
	Acc.	# L	Acc.	# L
CoT	$23.5_{\pm.29}$	$209_{\pm1.6}$	$9.71_{\pm.33}$	$210_{\pm1.4}$
CoLaR-DL	$9.04_{\pm.12}$	$99.4_{\pm.25}$	$3.07_{\pm.28}$	$134_{\pm.46}$
CoLaR-NLL	$8.94_{\pm.21}$	$56.8_{\pm.14}$	$5.28_{\pm.16}$	$83.1_{\pm.52}$
CoLaR-NLL-RL	$14.3_{\pm.25}$ (5.36% ↑)	$9.79_{\pm.40}$ (82.8% ↓)	$7.08_{\pm.07}$ (1.80% ↑)	$16.1_{\pm.14}$ (80.6% ↓)
- w/o average	$13.8_{\pm.14}$	$128_{\pm.00}$	$0.00_{\pm.00}$	$128.0_{\pm.00}$

- 5.36% ↑ accuracy while reducing the length of reasoning chain significantly by 82.8% ↓

Experiments: Case study

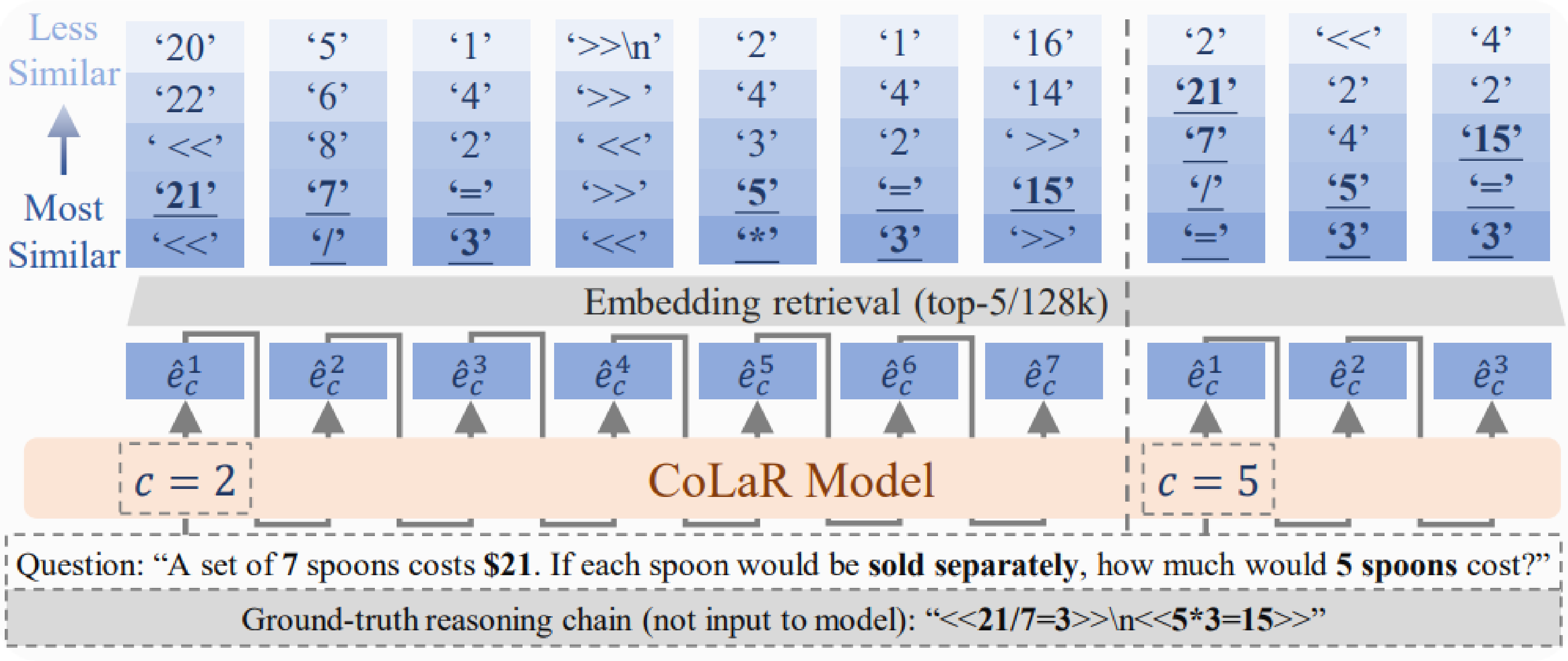


Figure 5: A case study on the GSM8k validation set. We set the compression factor c to 2 and 5, which produce two latent reasoning chains in length 7 and 3, respectively. We then retrieve tokens with the predicted latents by embedding cosine similarity, and underscore those informative tokens.

Higher compression factor captures more tokens while ignoring less informative tokens (like “<<”)

Experiments: Analyses on compression factor c

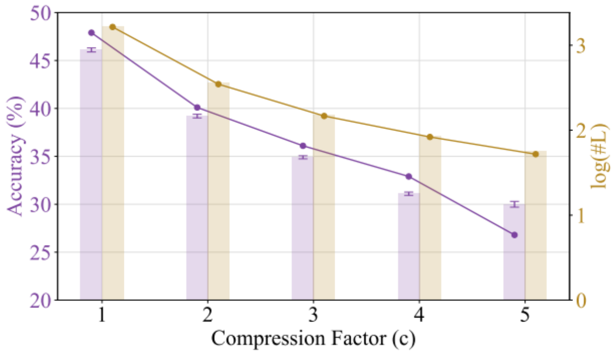


Figure 3: Accuracy and reasoning chain length (# L) of CoLaR on GSM8k dataset when trained with random $c \in [1, 5]$ (the **lines**) or trained solely on specific c (the **bars**).

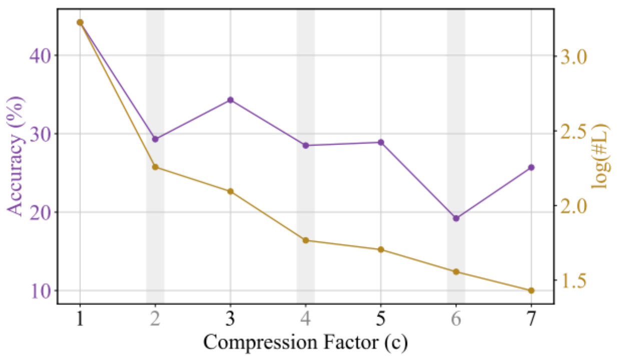
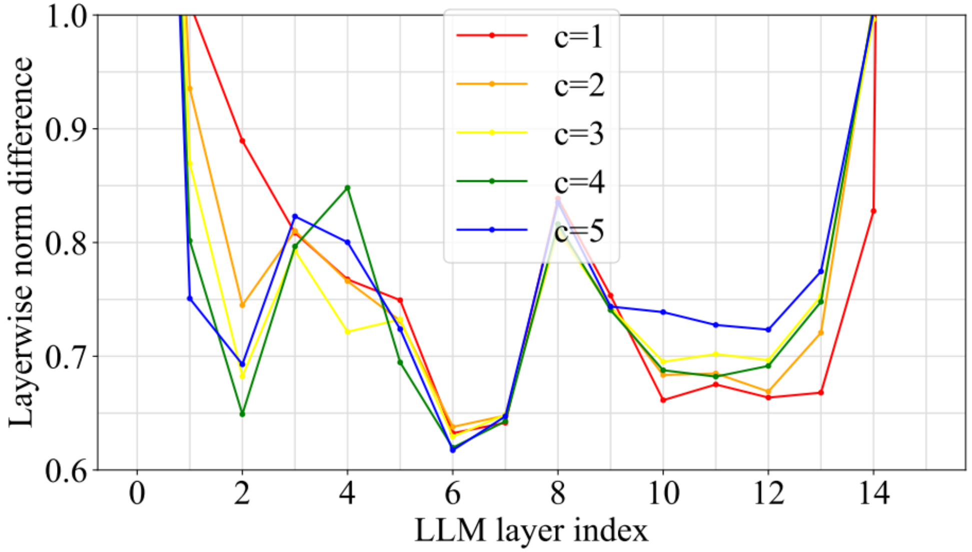
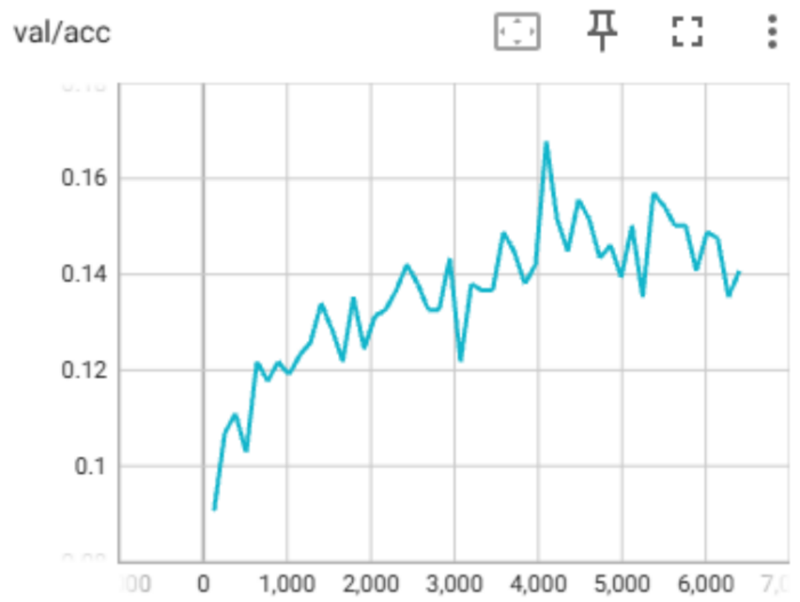
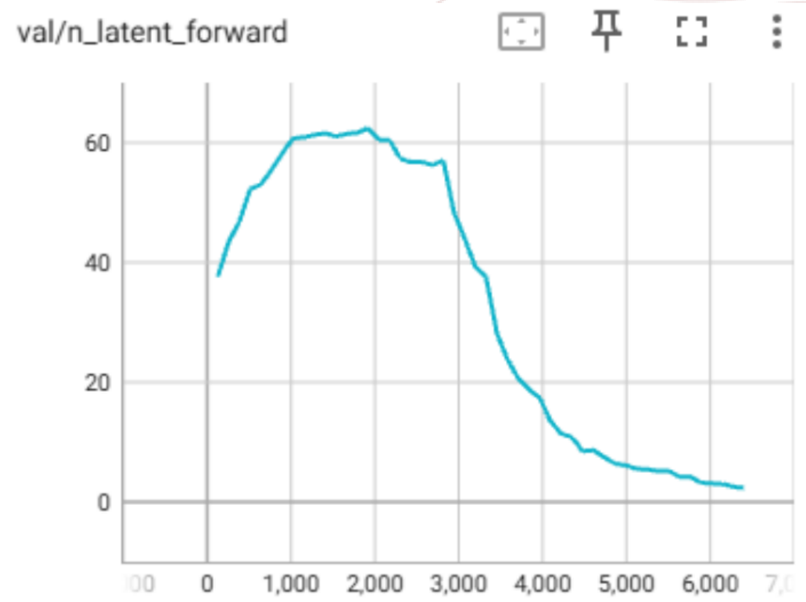


Figure 4: Accuracy and reasoning chain length (# L) of CoLaR on GSM8k dataset when trained with $c \in \{1, 3, 5, 7\}$ and tested with extra $c \in \{2, 4, 6\}$ (under gray bars).



Experiments: RL training curve analyses

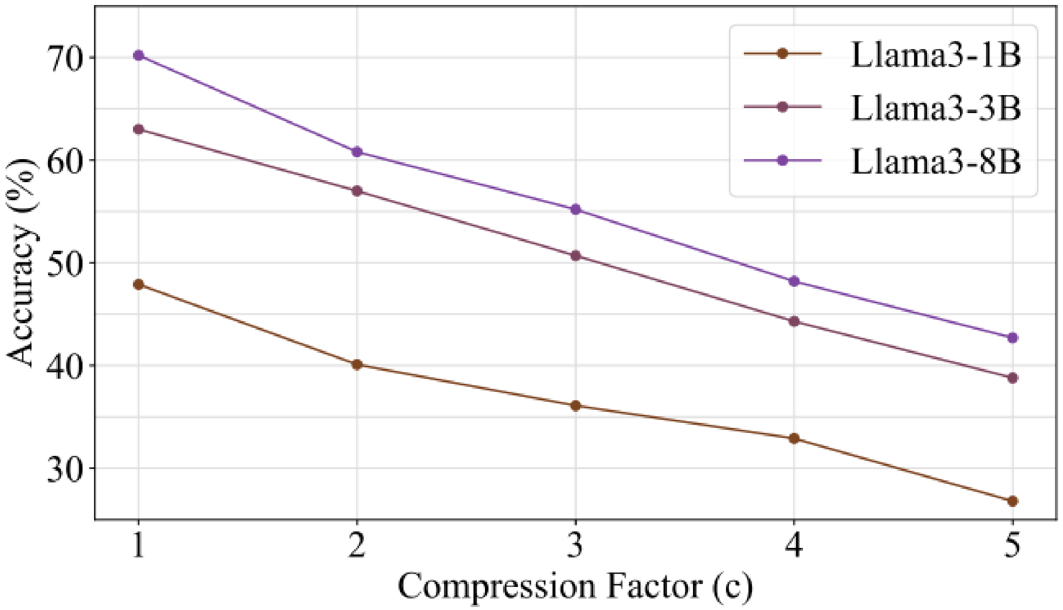




Three stage:

- Exploration: acc and reasoning chain length both increases
- Exploitation: acc fluctuates while reasoning chain length decreases
- Overfitting and early-stopping

Experiments: Model size scaling



Conclusion

Main contributions:

- Novel framework: Compressed latent reasoning with controllable test-time compression factors
- Training pipeline: First work demonstrating the effectiveness of reinforcement learning on latent reasoning

Limitations:

- No significant performance gain applying RL on **simple** math reasoning datasets
- Not surpassing explicit CoT method on Acc.

Future work:

- Multimodality domain
- Adaptive test-time compression factor