KaVa: Latent Reasoning via Compressed KV-Cache Distillation

Anna Kuzina*, Maciej Pióro*, Babak Ehteshami Bejnordi
akuzina@qualcomm.com

Qualcomm AI research

ICLR

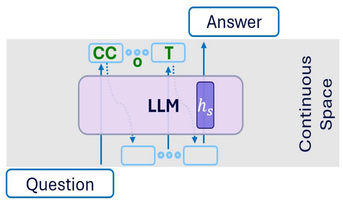
Summary

KaVa distills *compressed KV-cache* trajectories from a teacher model into a latent-reasoning student.

By aligning stepwise KV trajectories with continuous latent tokens, KaVa achieves:
- CoT-level accuracy
- Faster Inference
- Better scaling

Latent Reasoning

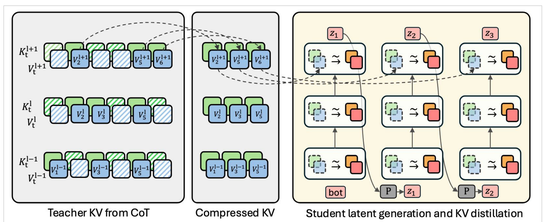
performs "reasoning" in the model's latent space



- More abstract thinking
- Compression of verbose reasoning traces

- No direct supervision signal for latent thought
- Lost interpretability

KaVa: Method Overview



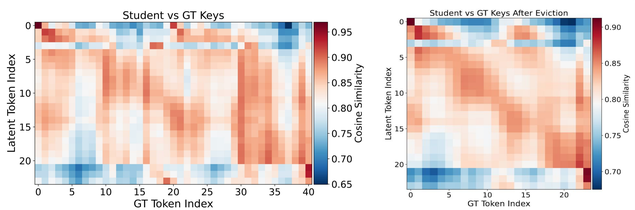
- Use Ground Truth CoT to get teacher's KV-cache
- Apply KV eviction method to compress the KV-cache
- Train the latent reasoning student to generate compressed KV-cache (KV cache distillation)
- Replace CoT with latent tokens at inference

The final training objective $\mathcal{L}_{\text{KaVa}} = \underbrace{-\frac{1}{N_A}\log p(\mathbf{A}|\mathbf{Z},\mathbf{Q})}_{\text{Student loss}} \underbrace{-\frac{1}{N_A+N_C}\log p(\mathbf{A},\mathbf{C}|\mathbf{Q})}_{\text{Teacher loss}} + \underbrace{\alpha_1\mathcal{L}_{\text{CODI}} + \alpha_2\mathcal{L}_{\text{KV}}}_{\text{CODI and KV distillation}}$

Student vs Teacher KV cache

We compare Keys (left) and Values (right) corresponding to the generated latent trace with compressed GT KV cache.

Student learned to generate *compressed KV-cache*



Results

KaVa demonstrates competitive performance requiring fewer forward passes than full CoT

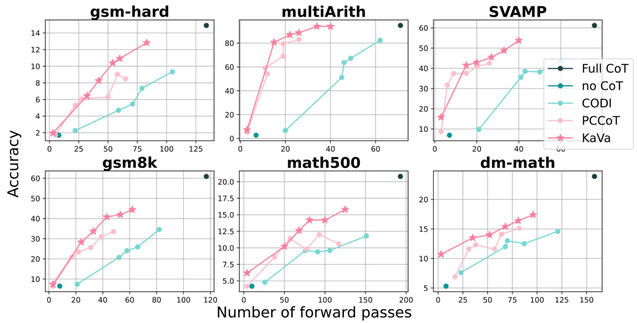
We use 2 training datasets:
- GSM8k-Aug
- GSM8k-Aug-NL

And 3 base models:
- Qwen0.5
- Llama-1b and Llama-3b

Method	GSM8k-AUG			GSM8k-AUG-NL		
	GSM8k	GSM8k-Hard	SVAMP	GSM8k	GSM8k-Hard	SVAMP
Test accuracy						
LLAMA3.2 - 3B - INSTRUCT						
FULL COT	73.2	21.6	78.0	68.4	20.5	77.6
NO-COT	41.7	10.5	56.9	41.7	10.5	56.9
CODI	61.0	15.0	72.4	55.9	13.6	**70.1**
PCCOT	54.7	13.5	69.5	47.6	11.0	65.2
KAVA (ours)	**65.7**	**15.2**	**72.7**	**60.0**	**14.8**	66.1
Efficiency (# forward passes)						
LLAMA3.2 - 3B - INSTRUCT						
FULL COT	31.6	40.3	17.0	75.2	32.9	38.3
NO-COT / ICOT	**6.1**	**7.4**	6.1	6.1	**7.4**	6.1
CODI / COCONUT	11.5	14.2	11.0	11.1	13.1	10.7
KAVA (ours) / PCCOT	6.4 (-80%)	8.2 (-80%)	**6** (-65%)	**6** (-92%)	7.9 (-76%)	**5.7** (-85%)

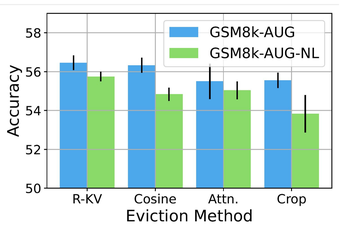
Scaling to Longer Traces

On MetaMathQA dataset, replacing 20–100% of CoT tokens with 24 latent tokens maintains high accuracy, demonstrating KaVa's scalability.



Ablations

Better eviction strategy improves KaVa



KV loss improves method robustness to data processing

Table 4: Test accuracy on GSM8k dataset when the teacher is trained on all the steps.

$\mathcal{L}_{KD}$	$\mathcal{L}_{KV}$	Drop Last	All Steps
✓	✓	**56.5 (0.4)**	51.2 (0.8)
✓	✗	*53.35 (0.18)*	47.2 (2.9)

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.