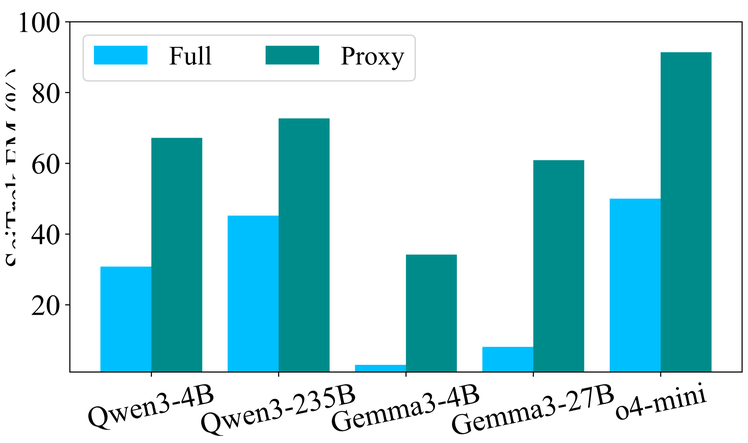
Long-Context Reasoning Through Proxy-Based Chain-of-Thought Tuning

Miao Li, Irina Saparina, Alexander Gurung, Mirella Lapata

The 64th Annual Meeting of the Association for Computational Linguistics

LLM reasoning capabilities can be transferred from short proxy contexts to full long contexts

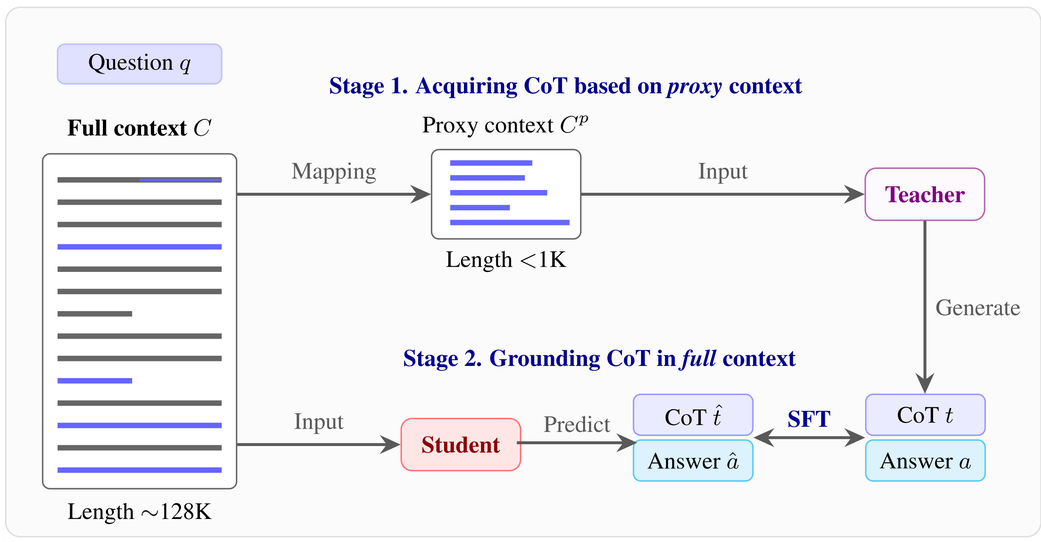
1. LLMs exhibit performance disparity on proxy and full contexts

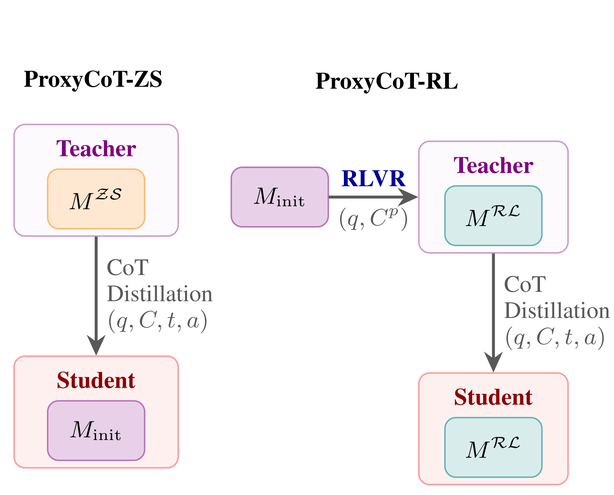


- Proxy Context, a subset of the long input, provides all evidence required to get correct output

- LLMs should essentially share the same underlying reasoning process when working on the two types of contexts

2. Can we transfer reasoning capabilities from short proxy contexts to full long contexts? Our solution: ProxyCoT, a two-stage training framework





General two-stage pipeline of ProxyCoT (left), and two instantiations (right): ProxyCoT-ZS and ProxyCoT-RL

M_{init}: the initial model, M^{ZS}: a large teacher model (zero-shot), M^{RL}: a small teacher model trained by RLVR on short proxy contexts, CoT distillation: supervised fine-tuning

3. Finding: ProxyCoT improves long-context reasoning across models (SciTrek in EM)

Model	Training Strategy	Proxy↑	Full↑
Qwen3-235B-Instruct	Zero-shot	72.7	45.2
Qwen3-235B-Thinking	Zero-shot	85.6	48.8
Qwen3-4B-Instruct	Zero-shot	67.2	30.8
	SFT on C	39.0	19.5
	RLVR on C	66.1	32.9
	SFT on C, CoT*	45.8	31.6
	ProxyCoT-ZS	67.8	38.8
	ProxyCoT-RL	88.5	46.5
Gemma3-4B-IT	Zero-shot	34.2	3.0
	SFT on C	19.1	12.7
	RLVR on C	39.9	5.5
	SFT on C, CoT*	53.1	36.9
	ProxyCoT-ZS	64.2	36.5
	ProxyCoT-RL	69.8	43.7

q	Luis Gianneo was teacher of which chief exponent of Argentine folk music?
C	{67 full-text Wikipedia articles}
C^p	Luis Gianneo (1897–1968) was an Argentine composer, pianist and conductor. As music educator, he was the teacher of composers Ariel Ramirez, Juan Carlos Zorzi, Virtú Maragno, Pedro Ignacio Calderón and Rodolfo Arizaga, among others.

Extended example for HotpotQA :
q: question,
C : abbreviated long context,
C^p : proxy context

$$|C^p| \ll |C|$$

4. Finding: ProxyCoT performs well across datasets (Extended HotpotQA in LLM scoring)

Model	Training Strategy	Proxy↑	Full↑
Qwen3-235B-Instruct	Zero-shot	92.1	60.8
Qwen3-235B-Thinking	Zero-shot	93.2	50.7
Qwen3-4B-Instruct	Zero-shot	91.3	44.5
	SFT on C	92.6	48.8
	RLVR on C	88.6	48.1
	SFT on C, CoT*	84.5	40.2
	ProxyCoT-ZS	91.4	50.3
	ProxyCoT-RL	92.1	52.7

5. Finding: ProxyCoT-RL reduces inference compute

Training Strategy	CoT Tokens↓
Zero-shot	1,744
RLVR on C	937
SFT on C, CoT*	6,683
ProxyCoT-ZS	5,520
ProxyCoT-RL	617

6. Finding: ProxyCoT pushes LLMs better out-of-domain performance on Loong

Model	Training	Financial↑
Qwen3-4B-Instruct	Zero-shot	37.76
	ProxyCoT-RL	40.83
Gemma3-4B-IT	Zero-shot	25.85
	ProxyCoT-RL	32.05

7. Ablation: Reasoning trace generation, the first stage of ProxyCoT-RL, is robust to noise incorporated into the proxy context

Dataset	Oracle: Noise	#Tokens	Performance↑
SciTrek	1:5	5,471	85.3
	1:2	2,588	89.9
	1:1	1,622	91.3
	1:0	659	91.5
HotpotQA	1:5	1,853	83.7
	1:2	922	88.4
	1:1	607	90.1
	1:0	301	92.2