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

Miao Li, Irina Saparina, ***Alexander Gurung,*** *Mirella Lapata*

Reasoning in Long-Context QA

Given question and long-context input, step-by-step reason and give answer

Luis Gianneo was teacher of which chief exponent of Argentine folk music?

Post-Training Methods for Reasoning Models

Chain-of-Thought Distillation

Reinforcement Learning

Long-Context Post-Training is Challenging

Performant teacher reliance

Even large models perform poorly on long contexts, decreasing distilled reasoning quality

Sample inefficiency

Long contexts substantially increase rollout cost and exacerbate credit assignment

However, Not All of the Context Is Needed

What is the total number of words in the titles of all articles that have exactly 60 references?

Full-text long context (128K tokens)

Article 1: "Attention Is All You Need"
Vaswani et al.
The dominant sequence transduction models are based on complex recurrent. . .
Article 2: "BERT: Pre-training. . . "
Devlin et al.
We introduce a new language representation model called BERT. . .
[+ 50 more full-text articles]

Meta-data proxy context (<1k tokens)

articles table

id	title	words	authors	refs
A1	Attention Is. . .	4	8	31
A2	BERT: Pre-tr. . .	9	4	60
A3	Transformer. . .	5	3	45
A4	GPT-3 Paper	2	12	60
A5	Deep Learn. . .	6	5	28

citations table

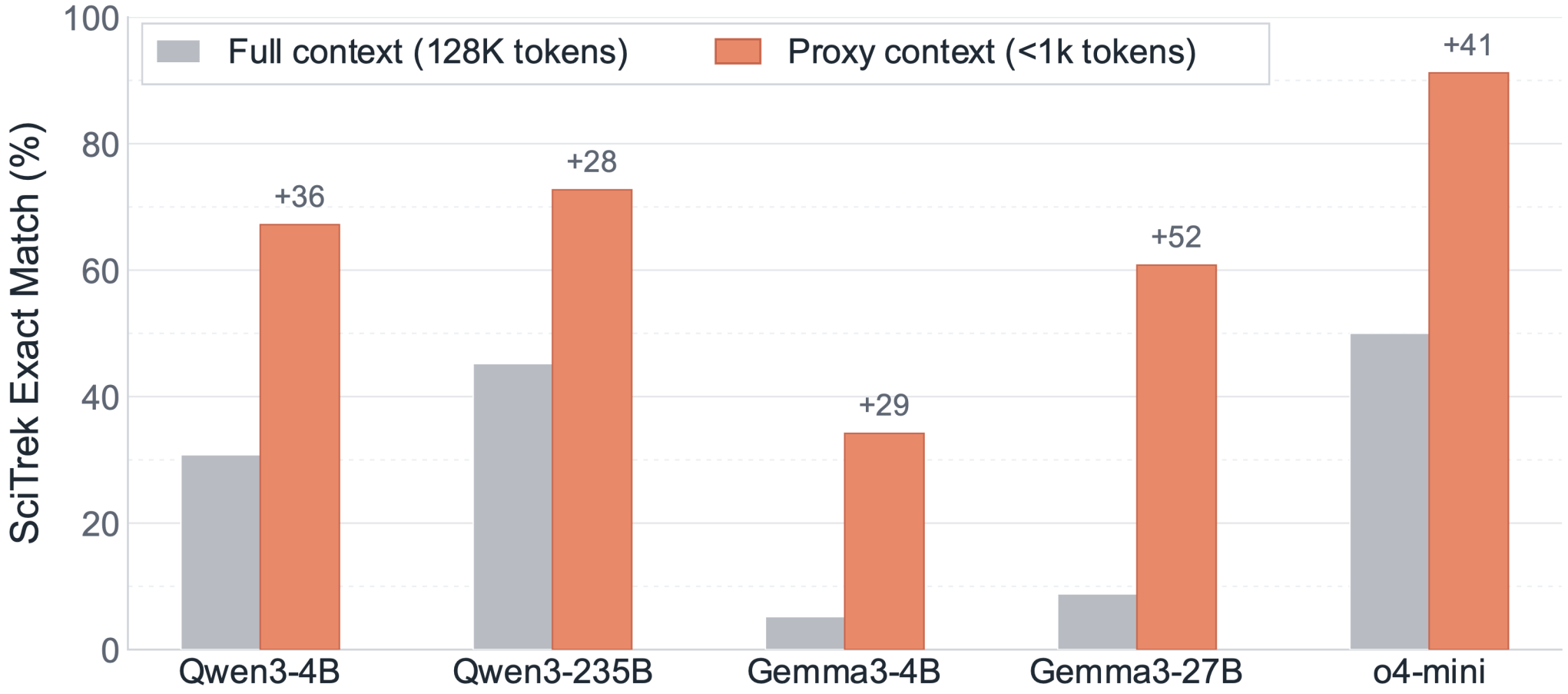
citing	cited	#
A1	A3	2
A2	A1	3
A2	A5	1
A3	A1	1
A4	A2	2

+ article-author table

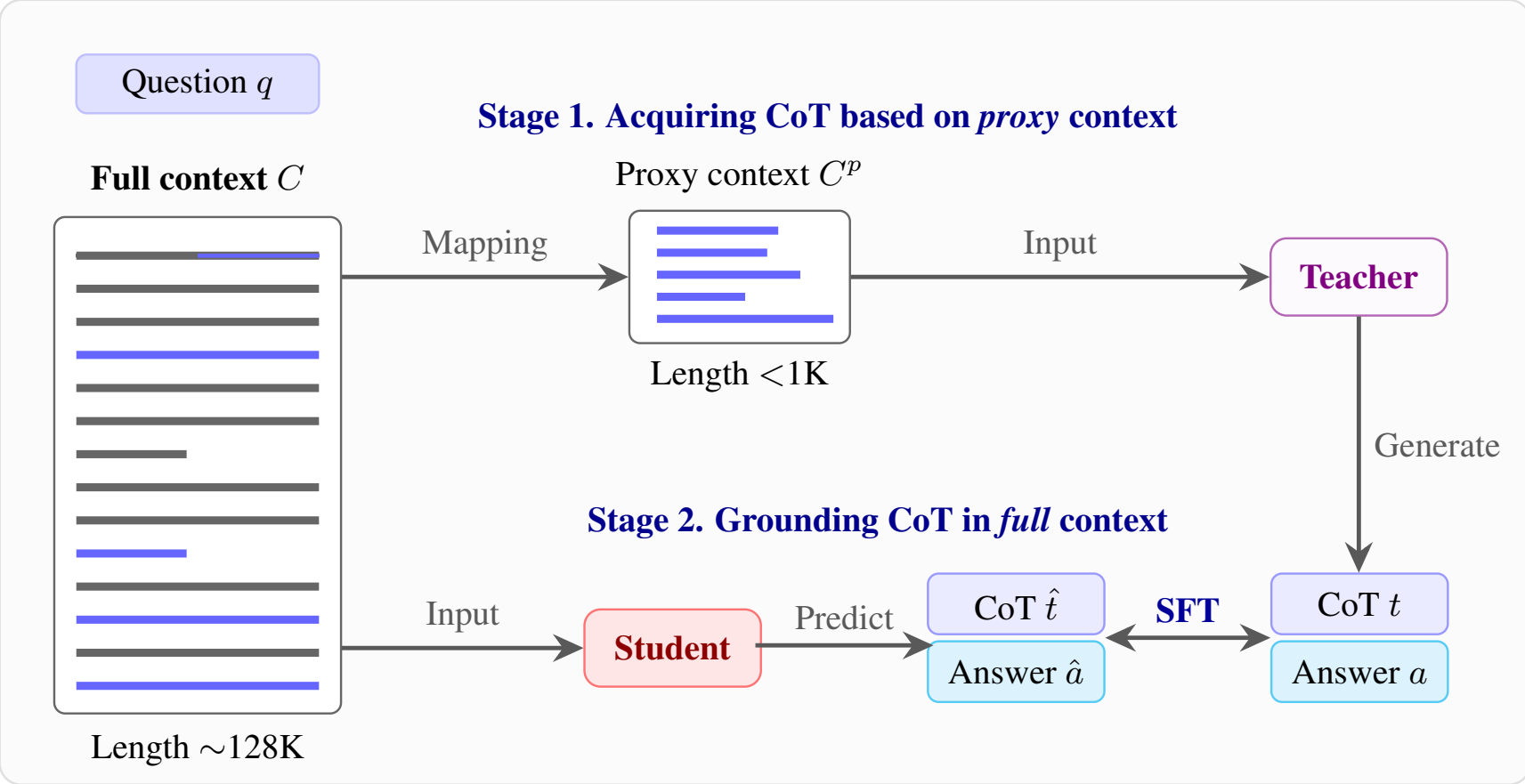
Example from SciTrek

Li et al., 2026. Who Gets Cited Most? Benchmarking Long-Context Numerical Reasoning on Scientific Articles

Performance on Proxy Context Is Much Better!

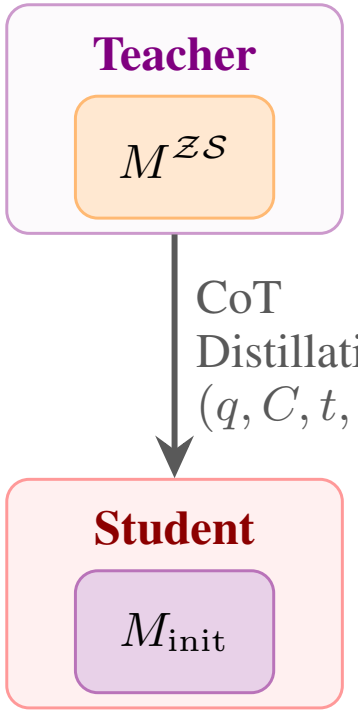


ProxyCoT: Transfering Proxy-Context Capabilities to Long Contexts

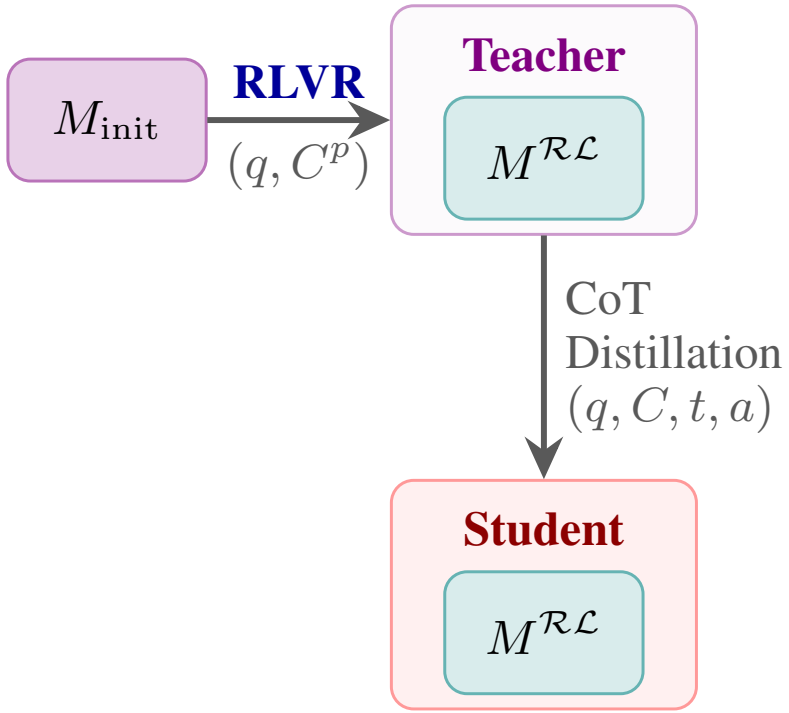


Our post-training framework for long-context reasoning

Two Teacher Instantiations



ProxyCoT-ZS: using zero-shot large model as the teacher



ProxyCoT-RL: RL train a small model as the teacher *with proxy contexts,* init student from teacher

Experiments

Baseline methods (given long-context):

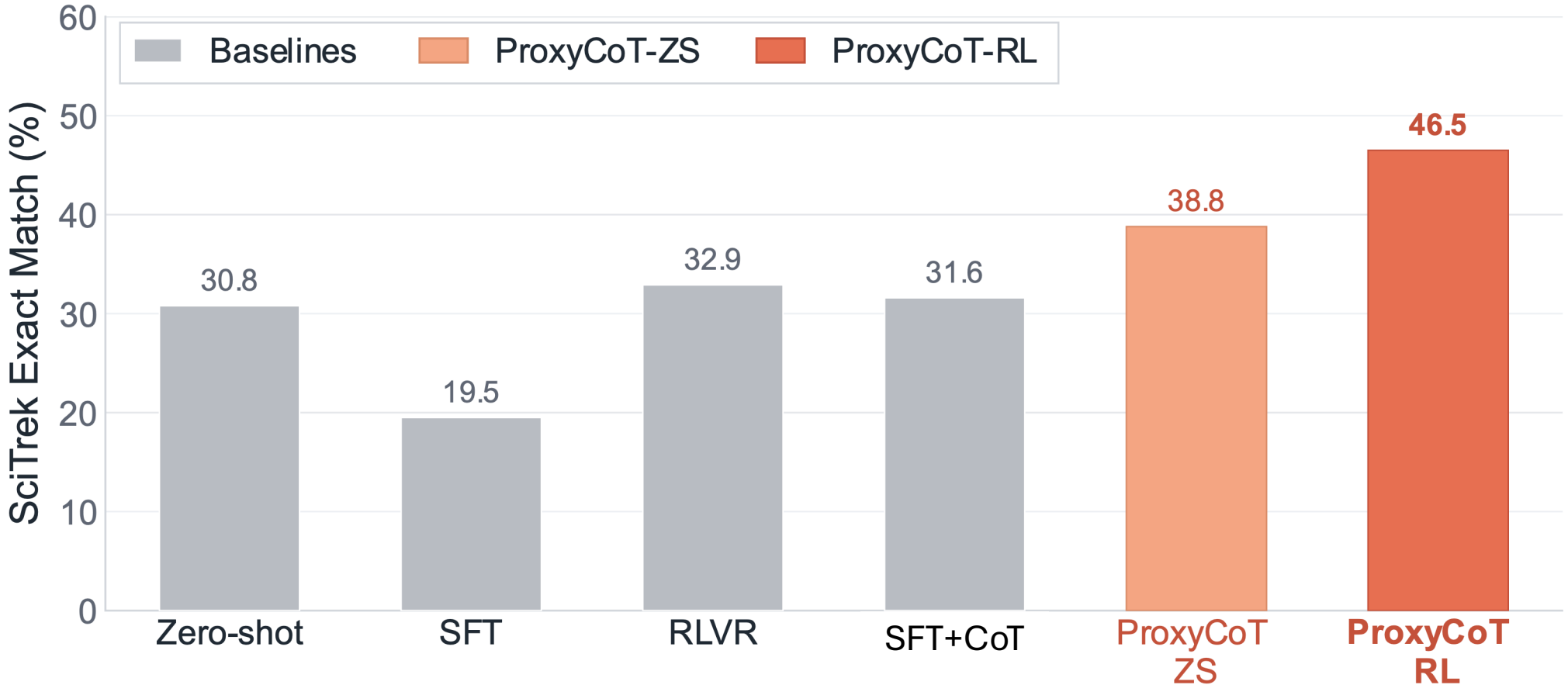
- **Zero-Shot**
- **SFT** - question → answer
- **SFT+CoT** - distillation from a strong teacher (Qwen3 235B)
- **RLVR** - outcome-based GRPO

Initial model: Qwen3 4B Instruct

Dataset: SciTrek

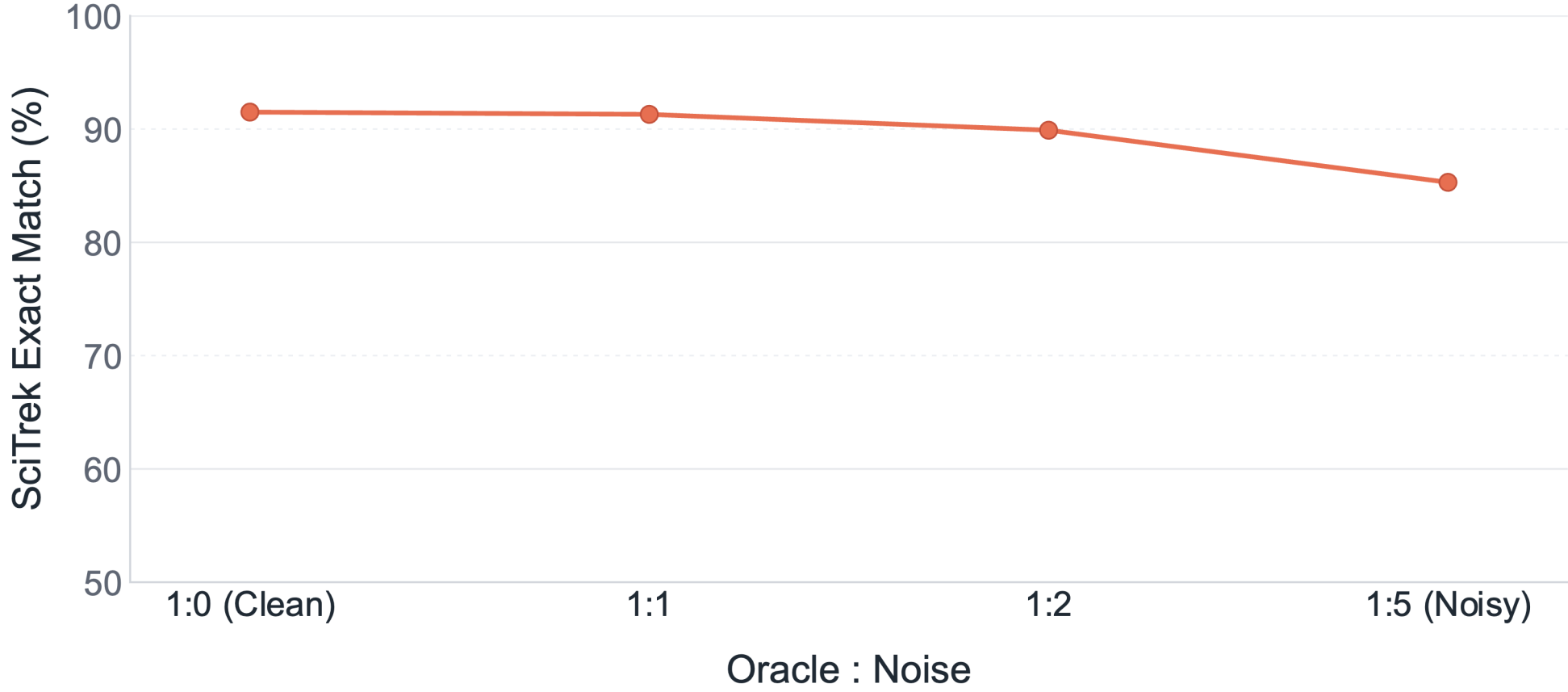
*See our paper for HotpotQA experiments

ProxyCoT Outperforms Strong Long-Context Baselines



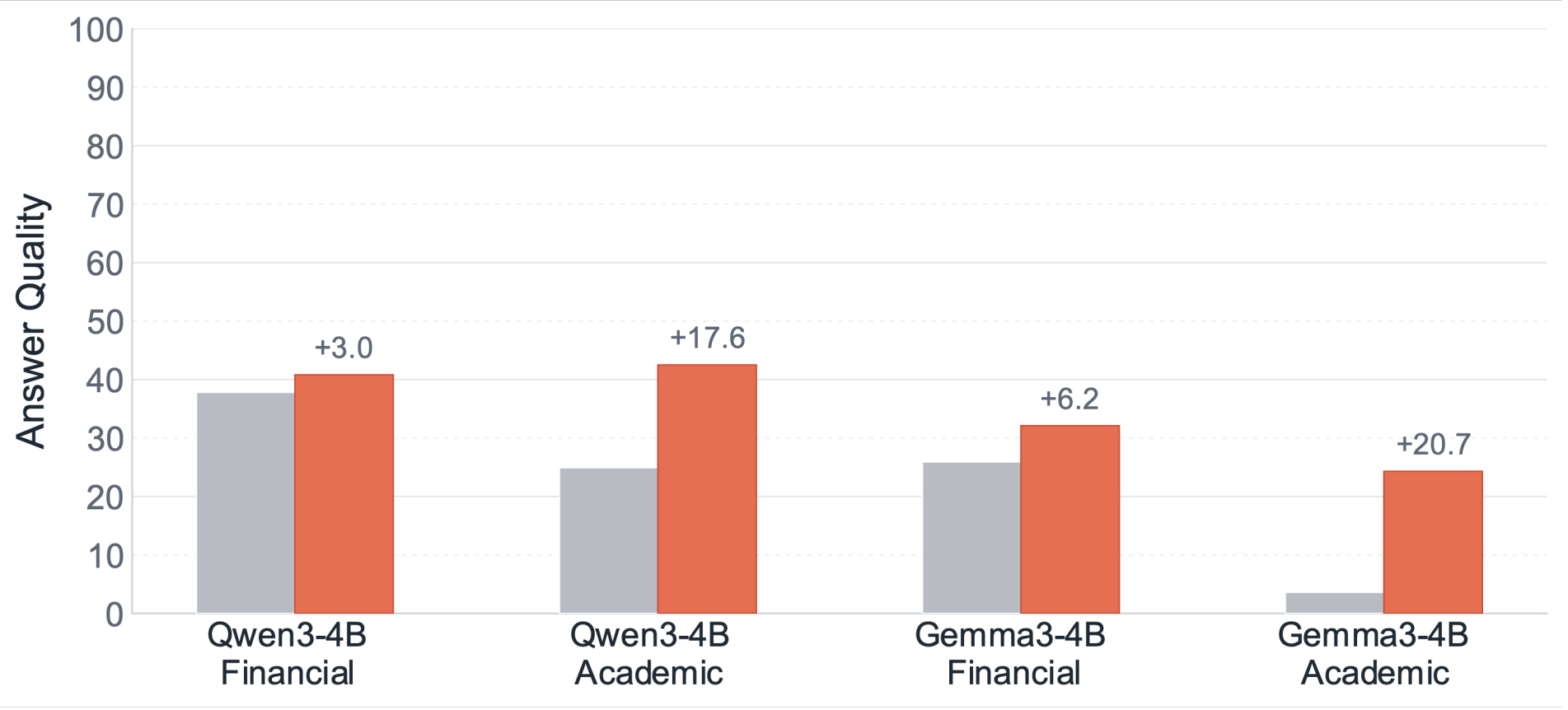
Proxy Performance Is Robust to Noise

- We add random sentences from other datapoints to proxy
- Silver-standard (high noise) performance is stable



ProxyCoT Long-Context Capabilities Generalize

Trained on SciTrek, evaluated on **Loong**



Wang et al., 2024. Leave No Document Behind: Benchmarking Long-Context LLMs with Extended Multi-Doc QA.

Conclusion

Proxy-based reasoning improves long-context capabilities

Also in the paper:

- **More experiments on HotpotQA** (long-context wiki articles, proxy is gold evidence paragraphs)
- **Ablations** on ProxyCoT variants
- **Efficiency** analysis of reasoning traces

Paper