PhD Symposium (Poster session)

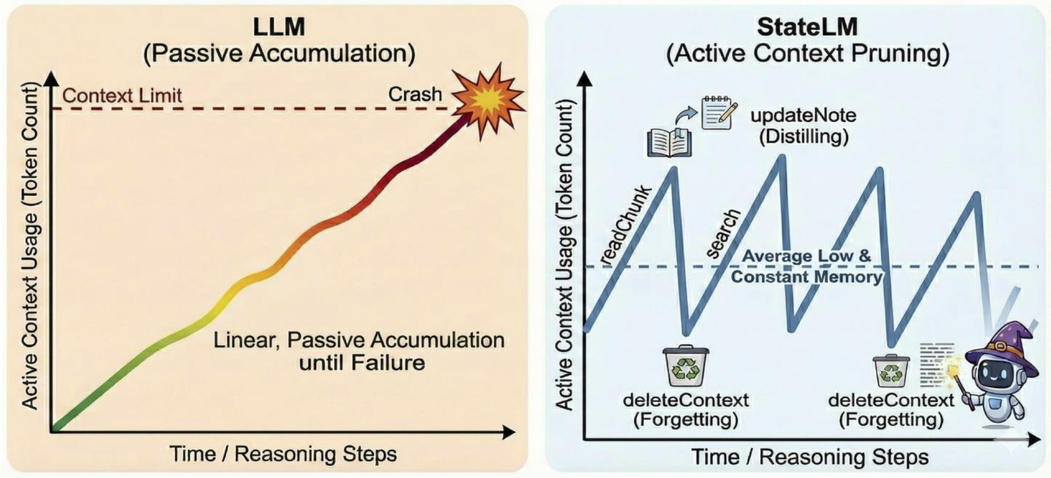
The Pensieve Paradigm: Stateful Language Models with Learned Memory Management

Xiaoyuan Liu[1,2], Tian Liang[2], Dongyang Ma[2], Deyu Zhou[2], Haitao Mi[2], Pinjia He[1], Yan Wang[2]

[1]The Chinese University of Hong Kong, Shenzhen

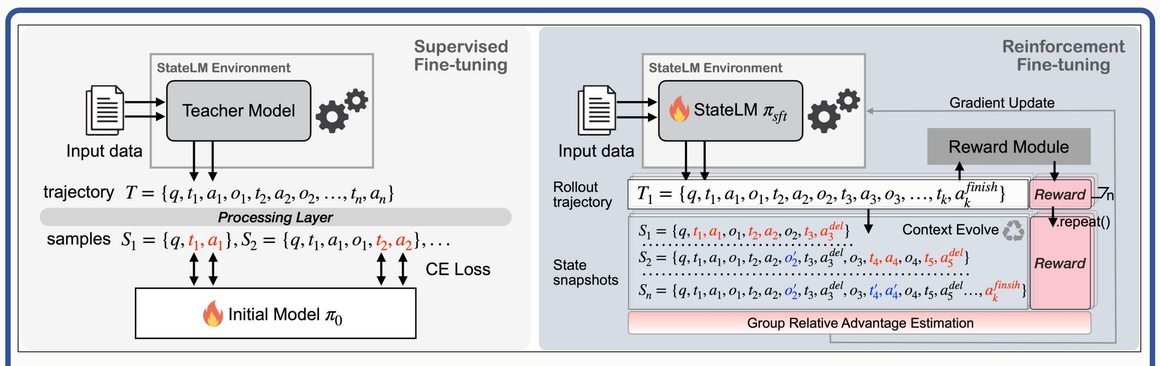
[2]Tencent AI Lab

Motivation



Current LLMs treat their context window as a passive buffer, retaining all tokens without discrimination. This unmanaged accumulation creates two critical issues: (1) **context grows rapidly** and soon exceeds native context limits; (2) **reasoning capability degrades**, since important information becomes buried in irrelevant or outdated content, making it increasingly difficult for the model to distinguish signal from noise.

Method

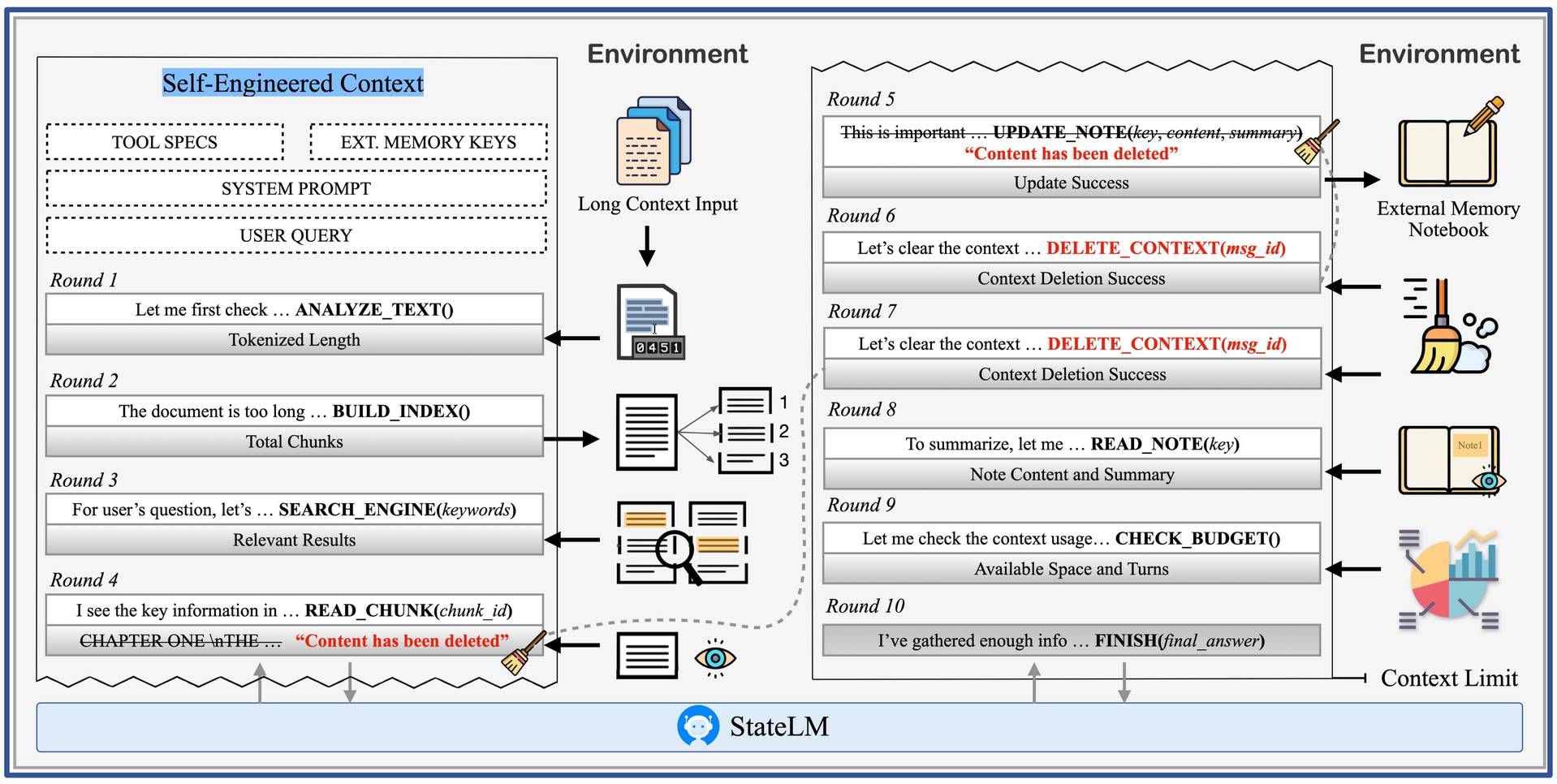


SFT: Supervised Learning from Expert Trajectories (35.7K)

- Guideline Engineering: design system prompt for the teacher model
- Outcome-based Reject Sampling: filter the incorrect final answers
- Process-based Reject Sampling: filter the pre-defined bad behaviors
- Action (tool-call) Balancing: downsample the over-represented samples

RL: Reinforcement Learning for Self-Improvement (488)

- Trajectory Snapshotting: record the traj before each context change
- Controlled Batching: keep a fixed number of training samples per rollout
- GRPO-based Optimization: correctness rewards with group-based adv.



Tool Set

Context Perception
analyzeText()
checkBudget()
loadDocument()

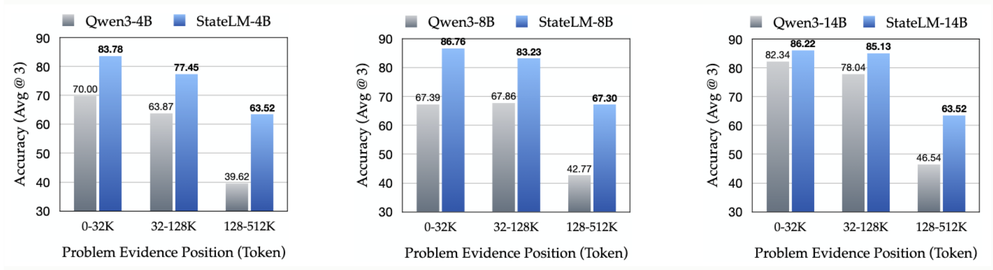
Info Acquisition
buildIndex(size, overlap)
searchEngine(keywords)
readChunk(chunk_id)

Memory Management
note(key, summary, content)
updateNote(mode, key, new_content)
readNote(key)
deleteContext(msg_id)

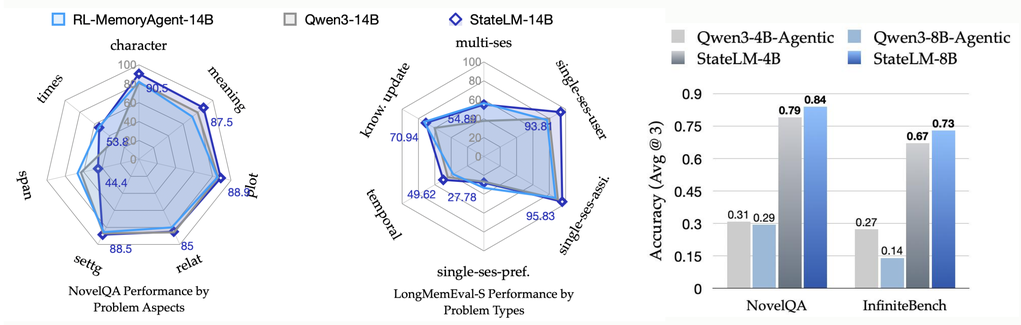
Termination
finish(answer)

Experiment and Analysis

Model	Context	LongDoc QA (135K)		Chat Memory (115K)	*BrowseComp+ (552K)
		NovelQA	∞Bench		
Claude-4-Sonnet	200K	85.60	85.59	-	-
Qwen3-235B-A22B-Ins.	256K	80.85	72.37	-	-
RL-MemoryAgent-7B	32K	30.38	34.93	40.60	-
RL-MemoryAgent-14B	32K	39.50	45.85	59.00	-
ReadAgent-8B	32K	16.38	24.02	0.00	-
ReadAgent-14B	32K	23.12	34.06	14.60	-
Qwen3-4B	128K	65.17 ± 0.53	59.97 ± 0.50	39.53 ± 1.36	2.89 ± 1.02
Qwen3-8B	128K	65.87 ± 1.42	66.81 ± 1.16	45.40 ± 1.56	5.56 ± 0.77
Qwen3-14B	128K	77.94 ± 0.26	74.96 ± 0.25	54.07 ± 0.76	5.46 ± 0.04
StateLM-4B	32K	**79.57** ± 0.93	**67.25** ± 0.76	**59.33** ± 0.23	**35.11** ± 1.39
StateLM-8B	32K	83.84 ± 0.42	70.16 ± 1.33	58.93 ± 2.53	**39.33** ± 2.67
↳StateLM-8B-RL	32K	**84.15** ± 1.00	**73.07** ± 1.33	**59.73** ± 2.20	**39.33** ± 2.00
StateLM-14B	32K	84.15 ± 0.82	77.44 ± 1.40	64.40 ± 1.40	42.67 ± 2.00
↳StateLM-14B-RL	32K	**84.85** ± 0.42	**78.46** ± 0.67	**64.47** ± 0.50	**43.78** ± 2.77



Model	Length							
	32K	64K	128K	256K	512K	768K	1M	2M
Qwen3-8B	**100.00**	**100.00**	88.33	41.67	23.33	16.67	3.33	1.7
StateLM-8B (w/o search)	99.44	98.33	**98.33**	**99.44**	**95.55**	**88.89**	**88.89**	**67.22**
Qwen3-14B	**100.00**	**100.00**	88.33	41.67	23.33	16.67	3.33	1.7
StateLM-14B (w/o search)	**100.00**	**100.00**	**99.44**	**97.78**	**89.45**	**94.44**	**95.00**	**83.89**



- **StateLMs outperform instruct model baselines** on the Needle-in-a-Haystack benchmark under **extreme context pressure.**
- On **long-context QA** tasks, StateLMs surpass instruct baselines while using **only 1/4** of the context window, and also outperform other agentic methods, especially under longer input scenarios.
- StateLMs substantially exceed their **untrained** agentic (tool-use) counterparts, showing the necessity of explicit agentic training.