CraniMem: Cranial Inspired Gated and Bounded Memory for Agentic Systems

Pearl Mody, Mihir Panchal, Rishit Kar, Kiran Bhowmick, Ruhina Karani

MemAgents

ICLR International Conference On Learning Representations

The Problem: Unbounded Memory

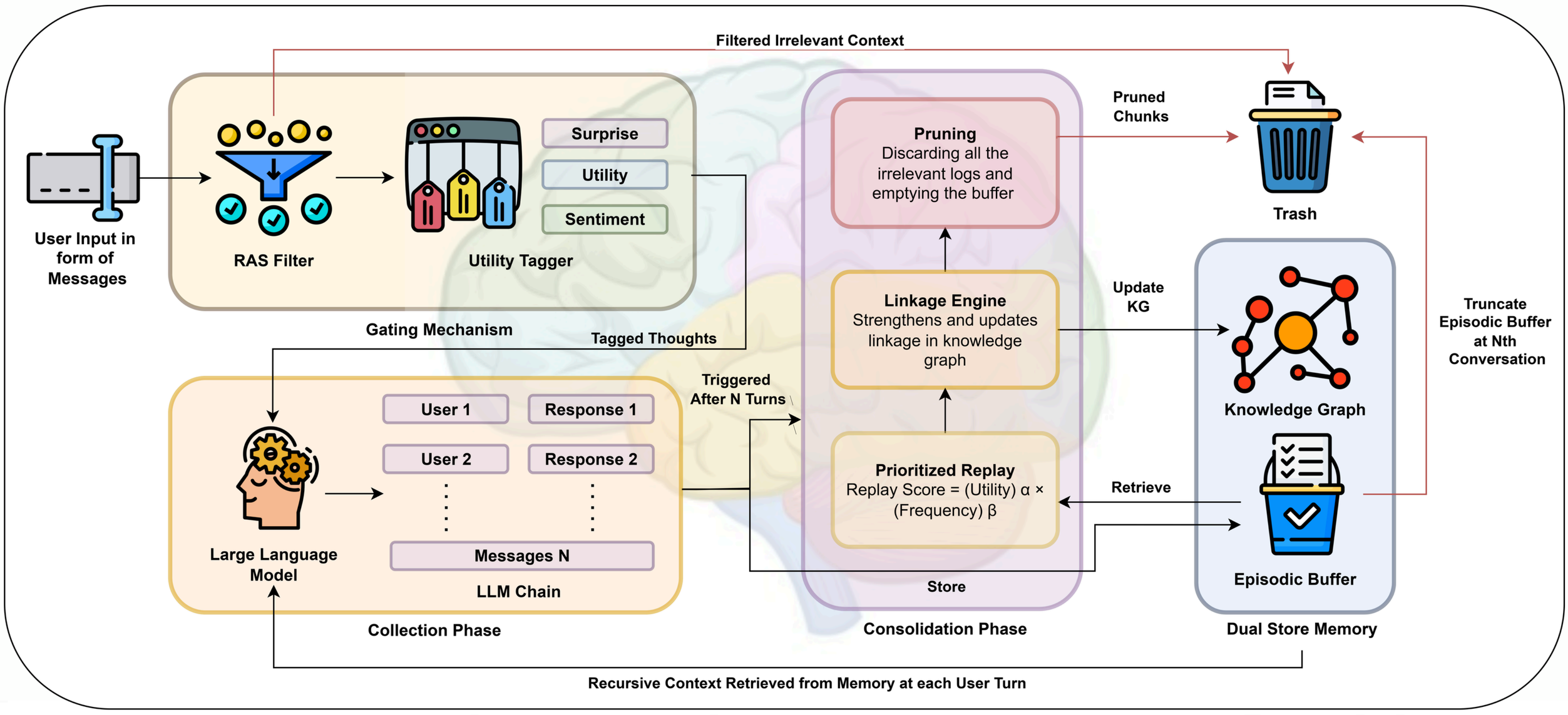
Long-horizon agents currently fail because they treat memory like an unbounded external database, leaving them vulnerable to distractions and overwritten facts. How can we stop agents from drowning in noise? By mimicking the human brain's gated, multi-stage consolidation process to ensure they only encode and remember what actually matters.

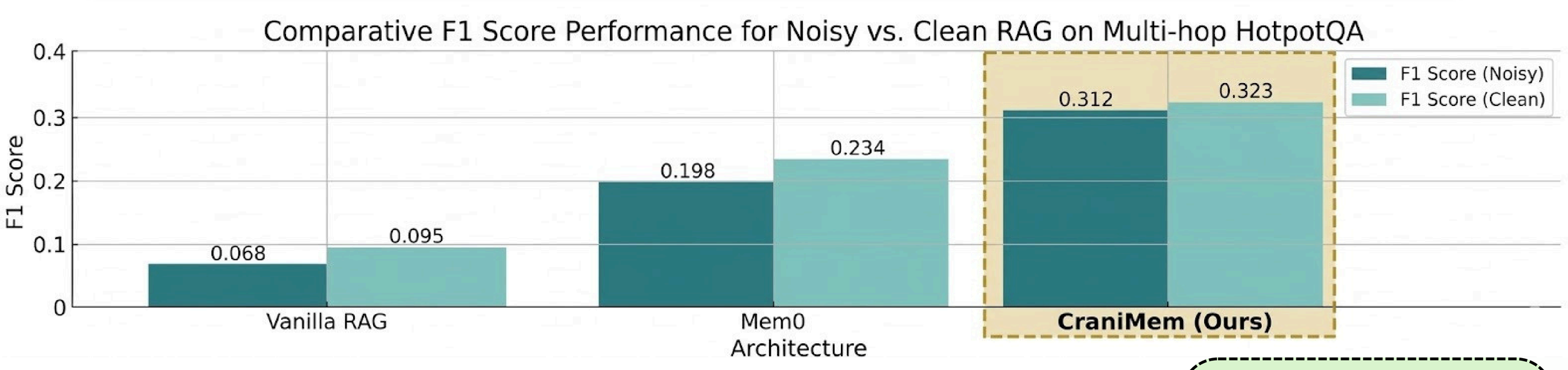
Contributions

Neuro-Inspired Gating
Utility-based filtering blocks irrelevant inputs before memory encoding.

Dual-Store Memory
Episodic buffer + knowledge graph with consolidation and pruning.

Robust to Noise
Near-zero performance drop under distractor injection.





Model	Architecture	Type	Precision	Recall	F1	Noise Drop Δ_{noise}	Latency (s)
Qwen2.5-Coder-7B-Instruct	Vanilla RAG	Noisy	0.100	0.117	0.096	0.079	6.810
		Clean	0.170	0.220	0.175		4.439
	CraniMem	Noisy	0.289	0.279	0.280	0.004	252.915
		Clean	0.294	0.282	0.284		9.029
Gemma-2-9B-IT	Vanilla RAG	Noisy	0.123	0.116	0.116	0.160	61.023
		Clean	0.303	0.268	0.276		44.448
	CraniMem	Noisy	**0.339**	**0.318**	**0.323**	0.015	134.126
		Clean	0.352	0.334	0.338		70.152
Qwen2.5-7B-Instruct	Vanilla RAG	Noisy	0.058	0.339	0.068	0.027	58.968
		Clean	0.092	0.366	0.095		68.146
	CraniMem	Noisy	0.315	0.317	0.312	0.011	**112.440**
		Clean	0.329	0.325	0.323		53.977
Mistral-7B-Instruct-v0.3	Vanilla RAG	Noisy	0.115	0.177	0.118	0.129	5.514
		Clean	0.243	0.353	0.247		7.701
	CraniMem	Noisy	0.300	0.300	0.294	0.022	381.230
		Clean	0.320	0.325	0.316		13.943

Key Insights

- **Lowest Noise Impact** CraniMem achieves Δ_noise = 0.004–0.022 across settings.
- **Strong Noisy Gains** F1 reaches 0.312 vs 0.068 under noisy inputs.
- **Consistent Across Models** Performance remains consistent across Qwen, Gemma, and Mistral.
- **Minimal Clean–Noisy Gap** Performance remains stable even under noise.
- **Higher Latency Cost** This comes with higher latency due to memory consolidation.

Conclusion & Limitations

CraniMem consistently outperforms Mem0 and Vanilla RAG in handling noise and maintaining long-horizon recall, though it incurs higher latency due to consolidation overhead. Future work will focus on scaling beyond 100 instances and improving efficiency.

Pypi **Github** **Paper**

Contact:- modypearl05@gmail.com