PANINI: Continual Learning in Token Space via Structured Memory

Shreyas Rajesh*, Pavan Holur*, Mehmet Yigit Turali, Chenda Duan, Vwani Roychowdhury

Department of Electrical and Computer Engineering, UCLA

Code: https://github.com/roychowdhuryresearch/gsw-memory

Continual Learning: The Question

Motivation

Models meet new information after training

New documents, changing facts, user-specific data, and new task skills arrive after pretraining.

Continual learning

Adapt without breaking old behavior

The goal is to learn new skills and information while avoiding catastrophic forgetting.

LLM setting

Parametric CL is promising, but delicate

Recent methods make progress, but post-trained weights entangle knowledge, skills, instruction following, and safety behavior.

New experience
documents, facts, user data

→

Learning mechanism
parametric or non-parametric

→

Future query
use new knowledge without forgetting

Why Study the Non-Parametric Route?

Complementary route

Keep the base model fixed

Parametric methods such as on-policy self-distillation (OPSD) still require training access and feedback. NPCL asks what memory alone can carry.

Practical

Cheap and black-box compatible

No model update is required. The same memory layer can sit in front of API or otherwise fixed models.

Safer update surface

Lower forgetting risk

The base model remains fixed; memory can be added, audited, edited, or removed separately.

Fixed LLM
no parameter update

+

External memory
accumulates experience

→

Non-parametric CL
learn by writing and reading memory

From Passages to a Reasoning Substrate

New experience: a document describes entities, events, and relationships that may be needed across future questions.

Standard RAG memory

Store text chunks; reason again at query time

Write: split the document into passages and embed them.

Read: retrieve semantically similar chunks for each query.

Problem: the model must rediscover entities, relations, and intermediate hops each time.

Cost: larger contexts, repeated computation, and no explicit reusable trace.

GSW / PANINI memory

Write reusable reasoning support

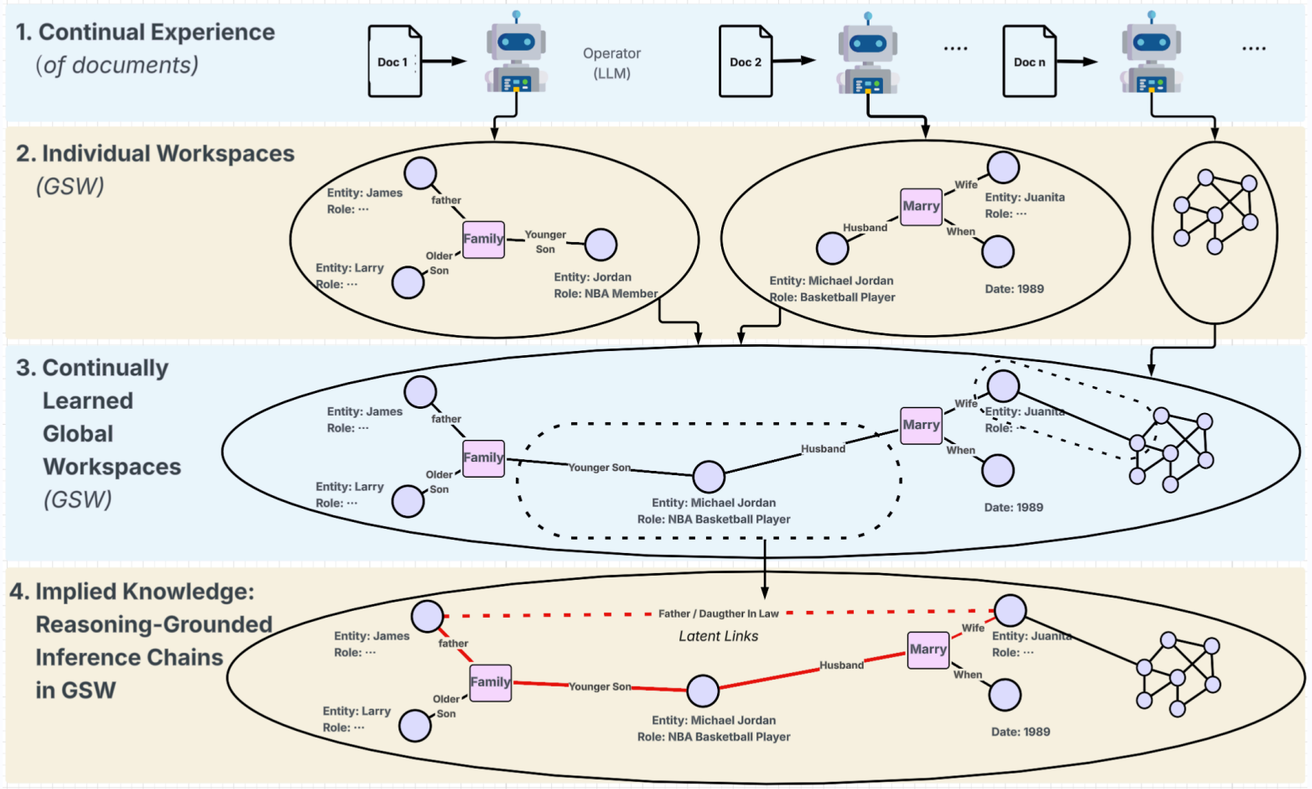
Write: convert the document into QA edges over entities, events, and claims.

Substrate: store intermediate facts as a graph-like token-space memory.

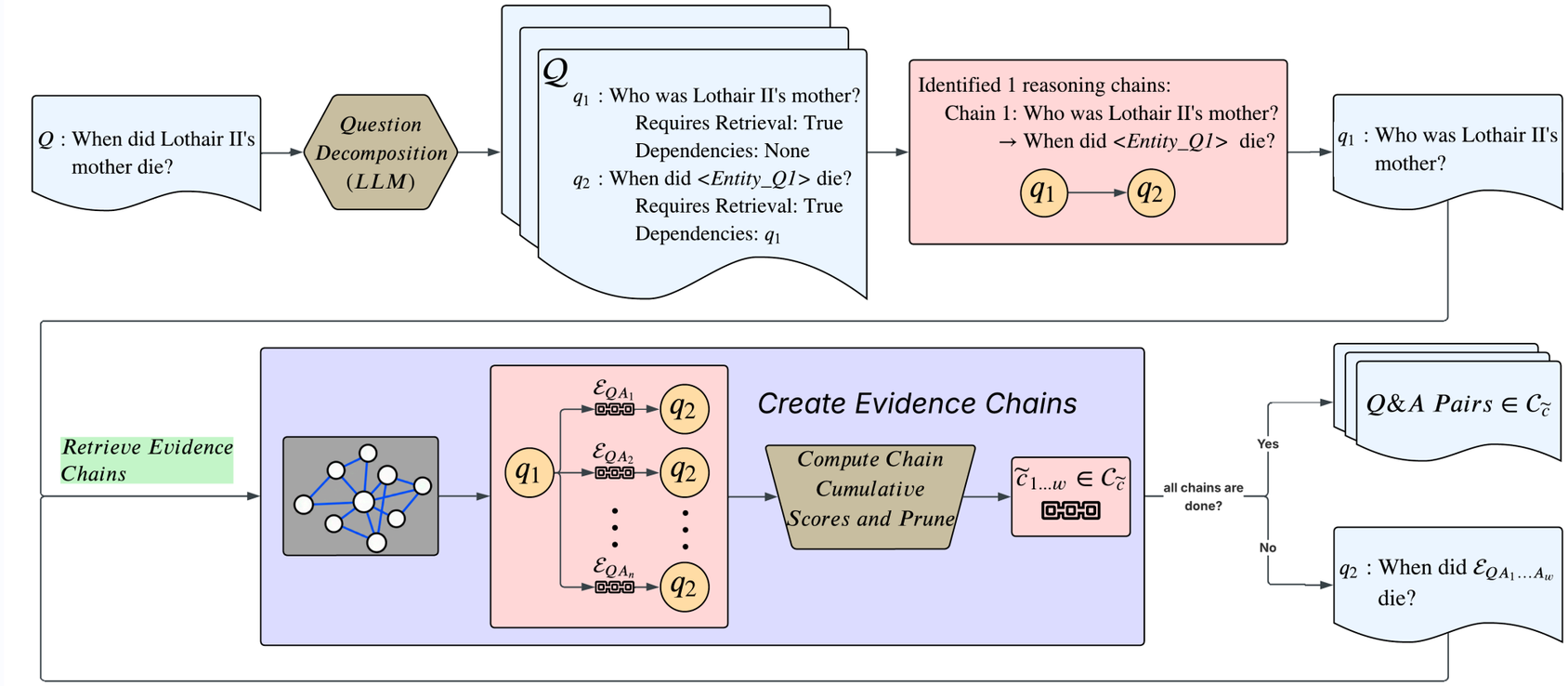
Read: retrieve a compact chain of QA edges that supports the question.

Benefit: future queries reuse prior reasoning instead of redoing it from raw text.

PANINI: Write Experiences Into Structured Memory



Intelligent Retrieval: Read Reasoning Chains



Results: Better QA With Less Context

Performance

Selected QA results

Retrieval	NQ	MuSiQue	2Wiki	Avg
Dense + reranker	61.4	43.7	57.9	50.5
GraphRAG	55.5	42.0	61.0	48.1
HippoRAG 2	60.0	49.3	69.7	53.3
PANINI	**67.44**	**52.27**	**72.37**	**56.06**

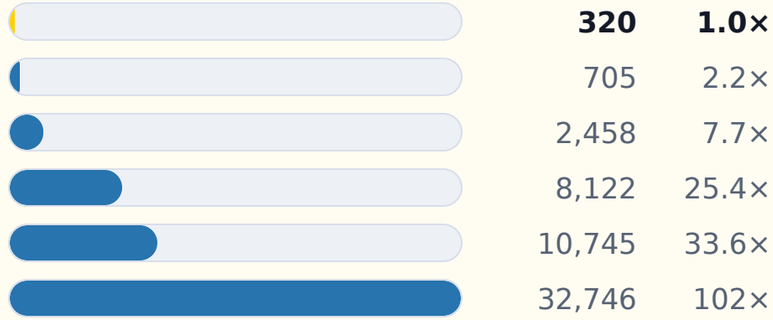
Full table covers NQ, PopQA, MuSiQue, 2Wiki, HotpotQA, and LV-Eval.

+2.8
Avg F1 over previous best

+6.0
NQ F1 over previous best

Efficiency

Average answer-context tokens



2.2x
fewer tokens than next-lowest baseline

-385
tokens vs standard retrieval

Agentic Retrieval Also Benefits From the Substrate

Appendix: same Search-R1 agent

Only the retrieval interface changes

Retrieval	MuSiQue	2Wiki	Avg
Search-R1: BM25 chunks	41.1	64.9	47.3
Dense chunks	45.1	64.6	49.0
Dense QA pairs	44.6	65.4	48.1
QA + PANINI retrieval	**46.6**	**67.6**	**49.4**

Appendix Table 17. All rows keep the Search-R1 agent fixed and vary retrieval only.

Previous best among Search-R1 variants: Dense chunks for Avg and MuSiQue; Dense QA pairs for 2Wiki.

What this says

Structured memory complements agentic retrieval

- Search-R1 is already an agentic retrieval baseline.
- Replacing chunk search with QA/RICR retrieval gives the best Search-R1 configuration.
- The result supports the broader claim: the memory substrate matters even as retrieval becomes more agentic.

+0.4
Avg F1 over previous best

+1.5
MuSiQue over previous best

+2.2
2Wiki over previous best

Reliability: Concentrated Context Helps

Retrieval	Avg answerable	Avg unanswerable
BM25	36.84	78.72
BM25 + reranker	45.78	75.60
Dense + reranker	62.50	66.57
HippoRAG 2	74.60	58.60
PANINI	**79.79**	**73.98**

Platinum split with GPT-4o-mini. Answerable measures supported QA; unanswerable measures abstention under missing evidence.

Takeaway

Better supported QA, strong abstention

PANINI improves answerable accuracy while maintaining high abstention accuracy when evidence is missing.

+5.2
Answerable accuracy over previous best

74.0
Unanswerable abstention accuracy

- GSWs amortize document understanding at write time.
- Compact chains help reduce unsupported answers.

Thank You