

DETERMINISTIC MEMORY FRAMEWORK

DMF

A Deterministic Memory Framework for Conversational AI Agents

Matteo Stabile · Enrico Zimuel

Roma Tre University · Italy



15-16 October | Berlin, Germany

Three questions drive the work

Can deterministic memory preserve quality while removing the LLM from memory management?

01

EFFECTIVENESS

Can a fully deterministic, non-generative memory framework achieve retrieval and answering performance comparable to LLM-based memory systems?

02

EFFICIENCY

Can conversational memory be managed without LLM calls while substantially reducing token consumption?

03

MEMORY FIDELITY

How does preserving source-level conversational information, rather than generating summarized memories, affect long-term recall and reasoning?

TAKEAWAY The evaluation must consider accuracy, token cost and source fidelity together.

Four design choices, one deterministic memory loop

Each choice removes a source of cost, drift or non-reproducibility.

01

INTERACTION-COUNT DECAY

SEQUENCE-STABLE

Age is measured in newer turns and not in elapsed time.

02

SCORE-DEPENDENT INERTIA

SALIENCE-AWARE

Salient records decay more slowly than marginal turns.

03

RECALL-TIME NLP

SOURCE-CANONICAL

Raw text remains canonical
interpretation happens on demand.

04

ZERO LLM CALLS IN MEMORY

TOKEN-FREE LOOP

Scoring, pruning and retrieval remain deterministic.

TAKEAWAY Memory state is a *reproducible* function of the message sequence and configuration.

From semantic divergence to *interaction-count decay*

Semantic novelty shapes the initial Survival Score; interaction count governs how that score evolves.

01 LOCAL CONTEXT

Normalize the last W embeddings into a moving centroid.

$$\vec{C}_n = \frac{\sum_{k=n-W}^{n-1} \vec{v}_k}{\left\| \sum_{k=n-W}^{n-1} \vec{v}_k \right\|}$$

Normalized centroid C

02

SEMANTIC NOVELTY

Cosine distance summarizes how far the turn is from recent context.

$D \in [0, 2]$
0 = aligned · higher = topic shift

Penalty, not hard gate: high information density can preserve a legitimate topic shift.

03

INITIAL RETENTION

D is penalized; all write-time signals then become Ω .

$\delta = -2.5$ (default)
Sets the strength of the divergence penalty.

$\Omega \in (0, 1)$
Initial retention priority after content, operational and provenance scoring.

04

TEMPORAL EVOLUTION

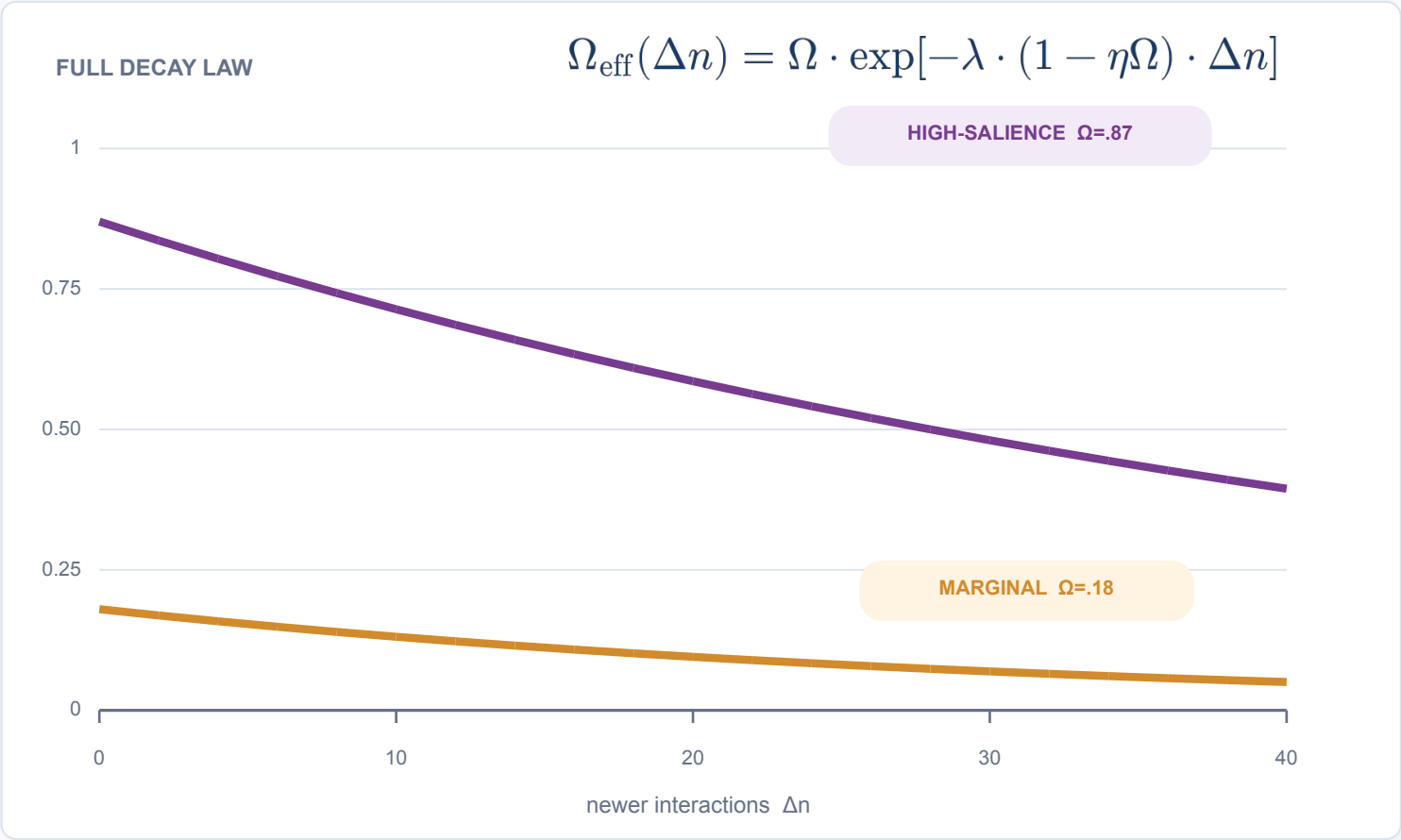
Once Ω is fixed, only newer interactions change its effective value.

$$\Omega_{\text{eff}}(\Delta n) = \Omega \cdot \exp[-\lambda \cdot (1 - \eta\Omega) \cdot \Delta n]$$

Δn = number of newer interactions

Score-dependent inertia

Illustrative trajectories (not benchmark results) using defaults $\lambda = 0.035$ and $\eta = 0.5$.



CONDENSED TABLE 4

Representative effective half-lives

Ω	λ_{eff}	$t_{1/2}$	tier
.87	.0198	35.0	healthy
.72	.0224	30.9	unstable
.50	.0263	26.4	unstable
.35	.0289	24.0	unstable
.25	.0306	22.6	critical
.18	.0319	21.8	critical

HALF-LIFE

$$t_{1/2}(\Omega) = \frac{\ln 2}{\lambda(1 - \eta\Omega)}$$

TAKEAWAY Inertia preserves ranking differences *without making any record permanent*.

Recall-time NLP

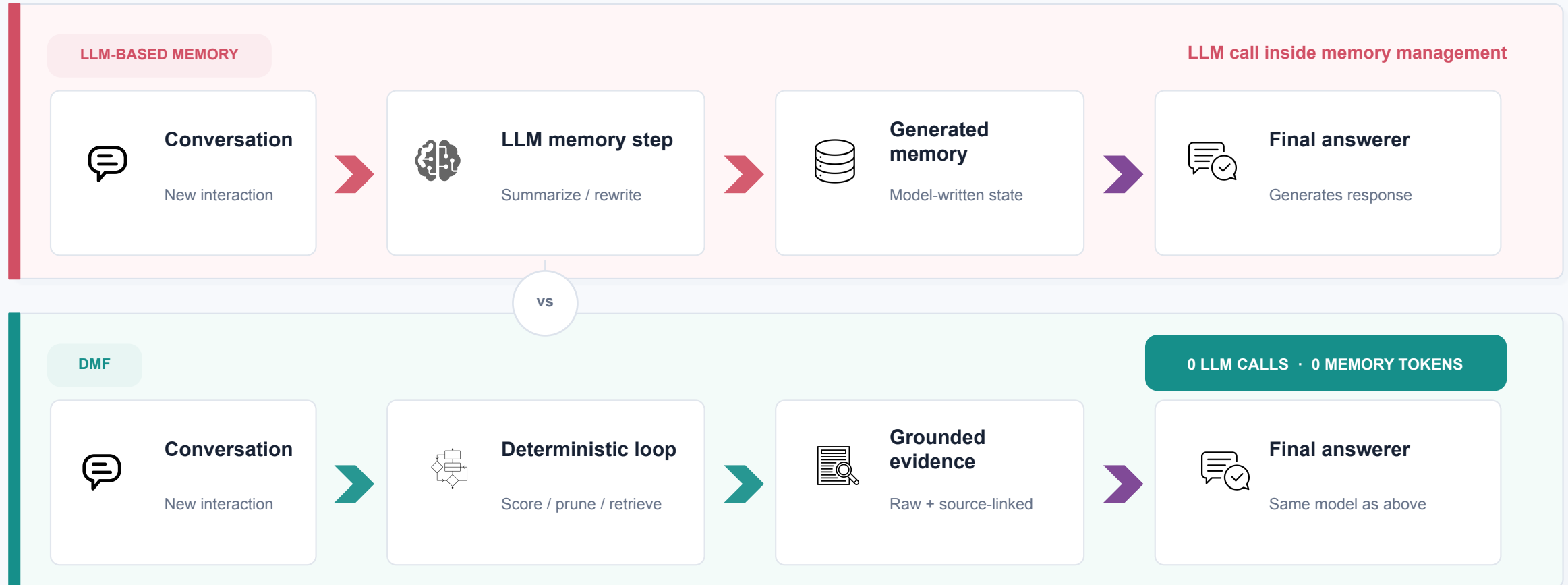
Store immutable evidence and derive query-specific meaning only when it is needed.



TAKEAWAY The retrieval pipeline can evolve *without rewriting and potentially distorting* stored memories.

Remove the LLM from the memory loop

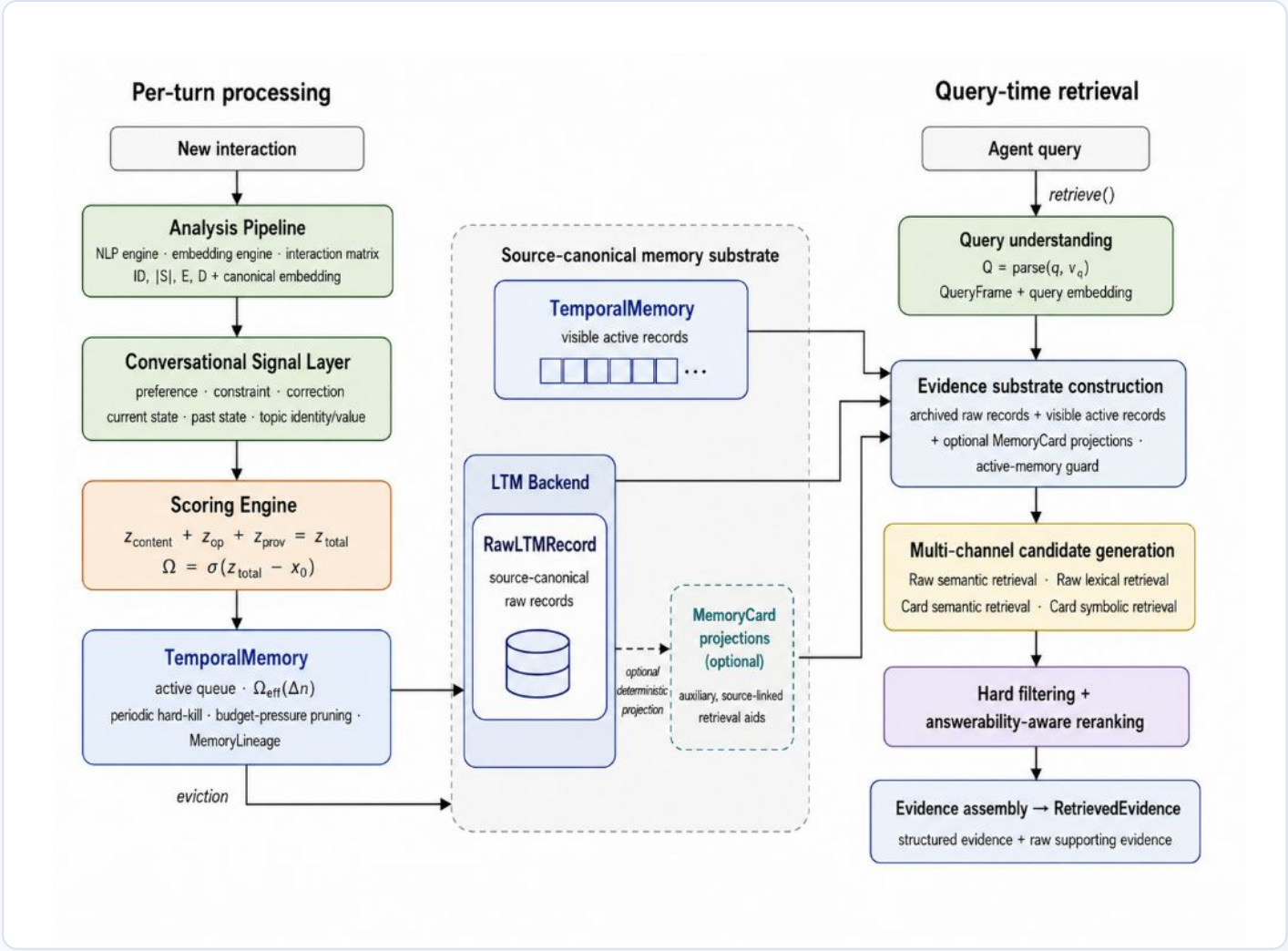
Same final answerer. Different memory architecture.



TAKEAWAY The answer model stays. The generative memory step disappears.

One pipeline, two operating modes

A lightweight update path runs per turn; richer interpretation runs only when queried.



01

PER-TURN PROCESSING

Signals → Survival Score → decay and pruning

Runs whenever a new interaction arrives.

02

SOURCE-CANONICAL CORE

Active queue + raw long-term records

MemoryCards remain optional and source-linked.

03

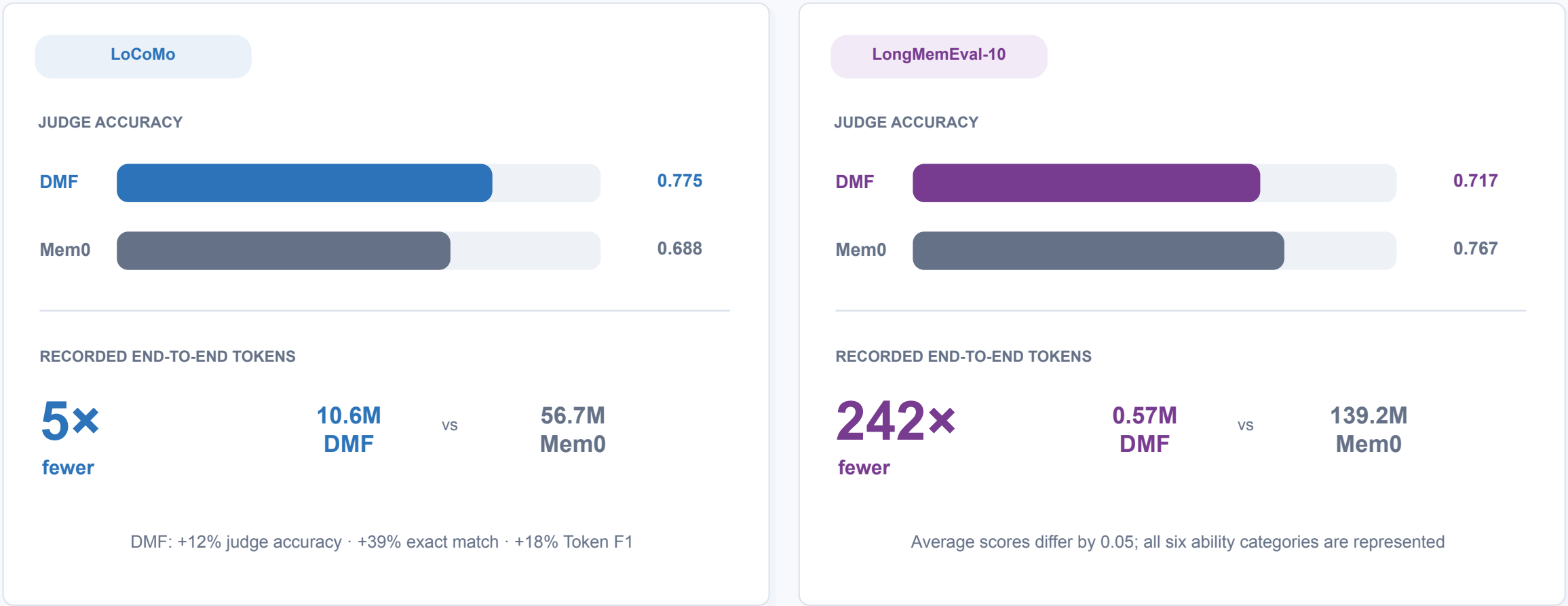
QUERY-TIME RETRIEVAL

Four retrieval channels → answerability reranking

Returns structured and raw supporting evidence.

Competitive accuracy, lower token cost

DMF vs Mem0 on LoCoMo (1,540 questions) and LongMemEval-10 (60 questions).



Future works

Extend coverage and adaptation without sacrificing provenance or auditability.

01

MULTILINGUAL SUPPORT

Language-specific POS, signals and keyword adapters.

Keep scoring and decay mathematics language-agnostic.

02

ADAPTIVE CALIBRATION

Tune selected weights from retrieval success and corrections.

Constrained, bounded and auditable updates.

03

EVIDENCE-GROUNDED CACHE

Reuse answers only when supporting records remain valid.

Key by query frame, evidence IDs and model version.

04

SHARED MEMORY

Reference or common namespaces across multiple agents.

Study access control, identity and conflict handling.

TAKEAWAY The invariant remains unchanged: raw source records stay canonical.

THANK YOU

Matteo Stabile

mat.stabile@stud.uniroma3.it

Enrico Zimuel

enrico.zimuel@uniroma3.it

References:

- <https://github.com/matstech/dmf>
- <https://github.com/matstech/dmf-benchmarks>



15-16 October | Berlin, Germany