DIALOGUE & DISCOURSE 16(2), 2025 · POSTER

LUFY

Long-term **U**nderstanding and identi**FY**ing key exchanges

Enhancing Long-term RAG Chatbots with Psychological Models of Memory Importance & Forgetting

Ryuichi Sumida · Koji Inoue · Tatsuya Kawahara

Graduate School of Informatics, Kyoto University

Code & Data on GitHub
Model, LUFY-Dataset & experiment scripts
github.com/ryuichi-sumida/LUFY

6 psychology-grounded memory signals

10% of memories kept — the rest forgotten

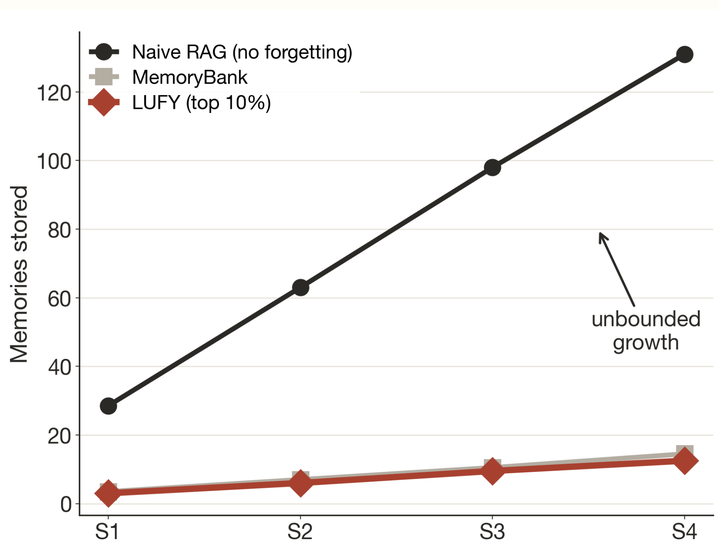
+17% retrieval precision vs. naive RAG

4.5× longer chats than any prior benchmark

01 The Challenge

RAG chatbots retrieve past memories to stay personal — but stored memory **grows without bound**, raising cost and burying useful facts in clutter that **degrades retrieval**.

Humans don't keep everything: we recall only about **10% of a conversation** (Stafford & Daly, 1984).



Naive RAG accumulates memory endlessly; LUFY keeps a steady ~10%.

> *Which 10% should the chatbot keep — and what should it forget?*

02 Six Psychological Signals

***A* Arousal**
emotion → memory

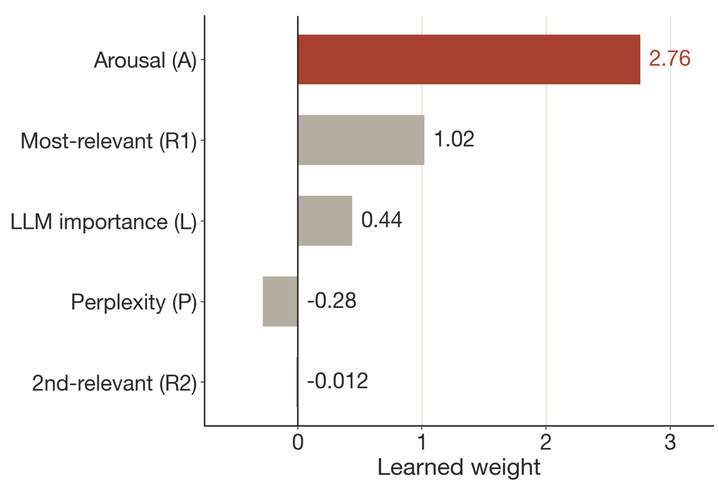
***P* Surprise**
perplexity

***L* LLM importance**
future usefulness

***R1* Top-retrieved**
reinforce (+)

***R2* 2nd-retrieved**
forget (−)

***Δt* Recency**
time decay



Weights fit to human importance labels — **emotional arousal dominates**.

03 How LUFY Differs

SYSTEM	FORGETTING	RETRIEVAL
Naive RAG	*none*	cosine sim.
MemoryBank	recency & frequency	cosine sim.
LUFY (ours)	**6 psych. signals**	**sim. + importance**

Prior systems rank memories only by how often or how recently they were recalled. LUFY adds emotion, surprise & narrative salience — and folds the resulting importance **directly into retrieval**, not just forgetting. The same score decides what surfaces and what fades.

04 Why Keep Raw Exchanges?

Beyond a user profile, LUFY also stores session summaries and raw exchanges to preserve **episodic memories** — unique personal events that never surface in a profile.

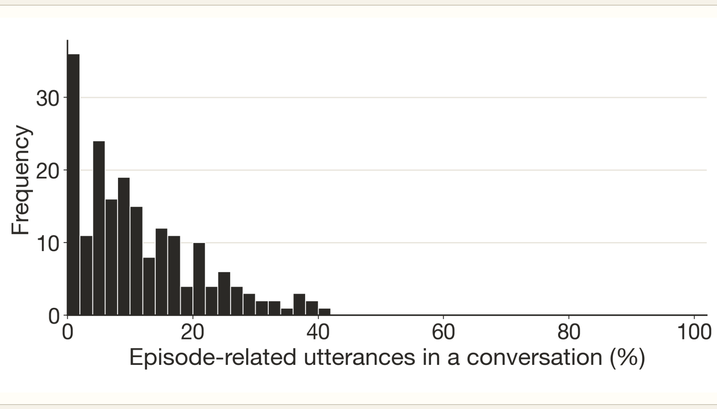


Fig. 8. *Episode-related utterances per conversation (N=170):* **~79%** *contain at least one.*

05 What LUFY Remembers

PROFILE	Stable user facts — "likes sushi", "has a dog named Luke". **Always kept.**
SUMMARIES	The gist of each 30-minute session.
EXCHANGES	Raw user–bot turns — episodic detail. **Pruned to the top 10%.**

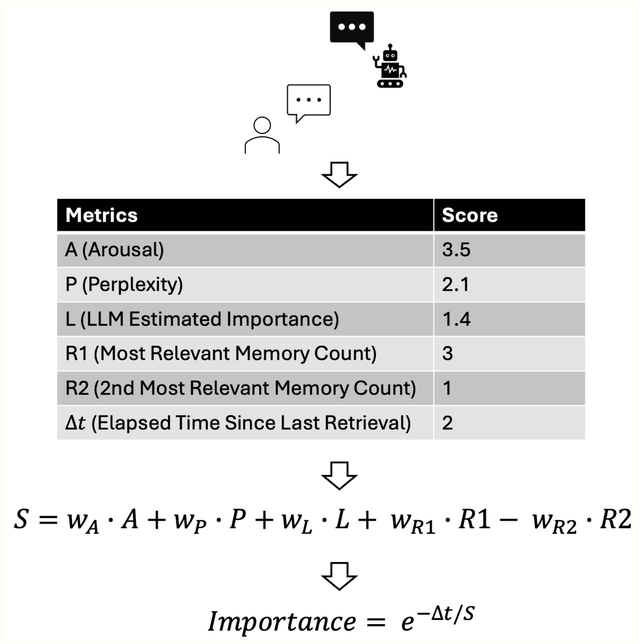
Forgetting applies to summaries & exchanges; the profile is refined and retained over time.

06 Scoring a Memory

Every user–bot exchange becomes a **memory**. Its six signals are combined with learned weights into a strength **S**, then decayed over time into an importance score:

$$S = w_A A + w_P P + w_L L + w_{R1} R1 - w_{R2} R2$$

$$\textbf{Importance} = e^{-\Delta t / S}$$



Metrics	Score
A (Arousal)	3.5
P (Perplexity)	2.1
L (LLM Estimated Importance)	1.4
R1 (Most Relevant Memory Count)	3
R2 (2nd Most Relevant Memory Count)	1
Δt (Elapsed Time Since Last Retrieval)	2

$$S = w_A \cdot A + w_P \cdot P + w_L \cdot L + w_{R1} \cdot R1 - w_{R2} \cdot R2$$

$$Importance = e^{-\Delta t / S}$$

Fig. 2. *Each exchange is scored, weighted, and turned into an importance value via an Ebbinghaus-style curve.*

PROMPT · LLM-ESTIMATED IMPORTANCE (L)

Rate this exchange 1–10 — where **1** is mundane (brushing your teeth) and **10** is pivotal (a breakup or a college acceptance) — by whether it will matter in later talks.
summary: {key_summary} · exchange: {content}

The **L** *signal — inspired by Generative Agents (Park et al., 2023) — captures salience that arousal & surprise miss.*

07 Retrieve & Forget

1 · Retrieval ranks by similarity + importance

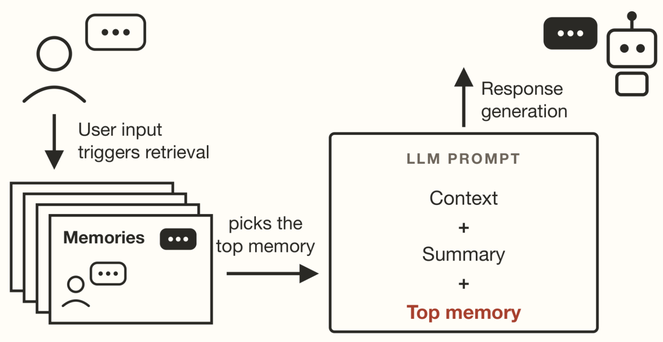


Fig. 3. *The top memory joins recent context & a summary in the prompt.*

2 · Forgetting keeps only the top 10%

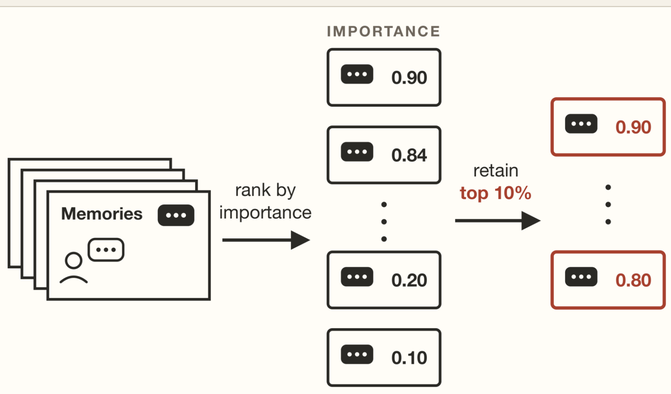


Fig. 5. *After each conversation, memories are ranked and pruned.*

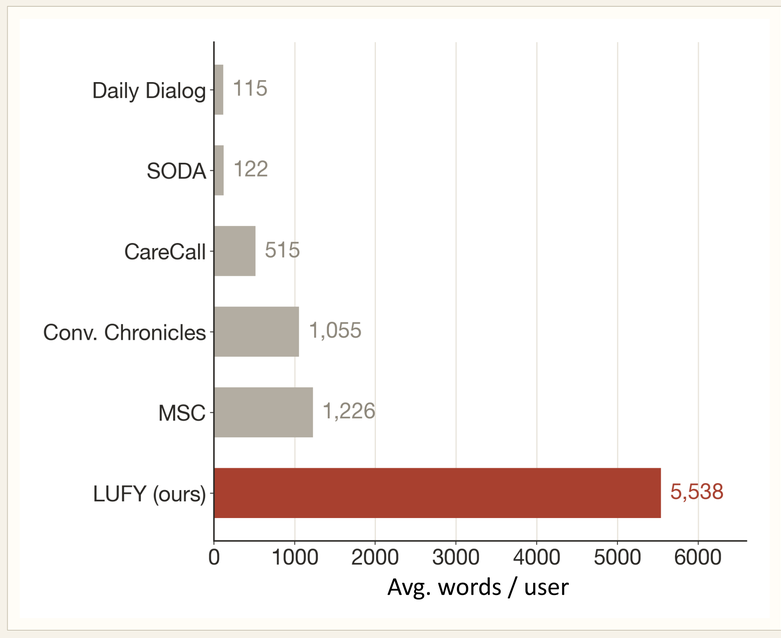
PROMPT · RESPONSE GENERATION

Using the key summary, recent utterances and the single most relevant memory, write a reply — casual, like a genuine friend.
summary: {key_summary} · recent: {recent} · memory: {related}

Only the **single most important & relevant** *memory is injected — not the whole history — keeping prompts short and on-topic.*

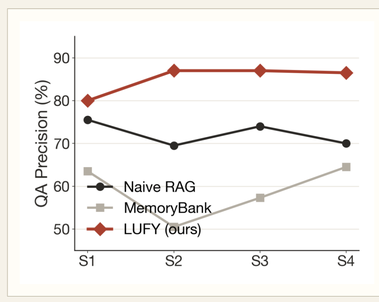
08 Largest Long-term Study

- **17 participants** × ten 30-min sessions over 4 days (≈5 h each)
- Three chatbots: **Naive RAG**, **MemoryBank**, **LUFY**
- Judged by annotators, GPT-4o & sentiment analysis

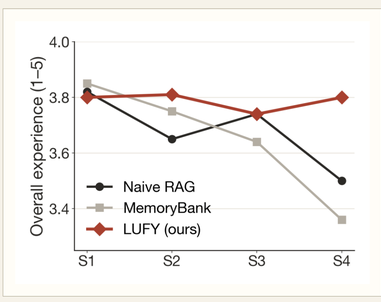


The public **LUFY-Dataset***: conversations 4.5× longer than the prior benchmark (MSC).*

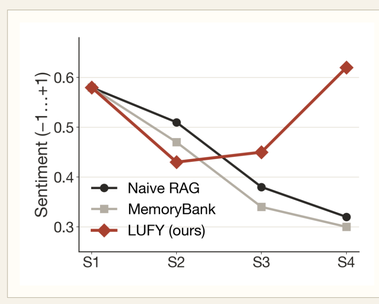
09 Results: LUFY Wins



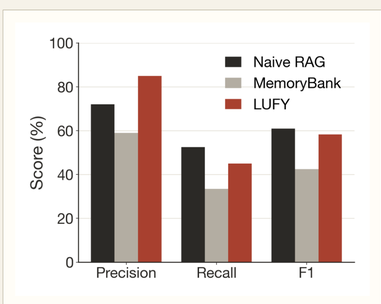
Retrieval precision



Overall experience



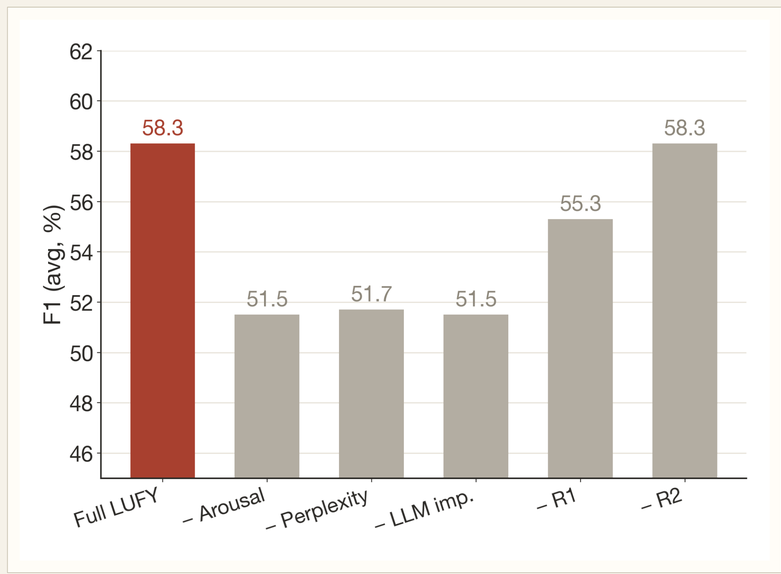
User sentiment



Precision / Recall / F1

- Highest **personalization, flow & overall** — gap widens in the longest sessions (S4)
- **85% retrieval precision** (vs. 72% / 59%); S4 sentiment **0.62** vs. 0.32 / 0.30
- Matches human importance judgments **17.6%** vs. 14.4% (MemoryBank) — beating human–human agreement in 2 of 4 sessions
- Trade-off: naive RAG keeps everything, so it edges out recall — but at far lower precision

10 Ablation & Takeaways



F1 drops most when **Arousal, LLM-importance** *or* **R1** *reinforcement are removed; R2 (tiny weight) barely matters.*

> *Forgetting ~90% of a conversation can improve a chatbot — if you keep the emotional 10%.*

- Importance-aware retrieval + forgetting beats recency/frequency heuristics
- Next: multimodal emotion & structured / relational memory

Kyoto University · Graduate School of Informatics | sumida@sap.ist.i.kyoto-u.ac.jp
Sumida, Inoue & Kawahara (2025). **Dialogue & Discourse 16(2), 74–110.**

Read the paper
doi.org/10.5210/dad.2025.203