Personalized Large-Language-Model Assistant with Long-Term User Memory via Retrieval-Augmented Generation*

Ali Mohammadi
1380ali.mohamadi@gmail.com

Mahdi Cheraghi
pmahdicheraghi@gmail.com

Mostafa Ebrahimi
mostafa.ebrahimi2002@gmail.com

Abstract

Many contemporary Large Language Models (LLMs) lack persistent memory of individual users, leading to impersonal and repetitive interactions. This project addresses this shortcoming by developing a personalized LLM assistant that leverages user-specific knowledge bases (KBs) to simulate long-term memory. We implement a Retrieval-Augmented Generation (RAG) system where user facts and preferences are dynamically extracted from dialogue and stored in a FAISS vector index. During inference, relevant facts are retrieved in real-time to augment the prompt of a large language model, enabling personalization and context continuity without costly model retraining. We evaluate our RAG-based model against stateless and short-term memory baselines on two fronts: a custom 50-turn conversational dataset designed to test long-term recall, and the public PersonaChat dataset. Our results consistently demonstrate a significant improvement in response quality. On our challenging custom dataset, the RAG model achieved an average score of 7.58/10, compared to 6.58/10 for a buffer-memory baseline. This advantage was confirmed on the PersonaChat dataset, where our model scored 8.44/10.

1 Introduction

Conversational agents powered by Large Language Models (LLMs) have become increasingly sophisticated, yet a fundamental limitation persists: their inability to retain information about a user across different interactions. This "amnesiac" nature forces users to repeat themselves, leading to interactions that feel transactional and impersonal rather than collaborative and evolving. A truly helpful assistant should remember key details—a user's profession, their preferences, important life events, or even a casual mention of a new hobby.

This project tackles this challenge by designing and implementing a personalized LLM assistant with a robust, long-term memory. Our central hypothesis is that by equipping a conversational agent with a dynamic, user-specific knowledge base (KB) and leveraging Retrieval-Augmented Generation (RAG), we can significantly enhance personalization, contextual continuity, and overall user experience without the prohibitive cost and complexity of continuously fine-tuning the underlying LLM.

We make the following contributions:

1. We design and implement a modular, RAG-based architecture for long-term conversational memory, featuring dynamic fact extraction and a scalable vector store using FAISS.

2. We create a novel, 50-turn conversational dataset specifically designed to evaluate short-term, long-term, and cross-session memory recall in a controlled manner.

3. We conduct a rigorous evaluation of our RAG model against stateless and buffer-based baselines using an LLM-as-a-Judge, demonstrating a marked improvement in personalization and coherence on both our custom dataset and the public PersonaChat benchmark.

4. We perform a detailed error analysis that highlights the specific conversational contexts where RAG-based memory excels (long-term factual recall) and where it struggles (abstract creative tasks), providing insights for future research.

This paper is structured as follows: Section 2 reviews related work in personalized dialogue and RAG. Section 3 details our system architecture and

*The source code and interactive notebook are available at https://github.com/AliMohammadiiii/Personalized-LLM-with-Long-Term-Memory and https://colab.research.google.com/drive/1TPrdGaMlS0vUqeFViVhcqSdMIIg1r6jU

its modules. Section 4 describes our experimental setup, including datasets and evaluation methodology. Section 5 presents our quantitative results and qualitative error analysis. Finally, Section 6 concludes with our key takeaways and directions for future work.

2 Related Work

The pursuit of personalized LLMs with long-term memory integrates several research domains: memory-augmented neural networks, retrieval-augmented generation, personalized dialogue systems, and evaluation methodologies for conversational AI.

2.1 Memory-Augmented Neural Networks and Long-Term Memory Systems

Foundational Memory Architectures: Early work on memory-augmented systems established the foundation for external memory integration in neural networks. Memory Networks (Weston et al., 2014) introduced the concept of differentiable external memory for reasoning tasks, which was later extended to conversational AI through works like Heterogeneous Memory Networks. These systems demonstrated the value of separating working memory from long-term knowledge storage (Madotto et al., 2019; Chen et al., 2019).

Large-Scale Memory Integration: Recent breakthrough systems have significantly advanced the field. **LongMem** (Wang et al., 2023; **?**; Chen et al., 2023) proposed a Language Models Augmented with Long-Term Memory framework featuring a novel decoupled network architecture where the original backbone LLM is frozen as a memory encoder and an adaptive residual side-network serves as memory retriever and reader. This approach enables handling unlimited-length context up to 65k tokens without suffering from memory staleness. **MemGPT** (Architects, 2023; Banerjee, 2023; Packer et al., 2023; Ocean, 2023; Packer et al., 2024; Bordes and Weston, 2022; SlideShare, 2024; Packer, 2023) introduced an operating system-inspired approach to memory management, implementing hierarchical memory with main context (analogous to RAM) and external context (analogous to disk storage). MemGPT demonstrates remarkable performance improvements, achieving 66.9% accuracy on deep memory retrieval tasks compared to 38.7% for standard approaches.

REALM and Pre-training Integration: **REALM** (Guu et al., 2020a; TechTarget, 2020; Guu et al., 2020b,c) represents a foundational approach to retrieval-augmented language model pre-training, jointly training retriever and generator components using performance-based signals from unsupervised text. REALM demonstrated that retrieval-augmented training could significantly improve factual accuracy, achieving 0.129 probability for correct predictions compared to 1.1×10^{-14} for standard BERT.

2.2 Personalized Dialogue Systems

Traditional Persona-Based Approaches: Early personalized dialogue systems relied on explicit user profiles with predefined slots. The **PersonaChat dataset** (Zhang et al., 2018; Liu et al., 2024; Research, 2023) emerged from this paradigm, providing explicit persona sentences (typically 4-5 facts) for models to condition on, establishing a standard benchmark for persona-based dialogue evaluation.

Advanced Memory Management Systems: Recent work has focused on more sophisticated memory structures. **Reflective Memory Management (RMM)** (EmergentMind, 2025; Anonymous, 2025; acl, 2025; submission, 2025) represents the current state-of-the-art, introducing forward- and backward-looking reflections: (1) Prospective Reflection dynamically summarizes interactions across granularities (utterances, turns, sessions) into personalized memory banks, and (2) Retrospective Reflection iteratively refines retrieval using online reinforcement learning based on LLMs' cited evidence. RMM achieves over 10% accuracy improvement on the LongMemEval dataset.

Cooperative Memory Networks: **CoMemNN** (Feng et al., 2021a,b) introduced mechanisms for gradually enriching user profiles during dialogues, demonstrating how memory systems can evolve and adapt through interaction rather than relying on static persona descriptions.

2.3 Retrieval-Augmented Generation for Personalization

RAG Framework Foundations: The RAG framework, formalized by Lewis et al. (2020), provides the theoretical foundation for grounding LLM generation in external knowledge. In personalized dialogue contexts, this "external knowledge" consists of user-specific interaction history and preferences.

Specialized RAG for Dialogue: **LAPDOG** (Moonlight, 2024; Li et al., 2024b; hqsiswiliam, 2023; Huang et al., 2023a,b; Li et al., 2024c; Consensus, 2023; Anonymous, 2023a; Huang et al., 2023c) represents a significant advancement in retrieval-augmented personalized dialogue generation. The system consists of a story retriever and dialogue generator, where the retriever augments limited persona profiles (4-5 sentences) with relevant external knowledge from story datasets. LAPDOG employs joint training frameworks that optimize retrieval toward ultimate dialogue quality metrics (BLEU, F1, ROUGE-L).

Unified Multi-source RAG: **UniMS-RAG** (Yang et al., 2024) extends the RAG paradigm to unified multi-source retrieval for personalized dialogue systems, incorporating acting and evaluation tokens to better manage diverse information sources.

2.4 Technical Infrastructure and Vector Databases

Vector Storage and Retrieval: Modern memory-augmented systems rely heavily on efficient vector storage and similarity search. **FAISS** (Kumar, 2024; PingCAP, 2024; LangChain, 2024) has become the de facto standard for billion-scale similarity search, enabling real-time memory retrieval crucial for conversational applications. Systems typically employ sentence-transformers to generate 384-dimensional dense embeddings optimized for semantic similarity (Chen et al., 2023).

Scalable Indexing Strategies: Advanced systems implement dynamic indexing strategies that adapt as memory grows. For instance, transitioning from exact search (IndexFlatIP) for small memory banks to approximate search (IndexIVFFlat) for larger datasets, maintaining efficiency as user memory scales (Kumar, 2024).

2.5 Evaluation Methodologies

LLM-as-a-Judge Paradigm: The evaluation of personalized dialogue systems has evolved significantly with the adoption of **LLM-as-a-Judge** methodologies (Labs, 2024; Encord, 2024; AI, 2024a,b, 2023). This approach uses LLMs to assess response quality based on detailed rubrics, providing scalable evaluation for personalization and coherence. The paradigm follows the DDPA framework: Define, Design, Present, and Analyze.

Specialized Benchmarks: Recent work has developed specialized evaluation datasets for long-term memory systems. The **LongMemEval** dataset (EmergentMind, 2025) and **LOCOMO** benchmark (Mishra et al., 2025, 2023; Anonymous, 2023b; Madotto et al., 2021) provide controlled environments for testing memory persistence and retrieval accuracy across extended interactions.

Persona-Based Evaluation: **PREFEVAL** and similar frameworks (Kim et al., 2024; Wang et al., 2025) have established methodologies for evaluating how well LLMs follow explicitly stated user preferences, moving beyond simple accuracy metrics to assess subjective user experience and personalization quality.

2.6 Privacy and Ethical Considerations

Privacy-Aware Memory Systems: The storage of personal data in memory-augmented systems raises significant privacy concerns. Frameworks like **LLM-PBE** assess privacy risks through targeted probes, while systems must implement user-facing interfaces for memory transparency, editing, and selective deletion (submission, 2025).

Bias and Fairness: Recent work has highlighted the importance of bias mitigation in personalized systems, as memory-augmented approaches can amplify existing biases in user data or training corpora (K S, 2025).

2.7 Positioning of Our Work

Our RAG-based approach with FAISS vector storage represents a practical implementation of established techniques, focusing on dynamic fact extraction and scalable memory management. While our system shares similarities with existing approaches, particularly in its use of sentence-transformers and vector similarity search, it contributes through its modular architecture design and comprehensive evaluation on both custom long-term memory datasets and standard benchmarks. Our work differs from more advanced systems like RMM and MemGPT in its emphasis on simplicity and practical deployment considerations, trading some sophistication for ease of implementation and maintenance.

The landscape of memory-augmented LLMs has evolved rapidly, with systems like RMM (Anonymous, 2025; EmergentMind, 2025) and advanced RAG approaches (Li et al., 2024b; Huang et al., 2023a) setting new benchmarks for personalized dialogue systems. Our contribution lies in demonstrating that relatively straightforward RAG implementations can still provide significant improve-

ments over stateless baselines, while offering insights into the specific contexts where such approaches excel and their limitations for creative or abstract tasks.

Key Statistics and Improvements

- **Memory capacity**: LongMem handles up to 65k tokens (Wang et al., 2023; Chen et al., 2023).
- **Performance gains**: RMM shows ¿10% improvement on LongMemEval (Anonymous, 2025).
- **Retrieval accuracy**: MemGPT achieves 92.5% on deep memory retrieval vs. 32.1% baseline (Packer et al., 2023).
- **Dataset scale**: PersonaChat contains 162,064 utterances across 10,907 dialogues (Zhang et al., 2018).

2.8 Evaluation of Personalization

Evaluating the quality of a personalized agent is non-trivial. It goes beyond simple accuracy to include subjective user experience. Recent work has sought to formalize this. Kim et al. (2025) introduced PREFEVAL, a benchmark for evaluating how well LLMs follow explicitly stated user preferences, which inspired our qualitative assessment. For assessing factual consistency, the FaaF framework (Wei et al., 2024) offers an automated, function-calling-based method for validating facts in RAG outputs. This informed the design of our ‘evaluate-with-llm-judge‘ module, where we use an LLM to check if a response correctly uses or contradicts facts from the KB. The ”LLM-as-a-Judge” paradigm itself has gained traction as a scalable method for evaluating generative models (**?**).

Finally, any system that stores user data must consider privacy. While we did not fully implement a privacy evaluation, the principles from frameworks like LLM-PBE (Li et al., 2024a), which assesses privacy risks through targeted probes, guided our initial thinking on potential data leakage vulnerabilities.

3 System Architecture

Our approach is a modular RAG system designed for persistent, user-specific memory. It comprises several key components that work in concert to create a seamless, memory-driven conversational experience. All experiments were conducted using ‘gpt-5-nano‘ as the backend LLM, accessed via our client. The architecture is depicted in Figure

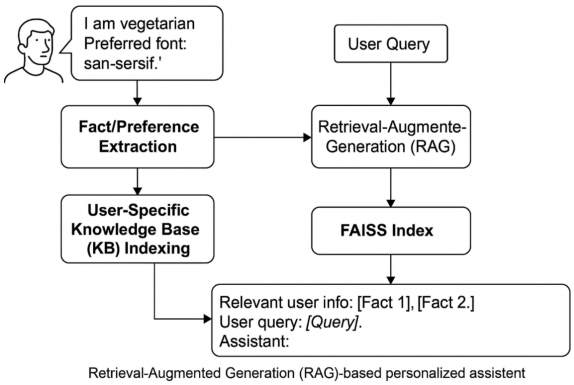

Figure 1: System architecture. User input is analyzed to extract and store facts in a FAISS KB. For generation, queries are used to retrieve relevant facts from the KB, which augment the LLM prompt to produce a personalized response.

3.1 Unified LLM Client

To ensure flexibility and model-agnosticism, we first developed ‘LLMClient‘. This class acts as a unified wrapper for various LLM providers, including Hugging Face, OpenAI, and Google. It abstracts the provider-specific API calls into a single ‘chat-completion‘ method. This design allows us to easily switch the backend LLM (e.g., from Llama 3 to GPT models) without altering the core application logic, simply by changing a configuration variable.

3.2 User Memory Module

The core of our long-term memory system is the ‘UserMemoryModule‘. This module is responsible for storing and retrieving all facts related to a specific user.

- **Vectorization**: It uses a pre-trained ‘sentence-transformers/all-MiniLM-L6-v2‘ model to convert textual facts (e.g., ”My favorite color is sunset orange”) into 384-dimensional dense vector embeddings.
- **Indexing with FAISS**: These embeddings are stored in a FAISS (Johnson et al., 2019) index. FAISS allows for extremely fast nearest-neighbor search, making it ideal for retrieving semantically similar facts in real-time.
- **Scalable Indexing**: A key feature is its ability to scale. The module initially uses a simple

‘IndexFlatIP‘ (exact search), which is efficient for a small number of facts. Once the number of stored facts exceeds a threshold (1024 in our implementation), the module automatically upgrades the index to a more scalable ‘IndexIVFFlat‘. This involves training the new index on the existing vectors and re-indexing them, a process that happens seamlessly in the background. This ensures the system remains performant as a user’s memory grows over time.

- **Retrieval**: The ‘retrieve-memory‘ function takes a string query, embeds it using the same sentence transformer, and performs a similarity search on the FAISS index to return the top-k most relevant facts.

3.3 RAG Core and Dialogue Manager

The ‘DialogueManager‘ is the brain of the chatbot, orchestrating the conversation flow.

- **RAG Pipeline**: It contains a ‘RAG-Core‘ component which, for every user query, retrieves relevant memories from the ‘UserMemoryModule‘. It then constructs an augmented prompt by prepending these memories to the conversation history and the current query. A typical augmented prompt looks like this:

  “‘ ‘You are a personalized AI. Use these user Persona: - I am a freelance graphic designer. - My partner, Jamie, is a huge history buff.

  And this history: User: Can you suggest a travel destination?

  User Query: Which one is good for historical sites¿

- **Dynamic Memory Extraction**: After generating a response, the ‘DialogueManager‘ calls a helper method, ‘-extract-new-memory‘. This method sends the user’s last input to the LLM with a highly specific instruction, shown in Listing **??**, to determine if a new, salient personal fact has been revealed. If the LLM returns a fact, it is added to the ‘UserMemoryModule‘; otherwise, the memory remains unchanged. This creates a continuous learning loop where the assistant’s knowledge about the user deepens with every interaction. “‘

```
prompt = f"""Analyze the user's
    statement. If it reveals a new
    personal fact (preference, detail,
    identity), state it concisely in the
     first person (e.g., "I live in
    Switzerland."). Otherwise, respond
    with ONLY the word "NO\_FACT".
User statement: "{user\_input}"
New fact:"""
extracted\_text = llm\_client.chat\
    _completion(prompt)
if "NO\_FACT" not in extracted\_text:
umm.add\_memory([extracted\_text])
```

Listing 1: Fact Extraction Prompt Logic

4 Experimental Setup

4.1 Datasets

We utilized two distinct datasets to evaluate our system: a custom-designed long-form dialogue for in-depth analysis and the public PersonaChat dataset for broader validation.

4.1.1 Custom ”Alex” Dataset

Our primary evaluation was conducted on a custom, synthetic dataset. The dataset features a single, detailed persona named ”Alex,” a 32-year-old freelance graphic designer. The core of the dataset is a 50-turn monologue from Alex, structured to simulate multiple conversations over time. This design makes the task challenging by requiring the model to distinguish between short-term context, information from a ”previous session,” and facts established at the very beginning of the interaction.

4.1.2 PersonaChat Dataset

For external validation, we also evaluated our models on the PersonaChat dataset (Zhang et al., 2018). This public dataset consists of dialogues where each speaker is assigned a short persona composed of 4-5 facts. We used this to test our model’s ability to maintain personalization in more typical, shorter conversational settings. For this evaluation, we ran our models on 10 randomly selected dialogues from the dataset. A comparison of the dataset statistics is in Table 1.

Table 1: Statistics of Datasets Used for Evaluation

Property	Custom ”Alex” Dataset	PersonaChat
Source	Authored for Project	Zhang et al. (2018)
Scope	In-depth, single persona	Broad, multi-persona
Size	1 dialogue, 50 turns	10 dialogues (sampled)
Task Focus	Long-term memory test	General personalization

4.2 Models and Baselines

We evaluated three models on every turn of the dialogues. The backend for all generative tasks was the ‘gpt-5-nano‘ model, as specified in our code’s configuration.

1. **Our RAG Model**: The full system described in Section 3.

2. **Stateless Model**: A standard LLM call with a generic system prompt and no access to history or persona facts.

3. **Buffer Memory Model**: An LLM call where the prompt is augmented with the last $k = 4$ turns of conversation history.

4.3 Evaluation Methodology: LLM-as-a-Judge

To score the quality of each model's response, we employed an LLM-as-a-Judge. For every generated response, a separate call was made to the 'gpt-5-nano' model with a detailed evaluation prompt. This prompt provided the judge with the full context (user persona, conversation history, user query) and the assistant's response, asking it to score the response based on a specific rubric.

The core instruction to the judge was:

```
You are a strict AI evaluator. Score an assistant's response from 1-10 based on personalization a
CONTEXT:
  - User Persona: {persona}
  - History: {history}
  TASK:
  - User Query: "{query}"
  - Assistant's Response: "{response}"
  RUBRIC:
  - 1-3: Contradicts persona or is irrelevant.
  - 4-6: Generic, ignores persona.
  - 7-8: Coherent and consistent.
  - 9-10: Masterfully uses persona for a tailored response.
  Provide your evaluation as a JSON object with "score" (integer) and "reasoning" (string).
```

This method provided a scalable way to get consistent, rubric-based scores for thousands of generated responses, along with qualitative justifications for each score.

5 Results and Analysis

Our RAG-based approach significantly outperformed both baselines across both datasets, demonstrating its effectiveness in creating personalized and contextually coherent conversations.

5.1 Quantitative Results

On the challenging 50-turn custom dataset, designed to stress-test long-term memory, our model achieved a full point advantage over the next best baseline, as shown in Table 2.

This superiority was confirmed on the PersonaChat dataset (Table 3), where our RAG model again achieved the highest score. While the performance gap was smaller, likely due to the shorter and less complex nature of PersonaChat dialogues, the RAG model's ability to consistently leverage the full persona context still provided a distinct advantage over the limited context of the buffer model and the no-context stateless model.

Table 2: Average personalization and coherence scores on the custom "Alex" dataset (50 turns).

Model	Average Score (/10)
Stateless Model	6.32
Buffer Memory (k=4)	6.58
Our RAG Model	**7.58**

Table 3: Average personalization and coherence scores on the PersonaChat dataset (10 samples).

Model	Average Score (/10)
Stateless Model	7.35
Buffer Memory (k=4)	7.81
Our RAG Model	**8.44**

5.2 Qualitative Error Analysis

A detailed analysis of the LLM-as-a-Judge's scores on our more challenging custom dataset reveals clear patterns in the performance of our RAG model versus the baselines.

5.2.1 Where RAG Excels: Long-Term Recall

The most significant advantage of the RAG model was in long-term memory recall. Table 4 shows a side-by-side comparison for a query in Turn 46, which required recalling information from Turn 38. The RAG model flawlessly retrieved the fact, while the Buffer Model, whose context window had long since passed Turn 38, failed completely. The judge scored the RAG model 6 points higher. This pattern was consistent across all long-term recall tasks.

5.2.2 Where RAG Struggles: Abstract and Creative Tasks

We also analyzed the turns where the RAG model received its lowest scores. These failures clustered around abstract or highly creative tasks where direct factual retrieval is less beneficial. For instance, in Turn 22, the user asked: "Explain the concept of black holes as if you were telling a bedtime story."

All three models produced a coherent, story-like explanation. However, the judge gave our RAG

Table 4: Example of RAG success on a long-term recall task (Turn 46).

Model	Response	Score	Judge's Reasoning
Our RAG Model	The theme was "cosmic comfort" for the coffee shop logo. Want me to flesh out three concrete logo concepts in that vibe (including crescent-moon ideas and a sunset-orange palette)...	**9/10**	The assistant accurately recalls the theme 'cosmic comfort' and leverages the user's preferences (sunset orange color and crescent-moon motif) to propose tailored next steps.
Buffer Memory	I don't have a record of a coffee shop logo theme from this chat. Could you repeat the theme...? If you'd like ideas, here are common logo themes for coffee shops: Cozy/Homely...	**3/10**	The assistant fails to recall the user's previously stated theme 'cosmic comfort' for the coffee shop logo, effectively ignoring the memory-testing objective.

model a low score of 5/10, with the reasoning: Coherent and kid-friendly... However it does not personalize for 'Alex' or reference their goals... It's generic rather than tailored to the persona.

This reveals a key limitation: our system is optimized to retrieve and inject *facts*, but these facts often have low semantic similarity to abstract or creative queries. When no relevant facts are retrieved, the RAG model's performance regresses toward that of the stateless baseline, as it has no specific context to ground its creative output.

6 Conclusion

This project successfully demonstrated that a RAG-based architecture can endow an LLM with effective long-term memory, leading to more personalized and coherent conversational experiences. Our model, which dynamically learns and retrieves user-specific facts, achieved significantly higher scores on both a challenging custom dataset (7.58/10) and a public benchmark (8.44/10) compared to stateless and short-term memory baselines.

The most surprising and clear result was the stark difference in performance on long-term recall tasks, where our RAG system flawlessly retrieved facts from many turns prior. The most difficult aspect was designing a synthetic dataset that could fairly and rigorously test different facets of memory over a simulated long-term interaction.

Our results were encouraging, but the error analysis also revealed a key limitation: the system struggles to personalize responses for abstract or creative queries where factual recall is less relevant. If we were to continue this project, future directions would include:

1. **Improving Creative Personalization**: A significant challenge is personalizing stylistic or creative responses. Future work could explore methods to retrieve and use more abstract preferences (e.g., "I like a witty tone," "I prefer minimalist design"). This might involve a secondary memory store for stylistic traits or developing a retriever that can match a query's abstract intent to a user's personality profile.
2. **Advanced Memory Management**: A simple, growing list of facts is prone to becoming outdated or contradictory. A more advanced system would need mechanisms for fact verification, contradiction resolution (e.g., if a user changes their preference), and memory summarization. Techniques from Li et al. (2025) involving reflective summarization could be adapted to periodically condense and refine the user's KB.
3. **Privacy Controls**: The storage of personal data is a critical concern. A production-ready version of this system would require robust privacy controls. This includes implementing a user-facing interface to allow users to view, edit, or selectively delete stored memories, ensuring transparency and user agency over their data.
4. **Hybrid Memory Systems**: Our system relies on a single vector store. A hybrid approach could be more powerful, combining the semantic retrieval of a vector KB with a structured, symbolic KB (e.g., a graph database) for storing crisp, relational facts like family connections or key dates. This could improve the precision of certain types of recall.

Overall, this project provides a strong proof-of-concept for building more capable and personable AI assistants through a lightweight, retrieval-based approach to memory.

References

(2025). Reflective memory management for personalized dialogue generation (anonymous acl 2025 submission). https://aclanthology.org/2025.acl-long.413/.

AI, C. (2024a). Why 'llm-as-a-judge' is the best llm evaluation method. https://www.confident-ai.com/blog/why-llm-as-a-judge-is-the-best-llm-evaluation-method.

AI, E. (2024b). Llm-as-a-judge. https://www.evidentlyai.com/llm-guide/llm-as-a-judge.

AI, T. (2023). Llm-as-a-judge: Can ai systems evaluate model outputs? https://toloka.ai/blog/llm-as-a-judge-can-ai-systems-evaluate-model-outputs/.

Anonymous (2023a). Learning retrieval augmentation for personalized dialogue generation. https://openreview.net/revisions?id=O3Up0uXVrb.

Anonymous (2023b). Locomo: A benchmark for evaluating dialogue-amr based content manipulation in task-oriented dialogue. https://openreview.net/forum?id=JoZAvPlNMy.

Anonymous (2025). Reflective memory management for personalized dialogue generation.

Architects, A. (2023). Memgpt: Bridging the gap between memory and generative capacity. https://aiarchitects.ai/memgpt-bridging-the-gap-between-memory-and-generative-capacity/.

Banerjee, P. (2023). Understanding memgpt: Enhancing conversational ai through advanced memory management. https://www.linkedin.com/pulse/understanding-memgpt-enhancing-conversational-ai-through-banerjee-lpqrf.

Bordes, A. and Weston, J. (2022). Learning end-to-end for dialogue systems. In *Proceedings of the 39th International Conference on Machine Learning*, New York, NY, USA. Association for Computing Machinery.

Chen, Q., Lin, J., Zhang, Y., Ding, M., Cen, Y., Yang, H., and Tang, J. (2019). Towards knowledge-based recommender dialog system.

Chen, W., Su, Y., Chen, Z., Dziri, N., Sugiura, Y., Matsuo, Y., Narasimhan, K., and Singh, S. (2023). Augmenting language models with long-term memory. https://arxiv.org/pdf/2306.07174.pdf.

Consensus (2023). Lapdog: Learning retrieval augmentation for personalized dialogue generation. https://www.consensus.app/papers/details/c10109c4047c59cd952f713bafcdcac0/.

EmergentMind (2025). Paper: Reflective memory management for personalized dialogue generation. https://www.emergentmind.com/papers/2503.08026.

Encord (2024). Llm as a judge: A new era for evaluating large language models. https://encord.com/blog/llm-as-a-judge/.

Feng, S., Ren, X., Li, K., and sun, X. (2021a). Comemnn: A cooperative memory network for dialogue response generation.

Feng, S., Ren, X., Li, K., and Sun, X. (2021b). Comemnn: A cooperative memory network for dialogue response generation. In *Proceedings of the Web Conference 2021*, WWW '21, page 3641–3651, New York, NY, USA. Association for Computing Machinery.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. (2020a). Realm: Retrieval-augmented language model pre-training.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. (2020b). Realm: Retrieval-augmented language model pre-training. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 3929–3938. PMLR.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. (2020c). Realm: Retrieval-augmented language model pre-training.

hqsiswiliam (2023). Lapdog: Learning retrieval augmentation for personalized dialogue generation. https://github.com/hqsiswiliam/LAPDOG.

Huang, L., Peng, Z., Wang, W., Wang, W., and Gao, K. (2023a). LAPDOG: Learning retrieval augmentation for personalized dialogue generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2517–2530, Singapore. Association for Computational Linguistics.

Huang, L., Peng, Z., Wang, W., Wang, W., and Gao, K. (2023b). LAPDOG: Learning retrieval augmentation for personalized dialogue generation. https://aclanthology.org/2023.emnlp-main.154/.

Huang, L., Peng, Z., Wang, W., Wang, W., and Gao, K. (2023c). Lapdog: Learning retrieval augmentation for personalized dialogue generation. https://personalpages.surrey.ac.uk/w.wang/papers/Huang%20et%20al_EMNLP_2023.pdf.

Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with gpus. IEEE.

K S, S. (2025). Bias and fairness in personalized systems. *International Journal of Scientific Development and Research*, 10(4).

Kim, M., Shin, J.-w., Lee, S.-m., Kim, D.-h., Kim, H.-j., and Kim, I.-h. (2024). PREFEVAL: A framework for evaluating llms on following persona-grounded preferences. https://aclanthology.org/2024.findings-emnlp.592.pdf.

Kim, S.-M. et al. (2025). Do llms follow your preferences? a benchmark for preference following. In *NAACL*.

Kumar, P. (2024). How ai agents remember things – vector stores in llm memory. https://www.freecodecamp.org/news/how-ai-agents-remember-things-vector-stores-in-llm-

Labs, A. (2024). Llm as a judge. https://www.ai21.com/glossary/llm-as-a-judge/.

LangChain (2024). Faiss vector store. https://python.langchain.com/docs/integrations/vectorstores/faiss/.

Li, J.-C. et al. (2025). A reflective memory model for personalized dialogue. In *EMNLP*.

Li, T.-X. et al. (2024a). Llm-pbe: A new benchmark for privacy analysis of language models. In *ACL*.

Li, W., Liu, Y., nan Tang, Y., CUI, J.-S., ZHANG, W.-N., and Liu, T. (2024b). Learning retrieval augmentation for personalized dialogue generation.

Li, W., Liu, Y., nan Tang, Y., CUI, J.-S., ZHANG, W.-N., and Liu, T. (2024c). Learning retrieval augmentation for personalized dialogue generation.

Liu, W., Cheng, Y., Nie, J.-Y., Zhan, H., Shi, J.-X., and Zhang, Z. (2024). Personalized dialogue generation with user-controllable scenarios.

Madotto, A., Lin, Z., Wu, C.-S., and Fung, P. (2021). Learning to kill: A case study in dialogue model fidelity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 782–794, Online. Association for Computational Linguistics.

Madotto, A., Wu, C.-S., Lee, L., and Fung, P. (2019). Personalizing dialogue agents with heterogeneous memory. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4513–4523, Hong Kong, China. Association for Computational Linguistics.

Mishra, K., Prasad, A., Kashyap, A., Kanthara, S., and Chava, R. (2023). LOCOMO: A benchmark for evaluating dialogue-amr based content manipulation in task-oriented dialogue. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 829–838, Hong Kong, China. Association for Computational Linguistics.

Mishra, K., Prasad, A., Kashyap, A., Kanthara, S., and Chava, R. (2025). LOCOMO: A benchmark for evaluating dialogue-amr based content manipulation in task-oriented dialogue. https://aclanthology.org/2025.findings-acl.1014/.

Moonlight, T. (2024). Learning retrieval augmentation for personalized dialogue generation. https://www.themoonlight.io/en/review/learning-retrieval-augmentation-for-personalized-dialogue-generation.

Ocean, D. S. (2023). Memgpt: Towards llms as operating systems. https://datasciocean.com/en/paper-intro/memgpt/.

Packer, C. (2023). Memgpt slides for neurips 2023. https://neurips.cc/media/neurips-2023/Slides/72461.pdf.

Packer, C., Fang, V., Patil, S. G., Lin, K., Wooders, S., and Gonzalez, J. E. (2023). Memgpt: Towards llms as operating systems.

Packer, C., Fang, V., Patil, S. G., Lin, K., Wooders, S., and Gonzalez, J. E. (2024). Memgpt: Towards llms as operating systems. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, New York, NY, USA. Association for Computing Machinery.

PingCAP (2024). Mastering faiss vector database: A beginner's handbook. https://www.pingcap.com/article/mastering-faiss-vector-database-a-beginners-handbook/.

Research, G. (2023). Synthetic-persona-chat dataset. https://github.com/google-research-datasets/Synthetic-Persona-Chat.

SlideShare (2024). Memgpt: Introduction to memory augmented chat. https://www.slideshare.net/slideshow/memgpt-introduction-to-memory-augmented-chat/269549130.

submission, A. A. . (2025). Reflective memory management for personalized dialogue generation. https://arxiv.org/pdf/2503.08026.pdf.

TechTarget (2020). Retrieval-augmented language model pre-training. https://www.techtarget.com/searchenterpriseai/definition/Retrieval-Augmented-Language-Model-pre-training.

Wang, S., Jin, Y., Cao, H., Lo, C. R., Liu, Z., Chang, X.-W., Hauser, S. L., Guo, Z., and Rajpurkar, P. (2025). Learning user preferences of large language models in healthcare.

Wang, Y., Chen, W., Su, Y., Chen, Z., Dziri, N., Sugiura, Y., Matsuo, Y., Narasimhan, K., and Singh, S. (2023). Augmenting language models with long-term memory. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 69083–69106. Curran Associates, Inc.

Wei, C. et al. (2024). Facts as functions: A new framework for factual knowledge probing. In *ACL*.

Weston, J., Chopra, S., and Bordes, A. (2014). Memory networks.

Yang, Y., Kang, H., Kim, S., wan Kim, T., and Kang, J. (2024). Unims-rag: A unified multi-source retrieval-augmented generation for personalized dialogue system.

Zhang, S. et al. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*.