

Memoria: A Knowledge Graph based Framework

Presenter: Dhagash Mehta

Authors: Samarth Sarin, Lovepreet Singh, Bhaskarjit Sarmah,

Dhagash Mehta

BlackRock, Inc.

Disclaimer: All views presented in this talk are those of the presenter and not of BlackRock, Inc.

Problem Statement



Isolated Exchanges

While LLM-powered chat applications are becoming increasingly widespread, they often fall short in preserving the flow of conversation. A significant limitation is the absence of a reliable memory mechanism, leading to exchanges where each user input is treated in isolation.



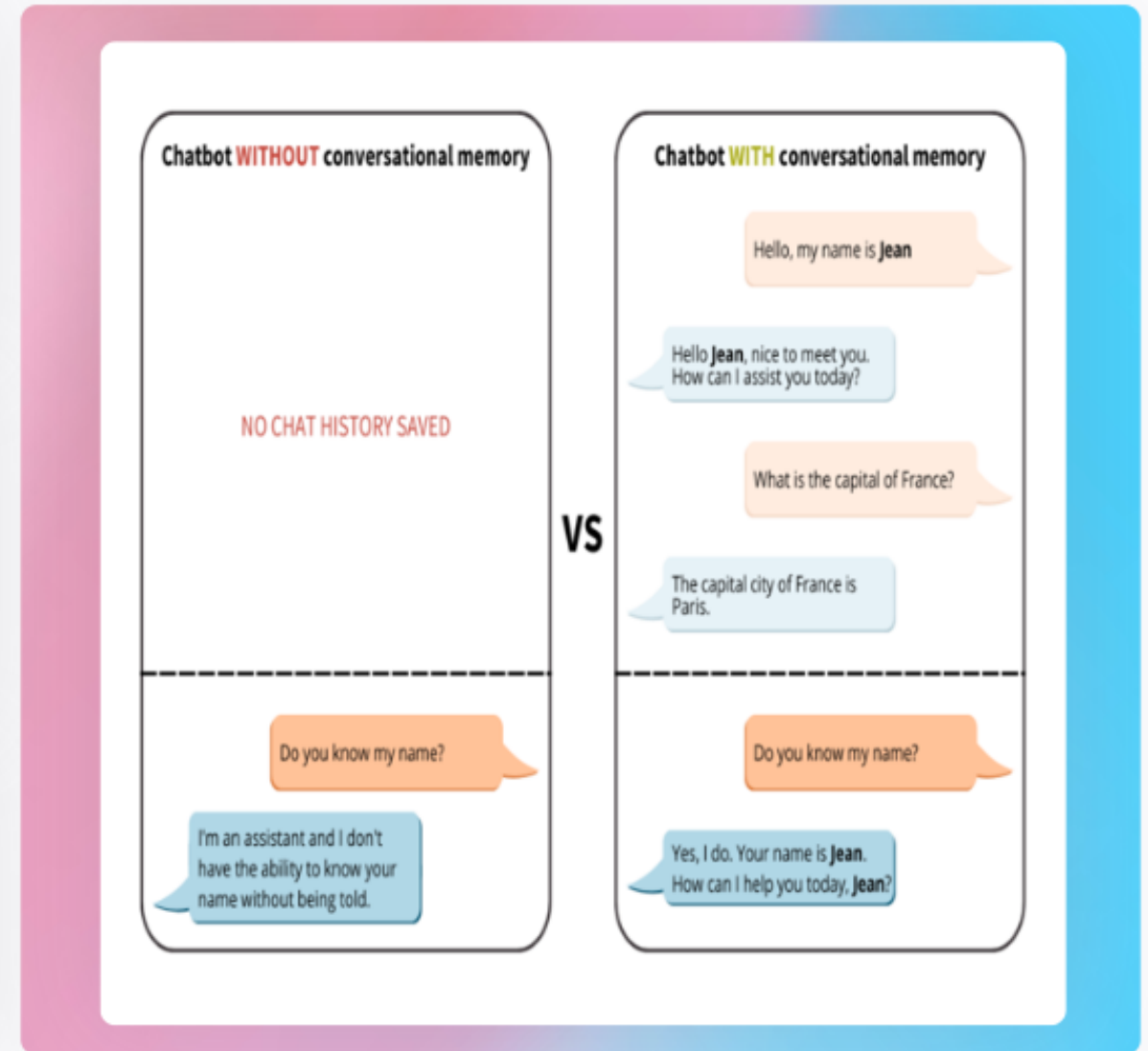
Session-Limited Memory

In instances where memory functionality is implemented, it is usually limited to the duration of a single session. As a result, any contextual information is lost once the session is refreshed, making it difficult to tailor responses to individual users over time.



Reduced Engagement

This reliance on ephemeral memory diminishes the system's ability to hold meaningful, context-rich conversations. Consequently, it hampers user engagement and reduces the potential for more personalized and interactive experiences.



Current Challenges in Existing LLM Systems

USER IMPACT



Frustrating Interactions

Without memory, users must constantly repeat themselves. This repetition disrupts the flow of conversation, leading to frustration and disengagement.



Lack of Personalization

Generic responses feel robotic. Without the ability to remember user preferences, the AI fails to offer tailored, meaningful interactions undermining trust and long-term engagement.

BUSINESS & TECHNOLOGY



Barriers to Adoption

Industries like customer service, education, and digital assistants rely heavily on context continuity. Without memory, these applications fall short of expectations, limiting real-world deployment.



Operational Inefficiencies

Redundant queries increase workload. Users re-explaining their context leads to higher support volume, driving up infrastructure and staffing costs.



Session-based Memory is Ephemeral

Most systems only retain memory within a single session. Once it ends, all context is wiped—breaking continuity across conversations.

Proposed Enhancements for Memory and Personalization



Structured Conversation History with SQL Database

✓ **What it does:** Stores all user interactions in an SQL database with key metadata such as timestamps, messages, summaries, and token usage.

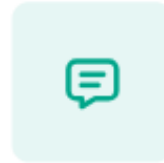
✓ **Why it matters:** Ensures historical context is always available, eliminating the need for users to repeat themselves.



User Persona via Knowledge Graph

✓ **What it does:** Builds and updates a user-level knowledge graph that captures preferences, discussed topics, and key entities.

✓ **Why it matters:** Enables highly personalized responses, making AI interactions more intuitive and engaging over time.



Session-Level Memory for Real-Time Context

✓ **What it does:** Generates session-level summaries to maintain coherence within ongoing conversations.

✓ **Why it matters:** Prevents short-term context loss, so users don't have to repeat information within the same session.



Seamless Retrieval for Context-Aware Responses

✓ **What it does:** Dynamically retrieves relevant past discussions and knowledge graph data for generating responses.

✓ **Why it matters:** Enables a smooth, continuous experience across multiple sessions, making the chatbot feel intelligent, responsive, and personalized.

Memoria: Key Capabilities



PERSISTENT MEMORY

Structured SQL Database as the foundation. Stores conversation history, summaries, metadata, token usage, and costs. Enables efficient retrieval of past interactions, ensuring seamless conversation flow without context loss.



DUAL MEMORY

Session-Level Summaries – Maintain real-time session continuity by dynamically summarizing conversations.

User-Level Knowledge Graph – Captures user persona across multiple interactions, including preferences, interests, and past discussions.



CUSTOMIZABLE PROMPTS

Stores system prompts, summary prompts, and knowledge graph prompts in database. Allows developers to fine-tune responses and control system behaviour.



USAGE TRACKING

Logs token consumption per conversation to help optimize LLM costs. Provides insights into the efficiency and cost-effectiveness of interactions.

Experiment Results

DATASET

LongMemEval (70 single-session user and 78 knowledge update sessions with each conversation up to 115K tokens)

LLM MODEL

gpt-4.1-mini

EMBEDDING MODEL

text-embedding-ada-002

TYPE	FULL CONVERSION	A-MEM (ST)	A-MEM (OA)	MEMORIA
Single-session	85.7%	78.5%	84.2%	87.1%
Knowledge-update	78.2%	76.2%	79.4%	80.8%

APPROACH	QUESTION TYPE	INFERENCE TIME	AVG. TOKEN LENGTH
Full Context	single-session	391 secs	115K
Memoria	single-session	260 secs	398
A-Mem (ST)	single-session	290 secs	958
A-Mem (OA)	single-session	252 secs	934

APPROACH	QUESTION TYPE	INFERENCE TIME	AVG. TOKEN LENGTH
Full Context	knowledge-update	522 secs	115K
Memoria	knowledge-update	320 secs	400
A-Mem (ST)	knowledge-update	364 secs	933
A-Mem (OA)	knowledge-update	328 secs	928

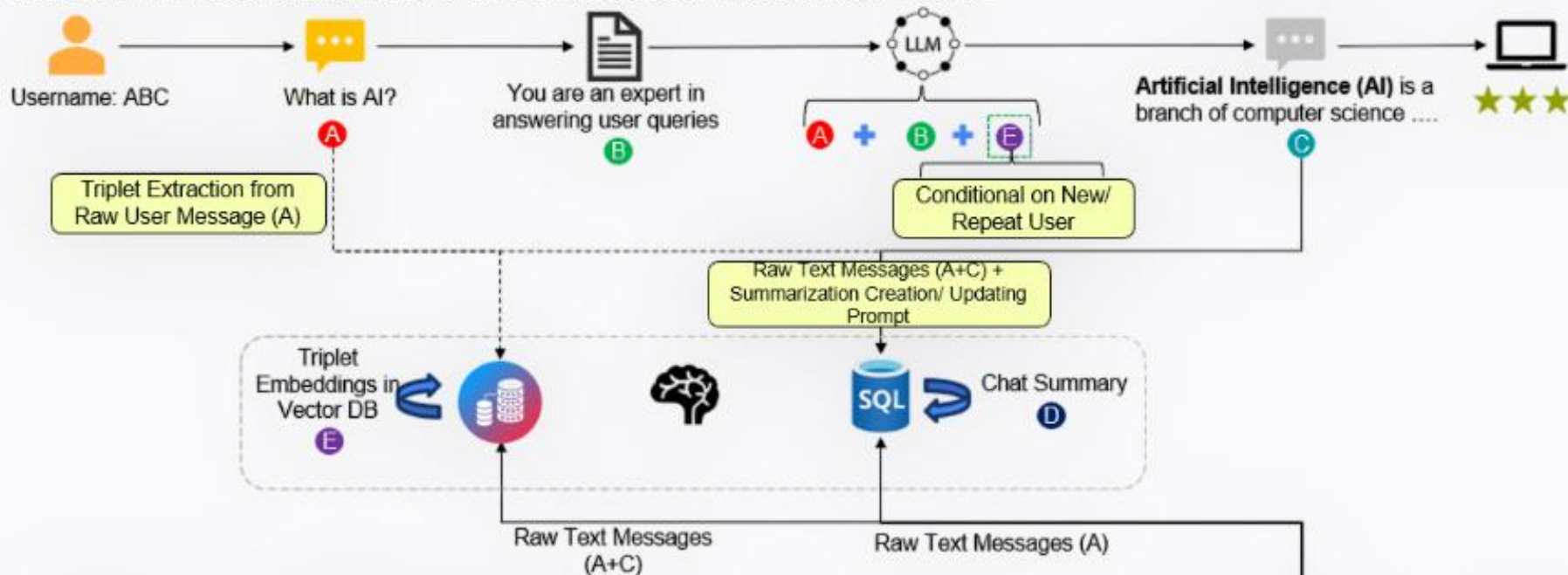
Thank you

For any further questions or feedback, please feel free to reach out.

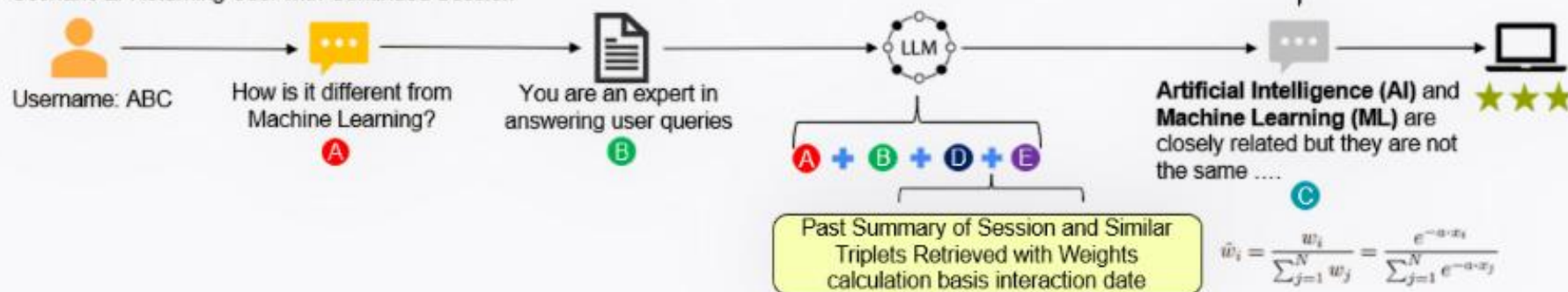
ALL THE VIEWS IN THIS PRESENTATION ARE THOSE OF THE PRESENTERS, NOT OF BLACKROCK INC.

Memoria Architecture

Scenario 1: When the user is either visiting for the first time or is a returning user starting a new chat.



Scenario 2: Returning User with Continued Session



Sample Data

single-session-user



I've been listening to audiobooks during my daily commute, which takes **45 minutes each way**.

? How long is my commute to work?

✓ 45 minutes each way

knowledge-update



I'm doing a century ride soon, do you think my **road bike** is ready for that distance, considering I've done 2,000 miles on it? I've been using it with my other two bikes, **a mountain bike and a commuter bike**.

? How many bikes do I currently own?



I just got a new **hybrid bike** recently.

✓ 4