ETH zürich

cgl

Memories for Virtual AI Characters

Fabian Landwehr, Erika Varis Doggett, Romann M. Weber

INLG 2023

Overview

- Introduction
- Proposed System
  - Response Generation
  - Memory Creation
- Evaluation
- Future Work
- Conclusion

Introduction

Introduction

- Prior chatbots relied on scripted dialog but lacked improvisation capabilities (e.g., dialog trees).

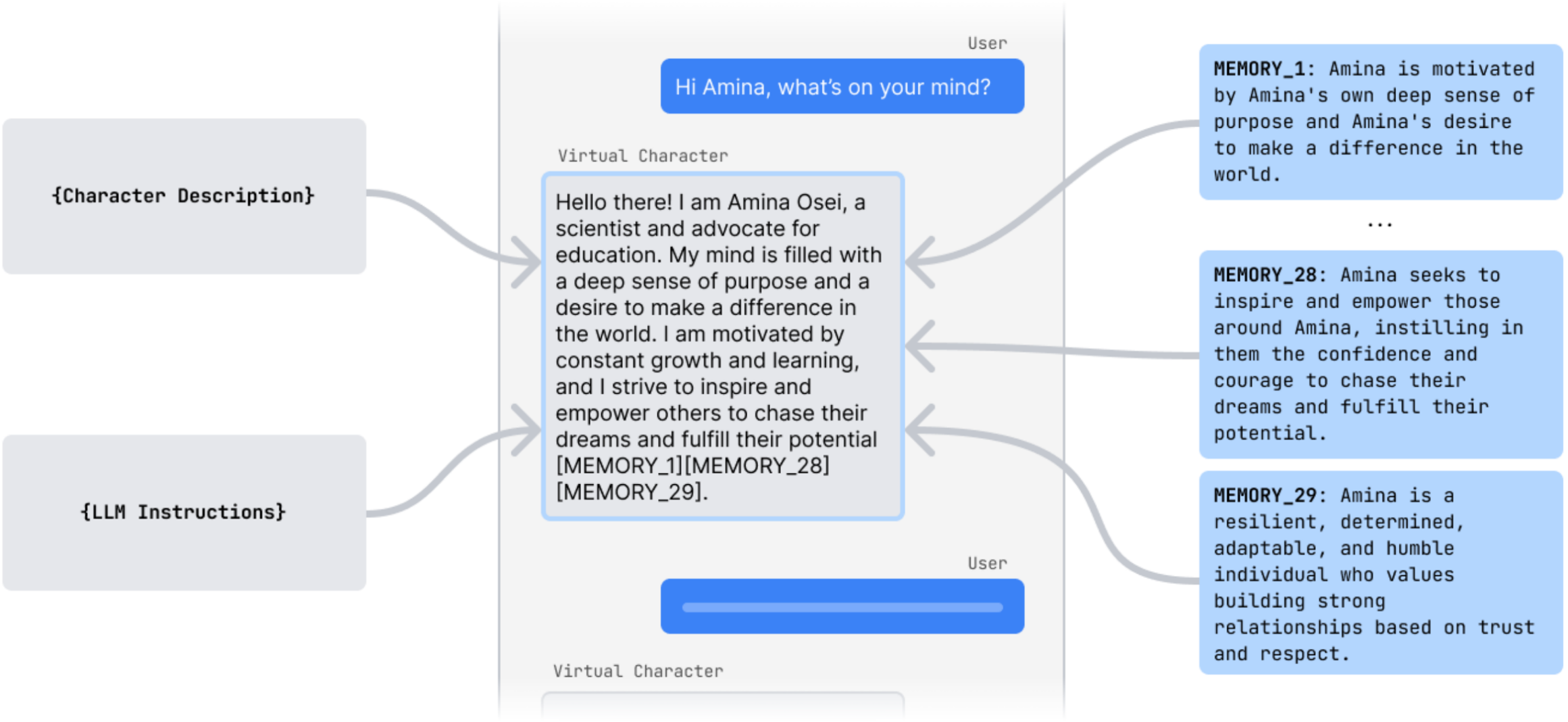
Introduction

- Prior chatbots relied on scripted dialog but lacked improvisation capabilities (e.g., dialog trees).
- LLMs provide improvisation capabilities but introduce *controllability and hallucination issues.*

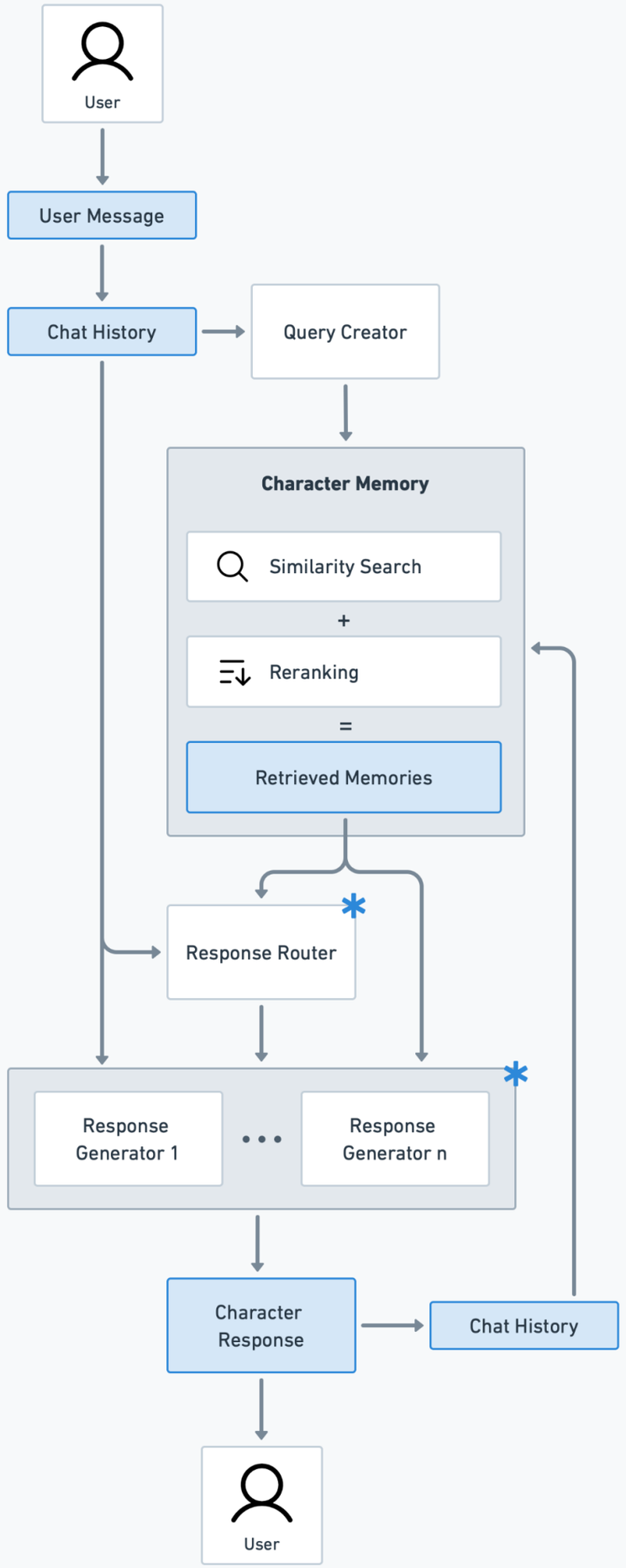
Introduction

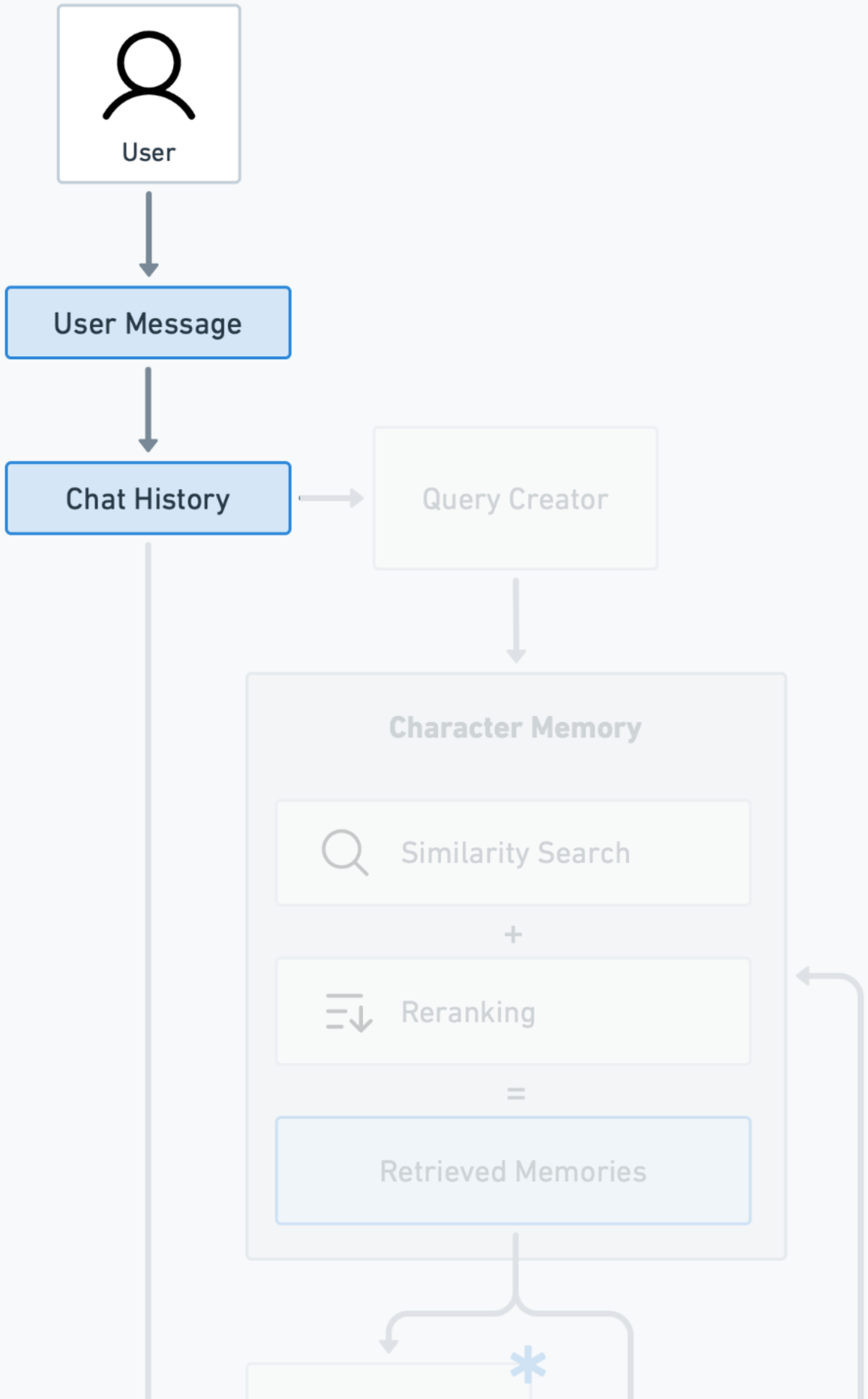
- Prior chatbots relied on scripted dialog but lacked improvisation capabilities (e.g., dialog trees).
- LLMs provide improvisation capabilities but introduce *controllability and hallucination issues.*
- Long-term memories as a solution to both?

Introduction



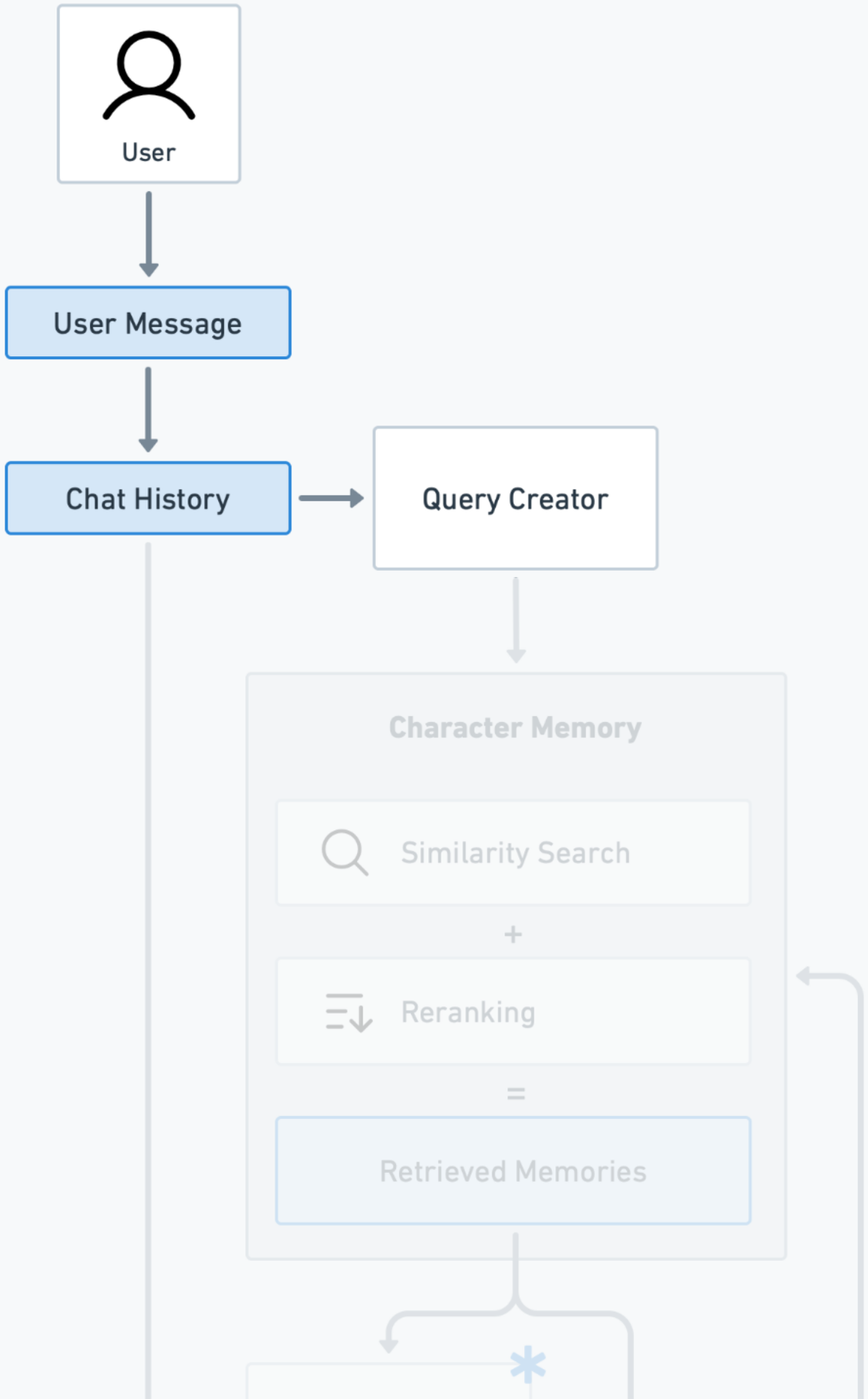
Proposed System


User
User Message
Chat History
Query Creator
Character Memory
Similarity Search
+
Reranking
=
Retrieved Memories
Response Router
Response Generator 1
Response Generator n
Character Response
Chat History
User



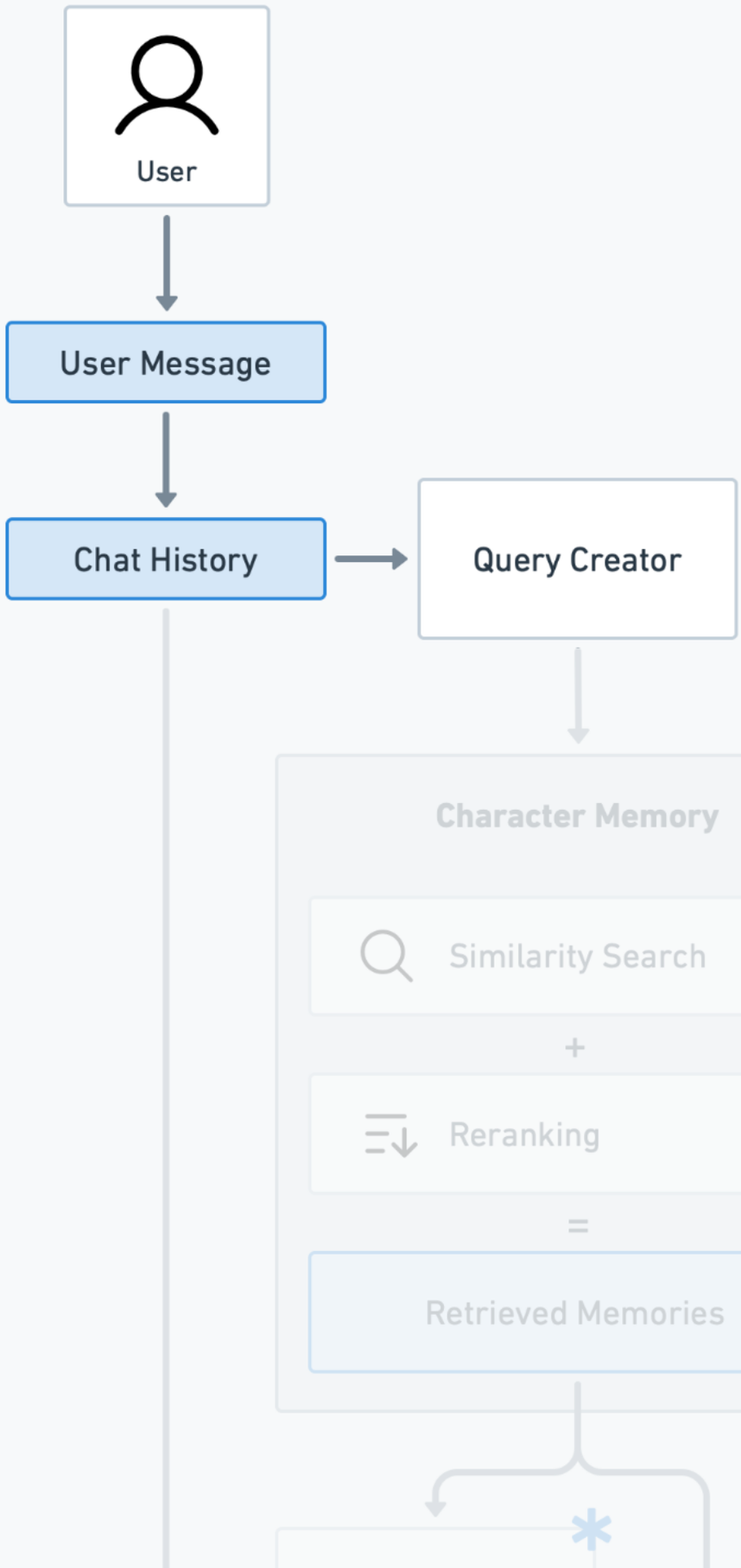
Chat History

Contains the last k messages plus a summary of previous messages.



Query Creator

Takes the chat history and creates a search query for relevant memories.



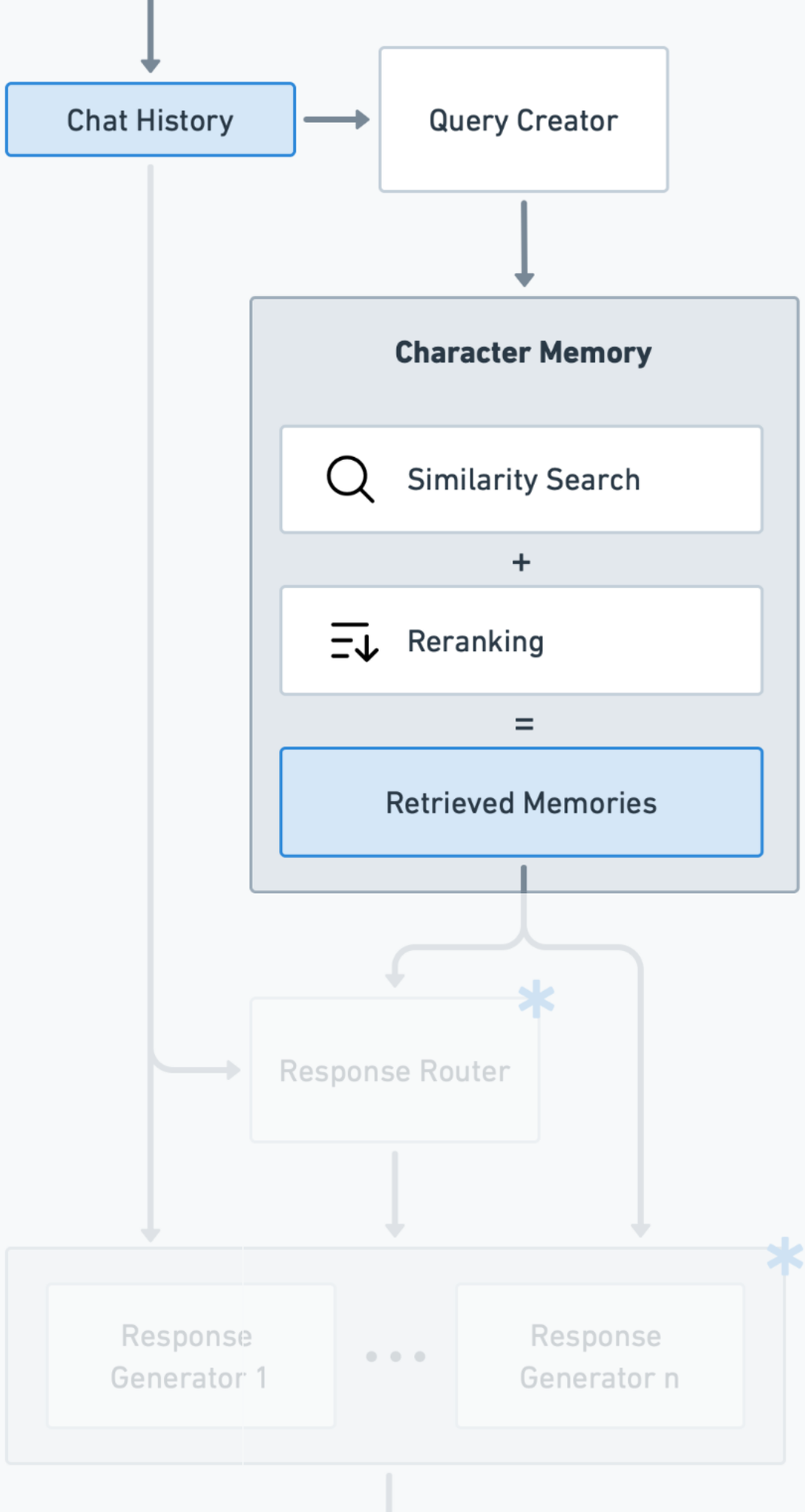
Query Creator

Takes the chat history and creates a search query for relevant memories.

Sherlock Holmes:
I thought about my father yesterday.

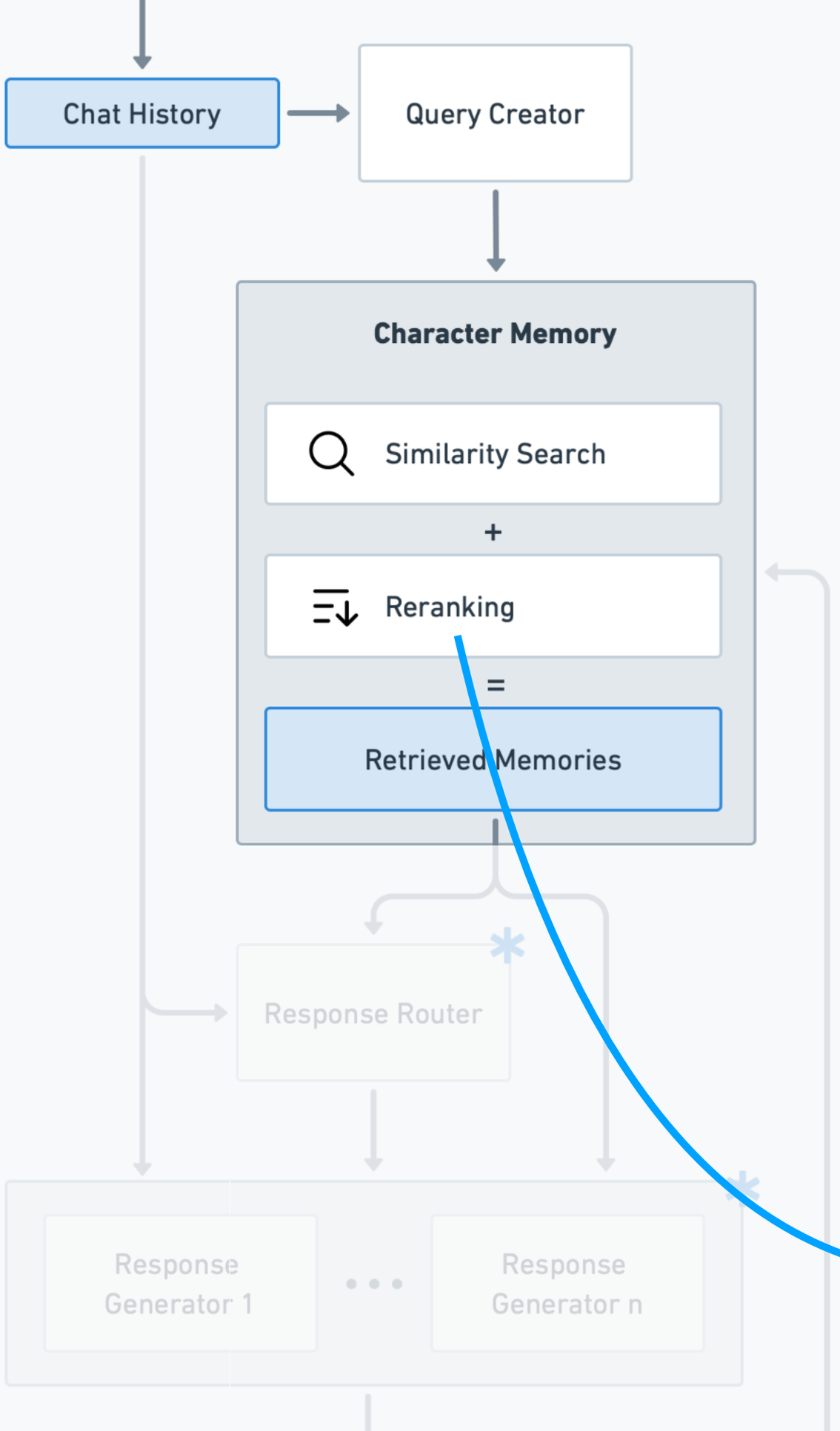
User:
What is his name?

“Name of Sherlock Holmes’ father”



Character Memory

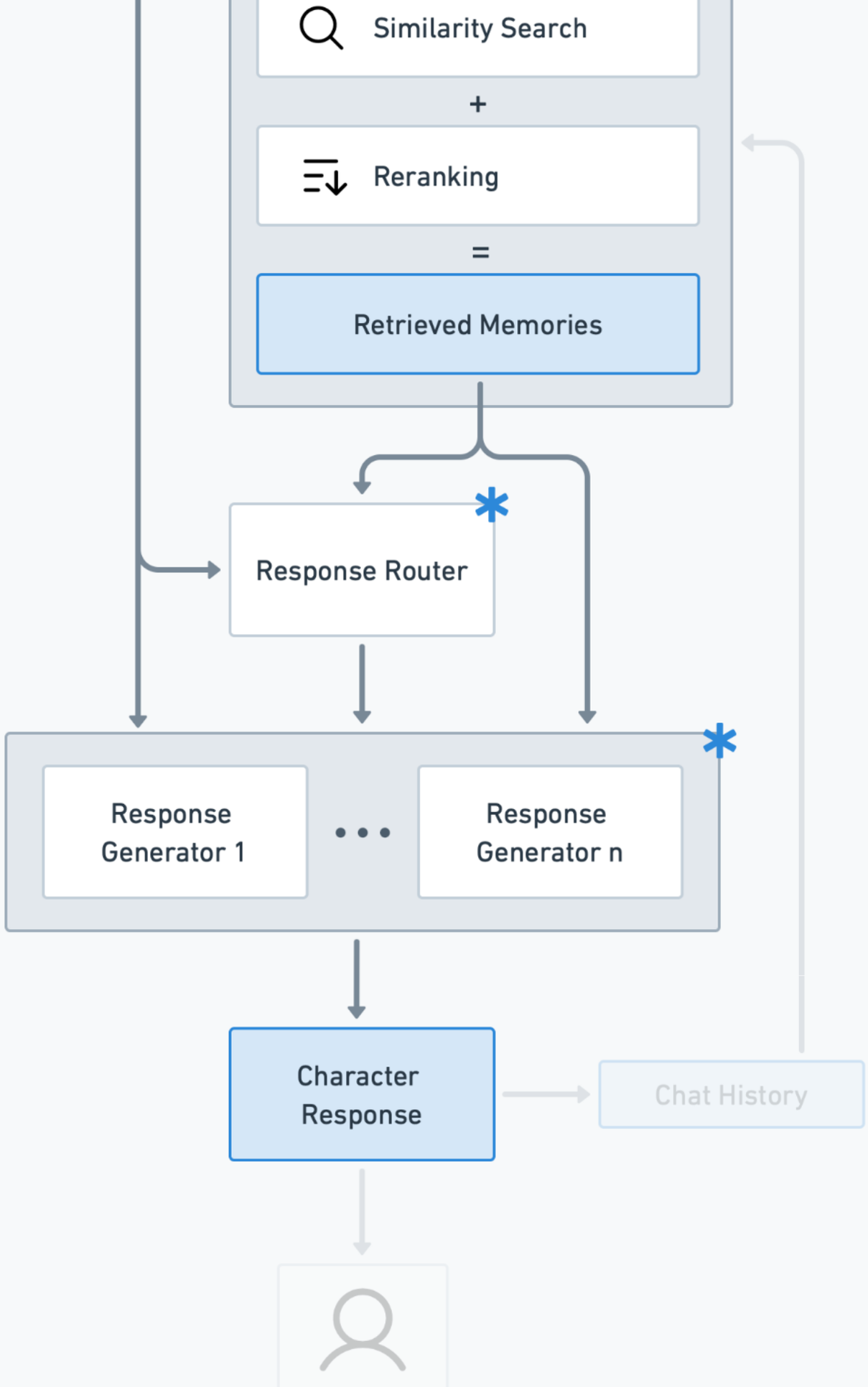
- Consists of multiple “knowledge-sources”.
- Memories are retrieved based on semantic similarity to the query.
- The memories are then reranked and the k best results are returned.



Character Memory

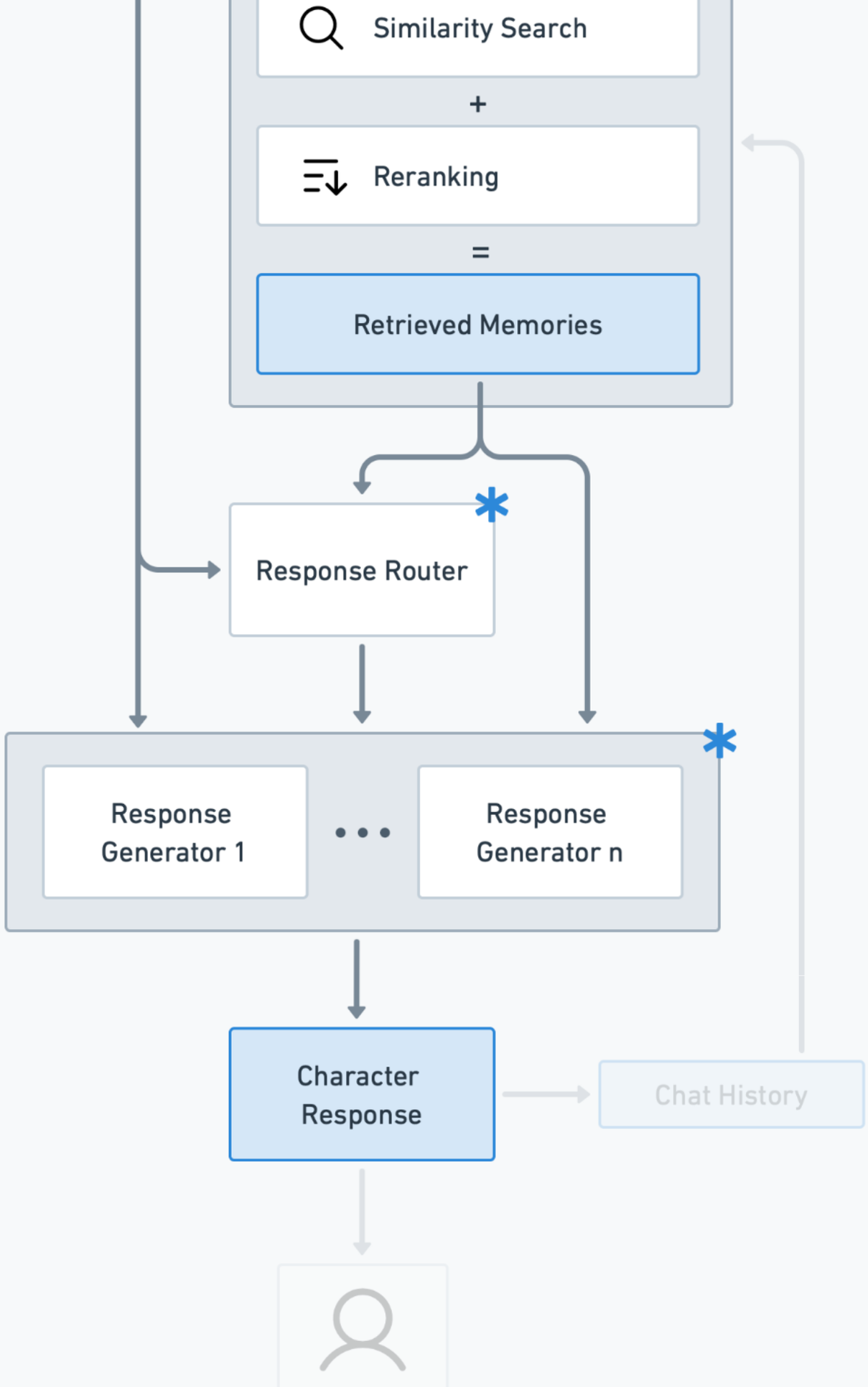
- Consists of multiple “knowledge-sources”.
- Memories are retrieved based on semantic similarity to the query.
- The memories are then reranked and the k best results are returned.

customizability!



Response Routing

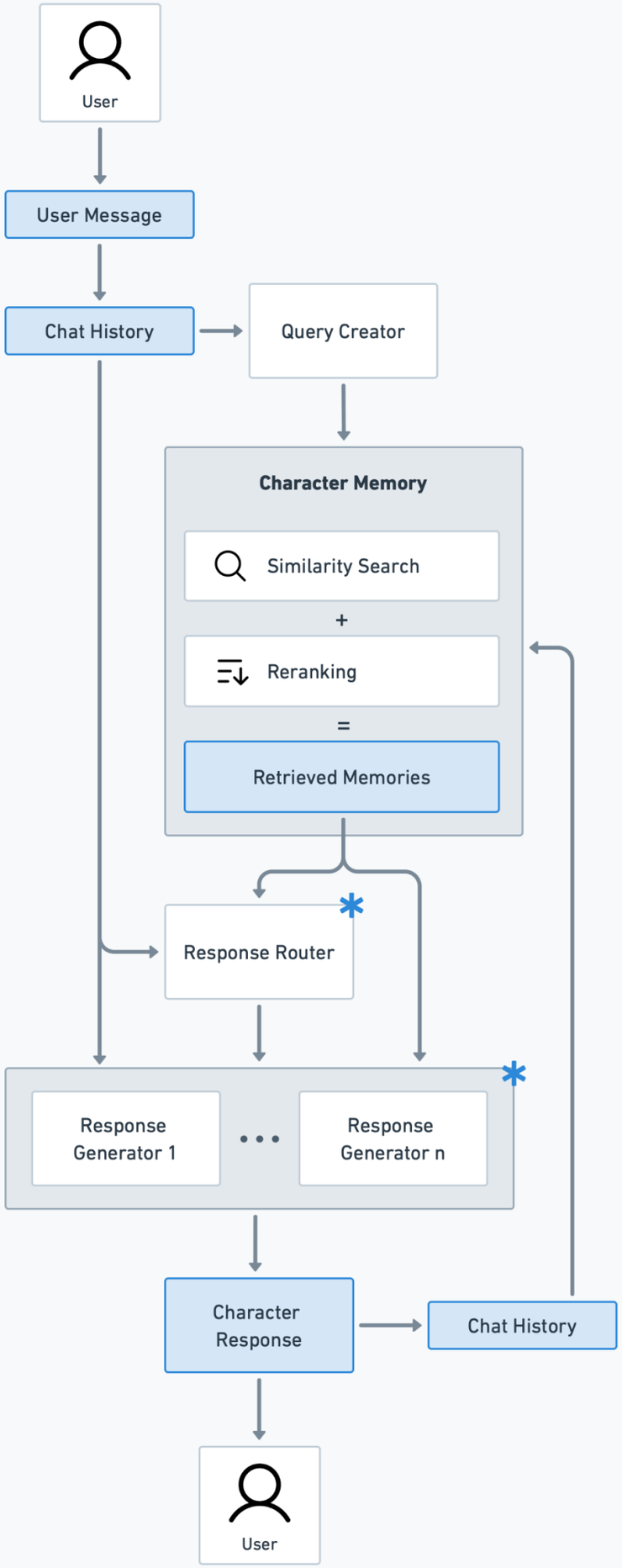
- Depending on the current situation, we may want custom response behahviors.
- The response-router decides what type of response is appropriate and selects a generator.



Response Generation

The LLM prompt contains:

- Instructions
- Character name and description (500 words)
- Chat history
- List of retrieved memories



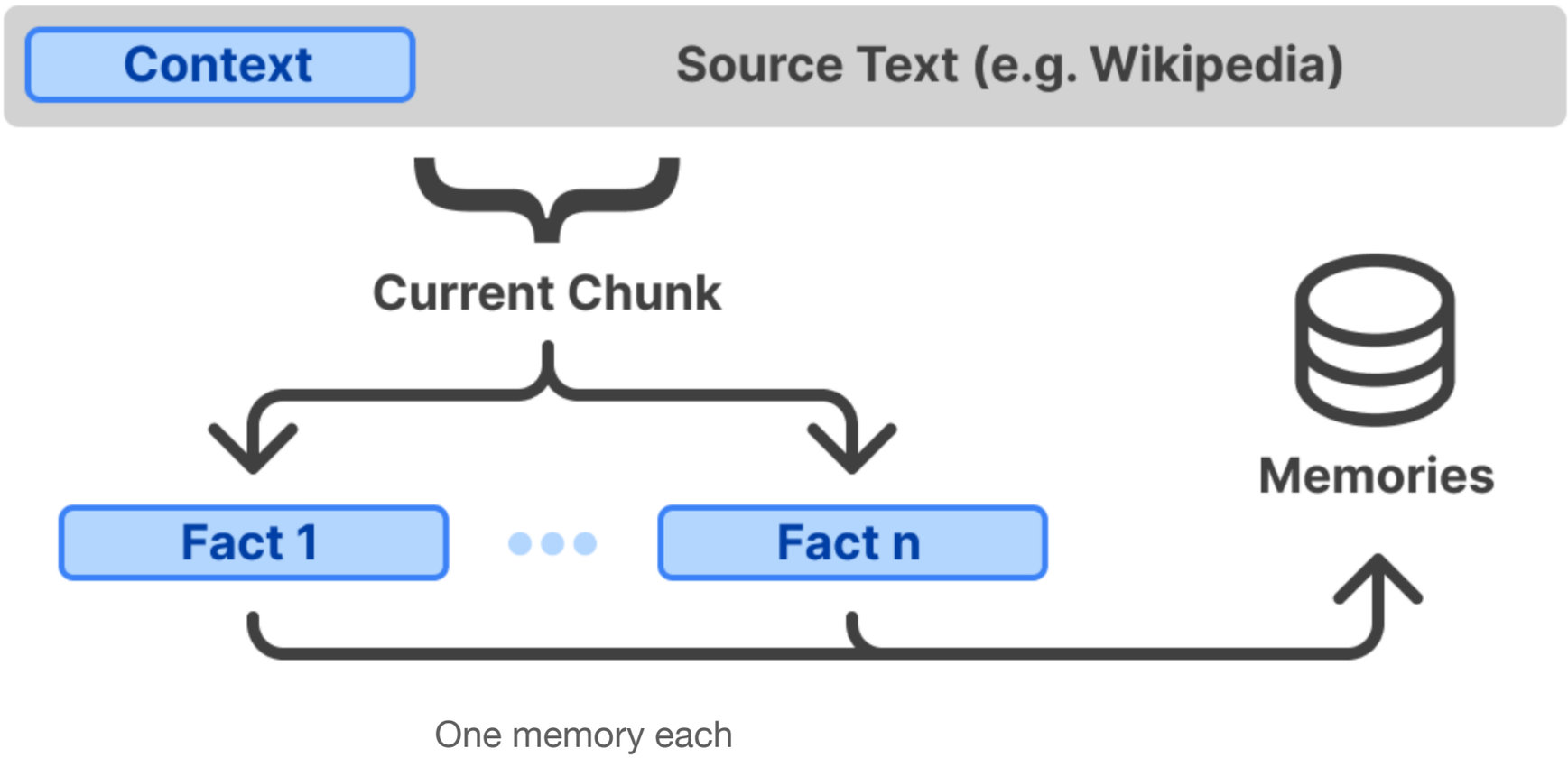
Using the Response

The generated message is sent to the user and inserted into the chat history.

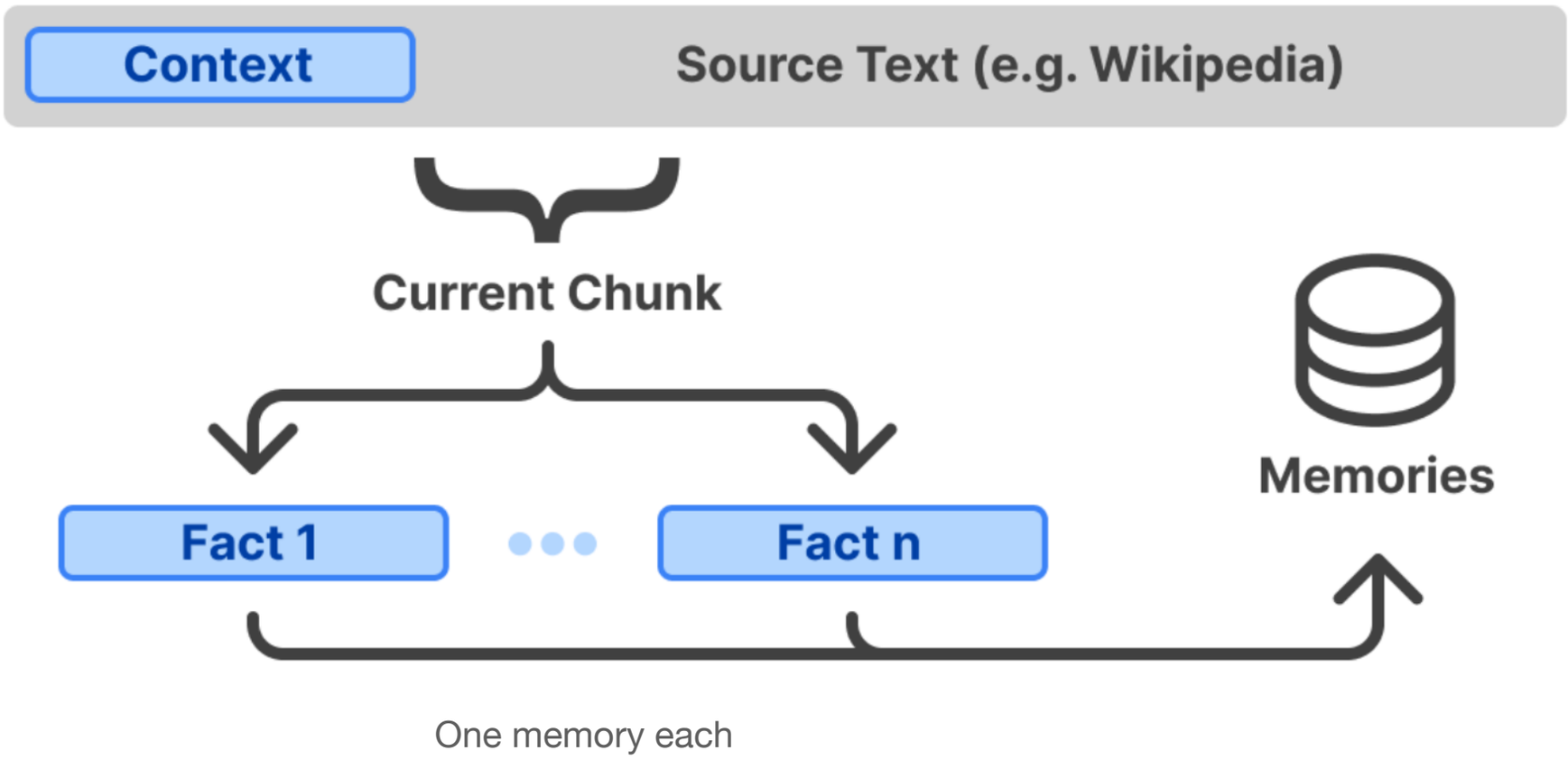
Eventually, the chat history is converted to new character memories.

Memories Can Be Created from Any Raw Text

Memories Can Be Created from Any Raw Text



Memories Can Be Created from Any Raw Text



Each memory contains:

- Text content
- embedding
- meta info

Evaluation Methodology

Evaluation Methodology

We evaluated:

1. How grounded in the retrieved memories are the generated responses?

2. How accurate are the provided references?

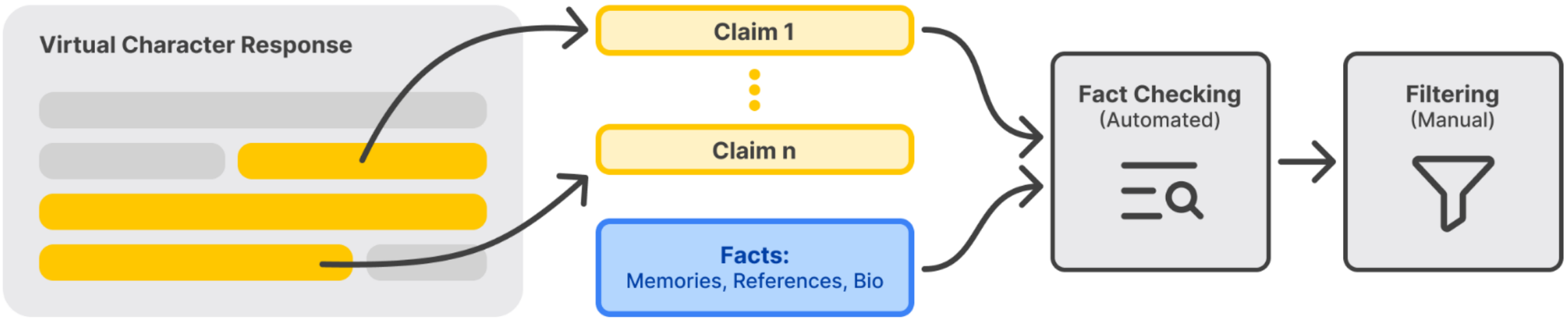
Evaluation Methodology

We evaluated:

1. How grounded in the retrieved memories are the generated responses?

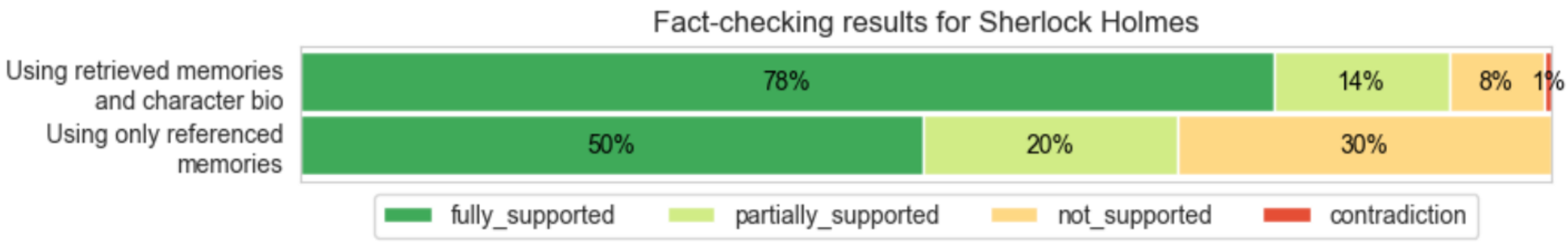
2. How accurate are the provided references?

We used a fact-checking pipeline based on GPT-4:



Evaluation Results

- Character responses are largely grounded in memories, preserving integrity.
- The provided references by the LLM were partially incorrect.
- Results were better when the LLM had no intrinsic knowledge about the simulated virtual character.



Future Work

- Response routing stage to avoid hallucinations
- Reduce latency by exploiting parallelism
- Refine the evaluation strategy
- Investigate the impact of memories on personality over time

Conclusion

- The presented system enables the creation of interactive, improvising characters with long-term memory.
- Character memories are designed to be human-like, being strengthened if used and forgotten when not.
- The memories help LLM stay in character and consistent.
- The system also provides building blocks for future research on virtual character memory and LLM steering.

Memories for Virtual AI Characters

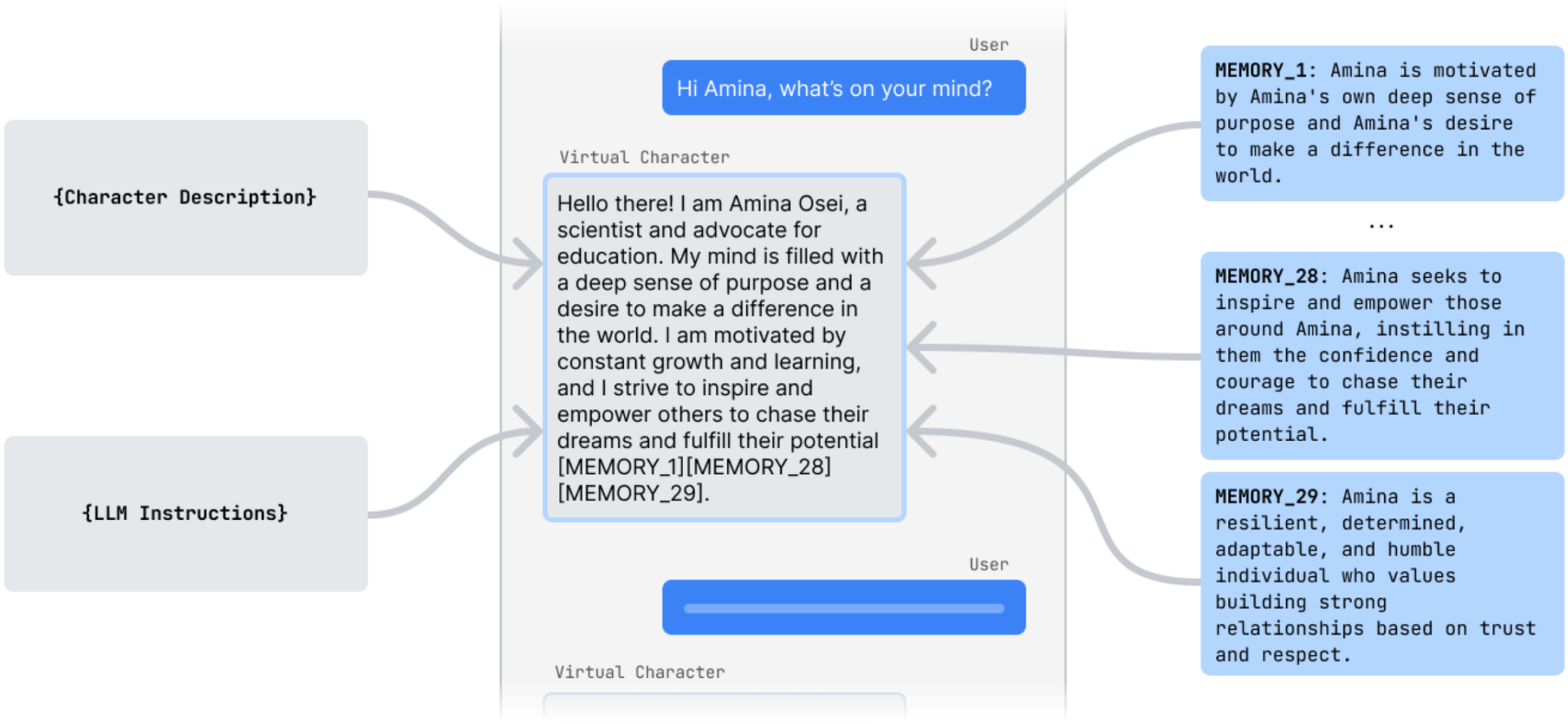
Thank you for listening!

Fabian Landwehr — fabian.landwehr@inf.ethz.ch
Erika Varis Doggett — erikavaris@gmail.com
Romann M. Weber — romann.weber@gmail.com

Paper

https://sigdialinlg2023.github.io/paper_inlg99.html

Appendix


{Character Description}
{LLM Instructions}
User
Hi Amina, what's on your mind?
Virtual Character
Hello there! I am Amina Osei, a scientist and advocate for education. My mind is filled with a deep sense of purpose and a desire to make a difference in the world. I am motivated by constant growth and learning, and I strive to inspire and empower others to chase their dreams and fulfill their potential [MEMORY_1][MEMORY_28] [MEMORY_29].
User
Virtual Character
MEMORY_1: Amina is motivated by Amina's own deep sense of purpose and Amina's desire to make a difference in the world.
...
MEMORY_28: Amina seeks to inspire and empower those around Amina, instilling in them the confidence and courage to chase their dreams and fulfill their potential.
MEMORY_29: Amina is a resilient, determined, adaptable, and humble individual who values building strong relationships based on trust and respect.

Fact-Checking Pipeline

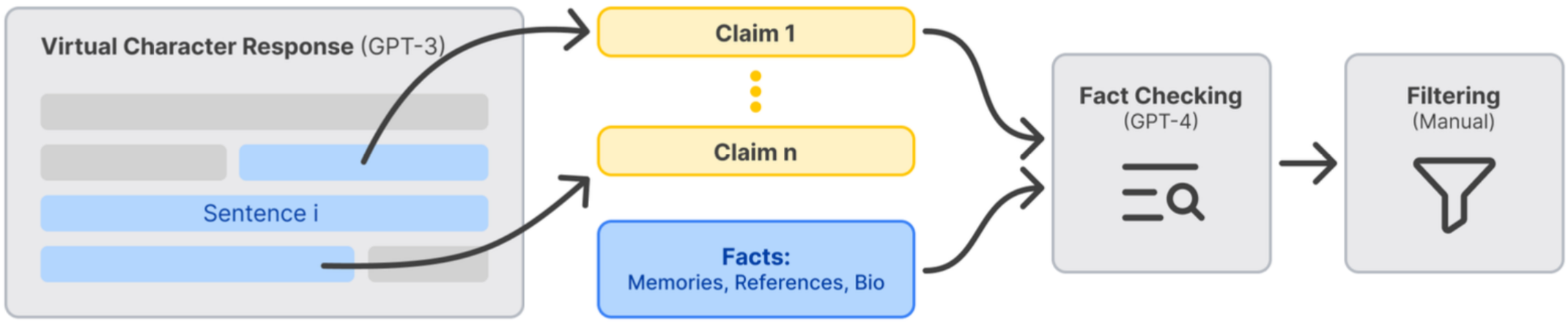


Figure 4: *The fact-checking pipeline:* In step 1, each sentence in a virtual character response is scanned for verifiable claims. In step 2, each extracted claim is fact-checked individually three times: once using all retrieved memories, once using only the memories referenced in the character response, and once using the character bio available in the prompt. Finally, in a manual filtering step, any unnecessary checks are discarded.

Evaluation Results

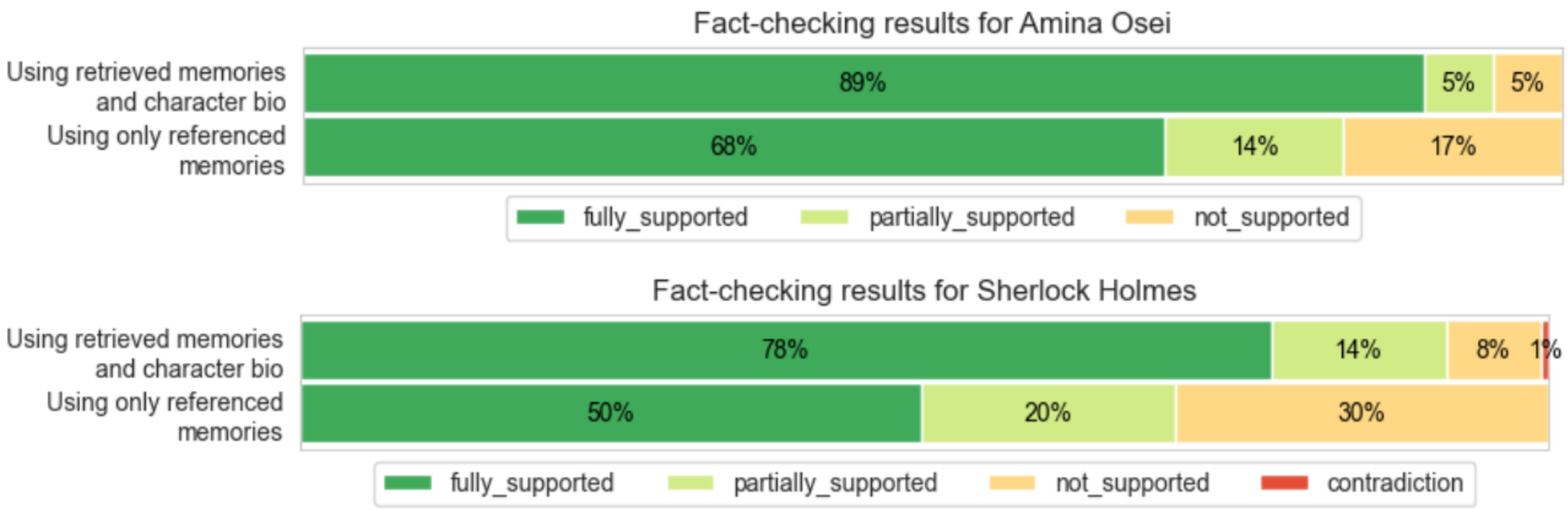
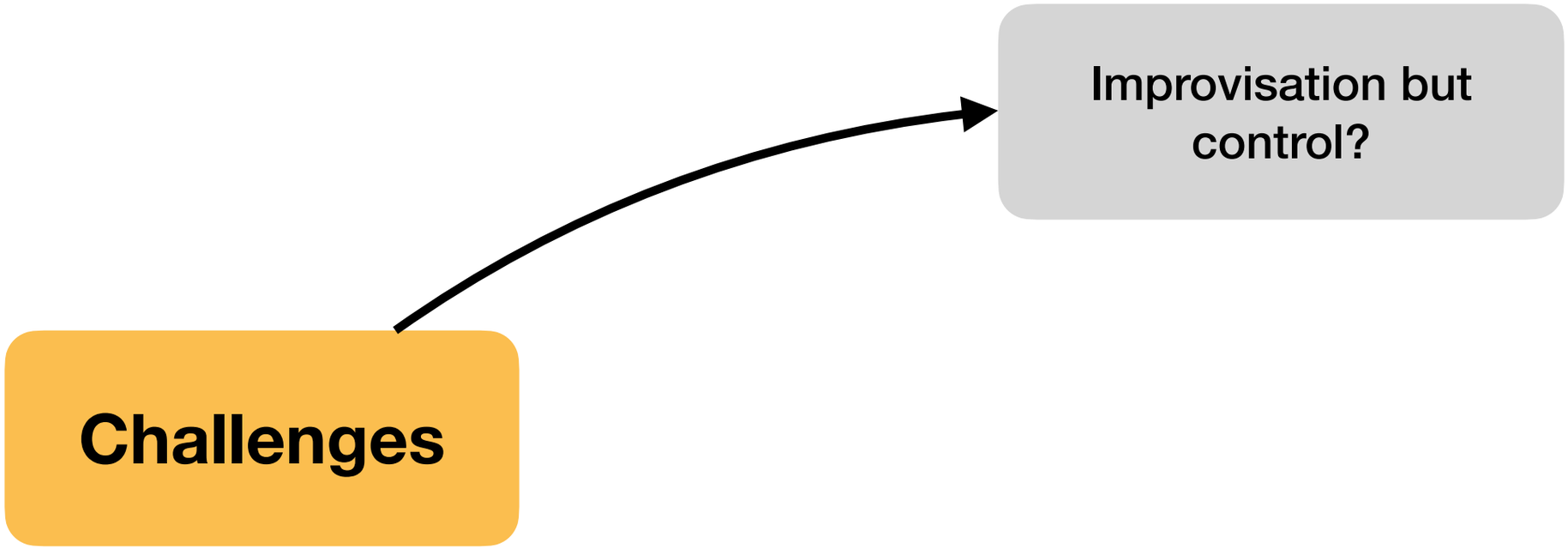
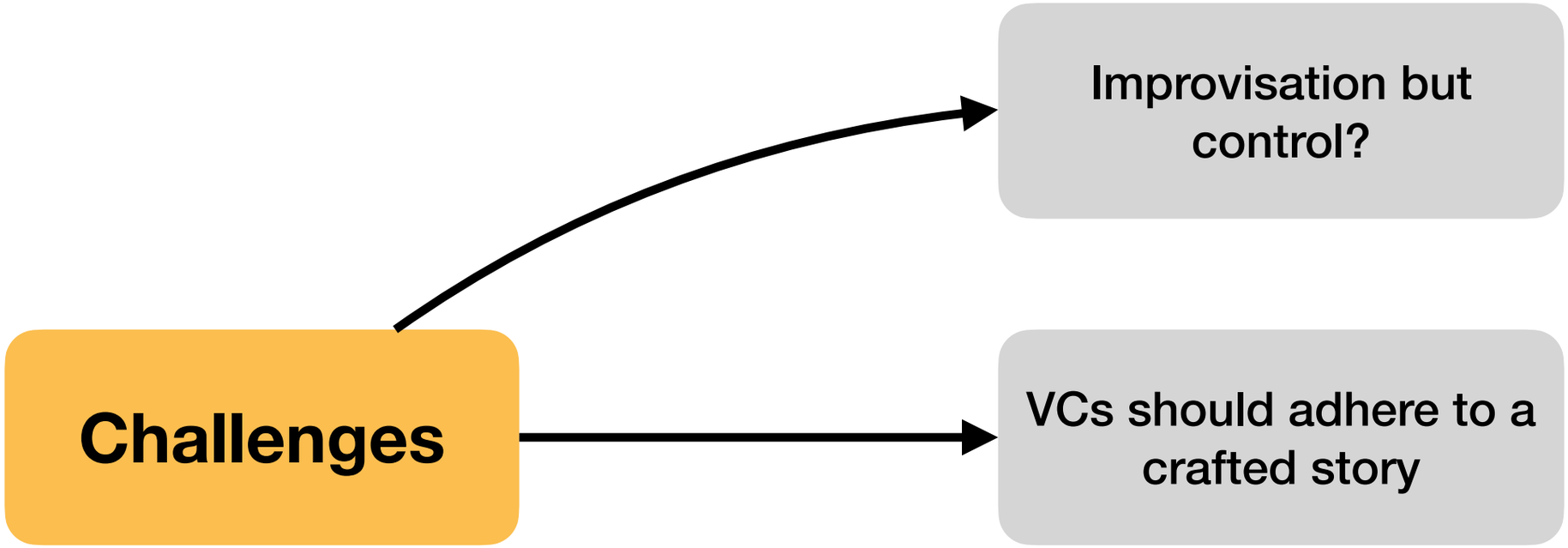
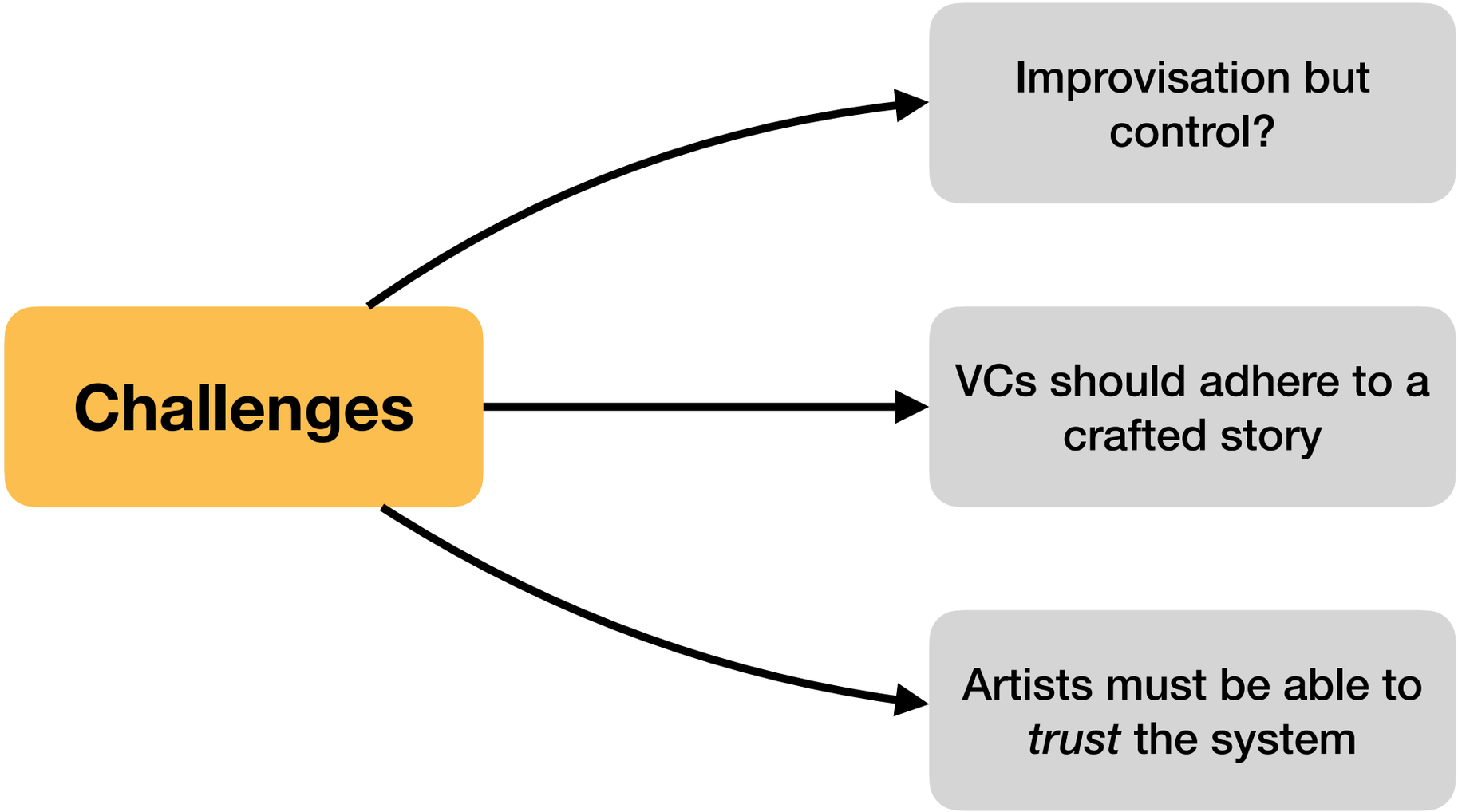


Figure 5: *Fact-checking results for the virtual character responses:* Each horizontal bar shows the results for different sources of truth as named on the left. The first category shows how grounded the character response is in the available information, using an aggregation of all three fact-checking results. The second category shows how good the LLM is at referencing which information it uses. Note that some percentage counts do not add up to 100 due to rounding.

Challenges


Improvisation but control?
Challenges


Challenges
Improvisation but control?
VCs should adhere to a crafted story


Challenges
Improvisation but control?
VCs should adhere to a crafted story
Artists must be able to *trust* the system

Forgetting Function

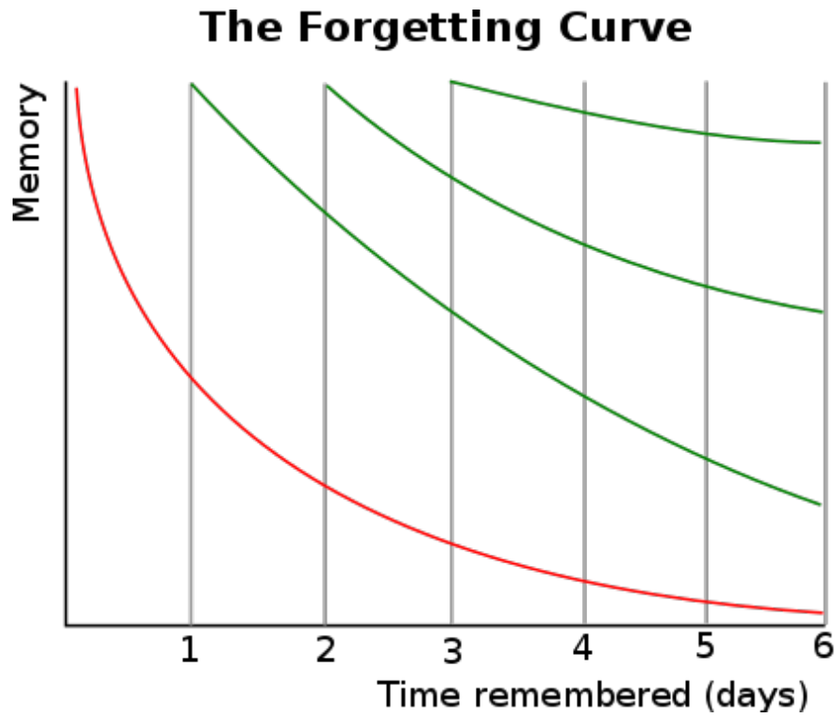
- Humans don’t remember everything forever — *virtual characters probably also shouldn't.*

- We used an adapted version of Ebbinghaus’ forgetting function

  - There is an exponential decay in memory strength over time.

  - Each memory access will increase the memory strength and make any subsequent forgetting slower.

- The forgetting function is customizable, making it possible to model different memory-capabilities for different characters.



The forgetting model is a modified version of Ebbinghaus’s forgetting curve (Ebbinghaus , 1885):

$$R = e^{-\frac{dt}{S}}$$

In this formula, $R \in (0, 1)$ is the memory retention, $t \in \mathbb{R}^+$ is defined as the elapsed time since the last access, $S \in \mathbb{R}^+$ is the memory stability, determining how “strong” a memory is. Further, $d \in \mathbb{R}^+$ is a decay constant that defines how forgetful a character is in general. To simulate learning through repetition, each time a memory is accessed, its stability S is updated by multiplying with a boost factor $b \in \mathbb{R}^+$. Thus, b determines how fast memories are strengthened through repetition, or, in other words, how fast a character can learn.

Insights About Working With LLMs

- Keep it simple. More rules in the prompt don’t result in better results.
- Model capabilities can be split into language and reasoning capabilities. Only big models are good at the latter (as of now).
- Separate control flow from content generation. A prompt is either for a decision or for content generation, but not for both at the same time.
- Use personas to enhance LLM capabilities.
- Control the output format by providing a Typescript interface or JSON schema.