LLM - RAG algorithm

Introduction

This document outlines the design choices behind a cutting-edge Large Language Model (LLM) utilizing a Retrieval-Augmented Generation (RAG) algorithm. The architecture is specifically tailored to enhance conversational AI applications, emphasizing persistent memory, contextual awareness, and efficient data compression. Key components include PersistentStore, HazyMemory, and the Conversational Memory Buffer.

PersistentStore

Purpose and Functionality

PersistentStore serves as the cornerstone of our architecture, designed to store and manage the hierarchical relationships and current states of entities known to the AI. This component uses XML format for data storage, chosen for its graph-like compression capabilities and flexibility in representing complex relational data.

XML Format Advantages

- **Flexibility**: XML's tag-based structure allows for the representation of diverse entities such as Users, Objects, Tasks, Persons, Places, and Animals with unique identifiers and timestamps.
- **Compression**: Leveraging XML enables LLM-driven compression, ensuring that important relationships and data are maintained efficiently without sacrificing accessibility.
- **Query Capability**: XML's structure facilitates querying for specific information, crucial for real-time interaction adjustments and internal monologue architectures in gaming and simulation environments.

HazyMemory

HazyMemory complements PersistentStore by maintaining a vector database of similarity search results. This component provides a broader context of past interactions, offering generalized conversational insights without the detailed temporal and relational information stored in PersistentStore. It enhances the AI's ability to draw on previous conversations to generate relevant and coherent responses.

Conversational Memory Buffer

The Conversational Memory Buffer temporarily holds the 50 most recent dialogues, ensuring smooth continuity in ongoing conversations. This short-term memory buffer plays a role in maintaining the relevance and coherence of the AI's responses.

Compression Strategy

Given the finite memory space, an LLM-driven XML compression strategy was implemented, prioritizing the retention of current state information while allowing older, less relevant data to be compressed more aggressively. This approach ensures the AI's memory remains efficient and responsive over time, with a particular emphasis on:

- **Time-Weighted-Memory-Decay**: Using the `last_updated_timestamp`, older memories are identified for compression or deletion, preserving the most recent and relevant information.
- **Selective Memory Retention**: Critical information, such as recent tasks or significant changes in entity states, is prioritized for retention, ensuring the AI remains up-to-date with the latest context.

Application in Gaming and Simulation Environments

The flexibility of the XML format and the strategic memory management approach make this architecture particularly suited for gaming and simulation environments. Here, the AI can dynamically interact with a finite number of entities, with the ability to query and update the game state based on inferred relationships and past interactions. This capability opens new avenues for creating immersive and responsive gameplay experiences.

Conclusion

The design choices behind the LLM-RAG algorithm reflect a commitment to enhancing conversational AI through innovative memory management, data storage, and retrieval strategies. By integrating PersistentStore, HazyMemory, and the Conversational Memory Buffer, the architecture sets a new standard for responsive, context-aware AI interactions across various applications.

HOW TO RUN THE CODE:

1) **Populate “secrets.txt” in the root directory with your openai_api_key, e.g**

   secrets.txt

```
sk-shu83hnfsEXAMPLE_KEYndskh7yw3h
```

2) To install all the dependencies enter the following command into the root directory of the project

```
pip install -r requirements.txt
```

3) To start the chat bot run the following command in the root directory

```
python3 main.py
```

   Or if python3 in unavailable type in,

```
python main.py
```

APPENDIX

1) **Dependencies and Libraries**
   - Primary Dependencies: **LangChain** - Provides easy interface for the defining chains in a declarative manner
2) **Demonstration of LLM Driven Compression**

   Example XML file before compressions

```
<Person:user time='04/30/2024 18:30:00'>
    is planning a vacation
    <Place:Italy time='04/30/2024 18:31:00'>next month</Place:Italy>
</Person:user>
<Person:user time='04/30/2024 17:15:00'>
    is considering joining
    <Task:pottery class time='04/30/2024 17:16:00'>a pottery
class</Task:pottery class>
</Person:user>
<Person:user time='04/30/2024 16:00:00'>
    has started
    <Task:diet plan time='04/30/2024 16:01:00'>a new diet plan focusing on
plant-based foods</Task:diet plan>
</Person:user>
<Person:user time='04/30/2024 15:45:00'>
    mentioned feeling more
    <Mood:energetic time='04/30/2024 15:46:00'>energetic
lately</Mood:energetic>
</Person:user>
<Person:user time='04/30/2024 15:30:00'>
    is thinking about adopting
    <Animal:pet time='04/30/2024 15:31:00'>another pet</Animal:pet>
</Person:user>
<Person:Ronny time='04/30/2024 14:30:00'>
    is helping a friend
    <Task:move time='04/30/2024 14:31:00'>move this weekend</Task:move>
    <Person:friend time='04/30/2024 14:32:00'>Drake</Person:friend>
</Person:Ronny>
<Person:user time='04/30/2024 13:45:00'>
    enjoyed
    <Object:book time='04/30/2024 13:46:00'>a new book on
meditation</Object:book>
</Person:user>
<Person:user time='04/30/2024 13:30:00'>
    is looking for recommendations for
    <Place:coffee place time='04/30/2024 13:31:00'>a good coffee
place</Place:coffee place>
</Person:user>
<Person:user time='04/30/2024 12:00:00'>
    had a successful meeting with colleagues about
    <Task:project time='04/30/2024 12:01:00'>a project</Task:project>
</Person:user>
<Person:user time='04/30/2024 11:30:00'>
```

```
    is experimenting with
    <Hobby:home gardening time='04/30/2024 11:31:00'>home
gardening</Hobby:home gardening>
</Person:user>
<Person:user time='04/30/2024 10:15:00'>
    feels optimistic about the changes they are making in their life
</Person:user>
<Person:Ronny time='04/30/2024 09:45:00'>
    tried a new recipe for
    <Meal:breakfast time='04/30/2024 09:46:00'>breakfast</Meal:breakfast>
</Person:Ronny>
<Person:user time='04/30/2024 09:00:00'>
    is enjoying the
    <Weather:sunny time='04/30/2024 09:01:00'>sunny weather
today</Weather:sunny>
</Person:user>
```

PersistentMemory file after first compression:

```
<Person:user time='04/30/2024 18:30:00'>
    <Place:Italy time='04/30/2024 18:31:00'>next month</Place:Italy>
    <Task:pottery class time='04/30/2024 17:16:00'>a pottery
class</Task:pottery class>
    <Task:diet plan time='04/30/2024 16:01:00'>a new diet plan focusing on
plant-based foods</Task:diet plan>
    <Mood:energetic time='04/30/2024 15:46:00'>energetic
lately</Mood:energetic>
    <Animal:pet time='04/30/2024 15:31:00'>another pet</Animal:pet>
    <Object:book time='04/30/2024 13:46:00'>a new book on
meditation</Object:book>
    <Place:coffee place time='04/30/2024 13:31:00'>a good coffee
place</Place:coffee place>
    <Task:project time='04/30/2024 12:01:00'>a project</Task:project>
    <Hobby:home gardening time='04/30/2024 11:31:00'>home
gardening</Hobby:home gardening>
    <Weather:sunny time='04/30/2024 09:01:00'>sunny weather
today</Weather:sunny>
</Person:user>
<Person:Ronny time='04/30/2024 14:30:00'>
    <Task:move time='04/30/2024 14:31:00'>move this weekend</Task:move>
    <Person:friend time='04/30/2024 14:32:00'>Drake</Person:friend>
    <Meal:breakfast time='04/30/2024 09:46:00'>breakfast</Meal:breakfast>
```

```
</Person:Ronny>
```

Example Persistent Memory File after the second compression

```
<Person:user time='04/30/2024 18:30:00'>
    <Place:Italy time='04/30/2024 18:31:00'>next month</Place:Italy>
    <Task:pottery class time='04/30/2024 17:16:00'>a pottery
class</Task:pottery class>
    <Task:diet plan time='04/30/2024 16:01:00'>a new diet plan focusing on
plant-based foods</Task:diet plan>
        <Animal:pet time='04/30/2024 15:31:00'>another pet</Animal:pet>
</Person:user>
<Person:Ronny time='04/30/2024 14:30:00'>
    <Task:move time='04/30/2024 14:31:00'>move this weekend</Task:move>
    <Person:friend time='04/30/2024 14:32:00'>Drake</Person:friend>
</Person:Ronny>
```