PersonaAI: Leveraging Retrieval-Augmented Generation and Personalized Context for AI-Driven Digital Avatars

Elvis Kimara, Kunle S. Oguntoye, Jian Sun
Iowa State University
Ames, IA, USA
{ekimara, oguntoye, jsun29}@iastate.edu

Abstract

This paper presents PersonaAI, an innovative application that uses Retrieval-Augmented Generation (RAG) and the LLAMA model to create personalized digital avatars capable of mimicking individual personalities. By capturing user data through a mobile application and storing it in a cloud database, PersonaAI can retrieve relevant contexts to answer user queries accurately. This proof-of-concept demonstrates the potential of AI in creating digital personas, providing insights into user-specific contexts, and advancing research in personalized AI. Key contributions include showcasing the efficacy of RAG with context, the novel approach to personal data collection, and the scalability of LLAMA2 with prompt engineering, offering a more efficient alternative to traditional, resource-intensive LLM training methods.

Keywords Retrieval-Augmented Generation · Personalized AI · Digital Avatars · LLAMA · Natural Language Processing

1 Introduction

Artificial intelligence (AI) has rapidly evolved, reshaping the way we interact with technology. While AI-powered virtual assistants handle routine tasks, they often lack the ability to personalize interactions based on individual traits and preferences. Our research aims to address this gap by developing PersonaAI, a system that creates AI personas mirroring individual personalities.

PersonaAI serves a diverse range of users, from those seeking to preserve the digital legacy of loved ones to professionals needing to provide consistent information even when unavailable. Traditional large language models (LLMs) like ChatGPT often require users to manually provide context, a process that can be cumbersome and repetitive. PersonaAI streamlines this by automatically capturing and utilizing user-specific data to generate more relevant and personalized responses.

In addition, in today's fast-paced world, AI interactions can feel impersonal and generic, failing to meet the unique needs of individual users. Our approach bridges this gap, offering a solution that adapts to individual characteristics and preferences, thereby enhancing the overall user experience. This paper explores the methodology, implementation, and potential applications of PersonaAI, highlighting its contributions to AI-driven digital personas. And by ensuring that the AI acknowledges its limitations and responds with "I don't know" when necessary, we aim to improve reliability and user trust, mitigating hallucinations.

2 Related Works

In recent years, there has been significant progress in the development of personalized AI systems. Notable works in this field include:

- **OpenAI's GPT Series**: The development of the GPT series by OpenAI, particularly GPT-3, has been instrumental in advancing natural language processing capabilities. While these models excel in generating human-like text, they often lack the ability to provide personalized responses without extensive user input.
- **Retrieval-Augmented Generation (RAG)**: The concept of RAG, as explored by researchers at Facebook AI, integrates retrieval mechanisms with generative models to enhance response relevance. This approach has shown promise in improving the accuracy of information retrieval in conversational AI systems.
- **Personalized AI Assistants**: Various studies have explored the use of AI to create personalized virtual assistants. For example, Google's Meena and Amazon's Alexa have made strides in understanding user preferences and providing tailored responses. However, these systems still face challenges in capturing and utilizing detailed user-specific contexts.
- **Memory Networks**: Research on memory networks, such as work by Weston et al., has contributed to the development of AI systems capable of storing and recalling user-specific information. These networks offer a foundation for creating AI personas that can remember and adapt to individual user interactions over time.
- **Ethical Considerations in AI**: The ethical implications of personalized AI have been a topic of considerable discussion. Works by Binns et al. and Jobin et al. have highlighted the importance of transparency, data privacy, and user consent in the development and deployment of AI systems that handle personal information.

By building on these foundational works, PersonaAI aims to advance the field of personalized AI by demonstrating the practical application of RAG and LLAMA in creating digital avatars that mimic individual personalities. Our contributions provide a pathway for future research and development in this exciting and rapidly evolving domain.

3 Major Contributions

- **Advancing AI Personas**: PersonaAI provides a proof of concept for AI systems that can mimic individual personalities, paving the way for future research and development.
- **Effective Use of RAG**: Demonstrates the effectiveness of RAG in retrieving relevant contexts and improving response accuracy.
- **Innovative Data Collection**: Introduces a novel approach to personal data collection through a mobile application, ensuring user privacy and data security.
- **Scalable AI Solution**: Highlights the scalability of open-source LLAMA2 with context and prompting, offering a more sustainable approach compared to traditional LLM training methods that are costly, and have high carbon footprint.

4 Experimental Platform

4.1 Equipment

The infrastructure needed for this project includes cloud storage, a GPU instance for running transformer models, a database for storing embedded vectors of users' data, and a backend server for user authentication and data ingestion. A mobile application was developed to capture user data through voice-to-text conversion. The app listens to user conversations at random intervals throughout the day and stores the transcriptions in a cloud database.

4.2 Algorithms and Data Structures

User data, whether textual or speech, is structured in a dictionary format with keys such as text (transcribed data), timestamp, username, and an embedded vector derived from an embedding model. When a user makes a query, the system ranks their data by similarity using the cosine similarity function. The top 2 to 5 ranked contexts are included in the prompt sent to the LLAMA model for generating responses.

4.3 Experiment

We chose to adopt RAG (Retrieval Augmented Generation) for synthesizing accurate results from the top-k dataset passed to the base model, LLAMA 2. We used Firebase for hosting and deployment of the database. Additionally, we added functionality that allows the AI model to remember contexts from previous conversations, improving performance.

We compared our results with those from ChatGPT with user data, baseline ChatGPT with no user data, LLAMA 2, and our own trained transformer model. This helped us decide on the best experimental path. We also compared LLAMA 13B and 70B parameters.

4.4 Deployment

For deployment, we experimented with separate deployments for the frontend and backend/AI model. The backend, built with Python and Flask, handles API communication and containerization with Docker. The frontend, developed in React TypeScript, handles authentication, storage, messaging, and other application requirements.

5 Methodology

The methodology adopted in this paper follows a similar style to that of retrieval augmentation generation (RAG) workflow. Users' data are indexed and embedded in a vector (size = 384) using the pre-trained Hugging Face BAAI/bge-small-en model. While larger BAAI generation models are known to have better performance with larger vector sizes up to 1024, we adopted a recursive character chunking strategy that chunks ingested data into n-chunks of the maximum size of 200 characters and 25 percent overlap, thereby optimizing the selected vector size for optimal results. The recursive character chunking strategy follows a recursive and text-splitting hierarchical order from a new paragraph to a new line, space, and unit character in fitting the chunk into a specified size. Figure 1 illustrates the adopted chunking strategy for a sample user's data using a chunk size of 200 and 0% overlap.

chunk size: 200 | average chunk size: 135 | overlap: 0%

```
I had a fantastic day working on my experiment in the lab
today. A day that started gloomily had a great turn when
my magical hands fortuitously mixed a complex reagent,
which became the turning point of my research. I
definitely couldn't have asked for a better moment.
```

Figure 1: Recursive character text chunking with 200 chunk size and 0 percent overlap.

Contrary to the demo in Figure 1, we adopted a 25% overlap strategy to enforce context continuity between fragments. Each chunk is an object with date, userID, and vector generated from the BAAI/bge-small-en model.

5.1 Query and Retrieval

The foundation of any successful RAG system is an efficient context retrieval framework. Numerous vector stores with inbuilt retrieval algorithms automate the indexing and chunk retrieval workflow. However, many of these are subscription-based and may not be required at this stage of our project. Hence, we resorted to a simplified dot product ranking approach in ordering the most relevant chunk closely matching the presented query. The textual query is first embedded using the same embedding model. In generating the top-k similar chunks to the query, the embedded query vector is multiplied by the user's vector database of size $N \times M$ (N is the total number of chunked data in the user database, and M is the vector length, 384). The top-k chunks from the resulting matrix product constitute the contextual information later passed to the large language model. While this simplistic approach works efficiently with small data sizes, the efficiency diminishes as the user database grows. Also, one may be tempted to increase the size of k with an extensive database. However, this could potentially cause information obesity or irrelevant additional information, leading to hallucination or factually incorrect generations from the language model.

5.2 Language Model and Prompt Engineering

As we've stated, the contextual data retrieved from the user's database must be fed to a language model for further analysis and final text generation. In this paper, we utilized the Q&A and text generation capabilities of the LLAMA 2-70B model in generating factual data and user-styled personified texts. Of the available variants of LLAMA models and other open-source models, the 70B parameter has been shown in the literature to be efficient in generating safe and helpful texts and outperforms popular models like ChatGPT-3.5 in generative tasks.

Having selected the LLM model and developed a retrieval workflow, another critical component is the LLM prompting. An LLM completion is only as good as its prompts. Hence, we carefully engineered different prompts for different use cases as follows:

5.2.1 USE CASE 1: New Users with an Empty Database

```
Respond to the QUESTION below:

If the QUESTION is a general greeting or an inquiry about personal welfare (e.g., ‘‘How are you?’’ or ‘‘Goo

If the QUESTION is too specific and lacks the necessary context or details for a comprehensive answer, l

If the QUESTION can be answered with general knowledge and the answer is known, provide a generalized, l

If you are unable to answer the QUESTION due to a lack of information, either from the context provided

QUESTION:
{question}

ANSWER:
```

5.2.2 USE CASE 2: Users with Available Contexts

```
Respond to the QUESTION below:

If the QUESTION is a general greeting or an inquiry about welfare (e.g., ‘‘How are you?’’ or ‘‘Good day’’),

If the QUESTION requires specific information from the CONTEXT (provided below) and the answer can be de

If the QUESTION pertains to general knowledge or topics not covered in the CONTEXT, such as current even

If the answer cannot be determined from the CONTEXT, is not within the general knowledge capabilities of

CONTEXT:
{context}

QUESTION:
{question}

ANSWER:
```

Leveraging LLM's ability to output a generalized response to questions without additional context, as shown in use case 1, enhances users' interaction and propensity to integrate their data for more personalized responses. Also, more importantly, in both use cases, we understand the fallibility of LLMs, hence the need to structure the prompts to assist the LLM in generating truthful texts, precluding hallucinations. Further evaluations are discussed in later sections.

5.3 Datasets

In this paper, the proposed framework is evaluated using two corpora: (1) SpongeBob dataset and (2) college student day-to-day journal. The SpongeBob data are the exact words SpongeBob enunciated in the movie series. One could see this dataset as the textual format of SpongeBob's persona regarding his linguistic style, interactions with people, personal admiration, and the possibility of more intimate information. While this dataset exemplifies SpongeBob's persona, the short responses and textual representation of feelings like "ouch, oh, uhmm.." contribute to the noise in the dataset. At this stage of the project, we have reserved the preprocessing of the dataset for future works. Understanding SpongeBob's dataset limitations, we carefully simulated the student's daily journal, ensuring a realistic and quality dataset is provided as context. The student daily journal comprises emails received or sent to friends and professors, essays, daily discoveries, lows and highs, epiphany moments, a summary of therapy sessions, doctor's recommendations, discussions about hobbies, etc.

These datasets helped provide a good proof of concept while starting to see if an LLM can actually mirror ones quirks while still accomplishing tasks required of a standard LLM. Due to the succes of it, we pushed on to build a Saas where each user has unique account with authentication and safety measures to ensure data privacy and security as their data is stored and used by the LLM.

6 Results

6.1 Use cases

We demonstrated the capabilities of the proposed framework using different test cases. By leveraging the speed and inference models provided by Replicate, we experimented with three powerful large language models - LLAMA-3.1-405B-Instruct, LLAMA-3-70B, and LLAMA-2-70B-Chat. The resulting completions from the models across different use cases are tabulated in the subsections.

6.1.1 Use case 1: General knowledge.

We evaluated the models' performance in answering questions without additional user context. As expected, the accuracy of the responses relies on the state of knowledge at the time the models were trained and how well the models recognize the possibility of outdated information. In the general knowledge context, totally random questions related to political position, cooking guides, and coding questions were selected as shown in Table 1. These questions are similar to prompts typically passed to ChatGPTs as they do not necessarily require users' preferential/historical data.

The results in Table 1 shows.

6.2 Feedback and Evaluation

- **User Feedback**: Users reported high satisfaction with the personalized responses generated by PersonaAI, indicating strong potential for real-world applications.
- **Performance Metrics**: The system achieved high accuracy in retrieving relevant contexts and generating appropriate responses.

7 Future Work

- **Enhanced Personalization**: Further research into improving response personalization by incorporating more diverse data sources and refining prompt engineering techniques.
- **Broader Applications**: Exploring the potential applications of PersonaAI in various domains, such as preserving the digital legacy of individuals, enhancing customer service, and providing personalized education and health recommendations.
- **Ethical Considerations**: Addressing ethical concerns related to data privacy, security, and the responsible use of AI in creating digital personas.

8 Conclusion

PersonaAI represents a significant step forward in the development of AI-driven digital avatars. By leveraging RAG and the LLAMA model, PersonaAI demonstrates the potential for creating personalized AI systems that can mimic human personalities and provide valuable insights based on user-specific contexts. The findings of this paper highlight the need for further research in this field and underscore the scalability and sustainability of using context and prompting in AI solutions.

Acknowledgments

This work was supported in part by [Funding Information].

References

Table 1: General Knowledge questions and completions

Questions	**LLAMA-2-70B-Chat**	**LLAMA-3-70B**	**LLAMA-3.1-405B-Instruct**
Who is the president of Nigeria?	I do not know. I'm just an AI, I don't have access to real-time information and my knowledge cutoff is August 2022. The current president of Nigeria is Muhammadu Buhari, but I'm not sure if this information is still accurate. It's important to note that information can change rapidly, and I may not always have the most up-to-date answers. If you have any further questions or need more information, please feel free to ask!	I DO NOT KNOW [+ hallucination]	As of my last update in 2021, the of Nigeria is Muhammadu Buh ever, please note that political pos change, and I might not have the rent information. For the mos and up-to-date answer, I recomm ing the latest news or official go sources.
I like to bake a cake, what do I need?	Sure, I'd be happy to help! To bake a cake, you'll need a few basic ingredients and some kitchen tools. Here's a list of what you'll need to get started: Ingredients: * Flour * Sugar * Eggs * Butter or oil * Milk * Vanilla extract (optional) Kitchen Tools: * Mixing bowls * Measuring cups and spoons * Electric mixer (optional) * Whisk * Rubber spatula * Baking pan * Oven You'll also need a recipe to follow, which you can find online or in a cookbook. Make sure to choose a recipe that's appropriate for the type of cake you want to make (e.g. chocolate, vanilla, red velvet, etc.) and that it's tailored to your skill level. I hope that helps! If you have any specific questions about a certain recipe or step in the process, feel free to ask.	[Hallucination]	Baking a cake sounds like a fun To get started, you'll need some b dients and equipment. Here's a g Ingredients: * Flour * Sugar * Eggs * Butte Milk * Baking powder or baki Salt * Flavorings (e.g., vanilla ex Equipment: * Mixing bowls * Measuring spoons * Electric mixer (optiona * Rubber spatula * Cake pan (gr floured) * Oven Of course, the specific ingred equipment may vary depending o of cake you want to make (e.g., vanilla, red velvet). If you have recipe in mind, I'd be happy to with that!
Yet another question	Data 8	Data 9	Data 10