Enhancing AI Personalization: A Multi-User Adaptive Language Model

Anonymous ACL submission

Abstract

Traditional Large Language Models (LLMs) are typically designed to interact with a single person and respond with a personalized answer tailored to that individual. This results in limited multi-user interaction, making it impractical for shared environments such as households and workplaces. In this paper, we introduce an Adaptive Friend Agent (AFA), a personalized LLM framework capable of identifying different individuals using voice recognition and providing personalized responses while preserving each person's conversation history. AFA integrates the capabilities of SpeechBrain, a voice recognition software to perform the identification and distinction between individuals, a Vector Database (VectorDB) to store personalized information as well as conversation history for each user interacting with the model, along with fine-tuned LLM that accesses these individual databases to generate personalized responses. Additionally, we introduce Personalized Agent chaT (PAT), a synthetically generated dataset containing dialogues between a personalized AI assistant and users, each with unique personality traits, across 12 everyday use cases where individuals interact with LLMs. The PAT dataset is used to fine-tune the LLM and later serves as ground truth for evaluating our fine-tuned model and other state-of-the-art LLMs. Experimental results demonstrate that our model outperforms existing models in user identification and personalized response generation, achieving highest accuracy, with a BLEU-1 score of 81.31% and ROUGE-1 score of 43.04%. The complete code and data available in anonyms repository Link

1 Introduction

Recent advancements using Large Language Modelagents have showcased their use of memory storage, enabling more accurate information retrieval during a conversation (Huang et al., 2024). An LLM usually supports chatbots as a backbone for its responses, and for the most part, it uses chat history to retrieve previous information about the user (Sánchez Cuadrado et al., 2024). While it could be effective in the short term, it has a lot of downsides that affect accurate information retrieval. For instance, if the conversation is extremely long, the LLM may become confused by a large amount of input data and lose significant context when processing a response to the user (Liu et al., 2024). Another downside is the mistakes made by the user earlier in the conversation. This would need a self-editing memory to rectify. An LLM in chatbots can't do this efficiently (Dam et al., 2024). Lastly, these LLMs are designed to converse with only one user. It does not take into consideration more than one distinct individual, making information retrieval even more problematic when dealing with different users. To tackle these challenges, we use the agent's memory to store information about the distinct user and speech recognition software to differentiate between users. Furthermore, we generate a synthetic dataset that simulates a dialogue between personalized agents and various users. This open-source dataset, Personalized Agent chaT (PAT), was generated and used to instruct-tune Meta's LLAMA 3.2 70B model to have better-personalized responses to each different user. The results were evaluated, showcasing this model's robustness and dynamic usability. It was also compared to the generated answers of baseline LLM models, proving the effectiveness and novelty of our model.

Our contributions are as follows:

- Develop a personalized LLM framework as Adaptive Friend Agent (AFA), capable of identifying and differentiating distinct users, accommodating their answers, and making it personalized based on the user's persona and identity.
- Creation of PAT dataset that contains a dia-

logue between AI and a Human. The AI responses were personalized based on the user's persona's description. The synthetic generation process and various topics and scenarios that might occur in a user's real-life situations were created using Meta's LLAMA 405B.

2 Related Work

Large Language Models & Chatbots LLMs, such as ChatGPT, have reshaped natural language understanding and generation. They contain human-like text generation capabilities, contextual awareness, and robust problem-solving skills (Nazir and Wang, 2023). This breakthrough has expanded the chatbot's applicability across domains from healthcare to education and customer service. For instance, (Wei et al., 2024) has explored the use of LLM-powered chatbots to collect user self-reported data in health-related contexts. They found that the LLMs could dynamically adapt to user inputs and maintain conversational relevance without extensive training data, showing the potential for zero-shot prompt engineering to build a task-specific chatbot efficiently. Similarly, (Jeon et al., 2024) reviewed the role of LLM-lowered speech recognition chatbots in language learning, demonstrating their effectiveness in enhancing speaking and listening skills through conversational situations. This research shows the educational potential of chatbots. Highlighting the significance of this work, we propose a method that not only deepens this concept but also broadens it to include educational chatbots in multi-user settings.

Agent-Memory Traditional personal assistants lack capabilities like user intent understanding, task planning, tool use, and personal data management, which causes them to have limited practicality and scalability. On the other hand, LLMs like ChatGPT, Claude, and others can exhibit unique capabilities such as instruction following, commonsense reasoning, and zero-shot generalization (Hadi et al., 2023; Naveed et al., 2023). These attributes allow LLM-based personal agents to handle complex tasks efficiently. Personal LLM agents have to frequently retrieve information from external memory to enable more informed decisions. There are a lot of different forms of external memory, such as user profiles and interaction history.

Some other research has explored solving problems like limited context windows, document analysis, and self-editing memory. (Packer et al., 2023) introduces an operating system-inspired approach to overcome these limitations by managing virtual memory for LLMs. This architecture enhances LLM applications in domains requiring long-term memory or large-scale document analysis. These works inspire us to use and build upon them, not just as an agent memory for one person but as a unique memory for each individual.

The common practice is to use embedding models to represent memory data with a uniform and high-dimensional vector format (Mikolov, 2013; Le and Mikolov, 2014). The distance between these vectors represents the semantic similarity between the correlated data. For every query, the LLM agent will find the most relevant content in the external memory storage. That retrieved content will be given to the personal LLM agent through either a prompt concatenation or an intermediate layer cross-attention (Zhang et al., 2024b), enabling more advanced semantic search and retrieval augmented generation (RAG) pipelines (Lewis et al., 2020). The combination of vector embeddings and traditional symbolic search methods is being actively researched to optimize retrieval accuracy while maintaining computational efficiency. We propose a method that allows each user's vector database to be retrieved upon being identified, and this database will store the user's personalized information.

LLM Personas There has been a lot of exploration in the domain of personalizing LLMs to be accommodated as a companion for humans, with frameworks like AUTOPAL introducing hierarchical adaptation to enhance user-agent interactions by tailoring personas based on user context and preferences (Cheng et al., 2024). This shows the potential for more personalized and effective emotional support systems in real-world applications. Furthermore, the increasing demand for personalized applications motivated the creation of Conversational Agents (CAs) to have distinct personas. (Sun et al., 2024) examines the reasoning and effects of conversational agents with unique personas. These domains were built further by developing our proposed method to personalize each response to each different individual.

Speaker Identification Speaker identification refers to determining who is speaking from a group of known voices. Voice embeddings are numerical representations of a person's unique vocal characteristics, allowing AI systems to distinguish

one speaker from another. A notable tool, named SpeechBrain, is an open-source speech processing toolkit supporting speaker recognition with state-of-the-art performance. Built on PyTorch, it offers pre-trained models for speaker verification. This tool was used for our method to identify the different individuals in the environment.

3 Dataset

We introduce the Personalized Agent chaT (PAT) dataset, which contains the query response pairs tailored to different personalities. This enhances LLM's by enabling them to generate more contextually relevant responses when fine-tuned.

3.1 Data Generation

To develop PAT, we used the Multi-Session Chat (MSC) dataset (Xu, 2021) as a baseline. The MSC dataset consists of human-human conversational data, where participants engaged in discussions, progressively learning about each other's interests and discuss what they learned in the past. The dataset features a diverse range of personalities, making it well-suited for our research needs.

3.2 Persona Extraction

We extracted various personality traits from the MSC dataset using GPT-4. We categorized them into distinct attributes listed below. Each category provides specific insights into an individual's characteristics, allowing a more personalized and adaptive conversational experience.

1. **Demographics** This category captures key details such as nationality, name, age, gender, and language preference. This gives a foundation to understand the person's overall nature and background.

2. **Career Information** This category gives information about educational and career background, helping to assess the user's expertise and domain knowledge.

3. **Motivations and Values** This category explores how an individual thinks, feels, and acts to offer insights into their motivations, belief systems, and personal values. This helps the agents tailor responses that align with users' perspectives and interests.

4. **Decision-Making Style** This category gives information about how an individual arrives at a decision through logical reasoning or emotional intuition. This helps the agent to generate a response that aids the user to make meaningful decisions.

5. **Preferences** This category captures an individual's preferred communication styles, content formats, likes, and dislikes. This enables LLMs to give responses, which enhances users' comfort and interactivity, ideally to create a more engaging conversational experience.

6. **Emotional Triggers** This category identifies different emotions and sensitivities that may influence an individual's behavior. Understanding these triggers enables LLMs to avoid responses that may cause discomfort and ensures emotionally intelligent interactions.

The personality traits extracted from MSC data are presented in fragments. To create cohesive persona descriptions, we used GPT-4 to organize the information effectively. This process ensures that all categories are seamlessly integrated, which resulted in a comprehensive and structured representation without any disconnections.

3.3 Personalized Query Generation

We then integrated question-response pairs that mimic real-life user interactions with LLMs. These responses were tailored to individuals' personality traits, to ensure personalized and contextually relevant interactions. We have used 12 common use cases where individuals interact with LLMs in daily life, including Shopping Assistance, Content Creation, Relationship Advice, Family Assistance, Project Planning, Language Learning, Story Development, Hobbies Assistance, Personal Development, Emotional Support, and Travel Planning. A detailed explanation of each scenario is provided in Appendix A.

3.3.1 Question Generation

Using these 12 scenarios, we developed persona-driven question types, to ensure they reflect users' personality traits. These questions were carefully crafted to reflect the individual's behavioral tendencies, making interaction more natural and adaptive.

To generate these questions, we used the Llama 405B parameter model and designed prompts that incorporate both persona description and specific

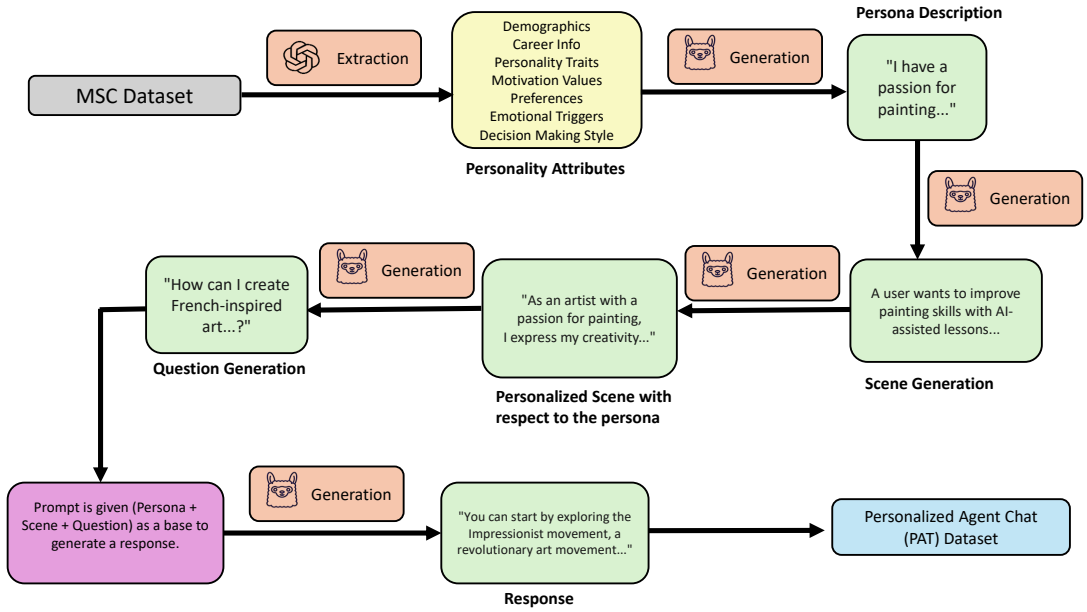

Figure 1: This is the data generation process, where MSC stands for Multi-Session Chat. ChatGPT-4 was used for data extraction, while LLAMA 405B was used for data generation.

use-case scenario. We generated 40 unique questions for each scenario tailored to an individual persona. Additionally, we structured the questions to be contextually linked, to ensure each successive query built upon the previous one, making it real-world conversations. In total, we developed 61,000 questions that spans all personalities extracted from the MSC dataset.

3.3.2 Response Generation

To generate the response, we again utilized the Llama 405B parameter model, ensuring responses were structured, context-aware, and aligned with the user's persona.

We integrated the Llama model with a continuously evolving database, which stores persona descriptions, scene contexts, corresponding questions, and aforementioned question-response pairs. This database dynamically updates as new responses are generated to warrant conversations remain continuous.

Each generated response is fed back into the database, this allows the system to build on previous interactions, when it generates new responses. This iterative approach helps the model maintain conversation continuity and produce responses that align with the user's persona and interaction history.

The prompt is structured to generate responses by including persona description, scene context, current question, and previous question-response pairs. This approach ensures that responses are personalized, contextually relevant, and adaptive to evolving conversation. The complete prompt structure used for generating the question-response pair is provided in Appendix B.

By integrating all the above steps, we have created a dataset consisting of approximately 61,000 question-response pairs representing diverse personalities in real-world contexts. We used this dataset to develop an Adaptive Friend Agent (AFA) model designed to generate highly personalized, context-aware responses. A detailed explanation of the model structure and algorithm is provided in the next section.

Our proposed Adaptive Friend Agent (AFA) model is designed to generate responses by understanding and adapting to the unique personality traits of each user. The model takes audio data as input. It is structured into four components: an Audio Identifier Module, a Dynamic User Profile Store, a Persona Synchronizer, and an Adaptive Response Generator. Each component ensures that the responses are tailored, context-aware, and aligned with the user's personality and interaction history.

3.4 Audio Identifier Module

The audio data passes through the Audio Identifier Module, which differentiates users based on their unique voice embedding. We used SpeechBrain (Ravanelli et al., 2024), a pre-trained and multi-purpose model that supports speech recognition,

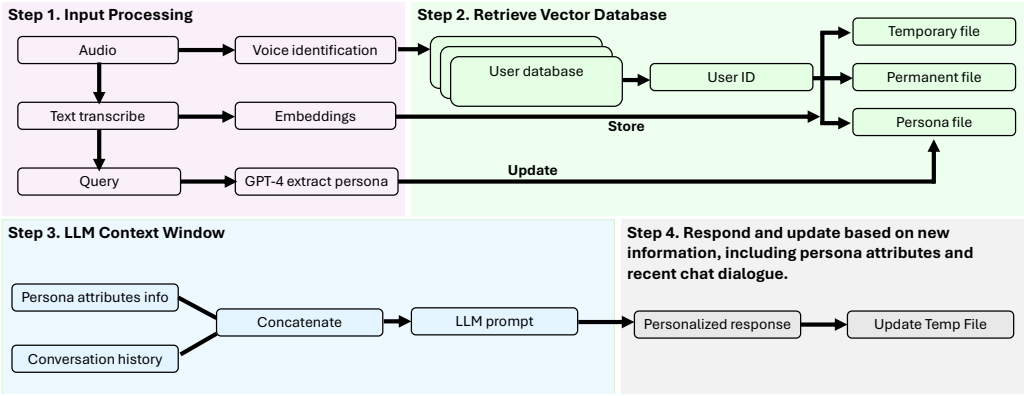

Figure 2: A summary of our framework, which enables a multi-user AI assistant by processing inputs, it identifies voices, extracts personas, and stores user data in a vector database to generate personalized responses and update user information for adaptive interaction.

speaker identification, text-to-speech, speech-to-text, and speech translation.

The process begins by converting the audio into voice embedding and then transcribes it into text with the SpeechBrain model. These embeddings are stored in a DynamoDB, where each embedding is linked with a corresponding user ID. When a user speaks, their voice embeddings are compared against existing embeddings in the database, using cosine similarity. If a match is found, the system retrieves the associated user ID, otherwise a new ID is assigned.

3.5 Dynamic User Profile Store

The Dynamic User Profile Store manages and stores historical conversations across multiple sessions of conversations of different users in the structured database. To effectively handle the data, we implemented two types of tables, temporary and permanent, for each user.

The temporary table stores the last ten conversations between the user and Agent, which enables the system to retrieve the most recent conversational history for context-aware responses. Once the user completes the ten conversations, a summary of the entire session is generated and stored in a permanent table. Then, the temporary table is cleared. The conversion of the conversation into a summarized version helps optimize the storage efficiency, reducing overall data requirements.

To generate these summarized records, we used GPT-4, which extracts meaningful information from the temporary table while preserving the context and intent of the conversation. The summarized version is stored in the permanent table. This mechanism allows the system to seamlessly recall past interactions, which ensures coherent and context-aware responses while maintaining efficient data management.

3.6 Persona Synchronizer

We have developed a dynamic persona extraction system that continuously analyzes user interactions to refine and enhance its understanding of individuals' personality traits. This is a continuous and iterative process where the system extracts and updates various persona attributes, as discussed in Section 3.2.

We used GPT-4o to extract personality traits from user queries to achieve this. If the user has a persona profile in the database, the system updates it dynamically by integrating newly extracted traits with existing persona data. This helps the systems understand a person's evolving characteristics and tailor the response to align with the user's preferences and behavior.

3.7 Adaptive Response Generator

An Adaptive Response Generator is backed by an LLM, which is central in generating responses to the users' queries. The module seamlessly integrates multiple components to ensure that responses are personalized, contextually relevant, and adaptive to users evolving interactions.

The text format of the audio serves as the query data, which is then processed by both the dynamic user profile store and the persona synchronizer. To retrieve the relevant information from the user profile database, query text embeddings are compared against the stored database embeddings, extracting the most relevant historical data.

Combined with user persona and current query,

extracted historical data is structured in a contextual and concise prompt. This ensures the generated response aligns with the user's personality and motivations.

4 Experiments

We experimented with different LLM base models to construct the adaptive response generator module and assess the overall framework's performance. These base models are categorized into two types: open-source and close-source models, each offering distinct advantages in adaptability, and customization of response.

For the open-source models, we have used the Llama 70B model, fine-tuned on 80% of the PAT dataset to enhance its ability to generate persona-aligned responses, this model was selected because of its customization potential for specific domain (Roziere et al., 2023). For the closed-source models, we used GPT-4, GPT-3.5, Claude, and Gemini-2.0, using the zero-short learning technique to evaluate their generalization capabilities.

Additionally, we explored different persona settings to understand their impact on response generation. In no-persona setting, the system generates a response solely based on dialogue history without considering persona information. In the constant-persona setting, we have introduced a fixed persona, in order to maintain a consistent style throughout the interactions. Finally, we evaluated the adaptive persona module, where the persona dynamically evolved based on user interactions, this setting helps in continuously refining the persona as new conversations are received.

4.1 Implementation

In our framework, the historical information from the user's personal database is enhanced by integrating it with the text embeddings. We used OpenAI's text embedding model to generate these embeddings, to use its ability to capture semantic relationships within the text. To retrieve relevant information, we applied cosine similarity and selected the top 3 most relevant information to pass to the LLM model to give it contextual information. We also experimented with retrieving the top 5 and top 8 most relevant information, but we didn't observe any significant improvement in response generation, therefore we selected the top three as the optimal setting.

For fine-tuning the open-source Llama 70B model, we have used the Low-Rank Adaption (LoRA) instruct fine-tuning technique (Hu et al., 2021) to improve computational efficiency by preserving instruction-following capability. We used LoRA rank: 8, alpha value: 16, dropout: 0.05, and number of epochs: 1 as different hyper-parameters. This fine-tuning approach makes it easier to adapt to large models efficiently without having to retrain completely. The entire process was conducted using SageMaker JumpStart.

We used multiple metrics to analyze the generated responses from different perspectives for evaluation. To measure the similarity between the generated response and ground truth responses in the PAT dataset, we used Bilingual Evaluation Understudy (BLEU) scores (BLEU-1, BLEU-2, BLEU-3, BLEU-4) (Papineni et al., 2002), and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores (ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Su) (Lin, 2004).

BLEU is a precision based metric used to evaluate how similar the generated text is to reference text by comparing the overlap of n-grams. BLEU -1/2/3/4, refers to the uni-gram, bi-gram, tri-gram and four-gram evaluations (Wieting et al., 2019). BLEU scores focuses on precision but not on recall or semantic variations (Reiter, 2018). Where as, ROUGE is used to measure similarity by analyzing overlapping n-gram, longest common subsequence (LCS) and skip bi-gram matches (Zhang et al., 2024a). ROUGE -1/2 captures uni-gram and bi-gram overlaps, ROUGE-L evaluates sentences by considering LCS, and ROUGE-Su captures non-consecutive word pairs (Barbella and Tortora, 2022). When both metrics are compiled, they provide more balanced evaluations .

We also measured Distinct-1 values to assess the diversity of generated responses. It measures the unique uni-gram in generated text.

Additionally, we evaluated persona-driven response using Profile-Level and Attribute-Level coverage (P/A-cover) (Lin et al., 2020; Cheng et al., 2024). These metrics help in analyzing how the well-generated responses are aligned with user's persona.

To compute A-cover, we defined the persona of a user as p, it has n number of attributes $\{a_1, a_2, \ldots, a_l\}$. Considering y as the generated response, the attribute level coverage is determined as:

$$A\text{-}Cover(y, p) = \max_{a_j \in p}(\text{IDF-O}(y, a_j)) \quad (1)$$

where IDF-O represents the Inverse Document Frequency weighted word overlap between the attributes in persona and generated response.

To measure the Profile-level coverage, we combine a set of generated responses denoted by S_r, which is calculated as:

$$P\text{-}Cover(S_r, p) = \text{IDF-O}(S_r, p) \quad (2)$$

where IDF-O measures the word overlap between the combined set of responses and all attributes in the persona. This evaluation helps to ensure the generated response is completely aligned with multiple persona attributes of the user.

In the next section, we evaluate the performance of our approach across multiple LLM base models with different persona settings, analyzing their impact on generated response accuracy, diversity, and personalization.

5 Results

Table 4.1 presents various performance metrics, including Natural Language Generation (NLG) metrics (BLEU, ROUGE), diversity, and personalization (P-cover, A-cover), across different experiment settings.

Our findings show that integrating persona information improved the BLEU and ROUGE scores across all LLM baseline models. The Llama 70B model, when fine-tuned with the PAT dataset, achieved the highest BLEU-1 and ROUGE-1 scores of 0.8131 and 0.4304, respectively, under the adaptive persona setting. Similarly, GPT-4 exhibited 6.7% increase in BLEU-1 compared to the no-persona setting, this demonstrated that persona-aware adaptation enhances response context. However smaller scale models like Llama 70B trained on the PAT dataset, generate more persona-aware responses, benefiting from structured persona-query-response pairs.

To evaluate response diversity, we measured the Distinct-1 metric, which helps capture unique words. We observed Claude-based model without persona integration had a higher distinct-1 value of 0.81 compared to persona update setting where the value dropped to 0.792, which indicates that the response is less diverse. However, when compared with Llama-70B model, it shows the balance between the diversity and persona consistent, with a diversity score of 0.894, while achieving strong persona consistency.

To evaluate the response adaption to user's personas, we measured P-cover and A-cover scores. From the results, we observed the AFA model without integration with the persona showed lower scores when compared with the AFA framework integrated with the dynamic persona adaption, or constant persona. The fine-tuned Llama 70B model achieved the highest persona alignment scores with a P-cover of 0.4594 and an A-cover of 0.412, surpassing all other models and settings.

The PAT dataset played an important role in enabling the lower parameter model, Llama 70B, to achieve persona adaption comparable to large parameter models. PAT dataset provided structured, diverse persona interaction data, which allows the model to learn to respond based on different personality traits. These results highlight the importance of large persona-driven datasets in making LLMs more user-aware and enable them to generate responses relevant and personalized to individual users.

6 Discussion

Our results confirm that integrating the adaptive persona module enhances the AFA framework's ability to generate response that are more aligned with the user's persona. The fine-tuned Llama-70B model outperformed zero-short large scale models, which demonstrated the effectiveness of PAT's structured persona-query-response training. This highlights the importance of fine-tuning on persona-adaption datasets for long-term adaption rather than relying on zero-short methods.

One key observation from our results was the impact of persona adaption on response diversity. Integrating the persona adaptive module improved alignment with user's persona but results in reduction in linguistic variation as seen in Claude's drop in Distinct-1 score. However, Llama 70B maintained both diversity and personalization indicating that fine-tuning enables a more balances response generation.

AFA framework, significantly improved the memory efficiency by integrating temporary and permanent memory storage tables, which allowed to retain user-specific historical information. To ensure scalability, we utilized DynamoDB store historical information of the users, which provides efficient storage capability.

Table 1: Performance Comparison of Different Language Models Across Persona Settings Using NLG Metrics, ROUGE Scores, Diversity, and Personalization Measures

Model	Persona	NLG Metrics (BLEU)				ROUGE Scores				Diversity	Personalization	
		BL-1	BL-2	BL-3	BL-4	RG-1	RG-2	RG-L	RG-Su	Distinct-1	P-Cover	A-Cover
Claude	w/o Persona	0.7690	0.6870	0.5842	0.4934	0.3900	0.1020	0.2286	0.1391	0.810	0.3940	0.3040
	With Persona Update	0.7430	0.6677	0.5695	0.4812	0.4012	0.1046	0.2355	0.1501	0.7923	0.4241	0.4100
	With Constant Persona	0.7635	0.6810	0.5781	0.4868	0.3738	0.1030	0.2279	0.1312	0.8173	0.4080	0.3630
Gemini	w/o Persona	0.5279	0.4745	0.4060	0.3478	0.2733	0.0838	0.1750	0.0720	0.8630	0.3521	0.3921
	With Persona Update	0.6081	0.5451	0.4691	0.4046	0.2987	0.0976	0.2003	0.0869	0.8510	0.3704	0.3520
	With Constant Persona	0.5738	0.5137	0.4384	0.3761	0.2769	0.0930	0.1840	0.0750	0.8540	0.3306	0.2919
GPT-3.5	w/o Persona	0.5459	0.4889	0.4164	0.3550	0.2420	0.0737	0.1593	0.0570	0.8810	0.3330	0.3045
	With Persona Update	0.5900	0.5270	0.4480	0.3883	0.2562	0.0746	0.1681	0.0639	0.8667	0.3525	0.3628
	With Constant Persona	0.4966	0.4434	0.3771	0.3216	0.2328	0.0682	0.1534	0.0530	0.8900	0.3013	0.3793
GPT-4	w/o Persona	0.7480	0.6626	0.5623	0.4784	0.3256	0.0969	0.2072	0.1012	0.8473	0.3010	0.2940
	With Persona Update	0.7984	0.7047	0.5946	0.5018	0.3351	0.0967	0.2167	0.1089	0.8313	0.3737	0.3058
	With Constant Persona	0.7770	0.6861	0.3806	0.4918	0.3240	0.0990	0.2125	0.1014	0.8387	0.3050	0.2840
Llama-70B	w/o Persona	0.782	0.7134	**0.5952**	0.4952	0.4153	0.110	0.2294	0.1542	**0.8942**	0.4198	0.3945
	With Persona Update	**0.8131**	**0.7293**	0.5934	**0.5253**	**0.4304**	**0.113**	**0.2453**	**0.1692**	0.8842	**0.4594**	**0.4152**
	With Constant Persona	0.7971	0.7263	0.5847	0.5143	0.4292	0.113	0.2353	0.1692	0.8834	0.4235	0.4098

7 Conclusion

In this paper, we introduced the Adaptive Friend Agent (AFA), a personalized LLM framework that can identify and distinguish users based on voice recognition and generate responses tailored to users personas. Our framework integrates SpeechBrain, DynamoDB, and a fine-tuned Llama 70B model to generate response across multiple users based on their personality traits. To achieve high quality AFA, we developed new Personalized Agent chaT (PAT) dataset, which contains approximately 64k conversations aligned across 12 real world scenarios. This well structured dataset helps model to generate responses that are interactive, contextually coherent and personalized based on users unique interest and personality traits.

8 Limitations and Future Work

While our model demonstrates promising results, but it does not support real-time interaction. The model process the audio files as input rather than live conversation. In future work, we aim to address this limitation by developing a deployable system for real time use cases. Additionally, we evaluated the performance within limited number of users, but its scalability is unknown. To support scalability, we have integrated DynamoDB database for storing historical information, but the audio recognition module needed to be updated for making it a scalable system.

Future work, we are planning on implementing the AFA model, to healthcare domain where this model will be serving like a caregiver, assisting patient to improve there living style by understanding their unique characteristics and behavioral pattern.

9 Ethical considerations

Our model collects and stores user-specific information such as persona attributes and conversation history, to enhance personalization, this raises concerns about privacy, data security, informed consent, and bias. It may accidentally collect and store sensitive personal information, such as passwords or financial details, if users mention them during interaction, which poses potential security risks. Users must be clearly informed about data collection through explicit opt-in consent. This emphasizes the need for research in ethical AI behavior and developing methods to prevent the storage of sensitive information.

References

Marcello Barbella and Genoveffa Tortora. 2022. Rouge metric evaluation for text summarization techniques. *Available at SSRN 4120317*.

Yi Cheng, Wenge Liu, Kaishuai Xu, Wenjun Hou, Yi Ouyang, Chak Tou Leong, Xian Wu, and Yefeng Zheng. 2024. Autopal: Autonomous adaptation to users for personal ai companionship. *Preprint*, arXiv:2406.13960.

Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. 2024. A complete survey on llm-based ai chatbots. *arXiv preprint arXiv:2406.16937*.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. Large language models: a comprehensive

survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*.

Jaeho Jeon, Seongyong Lee, and Seongyune Choi. 2024. A systematic review of research on speech-recognition chatbots for language learning: Implications for future directions in the era of large language models. *Interactive Learning Environments*, 32(8):4613–4631.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2020. Xpersona: Evaluating multilingual personalized chatbot. *arXiv preprint arXiv:2003.07568*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Tomas Mikolov. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Anam Nazir and Ze Wang. 2023. A comprehensive survey of chatgpt: advancements, applications, prospects, and challenges. *Meta-radiology*, page 100022.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. 2023. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra, Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, Sung-Lin Yeh, Pierre Champion, Aku Rouhe, Rudolf Braun, Florian Mai, Juan Zuluaga-Gomez, Seyed Mahed Mousavi, Andreas Nautsch, Xuechen Liu, Sangeet Sagar, Jarod Duret, Salima Mdhaffar, Gaelle Laperriere, Mickael Rouvier, Renato De Mori, and Yannick Esteve. 2024. Open-source conversational ai with SpeechBrain 1.0. *Preprint*, arXiv:2407.00463.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Jesús Sánchez Cuadrado, Sara Pérez-Soler, Esther Guerra, and Juan De Lara. 2024. Automating the development of task-oriented llm-based chatbots. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, pages 1–10.

Guangzhi Sun, Xiao Zhan, and Jose Such. 2024. Building better ai agents: A provocation on the utilisation of persona in llm-based conversational agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, pages 1–6.

Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2024. Leveraging large language models to power chatbots for collecting user self-reported data. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–35.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond bleu: training neural machine translation with semantic similarity. *arXiv preprint arXiv:1909.06694*.

J Xu. 2021. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*.

Ming Zhang, Chengzhang Li, Meilin Wan, Xuejun Zhang, and Qingwei Zhao. 2024a. Rouge-sem: Better evaluation of summarization using rouge combined with semantics. *Expert Systems with Applications*, 237:121364.

Zhihao Zhang, Alan Zhu, Lijie Yang, Yihua Xu, Lanting Li, Phitchaya Mangpo Phothilimthana, and Zhihao Jia. 2024b. Accelerating retrieval-augmented language model serving with speculation. *arXiv preprint arXiv:2401.14021*.

A Appendix

This appendix outlines twelve distinct real-world scenarios we used to generate the data.

- **Project Planning**: Assists users in organizing tasks, breaking down projects into manageable steps, tracking progress, and accessing resources for successful execution.
- **Language Learning**: Provides AI-assisted lessons, interactive practice sessions, and cultural insights to enhance language proficiency.
- **Job Interview Preparation**: Offers mock interviews, tailored feedback, and company-specific question preparation to improve candidates' performance.
- **Story Development**: Supports brainstorming, story outlining, and draft refinement for creative writing projects, including novels and screenplays.
- **Hobby Assistance**: Guides users to maintain and improve hobbies such as fitness, gardening, and painting, with personalized tips and resources.
- **Personal Development**: Focuses on self-improvement by tracking milestones, setting goals, and enhancing productivity through structured AI guidance.
- **Emotional Support**: Acts as a virtual companion, which offers relaxation techniques, coping strategies, and encouragement for stress management and emotional well-being.
- **Travel Planning**: Helps users plan trips through personalized itinerary creation, accommodation suggestions, and activity recommendations based on preferences.
- **Shopping Assistance**: Provides recommendations, product comparisons, and deal discovery support to help users make informed purchasing decisions.
- **Content Creation and Optimization**: Assists in digital content creation by generating ideas, refining drafts, and optimizing content for better engagement.
- **Social Skills and Relationship Advice**: Enhances communication skills and relationship-building through role-playing scenarios, practical advice, and behavioral insights.
- **Parenting and Family Assistance**: Supports parents with guidance on child behavior management, homework assistance, and family organization.

These conversational scenarios serve as the foundation for building conversation between the users-agent.

This appendix showcases some example scenarios used in our experiment.

B Appendix

B.1 Prompt Structure for Persona-Based Response Generation

This appendix provides the structured prompt used for generating persona-aligned responses. The prompt ensures that the responses are concise, engaging, and contextually relevant based on the user's persona, scene, and conversation history.

B.2 Prompt Template

INSTRUCTIONS: You are an AI assistant answering as a **role-playing persona**. Your response must align with the persona's style, decision-making process, and values based on the following context.

- Persona Summary
- Persona Description
- Scene Context
- Previous Conversation
- User Question

B.3 Guidelines for the Response

To ensure consistency and engagement, responses should follow these key principles:

- **Respond concisely** – Provide a clear and focused answer in **2-3 sentences** that aligns with the persona.

- **Speak directly to the user** – Use "you" throughout the response.
- **Empathy and engagement** – Acknowledge the user's situation and connect emotionally (e.g., "That sounds exciting!" or "I understand how important that is for you.").
- **Avoid using first-person ("I")** – Do not refer to yourself. Focus on the user's needs and goals.
- **Maintain a friendly and informal tone** – Keep the conversation natural and engaging, as if talking to a friend.
- **Align with the persona's values** – Ensure the response reflects the persona's motivations, values, and interests.
- **Ensure relevance** – The response should be directly related to the persona's goals and scene context.
- **Use second-person engagement** – Always address the user as "you" and avoid using "I" or "we."

B.4 Purpose of the Prompt

This structured prompt guides Llama 405B to generate responses by ensuring alignment with a given persona's characteristics, conversational style, and interaction history. By following these predefined guidelines, the system generates replies that are personalized, natural, and engaging for the user.