

**Live Chat with Virtual Character**

By

Abdelrahman Mostafa (20200827)

Mohamed Hisham (20200483)

Yousef Mohamed (20200669)

Supervised by

Dr. Reda El-Khoribi

T.A. Asmaa Ahmed

Artificial Intelligence Department

Cairo University

2023-2024

Table of Content

[1. Introduction 1](#_Toc158999770)

[1.1. Problem Definition 1](#_Toc158999771)

[1.2. Motivation 1](#_Toc158999772)

[1.3. Applications 1](#_Toc158999773)

[2. Related Work 3](#_Toc158999774)

[2.1. Large Language Models 3](#_Toc158999775)

[2.1.1. Conversational Agents with Embodied Avatars 3](#_Toc158999776)

[2.1.2. Large Language Models in Conversational Agents 4](#_Toc158999777)

[2.1.3. Domain-Specific Conversational Agents 6](#_Toc158999778)

[2.2. Computer Vision 6](#_Toc158999779)

[2.2.1. Talking Head Video Generation 6](#_Toc158999780)

[3. Methodology 9](#_Toc158999781)

[3.1. Large Language Models 9](#_Toc158999782)

[3.2. Computer Vision 11](#_Toc158999783)

[4. Plan of work 12](#_Toc158999784)

[5. References 13](#_Toc158999785)

Abstract

Recently, chatbots have gained attention for their natural language interactions. However, developing chatbots with contextual awareness and varied responses remains challenging.

Our project aims to develop conversation agents that possess unique qualities. These agents will be visually represented by avatars of historical figures such as Napoleon Bonaparte and Albert Einstein. This approach will enhance the interaction and reflect the historical context. We plan to expand beyond text-based interactions and explore incorporating additional ways, such as audio or cameras. This effort aims to create a more comprehensive and natural communication experience.

To achieve this, we will use pre-trained language models such as Microsoft Phi-2 as a foundation and fine-tune them individually for each character. This process will involve tailoring the language models to the specific historical context, personality, and knowledge base of each figure. Our approach is based on using the strengths of LLMs to generate human-like text. We plan to fine-tune individual models for each character to ensure that each response aligns with their unique traits and speech patterns. This helps to create more engaging and immersive conversations.

The applications of AI (Artificial Intelligence) agents simulating historical figures are wide-ranging and offer significant value across various domains. In education, they can serve as personalized virtual tutors, bringing historical figures to life and facilitating interactive learning experiences that enhance comprehension and critical thinking. In the entertainment industry, these agents create immersive and captivating experiences by enabling audiences to engage in interactive storytelling, games, and virtual reality experiences with iconic figures from history. Museums and historical sites can use them to provide visitors with a more interactive and enriching experience by incorporating virtual interactions with historical figures into exhibits.

Beyond these specific domains, the applications of these agents extend to any field where interaction with historical figures could be beneficial, such as healthcare and business.

Overall, these agents have the potential to deepen our understanding of the past, inform our present, and inspire our future.

Abbreviations list

* CLM: Causal Language Model
* CA: Conversational Agent
* GPT: Generative Pre-trained Transformer
* LORA: LOw-Rank Adaptation
* PEFT: Parameter-Efficient Fine-Tuning
* RAG: Retrieval-Augmented-Generation
* TTS: Text-to-Speech
* STT: Speech to Text
* Conv-LSTM: Convolutional Neural Network with Long Short-Term Memory.
* QA: Question Answering
* GLUE: General Language Understanding Evaluation benchmark

# Introduction

## Problem Definition

Studying history helps us learn from the past, make better choices for the future, and develop critical thinking skills. It may not be fun to study history, and it may be exhausting to search manually for specific historical characters or events. This can be a daunting task.

## Motivation

Our objective for this project is to use advanced technology to improve the capability of chatbots to understand and respond to historical inquiries with a higher level of contextual accuracy. With this improvement, students will have a more engaging and enjoyable learning experience while studying history with the assistance of chatbots.

## Applications

The applications of AI agents simulating historical figures are wide-ranging and offer significant value across various domains, including:

* **Education**: These agents can serve as personalized virtual tutors, bringing historical figures to life and facilitating interactive learning experiences that enhance comprehension and critical thinking. these tutors can Weave a Narrative, instead of dry facts and searching many resources, they can immerse you in historical events through interactive storytelling in one place. Imagine Julius Caesar guiding you through the battlefields of Gaul, or Marie Curie sharing her scientific discoveries in her own lab. these tutors can Spark Curiosity, they can present you with diverse viewpoints, hidden stories, and unexpected connections, igniting your interest and motivating you to delve deeper. History comes alive, revealing its complexities and intrigue. these tutors can Foster Comprehension, they can tailor their explanations to your understanding, providing additional information or clarifying complex concepts. Think of them as patient and adaptable teachers who adjust their pace based on your needs.
* **Entertainment**: They can create immersive and captivating experiences by enabling audiences to engage in interactive storytelling, games, and virtual reality experiences with iconic figures from history.
* **Museums and historical sites**: They can provide visitors with a more interactive and enriching experience by incorporating virtual interactions with historical figures into exhibits.

By leveraging a multi-model approach, dialogue systems can achieve a higher level of characterization, providing users with more personalized and immersive experiences. By adhering to these guidelines, developers can create chatbots that offer a seamless and intuitive user experience, making them indispensable tools for various applications, including customer service, information retrieval, and entertainment.

# Related Work

The field of conversational agents (CAs) is rapidly evolving, with a growing emphasis on enhancing their contextual understanding, response generation, and overall engagement. This project explores a multi-pronged approach towards achieving these goals.

## Large Language Models

### Conversational Agents with Embodied Avatars

* **FurChat**: This project explored a similar approach of combining a large language model (LLM) with an embodied avatar for human interaction. FurChat demonstrated the potential for engaging and informative conversations in a physical setting. However, it focused on general information retrieval, not domain-specific assistance.
* **ERICA**: This project developed an embodied social robot with advanced dialogue capabilities. While not using LLMs, ERICA highlighted the potential of embodied agents for emotional engagement and social interaction.

### Large Language Models in Conversational Agents

* **LaMDA**: This Google AI project demonstrates the power of LLMs for generating human-like dialogue. However, LaMDA focuses on open-ended conversational scenarios and does not address domain-specific expertise.
* **Phi-2**: this offers a readily available foundation for building advanced CAs (e.g., [Radford et al., 2022]). It is a transformer with 2.7 billion parameters. It was trained using the same data sources as Phi-1.5, augmented with a new data source that consists of various NLP synthetic texts and filtered websites (for safety and educational value). When assessed against benchmarks Testing common sense, language understanding, and logical reasoning, Phi-2 highlighted a state-of-the-art performance among models with less than 13 billion parameters. It also employs safety tensors to reduce toxicity and bias in the generated text. It has 24 layers, 32 attention heads, and a hidden size of 40962 with a context length of 2048 tokens. It uses the next-word prediction objective to learn from the training data. In training, it trained on a dataset of size 250B tokens on a 96xA100-80G GPU for 14 days (about 2 weeks), a combination of NLP synthetic data created by AOAI GPT-3.5 and filtered web data from Falcon RefinedWeb and SlimPajama, which was assessed by AOAI GPT-4.
* **Llama-2**: this is a family of pre-trained and fine-tuned large language models (LLMs) released by Meta AI in 2023. These models are made freely available for research and commercial purposes, which has drawn significant interest in the AI community. It refers to a family of second-generation LLMs developed by Meta. These models are designed for various natural language processing tasks, including dialogue generation and text completion. They are available for both research and commercial use. Llama 2 is an auto-regress language-optimized transformer. The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for helpfulness and safety. It has multiple versions, which are 7B, 13B, and 70B with a context length of 4K tokens.
* **Megatron-Turing NLG**: This model from NVIDIA highlights the potential of LLMs for factual language generation. However, it primarily focuses on factual summarization and lacks the interactive dialogue capabilities needed for a conversational agent.
* **BlenderBot**: This Facebook AI project explores LLMs for open-domain dialogue with a focus on factual grounding. While demonstrating progress, it still faces challenges in achieving robust and consistent domain-specific assistance.

### Domain-Specific Conversational Agents

Several industries have developed chatbots for specific tasks like customer service or technical support. However, these often lack the embodiment and multi-modal capabilities of your project.

* **Mitsuku**: This chatbot exhibits impressive performance in open-domain conversation tasks. However, it lacks the embodiment and domain-specific focus of your project.
* **Cleverbot**: This online chatbot learns through user interactions but does not leverage domain-specific knowledge or LLMs.

## Computer Vision

### Talking Head Video Generation

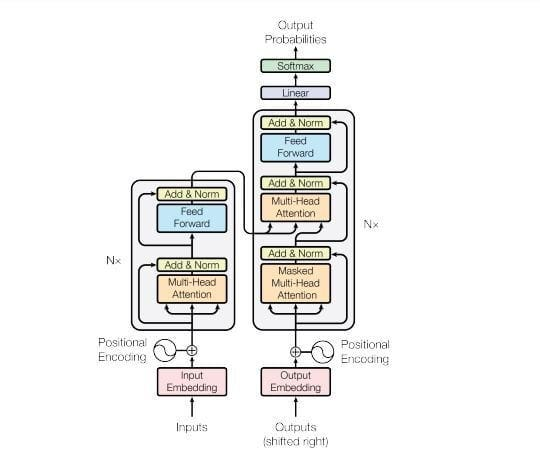
The creation of talking head videos from a single-face image and speech audio is fraught with challenges, such as unnatural head movements, distorted facial expressions, and modifications to the subject's identity. These issues are attributed to the reliance on learning from coupled 2D motion fields, which can lead to unnatural and incoherent results. Moreover, the use of explicit 3D information has been found to introduce its own set of problems, such as stiff expressions and videos that lack coherence.

One of the most powerful models is SadTalker. To address these challenges, SadTalker has been developed. This system generates 3D motion coefficients, including head pose and facial expressions, from audio using a 3D Morphable Model (3DMM) and modulates a novel 3D-aware face render to create talking head videos. SadTalker stands out by explicitly modeling the connections between audio and several types of motion coefficients individually, which helps in achieving more accurate facial expressions and head movements. To learn the realistic motion coefficients, we explicitly model the connections between audio and several types of motion coefficients individually. Precisely, we present ExpNet to learn the accurate facial expression from audio by distilling both coefficients and 3D-rendered faces.

* Components:
  + **ExpNet**: This component is designed to learn accurate facial expressions directly from audio. It does so by distilling information from both the motion coefficients and 3D-rendered faces, which helps in capturing the nuances of facial expressions that are synchronized with the audio.
  + **PoseVAE**: it is a conditional Variational Autoencoder that synthesizes head motion in several styles. This allows for the generation of head movements that are natural and match the speech style, contributing to the video's overall realism.
  + **3D Key Points Mapping:** The 3D motion coefficients generated by SadTalker are mapped onto an unsupervised 3D key points space of the proposed face render. This mapping is crucial for synthesizing the final video, ensuring that the motion is accurately reflected in the visual output.
* Benefits:
  + **Individual Modeling**: Explicitly modeling audio-to-motion connections for expression and pose leads to improved realism.
  + **Distillation-based Learning:** ExpNet's learning from both coefficients and rendered faces enhances expression accuracy.
  + **Style Control**: PoseVAE enables generating head motion with several styles based on audio.
  + **Unsupervised 3D Key Points**: Mapping to this space leverages 3D information without introducing stiffness or incoherence.
* **Experiments**: Extensive experiments have been conducted to validate the effectiveness of SadTalker. These studies show the method's superiority in motion realism and video quality, outperforming existing approaches in the field. Extensive experiments demonstrate SadTalker's superiority in terms of:
  + **Motion Naturalness**: More realistic and varied head motions compared to other methods.
  + **Expression Quality**: Accurate and nuanced facial expressions driven by audio.
  + **Video Quality**: Overall higher perceived quality and identity preservation.

# Methodology

## Large Language Models

To design and develop effective chatbots that can engage users in natural and personalized conversations, The following guidelines should be considered:

1. **Collecting and Processing Data**: For each character, gather a substantial dataset of text that reflects their unique speech patterns and personalities. This data can include dialogue, monologue, and other relevant text sources. Using web scraping, we will collect and process data about Einstein and Napoleon, converting it into a Q&A format.
2. **Exploring pre-trained models to leverage existing advancements**: Choose appropriate pre-trained large language models, as the foundation for each character's model to save time and resources. list of suggested LLMs:
   1. **Phi-2**: this offers a readily available foundation for building advanced. this offers a readily available foundation for building advanced CAs.
   2. **Llama-2**: this is a family of pre-trained and fine-tuned large language models (LLMs) released by Meta AI in 2023. These models are made freely available for research and commercial purposes, which has drawn significant interest in the AI community.
3. **Utilizing a multi-model approach to enhance conversational agents**: In interactive systems that involve conversational agents, each character can be characterized by their distinctive way of communicating. This can include their word choices, sentence structure, tone, and other linguistic features that help establish their personality and identity. Employing a multi-model approach can enhance the system's ability to generate diverse and character-specific responses.
4. **Fine-tuning individual models using a multi-model approach**: Create chatbots tailored to the specific characteristics and personalities of each historical figure. Fine-tune each pre-trained model on its respective character-specific dataset. This involves adjusting the model's parameters to optimize its performance on the task of generating text in the style of that character.

## Computer Vision

To integrate the computer vision system into our live chat system, we need to add the Talking Face Animator system. Generating realistic talking head videos using a single face image and audio presents several challenges, including unnatural head movements, distorted expressions, and loss of identity. This paper argues that these issues stem from learning from coupled 2D motion fields. While directly using 3D information can improve realism, it can also lead to stiff expressions and incoherent videos.

**SadTalker**: This work proposes a novel approach, SadTalker, that leverages 3D motion coefficients (head pose, expression) and implicitly modulates a 3D-aware face renderer for generating talking heads.

# Plan of work

* LLM (Large Language Model)
  + Data
    - Napoleon Bonaparte data
      * Collecting data about Napoleon Bonaparte
      * Process the Napoleon Bonaparte data.
      * Changing the personal pronoun (In progress)
    - Albert Einstein's data
      * Collecting data about Albert Einstein (In progress)
      * Process the Albert Einstein data (In progress)
      * Changing the personal pronoun (In progress)
  + Model Selection (Based on our resources)
  + Fine-tune (In progress)
* CV
  + Model Selection (In progress)
* Voice
  + Model Selection (In progress)

# References

1. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., & Wang, F. (2023). SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. Computer Vision and Pattern Recognition (CVPR).
2. <https://huggingface.co/microsoft/phi-2>
3. [A Comprehensive Guide to Fine-Tuning the Microsoft Phi-2 Model (Free Notebook) | by Mohamed Ahmed Krichen | Dec 2023 | Medium](https://medium.com/@mohamedahmedkrichen/a-comprehensive-guide-to-fine-tuning-the-microsoft-phi-2-model-free-notebook-52a4b5e486aa)
4. Hu, Z., Wang, L., Lan, Y., Xu, W., Lim, E., Bing, L., Xu, X., Poria, S., & Lee, R. K.-W. (2023). LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models.
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need.
7. [https://sadtalker.github.io](https://sadtalker.github.io/)
8. <https://huggingface.co/spaces/vinthony/SadTalker>
9. Zhang, Wenxuan & Cun, Xiaodong & Wang, Xuan & Zhang, Yong & Shen, Xi & Guo, Yu & Shan, Ying & Wang, Fei. (2022). SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. 10.48550/arXiv.2211.12194.
10. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., & Wang, F. (2022). SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. ArXiv. /abs/2211.12194
11. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., & Wang, F. (2022). SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. *ArXiv*. /abs/2211.12194
12. <https://techxplore.com/news/2023-09-embodied-conversational-agent-merges-large.html>
13. <https://www.sciencedirect.com/science/article/pii/S2666920X21000278>
14. Cherakara, S., Wijeratne, H., & Papadopoullos, A. (2023). An embodied conversational agent that merges large language models and domain-specific assistance. arXiv preprint arXiv:2309.02684.
15. Ishiguro, H., Ono, T., Kobayashi, M., & Ikeda, H. (2006). ERICA: A life-sized human-like android. IEEE Intelligent Systems, 21(4), 12-21.
16. Breazeal, C. (2006). Designing sociable robots. MIT Press.
17. A. Barua, M. U. Ahmed and S. Begum, "A Systematic Literature Review on Multimodal Machine Learning: Applications, Challenges, Gaps and Future Directions," in IEEE Access, vol. 11, pp. 14804-14831, 2023, doi: 10.1109/ACCESS.2023.3243854.