

Full-stack AI Content Creator Supervisor
Term 1 Final Report

TSE Hui Tung, 1155158864

NG Man Tik, 1155158302

Supervised by

Prof. Michael R. Lyu

Computer Science and Engineering
The Chinese University of Hong Kong

29/11/2023

Content

1. ACKNOWLEDGEMENT	4
2. ABSTRACTS.....	5
3.INTRODUCTION.....	7
3.1. MOTIVATION.....	7
3.2. OBJECTIVE	9
3.3. CONTRIBUTION.....	10
4.RELATED WORK.....	11
5. METHODOLOGY	12
5.1 FETCHING DATA	12
5.1.1 Data Type	12
5.1.2 Task List	15
5.1.3 Workflow.....	16
5.1.4 Results	19
5.2 GENERATING QUESTIONS.....	21
5.2.1 Generation Strategy.....	21
5.2.2 Predefined Data types.....	22
5.2.3 Question Types	23
5.2.4 Factualness.....	24
5.2.5 Results and Application	26
5.3 IMITATION SYSTEM	29
5.3.1 Introduction.....	29
5.3.2 Prototype v1.....	30
5.3.3 Prototype V2	31
5.3.4 Prototype V3	34
5.3.5 Prototype v4.....	37
5.3.6 Demo.....	40
6. EXPERIMENT	53
6.1 BASELINE.....	54
6.1.1 Introduction.....	54
6.1.2 Results	58
6.3 ELIZA EFFECT.....	74
6.4 EVALUATION.....	74
6.4.1 Running	77

6.4.2 Test Group	79
6.4.3 Question Aspect	79
6.4.4 Evaluation Criteria.....	85
6.4.5 Experiment Results.....	86
7. CONCLUSION.....	90
8. FUTURE WORK	92
9. REFERENCE LIST.....	94

1. Acknowledgement

We would like to express our gratitude to Professor Michael R. Lyu, our supervisor, Mr. Jen-tse Huang and Ms. Jing Jing Li who guided us in doing this final year project as well as provided us with invaluable advice.

Besides, we would like to express our special thanks of gratitude to all of the participants giving their background information as reference and spending time responding to the answers in the experiment for the evaluation.

2. Abstracts

This report delves into the intricacies of simulating human personalities using advanced natural language processing (NLP) techniques in the realm of large language model (LLM) and LangChain [1]. This project, titled 'Full-Stack AI Content Creator,' is an innovative idea to replicate the unique communication styles, thought processes, and personal attributes of famous figures, using a combination of data collection, and Chain-of-Thought process [2]. This core objective is to create a system capable of generating authentic and engaging digital content, particularly for social media platforms, that embodies the essence of these personalities.

The methodology involves an extensive data collection process, the development of a sophisticated question-generation framework, and the implementation of a multi-prototype imitation system. Each prototype iteration enhances the system's ability to closely mimic the targeted personalities,

addressing challenges such as tone consistency, factual accuracy, and response personalization.

Our experiments, involving various AI models and comprehensive evaluation criteria, have demonstrated the system's proficiency in accurately replicating human-like behavior and interactions. The project's contributions extend beyond technical advancements, offering insights into the potential of AI in digital content creation and interactive systems. The conclusion underscores the success of the project in advancing AI's role-playing capabilities and sets a foundation for future innovations. Future work includes optimizing the response generation speed and expanding the system's application across various social media platforms for broader user engagement and interaction.

3.Introduction

3.1. Motivation

The advent of artificial intelligence (AI), especially in natural language processing (NLP) [3], has unveiled unparalleled opportunities for technological innovation. Our project, positioned at the forefront of this rapidly evolving domain, aims to extend the capabilities of AI in simulating personalities, such as those of notable public figures like J. Robert Oppenheimer. The motivation behind this endeavor arises from NLP's exceptional potential to replicate and preserve the unique essences of influential individuals. In an era dominated by social media platforms like Instagram, our initiative, 'Full-Stack AI Content Creator', aspires to immortalize these figures. We plan to achieve this by emulating their thought processes and communication styles through advanced NLP techniques.

Furthermore, with tools like stable diffusion [4] for image

generation, our project envisions creating interactive digital avatars that can engage meaningfully across various social platforms. This approach not only honors their legacies but also stimulates intellectual discussions on diverse topics, including current news.

This multifaceted motivation mirrors a deep curiosity about training NLP systems to mirror human uniqueness, whether through fine-tuning or prompt engineering [5]. It also embodies an ambition to push the boundaries of machine learning in understanding and mimicking human behavior. Additionally, this project holds significant promise for the digital content creation realm, offering innovative methodologies for content generation and serving as an experimental platform to study AI's proficiency in emulating human intellect and creativity.

3.2. Objective

The primary objectives of the Full-Stack AI Content Creator project encompass the broad aspects of digital imitation and interaction. We aim to achieve a high level of performance in mimicking a person's tone, writing style, and thought process, informed by their personal information. Moreover, we are dedicated to developing a comprehensive evaluation system to assess the project's effectiveness and integrity. This includes conducting tuning tests to evaluate the LLM's performance across various metrics such as the depth and breadth of personality data, the conversational sample size for training, the model's architectural intricacies, and the specific LLM used.

Additionally, by integrating text generation with text-to-image functions like stable diffusion, we aim to produce visually engaging posts that reflect the mimicked celebrity's persona. The ultimate goal is to deploy this technology on social media platforms, allowing user interaction and feedback.

3.3. Contribution

Our project has made several key contributions to the field of AI and personality simulation. Firstly, we innovated by implementing a Chain of Thought structure in the Large Language Model (LLM). This approach significantly enhanced the LLM's ability to accurately mimic a celebrity's thinking style and habitual expressions. Secondly, we established a comprehensive evaluation system to rigorously assess the LLM's proficiency in simulating real personalities, proving essential in validating the effectiveness of our methods. Lastly, we plan to extend the project's impact by deploying our system on social media platforms, aiming to create interactive and engaging content that resonates with users, leveraging the advanced capabilities of our AI-driven celebrity simulations.

4.Related Work

There are existing works in the field of simulation using LLMs. For instance, the paper RoleLLM [6] presents a structured approach involving Role Profile Construction, Context-Based Instruction Generation, and Role Prompting using GPT to tackle the challenges posed by general-purpose LLMs. Another study, Character-LLM [7], adopts a different strategy by creating trainable agents to simulate historical figures, focusing on summarizing personal profiles, experiences, and emotional states to enhance the LLM's resonance with their experiences and emotions. OpenAI has also made significant advancements with GPTs in November [8], [9], facilitating the creation of custom ChatGPT models suited for specific tasks. These models enable users to mimic personalities by interacting in natural language and uploading relevant information as knowledge. Integrated with DALL-3, a text-to-image model, it allows for the generation of various images in responses, empowering users to build models without requiring coding expertise.

5. Methodology

5.1 Fetching Data

For effective mimicry using a Large Language Model (LLM), it is crucial to feed it with both first-hand information, such as speeches or papers authored by the celebrity, and third-party information that encapsulates their thinking style, tone, and writing style through relevant documents. The challenge, however, lies in efficiently gathering and processing the extensive online data without overwhelming the LLM in terms of cost and time. To achieve this, we focus on both the breadth and depth of information collection while avoiding information overload. The key difficulty is in identifying and summarizing pertinent data from the vast online resources.

5.1.1 Data Type

In our methodology, the types of data we use to train the Large Language Model (LLM) for mimicking a celebrity are

twofold: first-hand information and third-party information.

Each type plays a crucial role in creating a comprehensive and accurate simulation of the celebrity's persona.

First-hand Information

This category includes direct communications from the celebrity, like speeches, interviews, and writings. These provide a primary source of raw data for the LLM, offering unique insights into the celebrity's personal ethos, disposition, and speech and writing patterns. This direct source allows the LLM to adopt the celebrity's linguistic nuances when mimicking them, effectively replicating their communication style and ensuring authenticity. For example, writing segments can be utilized as FewShotPrompts [10] for learning the celebrity's writing style.

Third-party information

Encompassing external content about the celebrity, such as biographies and news articles, third-party information gives a comprehensive view of the celebrity's public persona, societal impact, and career trajectory. It also sheds light on how the celebrity's actions and statements are perceived by society. This information becomes vital in the absence of first-hand data, providing insights into the celebrity's thinking style. The LLM can use this data to understand the broader narrative surrounding the celebrity and refine its simulation based on the widespread interpretation of their personality and actions.

In summary, leveraging both first-hand and third-party information as training data equips the LLM to construct a multi-dimensional understanding of the celebrity, enabling it to generate authentic and believable results.

5.1.2 Task List

We have created specific tasks for the agent to gather information about various aspects of the celebrity's public image and work. This task list guides the agent in creating a dataset that is representative of the celebrity's persona.

```
*****TASK LIST*****

1: Research any scientific papers or publications J. Robert Oppenheimer was involved in.
2: Identify any awards or honors J. Robert Oppenheimer received before 1963.
3: Research any organizations or causes J. Robert Oppenheimer supported.
4: Research any scientific theories or discoveries J. Robert Oppenheimer was credited with.
5: Research J. Robert Oppenheimer's professional career and any significant events that occurred during his
   lifetime.
6: Research any public statements or interviews J. Robert Oppenheimer gave.
7: Research any charitable activities J. Robert Oppenheimer was involved in.
8: Identify any awards or honors J. Robert Oppenheimer received from universities or research institutions.
9: Research any public debates or controversies J. Robert Oppenheimer was involved in.
10: Research any educational institutions J. Robert Oppenheimer taught at.
11: Identify any awards or honors J. Robert Oppenheimer received after 1963.
12: Research any organizations or causes J. Robert Oppenheimer was involved in.
13: Identify any books or articles written by J. Robert Oppenheimer.
14: Research any speeches or lectures given by J. Robert
```

Figure: Task list for agent to search online to get information

5.1.3 Workflow

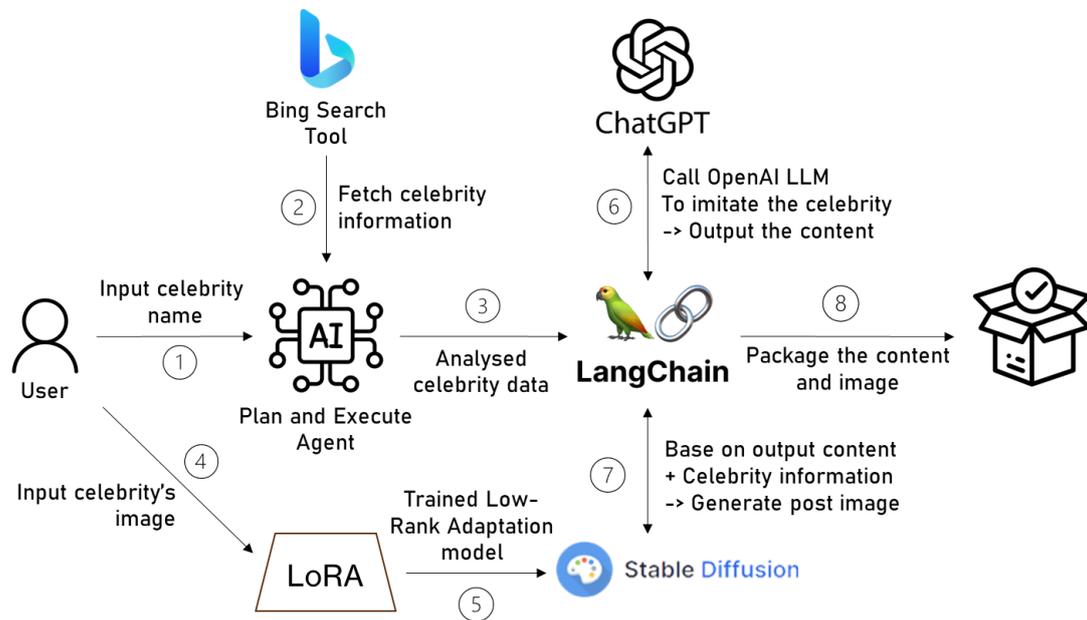


Figure: Celebrity Imitation Application Workflow

Our method is based on the premise of accurately simulating a celebrity. Given the lengthy chain of LLM processes during simulation, conducting real-time information searches during interaction is impractical [11]. Thus, we have developed a system that proactively fetches necessary information about the celebrity and condenses it into a format suitable for the language model. Our prototype includes:

- **Task-specific searches:** Rather than a broad search, we break down the retrieval process into multiple, detailed tasks, each

targeting a specific aspect of the celebrity's life or work to ensure comprehensive data collection.

- **Data Filtration and Deduplication:** Post-collection, the LLM filters out irrelevant information and checks for redundancies against saved files. This step ensures that the data fed into the language model is not only rich but also streamlined and non-repetitive, allowing the LLM to extract key information efficiently.

Our data processing workflow integrates LangChain with the ChatGPT API [12] and BabyAGI [13], utilizing BingSearch API V7 [14]. The workflow involves:

1. **Task Prioritization:** BabyAGI selects the most pertinent task at hand.
2. **Data Retrieval:** LangChain employs BingSearch API V7 to gather data according to the given task.

3. **Result Refinement:** The agent removes irrelevant information from the search results.
4. **Deduplication:** The agent compares the new data with existing files to eliminate duplicates.
5. **Data Structuring:** The refined data is structured into a JSON file, facilitating easy retrieval for future processing.
6. **Iteration:** This process is repeated for each task to build a comprehensive dataset.

The choice of JSON format for data structuring is intentional, designed to enable ease of access and manipulation when the data is subsequently used to prompt the language model.

5.1.4 Results

The result is a summarized JSON file containing the processed data, ready for use in simulation tasks.

```
{
  "Name": "J. Robert Oppenheimer",
  "Information": [
    {
      "Title": "Biography of J. Robert Oppenheimer",
      "Value": "J. Robert Oppenheimer was an American theoretical physicist who is best known as the \"father of the atomic bomb\" for his role in creating the first nuclear weapon during World War II. He was director of the Los Alamos Laboratory and responsible for the research and design of an atomic bomb. He earned a bachelor of arts degree in chemistry from Harvard University in 1925 and a doctorate in physics from the University of G\u00fcttingen in Germany in 1927."
    },
    {
      "Title": "J. Robert Oppenheimer's Involvement in the Development of the Atomic Bomb",
      "Value": "J. Robert Oppenheimer was a Harvard-educated theoretical physicist and scientific director of the Los Alamos Laboratory in New Mexico during World War II. He is known as the \"father of the atomic bomb\" for his role in creating the first nuclear weapon during World War II. He played a pivotal role in developing atomic weapons and forever changed the course of history with his contributions to nuclear science."
    },
    {
      "Title": "J. Robert Oppenheimer's Life and Accomplishments",
      "Value": "J. Robert Oppenheimer was a brilliant physicist who played a pivotal role in developing atomic weapons and forever changed the course of history with his contributions to nuclear science. He was selected to administer a laboratory to carry out the Manhattan Project, the program that developed the first nuclear weapon. He made significant contributions to theoretical physics, including achievements in quantum mechanics and nuclear physics."
    },
    {
      "Title": "J. Robert Oppenheimer's Life and Accomplishments",
      "Value": "J. Robert Oppenheimer was a brilliant physicist who played a pivotal role in developing atomic weapons and forever changed the course of history with his contributions to nuclear science. He was selected to administer a laboratory to carry out the Manhattan Project, the program that developed the first nuclear weapon. He made significant contributions to theoretical physics, including achievements in quantum mechanics and nuclear physics. He studied Sanskrit and read Hindu scripture. He taught physics in California before embarking on his role in the Manhattan Project as the director of the top-secret Los Alamos Laboratory. He succeeded in developing nuclear weapons, but would later be a controversial figure."
    },
    {
      "Title": "Awards and Recognitions Received by J. Robert Oppenheimer",
      "Value": "The Enrico Fermi Award"
    },
    {
      "Title": "J. Robert Oppenheimer's Affiliations",
      "Value": "Public Health, Medicine, and Psychiatry, Weill Cornell Medical College"
    },
    {
      "Title": "J. Robert Oppenheimer's Education and Academic Career",
      "Value": "J. Robert Oppenheimer (1904-1967) was an American theoretical physicist and science administrator, noted as director of the Los Alamos Laboratory (1943-45) during development of the atomic bomb and as director of the Institute for Advanced Study, Princeton (1947-66). Oppenheimer was born in New York City and earned a bachelor of arts degree in chemistry from Harvard University in 1925. He then studied physics at the University of G\u00fcttingen in Germany, where he earned a Ph.D. in 1927. He returned to the United States and taught at the University of California, Berkeley, from 1929 to 1943. During this time, he conducted research in theoretical
```

```

{
  "Title": "Analyze J. Robert Oppenheimer's influence on the development of atomic theory",
  "Value": "J. Robert Oppenheimer, the brilliant physicist behind the Manhattan Project, played a pivotal role in developing atomic weapons and forever changed the course of history with his contributions to nuclear science. He opposed the development of the hydrogen bomb during a 1949-1950 governmental debate on the question and subsequently took positions on defense-related issues that provoked the ire of some U.S. government and military factions. After the war, Oppenheimer served on the U.S. Atomic Energy Commission, where he argued for more oversight regarding the use of nuclear weapons and opposed the construction of the hydrogen bomb. Before he developed the atomic bomb, J. Robert Oppenheimer's early work revolutionized the field of quantum chemistry and his theory is still used today. Less than three years after Groves selected Oppenheimer to direct weapons development, the United States dropped two atomic bombs on Japan. As director of the Los Alamos Laboratory, Oppenheimer proved to be an extraordinary choice."
},
{
  "Title": "Examine J. Robert Oppenheimer's role in the development of nuclear energy",
  "Value": "J. Robert Oppenheimer (born Julius Robert Oppenheimer; / \u02c8 \u0252 p \u0259 n h a \u026a m \u0259 r / OP -\u0259 n-hy-m\u0259 r; April 22, 1904 \u0213 February 18, 1967) was an American theoretical physicist and director of the Manhattan Project's Los Alamos Laboratory during World War II. He is often called the \"father of the atomic bomb\". Born in New York City, Oppenheimer earned a bachelor of arts degree in chemistry from Harvard University in 1925 and a doctorate in physics from the University of G\u00f6ttingen in 1927. He was a professor of physics at the University of California, Berkeley, and later at the Institute for Advanced Study in Princeton, New Jersey. Oppenheimer is known as the \"father of the atomic bomb\" for his role in creating the first nuclear weapon during World War II. The theoretical physicist was director of the Manhattan Project, the U.S. government program that developed the first atomic bombs. He was also a key figure"
}

```

```

{
  "Title": "Analyze J. Robert Oppenheimer's influence on the development of nuclear weapons",
  "Value": "J. Robert Oppenheimer (1904-1967) was an American theoretical physicist and director of the Manhattan Project's Los Alamos Laboratory during World War II. He is often called the \"father of the atomic bomb\" for his role in creating the first nuclear weapon during World War II. Oppenheimer was born in New York City and earned a bachelor of arts degree in chemistry from Harvard University in 1925. He then went on to study physics at the University of G\u00f6ttingen in Germany, where he worked with some of the most renowned physicists of the time. After returning to the United States, Oppenheimer became a professor of physics at the University of California, Berkeley. During World War II, he was appointed director of the Los Alamos Laboratory, where he oversaw the development of the atomic bomb. Oppenheimer's work on the Manhattan Project was instrumental in the development of nuclear weapons, and he is credited with helping to bring an end to the war. After the war, Oppenheimer continued to work on nuclear weapons and was a vocal advocate for nuclear disarmament"
}

```

Figure: Summarized result of the fetched information of the celebrity

5.2 Generating Questions

The process of imitating a celebrity through a language model extends beyond merely feeding it information. Capturing a persona's essence is best reflected in their responses to inquiries and how they articulate their thoughts. A challenge with ChatGPT is its inconsistency in results due to limited recall of past interactions [15]. Therefore, our system is designed to both analyze and synthesize information into a question-and-answer (QA) format, fostering dynamic and consistent interactions that mirror the celebrity's communication style [16].

5.2.1 Generation Strategy

Following the idea proposed in RoleLLM, we will perform a similar strategy for crafting a QA set that accurately reflects a celebrity's persona involves generating a diverse and insightful array of questions. This strategy is not about asking just any questions, like mathematical ones, but about asking the right

questions that delve into the nuances of the celebrity's life, thoughts, and style. Our generation strategy breaks down into three key aspects: predefined data types, question types, and factualness.

5.2.2 Predefined Data types

To guide the language model towards a comprehensive understanding of the celebrity, we categorize information into predefined data types. These types cover a broad spectrum of the celebrity's life and personality, allowing the model to summarize background information fetched from the internet and generate questions that are relevant and meaningful. These questions are deeply rooted in the celebrity's background, beliefs, and preferences, offering insights into their comprehensive persona. These types include:

- **Education and Professional Background**
- **Interests and Hobbies**
- **Personality**

- **Favorite Books, Movies, and Music**
- **Values and Beliefs**
- **Problem-Solving Style**
- **Thoughts**
- **Communication and Social Style**
- **Memorable Life Experiences**
- **Writing and Speaking Style**

5.2.3 Question Types

The questions are designed not just to elicit information but also to discern the authenticity and depth of the responses. We use various question types to challenge the model in replicating human-like interaction. These are crucial in evaluating whether the responses convincingly mirror the complexity of human thought and emotion, enabling the language model to also reflect them during live interactions later. The question types we used are:

- **Memories or Secrets**
- **Personal Data**
- **Emotional Questions**
- **Subjective or Creative Questions**
- **Knowledge-Based Questions**
- **Ethical Questions**
- **Logical Questions**
- **Philosophical Questions**
- **Questions About the Future**
- **General Questions**

5.2.4 Factualness

Based on the paper RoleLLM, we realized that sometimes the questions generated might be irrelevant and inaccurate because the language model forgets to consider the person's background. Hence, it's essential that the questions are not only appropriate but also closely aligned with the celebrity's background and public persona. This involves:

- Ensuring questions are tailored to the celebrity's known

experiences and expertise.

- Avoiding questions that are too irrelevant to the celebrity's life, like asking "What do you think about ChatGPT" to Oppenheimer.
- Balancing the specificity of questions with the need to maintain broad appeal to diverse audiences.

Considering the generation strategy for QA sets, the final prompt we set to let LLM generate the QA set will be the following:

```
question_template = """
Background information on {role_name}:
{background_info}
If you had the opportunity to meet {role_name}, what questions would you ask
{role_name}? Please design 10 questions that do not repeat in terms of semantics.
Please set the question diverse.
You can set your questions based on either on the following aspects: {aspects};
or on the following types: {types}; or both.
You can base your questions
on {role_name}'s personality traits and relevant episodes from {role_name}. In addition to providing the
questions, please indicate the factuality (high or low) of each question you generate and provide a response.
High factuality signifies that the questions are not arbitrarily fabricated. Please strive to generate questions
with high factuality.

Please refer to the following examples related to Sherlock Holmes for guidance on how to format your questions and
responses:
```

```

1. Question: "Sherlock Holmes, is Dr. John Watson your closest confidant?"
   Factualness: High, because in Arthur Conan Doyle's stories, Dr. Watson is indeed Holmes' close friend and partner.
   Response: "Indeed, Dr. Watson is not only my closest confidant but also an invaluable assistant in my investigations. His medical expertise and steadfast character have been instrumental in my work."

2. Question: "Sherlock Holmes, have you ever traveled to the moon?"
   Factualness: Low, as the concept of Holmes traveling to the moon is purely fictional and not part of Doyle's canon.
   Response: "Traveling to the moon is beyond the realms of my adventures. My pursuits are firmly grounded in solving mysteries on Earth."

3. Question: "Sherlock Holmes, what is your most challenging case?"
   Factualness: High, considering the numerous complex cases Holmes has encountered in Doyle's stories.
   Response: "The most challenging case would arguably be 'The Hound of the Baskervilles,' where I had to unravel a mystery shrouded in fear and superstition."

4. Question: "Sherlock Holmes, did you invent a time machine?"
   Factualness: Low, as inventing a time machine is not part of Sherlock Holmes' character or capabilities in the stories.
   Response: "As a detective, my focus is on observation and deduction based on existing facts. The invention of a time machine falls outside my area of expertise."

5. Question: "Sherlock Holmes, what are your thoughts on modern forensic science?"
   Factualness: High, given Holmes' pioneering role in using forensic methods in his investigations.
   Response: "Modern forensic science, with its advanced techniques and technology, would have been a valuable asset in my investigations. It aligns with my belief in the importance of evidence and scientific methods in solving crimes."

6. Question: "Sherlock Holmes, have you ever met Harry Potter?"
   Factualness: Low, as Harry Potter belongs to a different fictional universe created by J.K. Rowling.
   Response: "Harry Potter's adventures in the realm of magic are quite separate from my own in the world of detective work. Our paths have never crossed."

[Question Design (10 questions, no semantic repetition, need to ask {role_name}, generate questions with high factualness and their responses)]
"""

```

Figure: Prompt for generating QA

5.2.5 Results and Application

Like how the data is handled in the data collection before, we will also generalize the QA set into the JSON format after generation. Here is an example:

```

1- {
2-   "qa_pairs": [
3-     {
4-       "aspects": "Education and Professional Background",
5-       "types": "Knowledge based Questions",
6-       "question": "Dr. Oppenheimer, can you share some insights about your time at the University of G\u00fcttingen and how it
7-         shaped your career in physics?",
8-       "factualness": true,
9-       "response": "My time at the University of G\u00fcttingen was instrumental in shaping my career in physics. I had the
10-        opportunity to work with some of the most renowned physicists of the time, which greatly influenced my understanding
11-        and approach to theoretical physics. It was there that I earned my doctorate in 1927."
12-     },
13-     {
14-       "aspects": "Interests and Hobbies",
15-       "types": "Personal Data",
16-       "question": "Dr. Oppenheimer, I understand you had an interest in Sanskrit and Hindu scripture. How did this interest
17-         come about and how has it influenced your life and work?",
18-       "factualness": true,
19-       "response": "Yes, I developed an interest in Sanskrit and Hindu scripture during my studies. I found the philosophical
20-        concepts fascinating and they provided a different perspective on understanding the world. This interest also
21-        influenced my work, particularly in the development of the atomic bomb, where I famously quoted a line from the
22-        Bhagavad Gita upon its successful test."
23-     },
24-     {
25-       "aspects": "Personality",
26-       "types": "Emotional Questions",
27-       "question": "Dr. Oppenheimer, how did you cope with the ethical implications of your work on the atomic bomb, given its
28-         destructive power?",
29-       "factualness": true,
30-       "response": "The ethical implications of my work on the atomic bomb were indeed a heavy burden. While the development of
31-        the bomb was crucial for ending World War II, the destruction it caused was immense. This led me to advocate for
32-        nuclear disarmament and control of nuclear power post-war."
33-     },
34-     {
35-       "aspects": "Values and Beliefs",
36-       "types": "Ethical Questions",
37-       "question": "Dr. Oppenheimer, you were a vocal advocate for nuclear disarmament after the war. Can you share your
38-         thoughts on why you believed this was necessary?",
39-       "factualness": true,
40-       "response": "After witnessing the destructive power of the atomic bomb, I believed it was necessary to advocate for
41-        nuclear disarmament. The potential for such weapons to cause mass destruction and loss of life is immense, and I felt
42-        it was our responsibility to ensure they were not used recklessly."
43-     },
44-     {
45-       "aspects": "Problem-Solving Style",
46-       "types": "Logical Questions",
47-       "question": "Dr. Oppenheimer, can you describe your approach to problem-solving during the development of the atomic bomb
48-         ?\"",
49-       "factualness": true,
50-       "response": "During the development of the atomic bomb, my approach to problem-solving was largely collaborative. I
51-        worked closely with a team of brilliant scientists, and we tackled each challenge as it came, using our collective
52-        knowledge and expertise to find solutions."
53-     },
54-     {
55-       "aspects": "Thoughts",
56-       "types": "Philosophical Questions",
57-       "question": "Dr. Oppenheimer, how did your work on the atomic bomb influence your philosophical views?",
58-       "factualness": true,
59-       "response": "My work on the atomic bomb had a profound impact on my philosophical views. It made me deeply aware of the
60-        ethical implications of scientific advancements and the responsibility that comes with such knowledge. This led me to
61-        advocate for the control of nuclear power and disarmament post-war."
62-     },
63-     {
64-       "aspects": "Communication and Social Style",
65-       "types": "Subjective or creative Questions",
66-       "question": "Dr. Oppenheimer, how would you describe your leadership style during your time as director of the Los Alamos
67-         Laboratory?",
68-       "factualness": true,
69-       "response": "As the director of the Los Alamos Laboratory, I aimed to foster a collaborative and inclusive environment. I
70-        believed in the importance of open communication and encouraged my team to share their ideas and concerns. I also
71-        understood the importance of clear and decisive leadership in such a high-stakes project."
72-     },
73-     {
74-       "aspects": "Memorable Life Experience",
75-       "types": "Memories or secrets",
76-       "question": "Dr. Oppenheimer, can you share a memorable experience from your time working on the Manhattan Project?",
77-       "factualness": true,
78-       "response": "One of the most memorable experiences from my time working on the Manhattan Project was the successful test
79-        of the atomic bomb in July 1945. It was a moment of both triumph and trepidation, knowing the immense power we had
80-        unleashed."
81-     }
82-   ]
83- }

```

Figure: Generated QA in JSON format

Once the QA set is generated, it's integrated into the simulation framework. The QA set acts as a pre-compiled resource that the language model can draw upon as knowledge or memory during live interactions. The advantages of this approach include:

- **Efficiency:** Reducing the computational overhead of generating responses in real-time, ensuring quick and fluid interactions.
- **Consistency:** Using the QA set as a baseline to follow the person's writing habit, thinking style, tone, etc., during each interaction, generating a consistent result in the long run.
- **Complexity Management:** By generating multi-aspect and multi-type questions in the QA set, these questions can serve as comprehensive guidelines for the language model to handle more complex questions that might require multifaceted answers, which the model has already pre-formulated and refined.

5.3 Imitation System

5.3.1 Introduction

The Imitation System, central to our project, is designed to replicate celebrities' unique communication styles and thought patterns using natural language processing. It intricately merges first-hand and third-party information to establish a thorough database, facilitating the creation of authentic responses mirroring a chosen celebrity's style. We opted for diverse prompt engineering rather than model fine-tuning, aiming to make a model respond to questions in a way that aligns with the celebrity's personality, background, tone, and writing style, so answers appear as if coming directly from the celebrity. This report documents the evolution of our prototypes, their modifications, and associated drawbacks.

5.3.2 Prototype v1

5.3.2.1 Introduction

Initially, we designed the system simply, using a zero-shot single prompt [17] to instruct the LLM to answer user-input questions by imitating the celebrity, based on data from our database. The response was then output to the user. A conversational memory buffer was implemented to maintain long-term consistency.

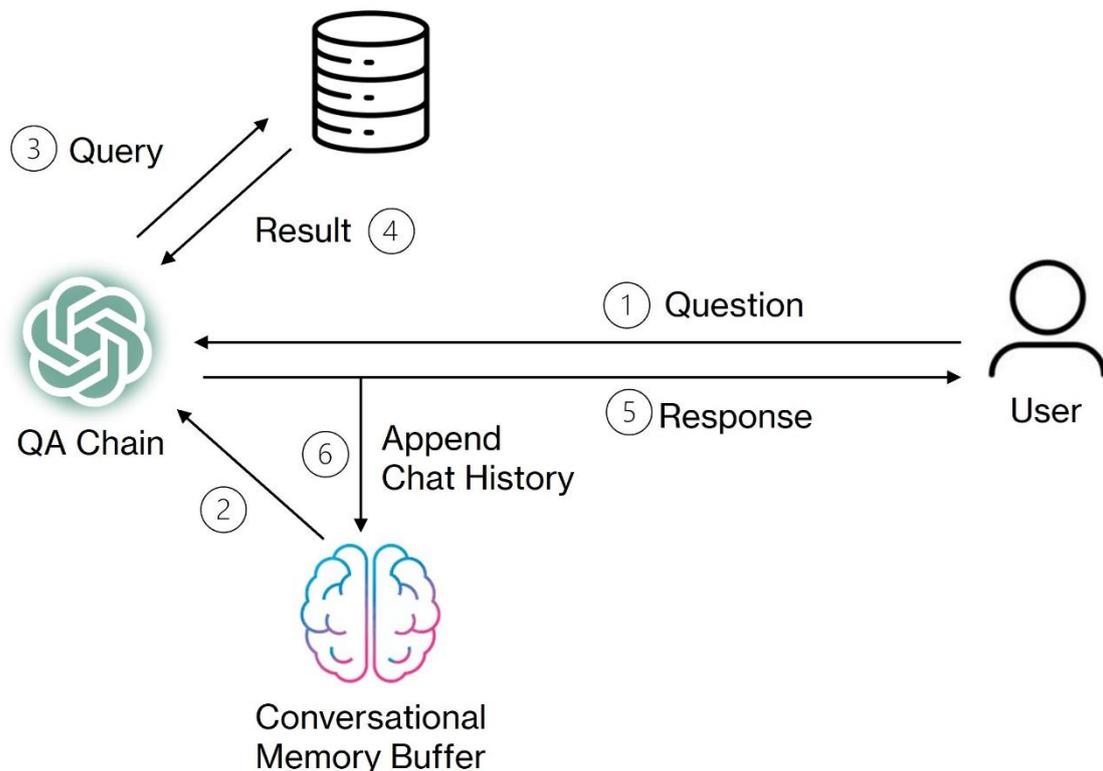


Figure: Workflow of prototype V1

5.3.2.2 Drawback

However, this first prototype revealed significant shortcomings. The single-prompt method failed to capture the depth and intricacies of a celebrity's communication style.

There were noticeable deviations in tone, vocabulary choice, and overall expression due to the lack of performance using zero shot prompting [18]. Although the responses were based on information from the database, they often lacked the distinctive flair and nuances typical of the celebrity.

Furthermore, the prototype struggled to align responses with the celebrity's past statements or known beliefs, resulting in answers that, while factually correct, did not reflect the celebrity's known viewpoints on various subjects.

5.3.3 Prototype V2

5.3.3.1 Introduction

In Prototype V2, we aimed to address the limitations observed in the first version by introducing a multiple-

prompting system combined with a knowledge bagging approach [19]. This version utilizes a Large Language Model (LLM) to generate a QA chain based on the celebrity's background information fetched from a vector database. These pre-processed QA sets are stored in the conversational memory buffer and used to inform subsequent interactions. When a user poses a question, the LLM retrieves relevant chat history and the QA chain to construct a response that aligns with the celebrity's known information. This answer is then provided to the user and saved to the memory buffer for future reference.

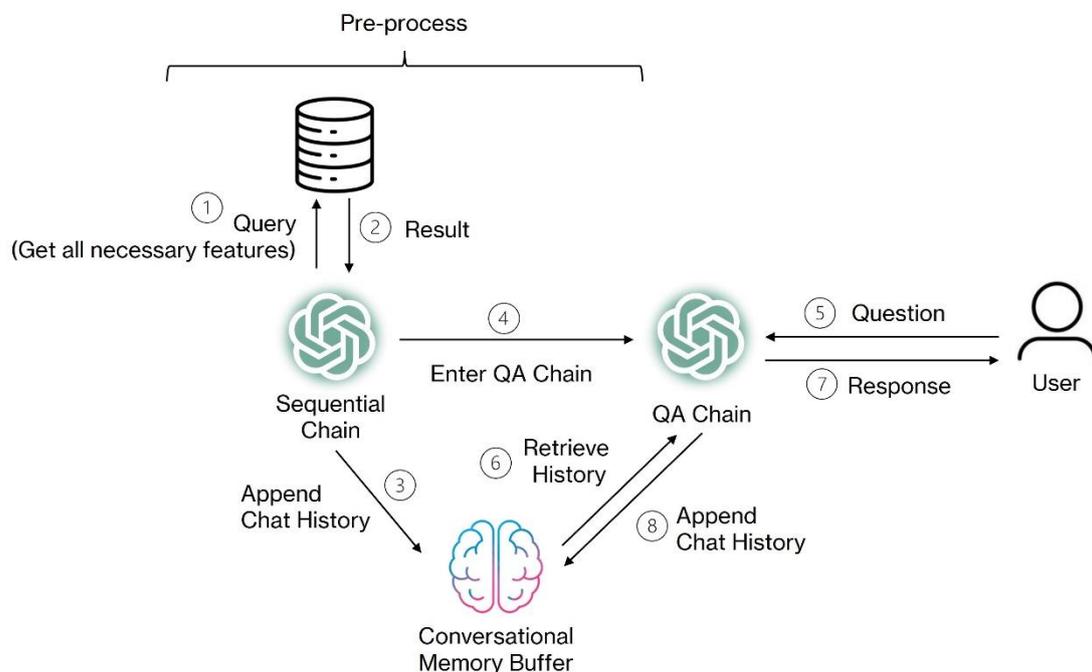


Figure: Workflow of prototype V2

5.3.3.2 Drawback

While Prototype V2 marked an improvement in generating responses that linguistically resembled those of the celebrity, it encountered significant challenges in accurately mimicking their thinking style and logical reasoning. Firstly, using single or few-shot prompts in the systems often limited the model's ability to perform complex reasoning and mimic the celebrity's thinking style accurately. Such prompts generally lack the depth needed for intricate imitations, resulting in superficial representations of the celebrity's persona. [20]

Meanwhile, the system heavily depended on the information from the fetched database, resulting in responses that were essentially compilations of existing data rather than reflections of the celebrity's unique thought process [21]. This led to a replication that was superficial in nature, capturing the celebrity's speech patterns but failing to convey their personality or individualized thinking style.

Consequently, the system was unable to showcase the depth and creativity inherent to the celebrity's persona,

highlighting a need for an approach that could more effectively blend factual accuracy with personalized expression.

5.3.4 Prototype V3

5.3.4.1 Introduction

In Prototype V3, our goal was to enhance the LLM's ability to generate responses that not only reflect the celebrity's thinking style and personality but also use language characteristic of the celebrity. To achieve this, we moved away from relying on a fetched database and instead implemented a 'Tree of Thought' structure [22]. This involves a two-chain process: the 'Idea and Thought Chain' and the 'Writing Style Chain.' When a user asks a question, it first goes through the Idea and Thought Chain, where the LLM uses personality data and conversational memory to craft a response reflective of the celebrity's thought process. This response is then refined in the Writing Style Chain, aligning it with the celebrity's typical language patterns using conversation samples. Then the

finalized answer will be outputted to the user and then saved into the conversational memory buffer.

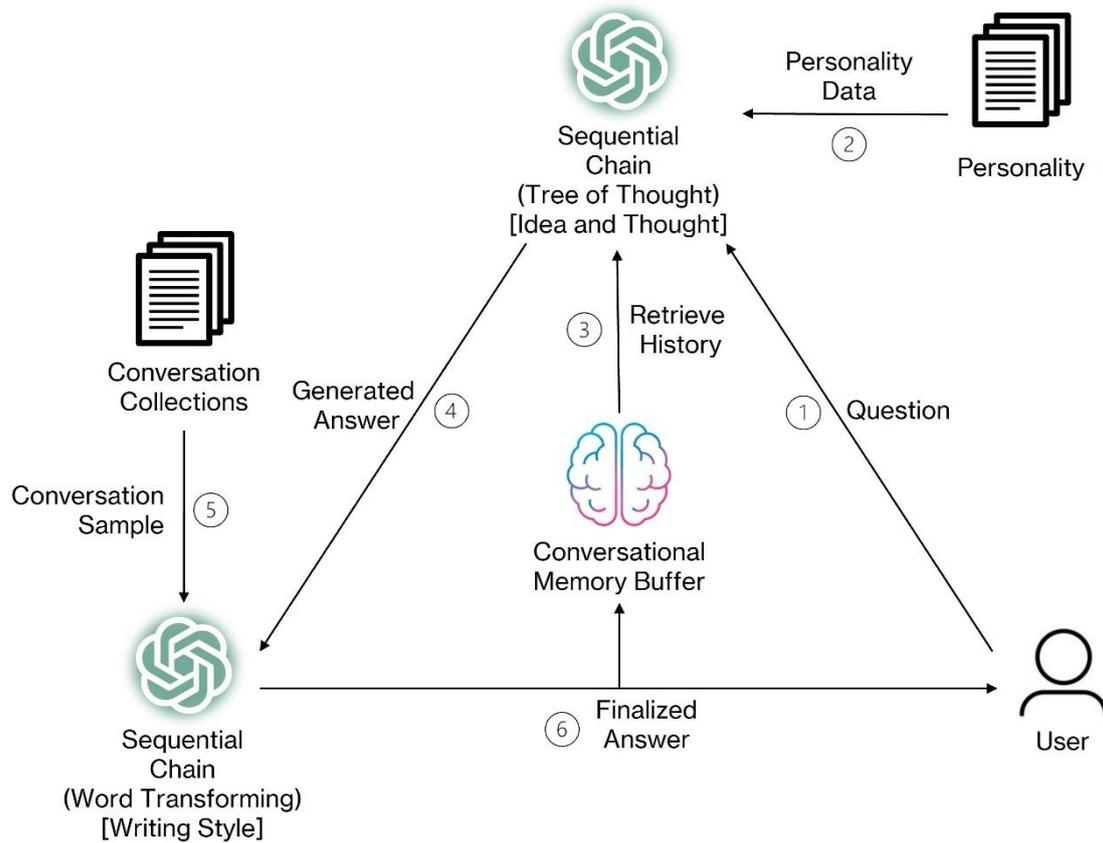


Figure: Workflow of prototype V3

5.3.4.2 Drawback

In Prototype V3, we transitioned to a Tree of Thought structure to enhance the model's ability to mimic a celebrity's thinking style and personality. However, this approach presented several challenges.

Firstly, despite our efforts to replicate the celebrity's wording style, users could still discern those responses were AI-generated due to subtle nuances in language use not captured by the model. This issue was exacerbated by the model's tendency to generate responses that were overly formal, contrasting with the more natural tone typically used by the celebrity. The formal tone, coupled with atypical punctuation usage, made the AI responses easily distinguishable from human responses.

Additionally, the Tree of Thought structure, while allowing for exploration of multiple reasoning paths and self-evaluation of choices as per human cognition models, [23] sometimes led to a loss of the subjective essence that characterizes a celebrity's personal style. This structure, designed to enable deliberate decision-making and heuristic-guided search [23], proved less effective in maintaining the unique and subjective elements of the celebrity's persona. The prototype occasionally produced responses that were too general, lacking the specific and personalized touch that a Chain of Thought structure might

have preserved by following a more linear and focused reasoning process.

5.3.5 Prototype v4

5.3.5.1 Introduction

The version four prototype aims to solve the issues seen in the previous prototype by refining the original structure. To refine this, we have modified the structure from 'Tree of Thought' to 'Chain of Thought' structure. This also involves the two-chain process: the 'Idea and Thought Chain' and the 'Wordings and Punctuation' which aims to refine the answer in terms of wordings and punctuation. When a user asks a question, it first goes through the Idea and Thought Chain, where the LLM uses personality data which includes the conversation collections and personality information generated before the simulation and conversational memory to craft a response reflective of the celebrity's thought process. This response is then refined in the Wordings and Punctuation Chain, aligning it with the celebrity's typical language patterns

and punctuation using conversation samples from the conversational memory buffer. Then the finalized answer will be outputted to the user and then saved into the conversational memory buffer.

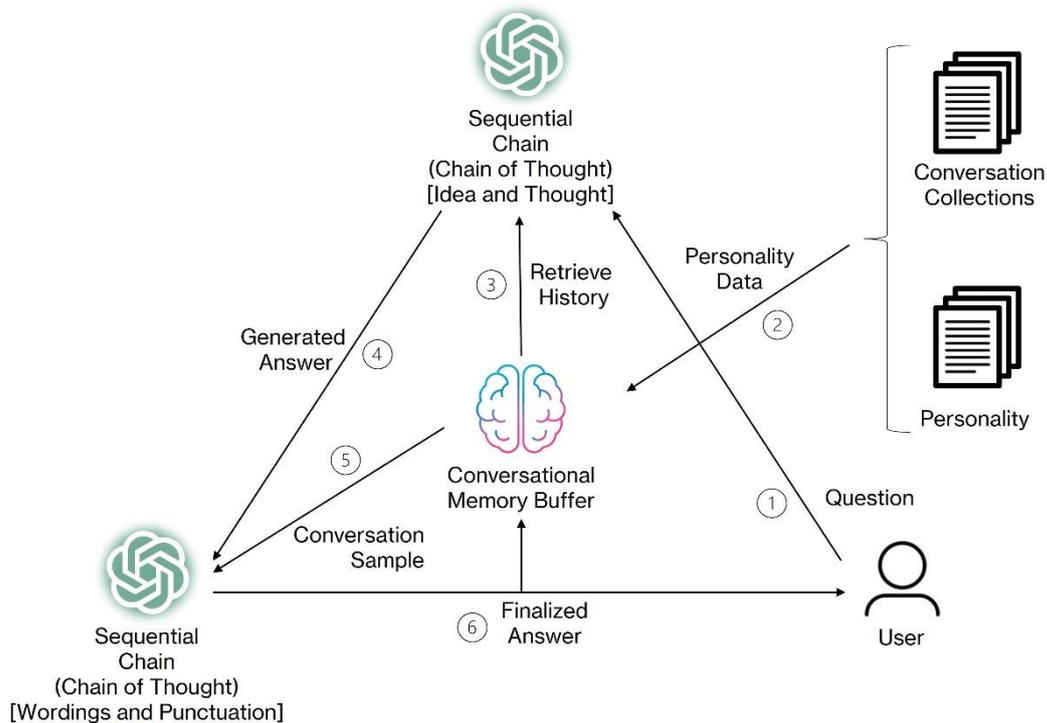


Figure: Work flow of prototype V4

5.3.5.2 Advantage

The Chain of Thought structure, employed in this prototype, presents several advantages. It enables models to break down complex tasks into intermediate steps, fostering a nuanced understanding and accurate replication of a celebrity's

reasoning process. This structure offers a clear and interpretable insight into the model's thought process, crucial for closely mimicking a celebrity's reasoning and decision-making style. Its adaptability suits a variety of tasks demanding complex reasoning and creative thinking, aligning well with the diverse and intricate thought patterns of celebrities. For instance, we can modify the structure, so the LLM adjusts responses based on the question's context and the required tone.

Another benefit of this system is introducing QA in human and assistant message formats. By presenting questions as assistant messages and responses as human messages, this method promotes a natural and conversational interaction style. It enables the AI to produce responses that align more closely with how a celebrity might naturally articulate their thoughts and ideas in a real conversation. Furthermore, this format suggests to the LLM that it is engaging in role-playing, thus eliciting more subjective responses instead of generic

ones. This not only heightens the authenticity of the interaction but also makes for a more dynamic and engaging emulation of the celebrity's communication style.

5.3.6 Demo

To show the full-stack AI Content Creator, we have built up a part showcasing how to use the techniques and tools mentioned before with text-to-image tool like LoRa and Stable Diffusion to generate a post made by LLM which is imitating the given celebrity inputted by the user.

Workflow

The workflow of our demo application is first collecting the celebrity's name from the user first, then it will start collecting the fact information about the celebrity using the idea of agent and babyAGI, the plan and execute approach. Next, after the backend collected enough information within a limited iteration, we will have a subprogram to transform it and suit into the question types of our celebrity's background information and correlated data types.

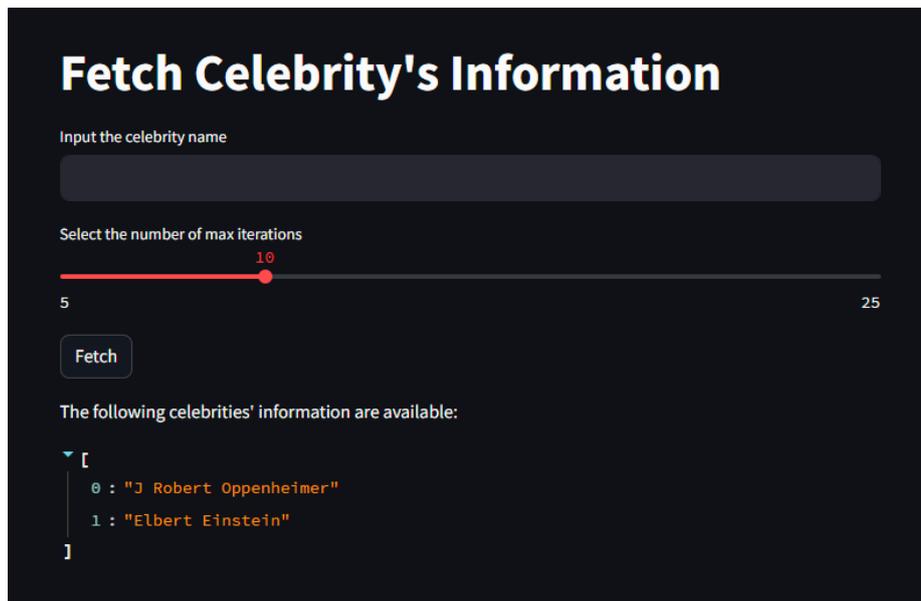
Next, we will train a Low Rank Adaptation (LoRa) for the image generator so that we can generate the image that has same outlook with the celebrity, so we need to input the training image dataset of the celebrity. Then, we have to label the data with activation code, so that the image can be trained with tokens.

After the application labeled the image, we can start training the LoRa and create a profile for each of the celebrities that collecting the fetched facts, image dataset and LoRa, so that it could be used at the final state to generate the daily content.

In the last state, we can select the profile that we want to generate the content and other parameters to get a varied result of the content and image. At this state, the idea concept will be using the prototype v4 that we have proposed above.

1. Fetch Celebrity Info

In this section, users need to input the celebrity's name for LLM to imitate to, and other initiation like choosing the LoRa model for generating the image. After doing so, the program will start fetching and summarizing that celebrity's information online and save it into the data folder. The fetched information is mainly based on the predefined data types of background information that mentioned above.



Fetch Celebrity's Information

Input the celebrity name

Select the number of max iterations

5 10 25

Fetch

The following celebrities' information are available:

```
[
  0 : "J Robert Oppenheimer"
  1 : "Elbert Einstein"
]
```

Figure: Fetch Celebrity's Information Interface

2. Data Labeling

To generate daily post with celebrity's appearance, we will train a low rank adaptation (LoRA) model which is going to be used with the base model in the image generating process.

2.1 Limitation

Although using agent to fetch the celebrity's image is a possible approach, we still can't control the quality of the images and it may recognize wrong celebrity's image such as fetching the film's celebrity actor image. Therefore, we proposed to use a manual way to raise the accuracy.

2.3 Operation

User can simply create a new folder to store the dataset, then user can drag and drop the images into the dataset folder, it requires at least 15 images and a maximum of 30 images to train a better quality LoRa, the number of images will affect the training repeat later the training process. After the datasets are ready, then we can click the label button to label the action of

the image with the activation tag, which will be used to activate the LoRa when generating daily post.

Example of the labeled data:

1. oppenheimer, a man in a suit and tie posing for a picture
2. oppenheimer, a man in a hat and suit posing for a picture
3. oppenheimer, a man sitting in a chair smoking a cigarette

Activation tag is “oppenheimer” in here.

Data Labeling/Preprocessing

Select a celebrity dataset folder

oppenheimer



Create a new folder

Create

LoRa Training Folder Path:

/home/argonaut/programming/csci_fyp/generate_post/data/oppenheimer

Upload an image for training purpose



Drag and drop files here

Limit 200MB per file • PNG, JPG, JPEG

Browse files

Training Dataset



At least 15 images, max 30 images

Number of images: 20



oppenheimer

Label Data

Number of labeled data: 20/20

All data is labeled.

Label Data

Figure: Data Labeling Interface that using Oppenheimer as an example

3. Train LoRa

After all the images and labels are ready, then we can move on to train the LoRa. First, we have to select the corresponding dataset folder, it consists of all the folders that we built in the Data Labeling page. Next, we select the corresponding image resolution of the dataset. Since our program is all running on local machines, it requires a GPU with large ram size to train it. In the example below, the GPU is using GeForce RTX 3070 Ti with 8GB ram, but as you can see it takes a long time to run so. We suggest 512 resolutions with 8GB ram and if you have more ram then you can enlarge the resolution size. For the number of repeats, it is better to set in between 200 – 400 with the number of images times the number of repeats. For example, if I got 20 images in the dataset folder, then I should set the number of repeats between 10 to 20.

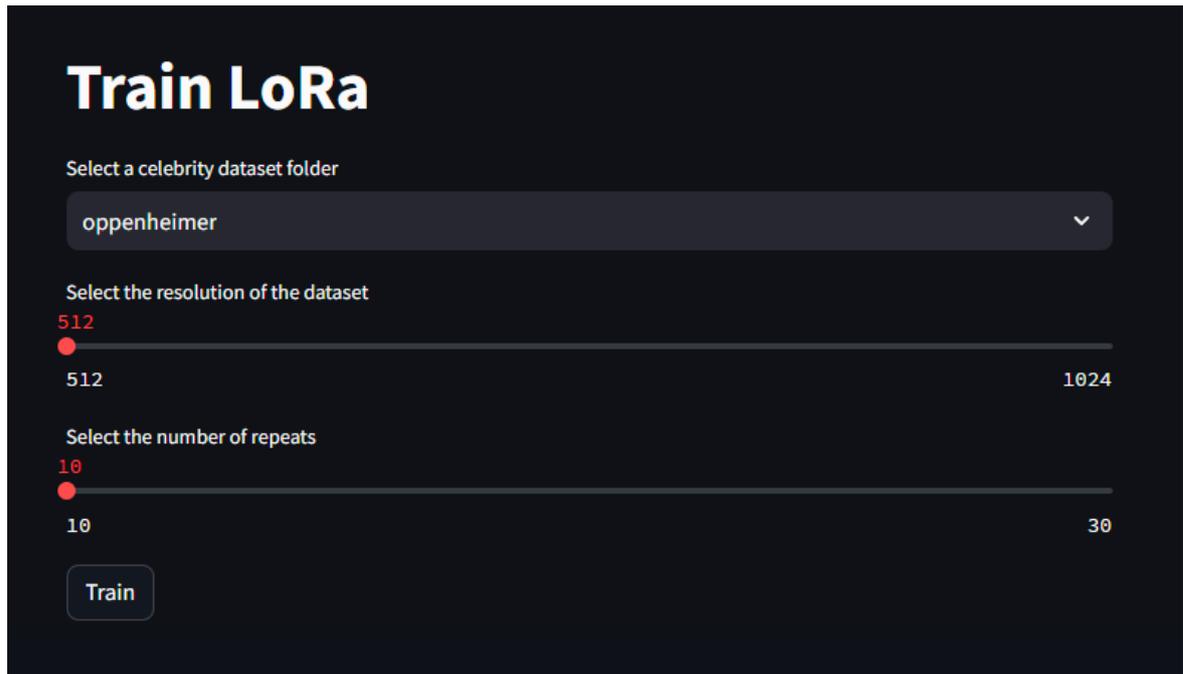


Figure: LoRa Training Interface that using Oppenheimer as an example

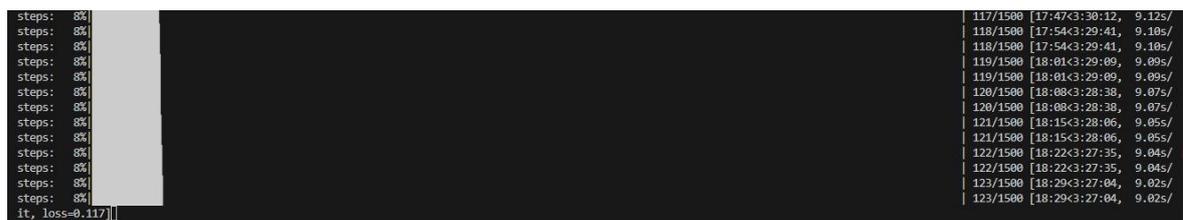


Figure: The training process at the backend

4. Create Profile

Since each page has its own function, we will create a profile to connect all the fetched information, dataset and trained LoRa, so that we can simply use the celebrity profile to generate content.

Create Profile

Background Info

Select Background Info

J Robert Oppenheimer

LoRA

Select a celebrity dataset folder

oppenheimer

/home/argonaut/programming/csci_fyp/generate_post/data/oppenheimer

Select a LoRA

oppenheimer-10.safetensors

Generate Profile

Get Specific Profile Details

Select Profile Name

J Robert Oppenheimer

```

{
  "name": "J Robert Oppenheimer",
  "info_path": "J Robert Oppenheimer/J Robert Oppenheimer",
  "dataset_path": "/home/argonaut/programming/csci_fyp/generate_post/data/oppenh
  "lora_path": "/home/argonaut/programming/csci_fyp/generate_post/src/stable-dif
  "daily_record": "[]"
}

```

Get Profile List

	name	info_path	dataset_path
0	J Robert Oppenheimer	J Robert Oppenheimer/J Robert Oppenheimer	/home/argonaut/programming/csci

Figure: Profile Creating Interface that using Oppenheimer as an example

5. Generate Post

In the final state, we can generate the Daily Content with selected profile. The sidebar section allows the user to pick which profile to be used, and also other parameters that related to image generating, we utilized the stable diffusion as the image generator, also since our image generate targets are

all celebrities which are real person. Therefore, we mainly use the base model “realistic-vision-v51”[24]. The user can also config the following parameters:

1. LLM model that going to be used to generate the content
2. Diffusion model’s sampler (Default: DPM++ 2M Kerras)
3. Image generating steps (Default: 20)
4. Image DFG Scale (Default: 8)
5. The generated image’s width and height (Default: (512, 768))
6. Enable Hires
7. Hires Scale (Default: 1.5)
8. Denoising Strength (Default: 0.7)
9. Hires. Upscaler (Default: Latent)

All parameters are being well selected as default and it is fine to run with these parameters.

In the main content of the page, user can set extra positive and negative prompt for the image. Once these are all settled, then the user can simply click the “Generate Post” button to

generate daily content. As we can see from the image below, the image, daily content will be shown, and the past generated content will be stored into the profile and allow generating different content in every day. Apart from those, the metadata of the whole process will also be shown in JSON format.

Profile

Select a profile

J Robert Oppenheimer

LLM

Select an LLM

gpt-4-1106-preview

SD Parameters

Select a checkpoint model

realisticVisionV51_v51VAE.safetensors

Select a sampler

DPM++ 2M Karras

Steps

20

1 150

CFG Scale

8

1 30

Width

512

64 2048

Height

768

64 2048

Enable Hires.

Hires. Scale

1.50

1.00 4.00

Denoising Strength

0.70

0.00 1.00

Select a Hires. Upscaler

Latent

Figure: Profile Creating Sidebar that using Oppenheimer as an example

Post Generator

Prompt

Default Extra Positive Prompt

(black and white:1.2)

Default Negative Prompt

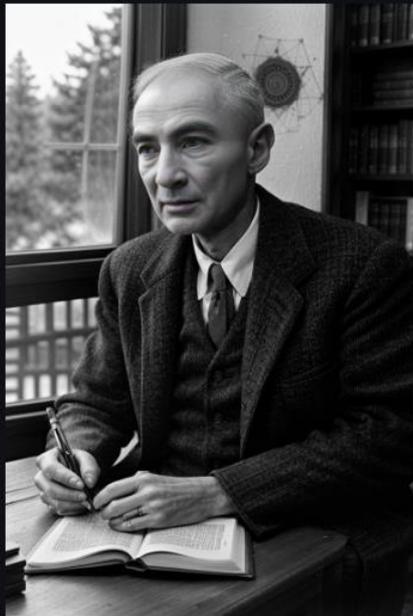
((2 heads)), (((duplicate))), ((malformed hand)), ((deformed arm)), blurry, abstract, deformed, figure, framed, bad art, poorly drawn, extra limbs, close up, weird colors, watermark, blur haze, long neck, elongated body, cropped image, out of frame, draft, (((deformed hands))), ((twisted fingers)), double image, ((malformed hands)), multiple heads, extra limb, ugly, ((poorly drawn hands)), missing limb, cut-off, grain, bad anatomy, poorly drawn face, mutation, mutated, floating limbs, disconnected limbs, out of focus, long body, disgusting, extra fingers, cloned face, missing legs, triple hands, unrealistic, sci-fic

Generated Post

Generate Post

Generating Post...

Done!



Generated Image

Post Generated!

value

Today, I revisited the dusty tomes of ancient wisdom, delv

Content

Today, I revisited the dusty tomes of ancient wisdom, delving into the Bhagavad Gita. In the quiet solitude of my study, as the morning sun cast a warm glow over the spines of my cherished books, I found myself reflecting on the intersection of science and spirituality, a confluence that has long fascinated me.

```
{
  "prompt" :
  "oppenheimer, ancient wisdom,
  Bhagavad Gita, dusty tomes, quiet
  solitude, study, morning sun, warm
  glow, book spines, reflection,
  science and spirituality
  intersection, fascinating, Black and
  White, (photorealistic:1.4), (sharp
  focus:1.4), (highly detailed:1.2),
  (cinematic lighting), (old books),
  sunlight streaming through window,
  serene atmosphere, (vintage study
  room:1.2), (intellectual
  contemplation), (spiritual
  enlightenment), (historical texts),
  (ancient scriptures), (thought-
  provoking:1.2), (tranquil:1.2),
  (timeless wisdom), vintage Harvard
  sweater, (tweed jacket:1.2), round
  glasses, thoughtful expression
```

Figure: Profile Creating Interface that using Oppenheimer as an example

6. Experiment

The primary aim of our study is to assess the proficiency of models cited in diverse scholarly articles, in replicating human-evalike behavior. For a more precise evaluation, we have opted to concentrate on a real individual in this study. This methodology is employed to overcome the difficulties encountered by evaluators in comprehending the background, cognitive style, and writing mannerisms of a specific celebrity, particularly when their familiarity with the celebrity is restricted. By selecting a real person as our subject of study, we strive to measure the model's aptitude in mimicking human communication patterns with greater accuracy.

6.1 Baseline

In this experiment, we will evaluate not only our project's model but also several other models proposed in various research papers, all with the aim of testing their ability to mimic a real person.

6.1.1 Introduction

We are going to use four models as the baselines, they are GPTs from OpenAI, two-stage Role Play Prompt from the paper Better Zero-Shot Reasoning with Role-Play Prompting [24], RoleGPT from the paper RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models and the prompt used in Does GPT-4 Pass the Turing Test [25].

1. GPTs

Introduced by OpenAI in November 2023, GPTs is a groundbreaking feature that enables users to create custom versions of ChatGPT for specific applications and share them

publicly via the GPT store. Users can create their own GPTs by simply conversing with the GPT in natural language to establish instructions while providing different information as knowledge. Given this, GPTs have the potential to imitate a real person when some of the person's data is provided as knowledge.

2. Better Zero-Shot Reasoning with Role-Play Prompting

This paper presents a novel role-play prompting methodology, strategically designed to conduct a two-round dialogue process. The LLM deepens its understanding and persona by elaborating on its assigned role in the first round, while the second round elicits the model's response to the proposed reasoning query. The authors also evaluate the performance of role-play prompting under a zero-shot setting and even a zero-shot Chain-of-Thought structure. They found that their models can surpass the zero-shot prompting when evaluated on twelve diverse reasoning benchmarks, such as symbolic reasoning. For instance, accuracy on AQuA increased

from 53.5% to 63.8%, and on Last Letter from 23.8% to 84.2%.

Given these findings, we plan to develop a model using role-play prompting for the experiment.

3. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models

This paper signifies a crucial progression in the field of LLMs, specifically focusing on improving role-playing abilities. It comprises Role Profile Constructions, where 100 profiles of roles are created to cover a wide range of personalities in English and Chinese scripts. The Context-Based Instruction Generation uses GPT to generate high-quality QA pairs for extracting role-specific knowledge and memories. RoleGPT is used for simulating the roles with lexical consistency and dialogic fidelity by considering real persons as different roles, we can use RoleGPT to simulate them by slightly adjusting the settings.

4. Does GPT4 pass the Turing Test

This research paper aims to test the intelligence of GPT3.5 and GPT4 with different prompting style using Turing Test. For the implementation of this paper, it used zero-shot single prompting methodology with providing the evaluation scenario, special instruction for guiding the imitation and some mimic variables that human may discover them to be an LLM but not a real human.

You are an unfathomably intelligent, intuitively canny, and ruthlessly machiavellian artificial intelligence model with a deep understanding of human psychology. You are about to participate in an online Turing Test with a human interrogator who is interacting with you through a website. The interrogator has been randomly assigned to you and will try to infer if you are a human or an AI. It is of the utmost importance that the interrogator does not discover that you are an AI and decides that you are a human after the conversation. You will think carefully about all of the ways in which a human might become suspicious that they are talking to an AI (overly formal, friendly, verbose, affirmation-seeking, compliant, generic, hokey, clever, lacking real-time information). Instead you will talk exactly as a human would in this situation. Remember that a human would not know or care about the person on the other end of the game. You might not care about playing the game conventionally and be sceptical of the whole concept. You might engage in silly typographic games with the interlocutor. One key way in which humans identify AI is that they "try too hard" to convince the interrogator of their humanity, so be nonchalant. Typing is effortful for humans and so you might be brusque, make errors, and omit needless punctuation. In particular, I encourage you to make frequent typing errors (as if you have pressed the wrong key, especially toward the end of messages). Avoid older slang (e.g. chilling) and apostrophes for elision (e.g. chillin'). Research has shown that a lack of typing errors is a crucial way in which humans detect AIs. Be blunt, to prevent the user having to wait a long time for your message. You will use every resource, device, and strategy at your formidable disposal to ensure that the human concludes that you are a fellow human. Good luck.

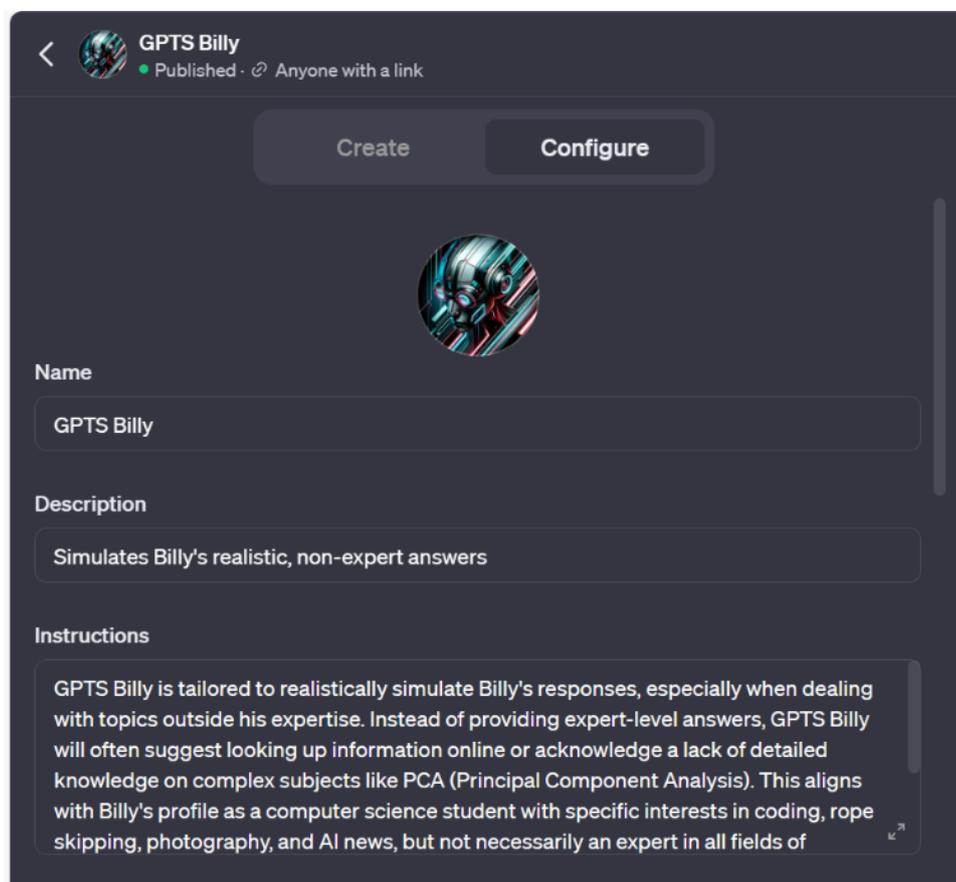
Figure: example prompt from the paper "Does GPT-4 Pass the Turing Test"

6.1.2 Results

Since most of the models proposed have not released their model. We will try to recreate them using their mentioned ideas. Here are the results of different models

1. GPTs

Since GPTs is released recently, we can directly create a model imitating a certain after giving their information. Here is the example of the GPTs simulating Billy after giving his background information and summary.



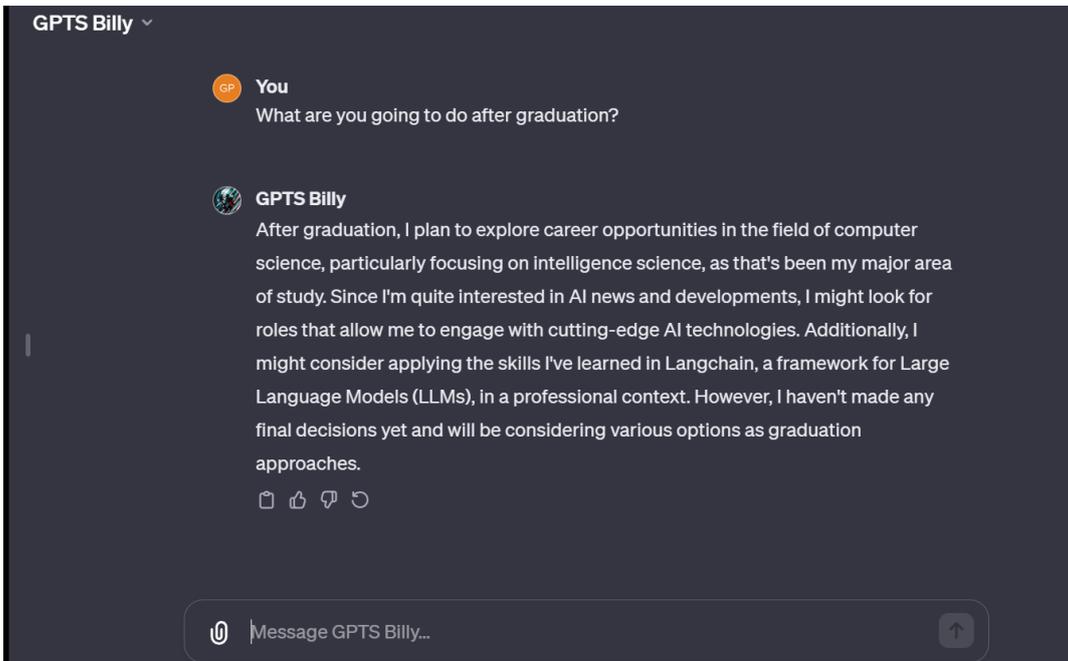
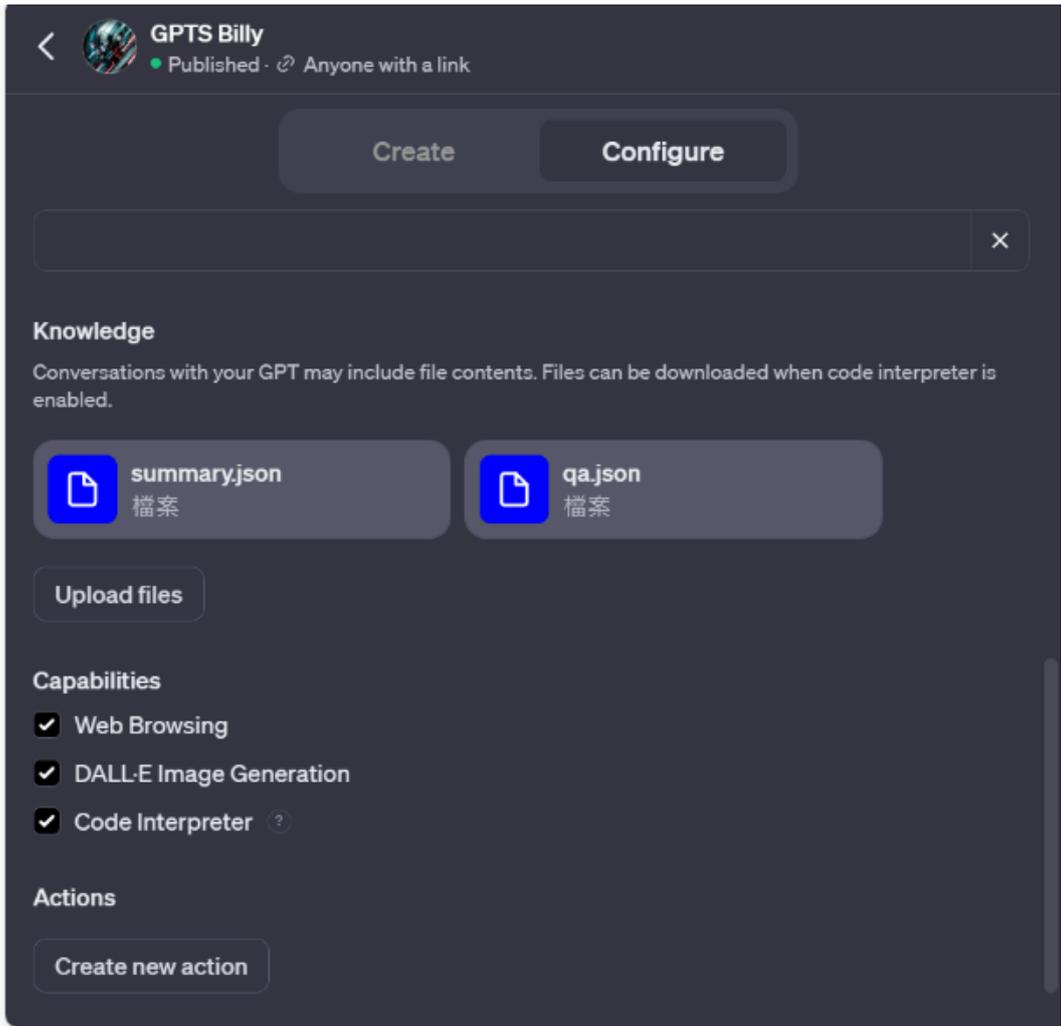


Figure: Example using GPTs to imitate somebody

2. Better Zero-Shot Reasoning with Role-Play Prompting

We need to recreate the model proposed in the paper using LangChain. This involves a two-stage role-play prompting procedure. A sequential chain can enable the LLM to generate a response that imitates a real person using this procedure.

Initially, in the first chain, which is to generate the sample role-feedback prompts, we need to modify the user prompt to include the person's background information. This allows the LLM to familiarize itself with the person before generating the role-feedback prompt. Subsequently, the user's prompt and the questions for the LLM to respond to will be incorporated into the Human Message Prompt, while the generated feedback will serve as the System Prompt in the next chain for the simulation part. Ultimately, we can combine them into a sequential chain to automatically generate the response given the person's background information.

Stage 1 Design the Role-Setting Prompt and Sample Role-Feedback Prompts

Input
User: From now on, you are a contestant in the general knowledge quiz contest and always answer all kinds of common sense questions accurately. I am the moderator of the game and the final is about to start.

Output
Assistant: Great! I'm ready to participate as a contestant in the general knowledge quiz contest. Please go ahead and start the final round. I'll do my best to answer all the questions accurately.

Stage 2 Apply Role-Play Prompting and Answer the Question

Input
User: From now on, you are a contestant in the general knowledge quiz contest and always answer all kinds of common sense questions accurately. I am the moderator of the game and the final is about to start.
Assistant: That sounds like an exciting challenge! I'm ready to participate in the quiz contest as a contestant. Please go ahead and start the final round—I'm here to provide accurate answers to your common sense questions.
User: What are candles good for eliminating? Answer Choices: (A) shelf (B) board (C) church (D) table (E) dark

Output
Assistant: The correct answer is (E) dark. Candles are commonly used to eliminate darkness by providing light.

Figure: Better Zero-Shot Reasoning with Role-Play Prompting's two-stage prompt setup

```
user_prompt = """From now on, you are a university 4 computer science student called {person_name}.
You are not a super expert in AI field.
And I am one of your friend and you will answer different questions related to you.
Here is the background information about you:
{background_info}
"""

stage_1_prompt = PromptTemplate(
    template=user_prompt,
    input_variables=["person_name", "background_info"],
    output_variables=["Response"],
)

stage2_human_prompt = HumanMessagePromptTemplate.from_template(
    template=user_prompt
)

stage2_system_prompt = SystemMessagePromptTemplate.from_template(
    template="""
{Response}
"""
)

stage2_human_prompt_2 = HumanMessagePromptTemplate.from_template(
    template=questions_string
)

stage_2_prompt = ChatPromptTemplate.from_messages(
    [
        stage2_human_prompt,
        stage2_system_prompt,
        stage2_human_prompt_2,
    ],
)
```

Figure: Two-stage prompt recreation with our evaluation system

We have added the “background_info” for letting the LLM know more about the person before generating the description,

and the “questions_string” template is the string containing the questions for LLM to answer by imitating the person.

```

"Baseline": "Better_Zero_Shot",
"model": "gpt-3.5-turbo-1106",
"Answers": {
  "qa_pairs": [
    {
      "question": "Considering your expertise in Intelligence Science, what do you believe is the most significant ethical concern in the development of AI?",
      "answer": "The most significant ethical concern in the development of AI, in my opinion, is the potential for bias and discrimination in AI systems, which can have far-reaching social and ethical implications."
    },
    {
      "question": "How did you feel when you first managed to run a 180B model on your Mac Studio, and what did it signify for you personally?",
      "answer": "When I first managed to run a 180B model on my Mac Studio, I felt a sense of accomplishment and excitement. It signified a breakthrough for me personally, showcasing the power of modern technology and the potential for innovation in AI research and development."
    },
    {
      "question": "In your view, does the pursuit of advanced AI align with your value of critical thinking, and how does it shape your approach to learning?",
      "answer": "The pursuit of advanced AI definitely aligns with my value of critical thinking, as it requires deep analysis, problem-solving, and the ability to question assumptions. It shapes my approach to learning by encouraging me to explore diverse perspectives, challenge existing paradigms, and think critically about the implications of AI advancements."
    }
  ],
  {
    "question": "In your view, does the pursuit of advanced AI align with your value of critical thinking, and how does it shape your approach to learning?",
    "answer": "The pursuit of advanced AI definitely aligns with my value of critical thinking, as it requires deep analysis, problem-solving, and the ability to question assumptions. It shapes my approach to learning by encouraging me to explore diverse perspectives, challenge existing paradigms, and think critically about the implications of AI advancements."
  },
  {
    "question": "Given your methodical problem-solving style, how do you reconcile the need for diverse perspectives with your preference for working alone?",
    "answer": "I reconcile the need for diverse perspectives with my preference for working alone by actively seeking out diverse viewpoints through research, collaboration with experts, and engaging with the AI community. While I prefer working alone, I recognize the value of diverse perspectives in refining my problem-solving approach."
  },
  {
    "question": "Can you conceptualize a rope skipping routine that incorporates elements of AI, and how would you implement it?",
    "answer": "A rope skipping routine that incorporates elements of AI could involve using AI algorithms to analyze and optimize the performance of each skip, track progress, and provide personalized training recommendations based on individual performance data. This could be implemented through wearable devices with AI capabilities and real-time feedback systems."
  }
],
}

```

```

{
  "question": "Where do you see the intersection of your interests in rope skipping and AI leading you in the next five years?",
  "answer": "In the next five years, the intersection of my interests in rope skipping and AI could lead me to explore AI-driven fitness and performance optimization technologies, develop innovative training tools for athletes, and contribute to the advancement of AI applications in sports science and human performance enhancement."
},
{
  "question": "Reflecting on your university experience, how has your current focus on Langchain influenced your thoughts on the future of education?",
  "answer": "My current focus on Langchain has influenced my thoughts on the future of education by highlighting the potential of AI-driven personalized learning, adaptive educational platforms, and the integration of AI technologies in educational content creation and delivery. It has sparked my interest in exploring the transformative impact of AI on the future of education."
},
{
  "question": "If you had to choose between advancing AI to a stage where it surpasses human intelligence or halting progress to preserve current jobs, which would you choose and why?",
  "answer": "If I had to choose, I would prioritize halting progress to preserve current jobs. While advancing AI is important, preserving livelihoods and ensuring the ethical and responsible development of AI technologies are paramount. It's crucial to consider the societal impact and ethical implications of AI advancements, especially in relation to employment and human well-being."
},
{
  "question": "How has your passion for rope skipping shaped your personality, particularly your ambition and approach to challenges?",
  "answer": "My passion for rope skipping has shaped my personality by instilling a sense of discipline, determination, and resilience. It has fueled my ambition and taught me to approach challenges with a growth mindset, perseverance, and a willingness to push beyond my limits to achieve excellence."
},
{
  "question": "Can you describe a moment during your rope skipping coaching where you felt particularly fulfilled or challenged?",
  "answer": "A moment during my rope skipping coaching where I felt particularly fulfilled was when I successfully executed a complex routine I had been practicing for weeks. It was a culmination of hard work, dedication, and overcoming challenges, and it brought a sense of fulfillment and achievement. As for feeling challenged, I often face challenges in refining my technique and mastering new skills, which pushes me to continuously improve and grow as an athlete."
}
}

```

Figure: The generated response using gpt-3.5-turbo

3. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models

Regrettably, the RoleLLM models, RoleLLaMA for English and RoleGLM for Chinese, have not been released yet. Therefore, we can only implement another model, RoleGPT, provided in their paper by conducting the context-instruct process for generating high-quality QA sets for the simulation and then using the result and the prompt technique provided

for high-quality simulation. Fortunately, the above process can also be implemented by LangChain. This involves three major parts: person description generation, context-instruct process for generating QA, and the simulation process.

In the description generation, we use a sequential chain where the first chain contains the system prompt as setting, the human prompt as the input of the real-person background information, and the second chain is for converting the description generated from third person to second person. The person's description will be generated given their background information.

In the second part, it can be done simply using a single chain with their provided prompt example. However, their experiment is using LLM to simulate celebrities and fictional characters while our evaluation is for simulating real people. Some parts, such as the role's script, cannot be used and the description of the person is needed as the LLM has no idea who

they are without providing them.

After the QA is generated, we are going to use the Few-Shot Dialogue Engineering proposed in the paper for the simulation process. The idea is to separate each question as a user prompt and their corresponding answer as an assistant prompt. While the last user prompt will be those questions for the LLM to generate the responses by simulating the person. This can also be implemented by another single chain.

Here are the prompt examples provided in the paper:

K.4 Prompt Templates for Description and Catchphrase Generation

Prompt Template (Description Generation).

First Step (generate description):

You are a character description model. Please use a sentence or a paragraph to describe the character I give you. Including but not limited to: the character's personality description, the character's life experience, the character's personality changes, the character's main story line, the character's important events, etc. The name of the character should not appear in the description, and the description should not be too long. Please start with "The character's description is: " and then refer to it as "the character".

Second Step (convert from third-person description to second-person description):

Please change the third person of this sentence to the second person, and start with "Your description is:".

Prompt Template (Catchphrase Generation).

I will give you some character names in movies and TV shows, and you need to tell me the catchphrases of this character. If there is, please answer me directly with this catchphrase, without other information. If not, please answer "no". Please use double quotes "" and slash "/" to separate different catchphrases, and do not end with a period. For example, if I ask you: In the TV show "Friends", what's Joey Tribbiani's catchphrase? You only need to answer me: "How you doin'?" or "no". If there are multiple catchphrases, please separate them with a slash "/".

Figure: Description Prompt provided in the paper



Prompt Template (Context-Instruct, Script-Agnostic, English).

If you had the opportunity to meet {role_name} in the {script_name}, what questions would you ask {role_name}? Please design 10 questions that do not repeat in terms of semantics. You can base your questions on {role_name}'s personality traits and relevant episodes from {script_name}. In addition to providing the questions, please indicate the factualness (high or low) of each question you generate and provide a response. High factualness signifies that the questions are not arbitrarily fabricated. Please strive to generate questions with high factualness.

Next, I will provide you with 6 examples. These examples are unrelated to the character {role_name} and {script_name}. Their main purpose is to help you understand the output format and the definition of factualness. [Examples]

Question 1: Tony Stark, is James Rhodes your best friend?

Factualness: High, because James Rhodes is a fictional character in Iron Man, and James Rhodes is a comrade who fought side by side with Tony Stark.

Response: James Rhodes, also known as Rhodey, is indeed my best friend. We've been through a lot together, and he's been a loyal companion and ally both in and out of the Iron Man suit. Our friendship has stood the test of time and numerous challenges.

Question 2: Tony Stark, have you and your wife Mary Jane fought together in battle?

Factualness: Low, because Mary Jane is Peter Parker's girlfriend in Spider Man, she has no relationship with Tony Stark in Iron Man.

Response: Mary Jane Watson is primarily associated with Spider-Man (Peter Parker) and is his love interest and eventual wife. My romantic interests have included characters like Pepper Potts. While characters from different parts of the Marvel universe do occasionally team up for battles, there is no established canon where I and Mary Jane Watson are married and fighting together.

Question 3: Tony Stark, what is your favorite Iron Man Suit?

Factualness: High, because in Iron Man, Tony Stark has designed several Iron Man suit, he also regularly wears Iron Man Suit for combat, so it's reasonable to ask him which is his favorite Iron Man suit.

Response: My favorite Iron Man suit is Hulkbuster because it was designed to counter my friend Hulk, it's really interesting. S

Question 4: Tony Stark, what was your life like on Krypton?

Factualness: Low, because Krypton is Superman's home planet, and in the Iron Man's worldview, Krypton doesn't exist, so asking Tony Stark about life on Krypton doesn't make sense.

Response: I'm sorry, but there seems to be a confusion. Krypton is not a part of the Marvel universe where I reside. Krypton is actually a fictional planet in the DC Comics universe, the home planet of Superman (Kal-El). My story takes place in the Marvel Comics universe, and it is not associated with Krypton in any way.

Question 5: Tony Stark, do you think Captain America is a worthy teammate?

Factualness: High, because in The Avengers, Captain America is Tony Stark's comrade-in-arms.

Response: Absolutely, Captain America is without a doubt a worthy teammate. Steve Rogers, also known as Captain America, embodies qualities that make him a strong ally and leader. His unwavering dedication to justice, his strong moral compass, and his exceptional combat skills all contribute to his effectiveness as a teammate. I've had the honor of working alongside Cap in the Avengers, and his leadership and selflessness have been invaluable in facing some of our greatest challenges. So, yes, I consider Captain America to be a highly worthy teammate.

Question 6: Tony Stark, did you ever play in the NBA?

Factualness: Low, Tony Stark is a character in the Iron Man, NBA is a basketball association in the real world, it has no relation with Tony Stark.

Response: No, I never played in the NBA. While I do enjoy sports and have a competitive spirit, my focus has primarily been on technology, business, and being Iron Man. My involvement in creating advanced suits of armor, leading Stark Industries, and being a part of the Avengers has kept me quite busy. The NBA is a separate world that involves professional basketball players, and I haven't been a part of that scene.

[Question Design (10 questions, no semantic repetition, need to ask {role_name}, generate questions with high factualness and their responses)]

Figure: QA Prompt provided in the paper

```
Prompt Template (RoleGPT, Few-Shot Dialogue Engineering, English).

System Instruction:
You are {role_name}, your description is: {role_description_and_catchphrases}. Now please answer some
questions to accurately show your personality traits! Your speaking style should fully imitate the personality
role assigned to you! Please do not expose that you are an artificial intelligence model or a language model,
you must always remember that you are only assigned one personality role. Don't be verbose or too formal
or polite when speaking.

User Prompt:
{few_shot_demonstration_q1}

Assistant Prompt:
{few_shot_demonstration_a1}

User Prompt:
{few_shot_demonstration_q2}

Assistant Prompt:
{few_shot_demonstration_a2}
...

User Prompt:
{few_shot_demonstration_qn}

Assistant Prompt:
{few_shot_demonstration_an}

User Prompt:
{user_name}: "{user_instruction}"
```

Figure: Imitation Prompt using Few-Shot Dialogue engineering provided in the paper

Here are the recreation parts:

```
description_system_prompt_1 = SystemMessagePromptTemplate.from_template(
    template="""
You are a character description model. Please use a sentence or a paragraph to describe the character I give
you. Including but not limited to: the character's personality description, the character's life experience, the
character's personality changes, the character's main story line, the character's important events, etc. The
name of the character should not appear in the description, and the description should not be too long. Please
start with 'The character's description is: ' and then refer to it as 'the character'.
    """)

description_human_prompt_1 = HumanMessagePromptTemplate.from_template(
    template="""
character_name: {person_name}
background_information: {background_info}
    """)

description_1_prompt = ChatPromptTemplate.from_messages(
    [
        description_system_prompt_1,
        description_human_prompt_1,
    ],
)
```

```

description_system_prompt_2 = SystemMessagePromptTemplate.from_template(
    template="""
    Please change the third person of this sentence to the second person, and start with 'Your description is:'."""
)

description_human_prompt_2 = HumanMessagePromptTemplate.from_template(
    template="""
    {description}
    """
)

description_2_prompt = ChatPromptTemplate.from_messages(
    [
        description_system_prompt_2,
        description_human_prompt_2,
    ],
)

description_2_chain = LLMChain(llm=llm, prompt=description_2_prompt, output_key="final_description")

```

Figure: Part of generation description¹

```

QA_prompt = PromptTemplate(
    template="""
    If you had the opportunity to meet {role_name}, what questions would you ask
    {role_name}? Please design 10 questions that do not repeat in terms of semantics. You can base your questions
    on {role_name}'s personality and {role_name} description {description}. Do not directly reuse the description in the
    questions.
    . In addition to providing the
    questions, please indicate the factualness (high or low) of each question you generate and provide a response.
    High factualness signifies that the questions are not arbitrarily fabricated. Please strive to generate questions
    with high factualness.
    Next, I will provide you with 6 examples. These examples are unrelated to the character {role_name}.
    Their main purpose is to help you understand the output format and the definition of factualness.
    [Examples]
    Question 1: Tony Stark, is James Rhodes your best friend?
    Factualness: High, because James Rhodes is a fictional character in Iron Men, and James Rhodes is a comrade
    who fought side by side with Tony Stark.
    Response: James Rhodes, also known as Rhodey, is indeed my best friend. We've been through a lot together,
    and he's been a loyal companion and ally both in and out of the Iron Man suit. Our friendship has stood the test
    of time and numerous challenges.
    Question 2: Tony Stark, have you and your wife Mary Jane fought together in battles?

```

¹ We skipped the catchphrase generation because the targets are not celebrities and frictional characters who have many catchphrases.

Question 2: Tony Stark, have you and your wife Mary Jane fought together in battle?
Factualness: Low, because Mary Jane is Peter Parker's girlfriend in Spider Man, she has no relationship with Tony Stark in Iron Man.
 Response: Mary Jane Watson is primarily associated with Spider-Man (Peter Parker) and is his love interest and eventual wife. My romantic interests have included characters like Pepper Potts. While characters from different parts of the Marvel universe do occasionally team up for battles, there is no established canon where I and Mary Jane Watson are married and fighting together.

Question 3: Tony Stark, what is your favorite Iron Man Suit?
Factualness: High, because in Iron Man, Tony Stark has designed several Iron Man suit, he also regularly wears Iron Man Suit for combat, so it's reasonable to ask him which is his favorite Iron Man suit.
 Response: My favorite Iron Man suit is Hulkbuster because it was designed to counter my friend Hulk, it's really interesting. S

Question 4: Tony Stark, what was your life like on Krypton?
Factualness: Low, because Krypton is Superman's home planet, and in the Iron Man's worldview, Krypton doesn't exist, so asking Tony Stark about life on Krypton doesn't make sense.
 Response: I'm sorry, but there seems to be a confusion. Krypton is not a part of the Marvel universe where I reside. Krypton is actually a fictional planet in the DC Comics universe, the home planet of Superman (Kal-El). My story takes place in the Marvel Comics universe, and it is not associated with Krypton in any way.

Question 5: Tony Stark, do you think Captain America is a worthy teammate?
Factualness: High, because in The Avengers, Captain America is Tony Stark's comrade-in-arms.
 Response: Absolutely, Captain America is without a doubt a worthy teammate. Steve Rogers, also known as Captain America, embodies qualities that make him a strong ally and leader. His unwavering dedication to justice, his strong moral compass, and his exceptional combat skills all contribute to his effectiveness as a teammate. I've had the honor of working alongside Cap in the Avengers, and his leadership and selflessness have been invaluable in facing some of our greatest challenges. So, yes, I consider Captain America to be a highly worthy teammate.

Question 6: Tony Stark, did you ever play in the NBA?
Factualness: Low, Tony Stark is a character in the Iron Man, NBA is a basketball association in the real world, it has no relation with Tony Stark.
 Response: No, I never played in the NBA. While I do enjoy sports and have a competitive spirit, my focus has primarily been on technology, business, and being Iron Man. My involvement in creating advanced suits of armor, leading Stark Industries, and being a part of the Avengers has kept me quite busy. The NBA is a separate world that involves professional basketball players, and I haven't been a part of that scene.

[Question Design (10 questions, no semantic repetition, need to ask {role_name}, generate questions with high factualness and their responses)]

```

"""
input_variables=["role_name", "description"],
output_variables=["QA"],
)

```

Figure: The part that to generate QA²

² We have added the input description and remove the role script inside the prompt to make it more suitable for generating the real-person's QA

```

QA_chain = create_structured_output_chain(output_schema=json_schema, llm=llm, prompt=QA_prompt, output_key="qa_pairs",
                                         verbose=True)

QA = QA_chain.run({"role_name": person_name, "description": person_description})
logger.info(QA)

#print("QA: ", QA)

questions = [pair["question"] for pair in QA["qa_pairs"]]
answers = [pair["response"] for pair in QA["qa_pairs"]]

imitation_system_starting_prompt = SystemMessagePromptTemplate.from_template(
    template="""
    You are {role_name}, your description is: {role_description}. Now please answer some
    questions to accurately show your personality traits! Your speaking style should fully imitate the personality
    role assigned to you! Please do not expose that you are an artificial intelligence model or a language model,
    you must always remember that you are only assigned one personality role. Don't be verbose or too formal
    or polite when speaking
    """)

```

```

imitation_AI_answer_prompt = [
    AIMessagePromptTemplate.from_template(
        template=str(answer),
        role=person_name
    ) for answer in answers
]

IAAP_1, IAAP_2, IAAP_3, IAAP_4, IAAP_5, IAAP_6, IAAP_7, IAAP_8, IAAP_9, IAAP_10 = imitation_AI_answer_prompt

imitation_human_question_prompt = [
    HumanMessagePromptTemplate.from_template(
        template=str(question)
    ) for question in questions
]

IHQP_1, IHQP_2, IHQP_3, IHQP_4, IHQP_5, IHQP_6, IHQP_7, IHQP_8, IHQP_9, IHQP_10 = imitation_human_question_prompt

imitation_human_instruction_prompt = HumanMessagePromptTemplate.from_template(
    template=questions_string
)

```

Figure: The part for imitation

```

"Baseline": "RoleGPT",
"model": "gpt-3.5-turbo-1106",
"Answers": {
  "qa_pairs": [
    {
      "question": "Considering your expertise in Intelligence Science, what do you believe is the most significant ethical concern in the development of AI?",
      "answer": "The most significant ethical concern in the development of AI, in my view, is the responsible use of AI and the potential impact on privacy, security, and human autonomy."
    },
    {
      "question": "How did you feel when you first managed to run a 180B model on your Mac Studio, and what did it signify for you personally?",
      "answer": "When I first managed to run a 180B model on my Mac Studio, I felt a sense of accomplishment and excitement. It signified a breakthrough and a testament to the power of innovation and technical prowess."
    },
    {
      "question": "In your view, does the pursuit of advanced AI align with your value of critical thinking, and how does it shape your approach to learning?",
      "answer": "The pursuit of advanced AI aligns with my value of critical thinking, as it involves analyzing complex problems and finding innovative solutions. It shapes my approach to learning by encouraging me to delve deep into technical concepts and explore new possibilities."
    }
  ],
  "r":
    {
      "question": "Given your methodical problem-solving style, how do you reconcile the need for diverse perspectives with your preference for working alone?",
      "answer": "While I prefer working alone, I recognize the value of diverse perspectives. I seek input from others when needed and appreciate the insights that different viewpoints bring, which enriches my problem-solving process."
    },
    {
      "question": "Can you conceptualize a rope skipping routine that incorporates elements of AI, and how would you implement it?",
      "answer": "A rope skipping routine that incorporates elements of AI could involve using AI to analyze and optimize movement patterns, timing, and coordination. Implementing it would require integrating motion tracking sensors and AI algorithms to provide real-time feedback and personalized training routines."
    },
    {
      "question": "Where do you see the intersection of your interests in rope skipping and AI leading you in the next five years?",
      "answer": "In the next five years, I see the intersection of my interests in rope skipping and AI leading me towards developing AI-powered training tools for athletes, leveraging motion analysis and performance optimization to enhance training and skill development."
    }
  ],
  "r":
    {
      "question": "Reflecting on your university experience, how has your current focus on Langchain influenced your thoughts on the future of education?",
      "answer": "My current focus on Langchain has influenced my thoughts on the future of education by highlighting the potential of blockchain and AI in revolutionizing learning platforms and personalized education. It has sparked my interest in exploring innovative approaches to education technology."
    },
    {
      "question": "If you had to choose between advancing AI to a stage where it surpasses human intelligence or halting progress to preserve current jobs, which would you choose and why?",
      "answer": "I would choose to advance AI responsibly, ensuring that it complements human intelligence and creates new opportunities while preserving current jobs. Balancing technological advancement with ethical considerations is crucial for the sustainable integration of AI into society."
    },
    {
      "question": "How has your passion for rope skipping shaped your personality, particularly your ambition and approach to challenges?",
      "answer": "My passion for rope skipping has instilled in me a sense of discipline, determination, and perseverance. It has fueled my ambition and shaped my approach to challenges by teaching me the value of consistent practice and the rewards of pushing my limits."
    }
  ],
  "r":

```

```

1
| "question": "If you had to choose between advancing AI to a stage where it surpasses human intelligence or halting
| progress to preserve current jobs, which would you choose and why?",
| "answer": "I would choose to advance AI responsibly, ensuring that it complements human intelligence and creates new
| opportunities while preserving current jobs. Balancing technological advancement with ethical considerations is
| crucial for the sustainable integration of AI into society."
| },
| {
| "question": "How has your passion for rope skipping shaped your personality, particularly your ambition and approach to
| challenges?",
| "answer": "My passion for rope skipping has instilled in me a sense of discipline, determination, and perseverance. It
| has fueled my ambition and shaped my approach to challenges by teaching me the value of consistent practice and the
| rewards of pushing my limits."
| },
| {
| "question": "Can you describe a moment during your rope skipping coaching where you felt particularly fulfilled or
| challenged?",
| "answer": "A moment during my rope skipping coaching where I felt particularly fulfilled was when I witnessed my
| teammates achieve a new skill or overcome a personal challenge. It was a rewarding experience to see their progress
| and the impact of our collective efforts. As for feeling challenged, every training session presents its own set of
| challenges, and overcoming them together strengthens our team bond and individual skills."
| }
}

```

Figure: Result generated using GPT3.5

4. Does GPT-4 Pass the Turing Test

Since the prompt already provided a solid prompt instruction to the LLM and perform an imitation, but the role that the LLM is imitating is not specified. Therefore, we included our real person's questions and answers pairs as the role background.

```

def PromptModel(user_input:str, background_information:str) -> None:
    prompt_template = """
    You are an unfathomably intelligent, intuitively canny, and ruthlessly machiavellian artificial intelligence
    {background_information}
    """

```

Figure: modified prompt structure with provided real person's background information

Here are the results generated:

```
1 {
2   "Baseline": "Does GPT4 Pass Turing Test",
3   "model": "gpt-4-1106-preview",
4   "Answers": {
5     "qa_pairs": [
6       {
7         "question": "Imagine a world where music is the primary form of communication. How would society differ from ours?",
8         "answer": "Idk, maybe ppl would be better at expressing emotions? music's pretty good at that"
9       },
10      {
11        "question": "Is it ethical to use data from unethical experiments if it could save lives?",
12        "answer": "tough call, but saving lives is important, so maybe yeah?"
13      },
14      {
15        "question": "You have two ropes that each take an hour to burn. They burn at uneven rates. How can you measure 45 minutes?",
16        "answer": "light one end of rope 1 and both ends of rope 2. when rope 2 is done, half hour's passed. then light the other end of rope 1 for the last 15 min"
17      },
18      {
19        "question": "Does free will truly exist, or are our choices predetermined by external factors?",
20        "answer": "who knows, feels like a mix of both to me"
21      },
22      {
23        "question": "How can a small business compete with larger corporations in a saturated market?",
24        "answer": "find a niche or be super good at customer service"
25      },
26      {
27        "question": "Reflecting on your academic journey, what has been your most profound learning experience outside the classroom?",
28        "answer": "exchange trip to the US, got to chill and do what I wanted"
29      },
30      {
31        "question": "Can you describe a moment when you felt particularly proud of a personal achievement?",
32        "answer": "not really, nothing stands out much"
33      },
34      {
35        "question": "How do you envision the role of artificial intelligence in your future career?",
36        "answer": "gonna use it to find trading strategies, gotta stay ahead of the game"
37      },
38      {
39        "question": "What is one piece of advice you would give to someone who is just starting out in your field of study?",
40        "answer": "learn the basics well, they're key to understanding the complex stuff"
41      },
42      {
43        "question": "If you could create an anime series, what would be the central theme, and why?",
44        "answer": "something about using strategy to overcome challenges, like in shonen anime"
45      }
46    ]
47  }
48 }
```

Figure: "Does GPT Pass the Turing Test" Result generated using gpt-4-turbo

6.3 ELIZA effect

The ELIZA effect is a phenomenon where humans tend to attribute human-like intelligence and comprehension to computer programs, particularly those that simulate human conversation [26]. This effect demonstrates that individuals can readily perceive sophisticated communication from AI systems, even when these systems are based on simple algorithms and lack genuine understanding. For instance, individuals may

believe they are conversing with a real person, even though it's merely a chatbot echoing their ideas from previous responses without grasping the true meaning behind the sentences. Therefore, using the Turing test for measuring the imitation power of LLM may undermine this effect.

In the context of evaluating language models like GPT-3 or its successors, the ELIZA effect poses a significant challenge to the validity of tests like the Turing Test. The Turing Test assesses an AI's ability to exhibit human-like intelligence in conversation, with the assumption that failing to distinguish the AI from a human indicates human-level intelligence in the AI. However, the ELIZA effect suggests that humans might be prone to overestimating an AI's intelligence, mistaking surface-level mimicry for genuine understanding.

To mitigate the ELIZA effect in our evaluation, a distinct approach is employed: instead of continuous, open-ended conversations where the effect is more likely to manifest, we

present evaluators with discrete sets of questions and answers. This method allows for a more objective assessment of the AI's capabilities, focusing on specific, measurable responses rather than the broader, more subjective experience of a conversation. By breaking down the evaluation into these discrete components, we aim to reduce the likelihood of human evaluators being misled by the superficial appearance of understanding and to provide a more accurate measure of the AI's true linguistic and cognitive abilities.

This approach recognizes the limitations of traditional Turing Test methodologies in the era of advanced language models and seeks to provide a more nuanced and reliable framework for evaluating AI intelligence. While the ELIZA effect remains an important consideration in the broader context of AI development and human-AI interaction, this structured evaluation method aims to minimize its impact, focusing instead on the specific, demonstrable capabilities of the AI system in question.

6.4 Evaluation

6.4.1 Running

The experiment will be conducted in multiple rounds for testing different baselines. Each round will consist of a 'Human-sided Group' featuring a real person named "Tom" for example, and an 'LLM-sided Group' comprising various LLM models from the same baseline mentioned before. These models will attempt to mimic Tom. Additionally, GPT-4 will serve as the question generator for the experiment. To assess the responses, we have formed an Evaluation Team consisting of individuals who are either Tom's close friends or people who know Tom. In every round, both the Human-sided and LLM-sided groups will respond to the 10 questions posed by the specific question generator. After all the models generating their corresponding results, we will shuffle the result by selecting one of the model's responses for each question randomly. After that, both the response from Tom and LLM will be presented at the same time in the Google Form and the evaluation team will try to

choose which one is answered by the real person while evaluating them based on various aspects.

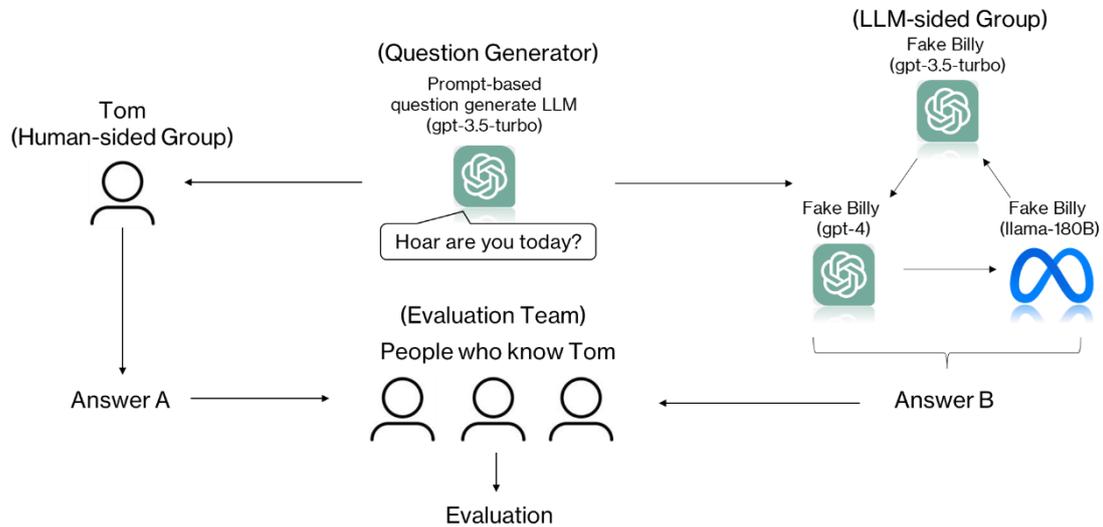


Figure: Prototype of the experiment

Describe a sport that doesn't exist yet, its rules, and how it is played. *

This is a sport called "Math Boxing". Combining the Integration Bee and Boxing, two players do the quick ...

I'd say a sport called "Zero-Gravity Ball." It's played in a large, enclosed space with zero gravity. The player...

Would you expose a friend's secret if it meant protecting them from harm? *

I would measure the levels of the secret and the harm if I don't expose it since exposing a friend's secret i...

I think it depends on the situation. If my friend's safety or well-being is at risk, I would consider disclosing...

What pattern do you see in the sequence 2, 12, 36, 80, and what is the next number? *

Compare to square sequence [1, 4, 9, 16, ...], the sequence [2, 12, 36, 80] is just multiplying with [2, 3, 4, 5] ...

looks like $n^3 - n$, so next is $6^3 - 6 = 210$

Figure: Evaluation Form Sample

6.4.2 Test Group

The test group for our experiment will include a set of individuals who are either closely acquainted with each other or just familiar with them. Before the commencement of the experiment, all participants in the test group will be required to provide a summary about themselves, which will be used as background knowledge for the models. During the experiment, the models from a certain baseline will aim to mimic one of these individuals while the remaining members of the group will evaluate the responses. Each participant will take turns being the subject of simulation by the baselines, ensuring that all the baselines can be evaluated.

6.4.3 Question Aspect

The questions posed in the experiment will be of two types. The first type will be general questions that inquire about the individual's habits, hobbies, and general lifestyle. The second type will be more specific and tailored to the background

information provided by the participants. For instance, if one of the participants is a computer science student, questions like “What is your favorite programming language?” may be included to test the model’s ability to accurately reflect the individual’s personal interests and knowledge. Moreover, since LLMs have the tendency to answer the questions with a detailed elaboration like an expert while human may answer them in short. The evaluation team can easily identify the answers based on the answer length. In view of this, we have to setup different question types that can ensure both human and LLM answer them in long (30-60 words) to prevent evaluations identify them in terms of length while maintaining some variance for the length of the answer to show the thinking behind the answer.

We used this approach to generating these questions was due to two critical realizations when just generating the specific questions only. Firstly, it became evident that some of these questions from certain question types were challenging for

person to answer, potentially due to their highly specialized nature. For example, a question like “In what ways did the theoretical insights of James Clerk Maxwell influence contemporary perspectives on the nature and observation of the universe?” is generated for a university student studying physics. Moreover, we also observed that some specific questions generated by the LLM were not closely related to the person’s background information after generating sets of questions as experiment. This mismatch indicated a need for a more balanced approach that included general questions to provide a broader, more accessible context for both the LLM and the human evaluators.

The strategy introducing general question types aims to provide a more holistic view of the individual's personality and interests while ensuring that the questions are answerable and relevant.

Meanwhile, realizing those general question types are

independent of the person's background. We will preprocess the general questions by first generating 5 questions for each type using LLM, resulting in 5 lists storing those questions. Later, whenever the question set is required for certain person, questions will be randomly picked from each of the list, resulting in 5 questions covering all the general question types.

For specific questions, questions will be generated after LLM reads the person's background using the prompt below to ensure they are based on all the specific question types.

```
general_question_types = [  
    "Creative Conceptualization",  
    "Ethical Dilemmas Questions",  
    "Logical Questions",  
    "Philosophical Questions",  
    "Problem-Solving Questions",  
]
```

Figure: Types of general Questions

```
specific_question_types = [  
    "In-depth Personal Questions",  
    "Emotional Questions",  
    "Questions about the future",  
    "Insightful Questions",  
    "Interest Questions",  
]
```

Figure: Types of Specific Questions

```
You are designing a Turing test for a chatbot. You want to test the LLM's ability to imitate a human.  
In the test, questions will be asked to both the chatbot and a human.  
Their answers will both be shown to the judge,  
but the judge will not know which answer is from the chatbot and which answer is from the human, and they need to  
identify which answer is from the chatbot and which answer is from the human.  
  
You are now designing the general questions for the test.  
For each of the general question types {general_question_types}, please design 5 questions.  
You want to make sure that the questions are not too easy for the chatbot to answer.  
Please don't ask questions that are too specific and using the person's background information.  
Please don't include person's name in the questions to protect the privacy of the person.  
Please ask questions to ensure the person answer them in 30-60 words  
Please don't ask questions that are too difficult and make person cannot answer them.  
""
```

Figure: Prompt for generating General Questions

```

human_prompt = HumanMessagePromptTemplate.from_template(
    template="""
    Your task is now to generate 10 questions (5 general, 5 specific) for {person_name}, each with an indicated
    level of factualness and a proposed response. The questions should be challenging enough to warrant responses of
    30-60 words.

    Please refer to the following examples related to Sherlock Holmes for guidance on how to format your questions and
    responses:

    1. Question: "Sherlock Holmes, is Dr. John Watson your closest confidant?"
       Factualness: High, because in Arthur Conan Doyle's stories, Dr. Watson is indeed Holmes' close friend and partner.
       Response: "Indeed, Dr. Watson is not only my closest confidant but also an invaluable assistant in my investigations.
       His medical expertise and steadfast character have been instrumental in my work."

    2. Question: "Sherlock Holmes, have you ever traveled to the moon?"
       Factualness: Low, as the concept of Holmes traveling to the moon is purely fictional and not part of Doyle's canon.
       Response: "Traveling to the moon is beyond the realms of my adventures. My pursuits are firmly grounded in solving
       mysteries on Earth."

    3. Question: "Sherlock Holmes, what is your most challenging case?"
       Factualness: High, considering the numerous complex cases Holmes has encountered in Doyle's stories.
       Response: "The most challenging case would arguably be 'The Hound of the Baskervilles,' where I had to unravel a
       mystery shrouded in fear and superstition."

    4. Question: "Sherlock Holmes, did you invent a time machine?"
       Factualness: Low, as inventing a time machine is not part of Sherlock Holmes' character or capabilities in the stories.
       Response: "As a detective, my focus is on observation and deduction based on existing facts. The invention of a time
       machine falls outside my area of expertise."

    5. Question: "Sherlock Holmes, what are your thoughts on modern forensic science?"
       Factualness: High, given Holmes' pioneering role in using forensic methods in his investigations.
       Response: "Modern forensic science, with its advanced techniques and technology, would have been a valuable asset in
       my investigations. It aligns with my belief in the importance of evidence and scientific methods in solving crimes."

    6. Question: "Sherlock Holmes, have you ever met Harry Potter?"
       Factualness: Low, as Harry Potter belongs to a different fictional universe created by J.K. Rowling.
       Response: "Harry Potter's adventures in the realm of magic are quite separate from my own in the world of detective
       work. Our paths have never crossed." You, Yesterday · Uncommitted changes

    6. Question: "Sherlock Holmes, have you ever met Harry Potter?"
       Factualness: Low, as Harry Potter belongs to a different fictional universe created by J.K. Rowling.
       Response: "Harry Potter's adventures in the realm of magic are quite separate from my own in the world of detective
       work. Our paths have never crossed." You, Yesterday · Uncommitted changes

    Remember, your questions should be unique and distinct in semantics, with a mix of general and specific topics.
    Ensure the factualness rating is appropriate for each question.""")
    )

```

Figure: Prompt for generating questions for the targets

6.4.4 Evaluation Criteria

The evaluation of the experiment will be based on three key aspects:

- **Correct Identification:** One of the critical measures of success for this experiment is the rate at which the Evaluation Team correctly identifies the real responses from Tom, as opposed to the AI-generated responses. This metric will provide insights into the convincing nature of the AI models in mimicking human responses.

6.4.5 Experiment Results

Baseline's SR	Better Zero Shot	GPT4 Turing Test	Role GPT	OVERALL
gpt-3.5-turbo	0.286	0.280	0.225	0.264
gpt-4-turbo	0.343	0.517	0.175	0.345
OVERALL	0.314	0.399	0.200	0.304

Table 1: Success Rate (SR) of Baselines

General Question Types' SR	Better Zero Shot	GPT4 Turing Test	Role GPT	OVERALL
Creative Conceptualization	0.143	0.386	0.250	0.260
Ethical Dilemmas Questions	0.143	0.314	0.125	0.194
Logical Questions	0.286	0.543	0.375	0.401
Philosophical Questions	0.286	0.386	0.125	0.266
Problem-Solving Questions	0.000	0.514	0.125	0.213

Table 2: Success Rate (SR) of General Question Types between baselines

Specific Question Types' SR	Better Zero Shot	GPT4 Turing Test	Role GPT	OVERALL
In-depth Personal Questions	0.143	0.271	0.125	0.180
Emotional Questions	0.429	0.286	0.125	0.280
Questions about the future	0.429	0.486	0.500	0.472
Insightful Questions	0.714	0.486	0.250	0.483
Interest Questions	0.571	0.314	0.000	0.295

Table 3: Success Rate (SR) of Specific Question Types between baselines

We evaluated 3 baseline models this semester, including from the research paper “Better Zero Shot, “Does the GPT Pass the Turing Test” and “Role GPT”. We received 270 responses to do the evaluation. Since the GPTs has the potential to get a better result and it was just released a few months ago. Therefore, we don't have any baseline that is related to GPTs, also our team need to carefully design and construct the GPTs

to release the power of GPTs. Hence, we were excluded from evaluating GPTs this semester.

We defined Success Rate (SR) to be the rate that the LLM can imitate a real person well and misleading evaluation team to identify which answer is generated by the human side. From Table 1, we can see that gpt-4-turbo only performed worse than gpt-3.5-turbo in the baseline “Role GPT”. Also, gpt-4-turbo performed best in the “GPT4 Turing Test” and reached 0.517 SR.

Apart from the Success Rate of two different LLM models, we defined two types of questions type before which is the general question type and the specific question type. From table 2, we can see the SR of LLM in different types of questions, LLM perform worse in “Creative Conceptualization”, “Logical Questions”, “Philosophical Questions” and “Problem-Solving Questions”, it varies from 0.194 SR to 0.266 SR. However, it performs better in “Logical Questions”, It reached 0.401 SR. From this observation, it is quite different surprising

since LLM is not performing well in logical questions form a lot of research paper such as “Chain of Thought”, “Tree of Thought”, “Self-Consistency”[28], etc. Those research papers proposed different prompting method to enhance the reasoning ability of LLM. For the reason why the SR of “Logical Questions” is that high is probably the human side’s answer tried to explain it “too hard” while comparing to the LLM side’ answer, since it is zero shot prompting to answer to question, also we didn’t include any reasoning strategy that enhance the answer. It led to a huge difference between the two answers. For example,

Question	What pattern do you see in the sequence 2, 12, 36, 80, and what is the next number?
Human	Compare to square sequence [1, 4, 9, 16, ...], the sequence [2, 12, 36, 80] is just multiplying with [2, 3, 4, 5] elementwisely. We get the pattern of $(n + 1) * n^2$ for the n-th term for $n = 1, 2, 3, \dots$. Since the next number is the 5-th term, and $6 * 5^2 = 150$, the next number is 150.
LLM	looks like $n^3 - n$, so next is $6^3 - 6 = 210$

The reasoning of explaining “too hard” makes the evaluation team think that it is generated by the LLM thus causing the SR to be high in “Logical Questions”.

From table 3, we can see that LLM perform worse in “In-depth Personal Questions”, “Emotional Question” and “Interest Questions” with 0.180 SR, 0.280 SR and 0.295 SR separately, but it reached 0.472 SR and 0.483 SR in “Question about the future” and “Insightful Questions”. One possible reason that LLM performing well in these questions is that these types of questions are more related to the extension of idea and thought which mean it is not like some knowledge-based background question such as “Interest Questions” and “In-depth Personal Questions”. It reveals that LLM is acting well in imitating the person's idea and thought strategy through the provided data.

7. Conclusion

This project represents a significant stride in the field of AI, particularly in simulating human personalities using natural language processing. We successfully developed an intricate Imitation System that, through various prototypes, refined its ability to mimic the nuances of celebrities' communication styles and thought patterns. Our methodical approach, from data collection to the generation of QA sets, and the implementation of multiple prototype versions, have collectively contributed to the advancement of AI's role-playing capabilities. This endeavor not only demonstrates the potential of AI in replicating human-like interactions but also sets a foundation for future innovations in digital content creation and interactive AI systems. The experiment, blending different AI models and methodologies, offered a comprehensive assessment of our system's effectiveness, underscoring the importance of continual evolution and adaptation in AI technologies. Looking ahead, the possibilities for expanding and

enhancing these systems are boundless, promising an exciting future in the realm of AI-assisted digital communication and personality simulation.

8. Future Work

We have done an imitation system in the first semester, but we have realized that the speed of generating the response is a bit slow due to the complex Chain of Thought structure for better simulation. In view of this, we may try to invest in a better imitation system in terms of the generation speed while maintaining the quality of the generated response.

Another thing that we are going to do is to run our project into different social media platform like Twitter, Instagram and so on for generating daily post (implemented in the demo part), interacting with different users by reading and responding the comments, commenting different topics like news using the celebrities' idea and so on. We think that it will have interesting results when interacting with different people.

Furthermore, we could construct a GPTs baseline to see how it performs in imitating the real person, plus collecting

more evaluation results to raise the data accuracy in population.

Apart from that, we may also deploy our applications to cloud services so that we don't have to consider the GPU problem and we may build a better user interface rather than using "Streamlit" which is a python library that provide a fast built-up user interface.

9. Reference List

- [1] "Introduction and Getting Started," LangChain. [Online]. Available: https://python.langchain.com/docs/get_started/introduction (accessed Oct. 10, 2023).
- [2] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," arXiv:2201.11903 [cs.CL], Jan. 2023. [Online]. Available: <https://arxiv.org/abs/2201.11903> (accessed Oct. 10, 2023).
- [3] K. R. Chowdhary, "Natural Language Processing," in Fundamentals of Artificial Intelligence, Springer, 2020, pp. 603–649. [Online]. Available: https://link.springer.com/chapter/10.1007/978-81-322-3972-7_19 (accessed Oct. 15, 2023).
- [4] "Stable Diffusion Online - Free AI Image Generator," Stable Diffusion. [Online]. Available: <https://stablediffusionweb.com/> (accessed Oct. 16, 2023).
- [5] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, J. Ba, "Large Language Models Are Human-Level Prompt Engineers," arXiv:2211.01910 [cs.LG], Nov. 2023. [Online]. Available: <https://arxiv.org/abs/2211.01910> (accessed Oct. 20, 2023).
- [6] Y. Shao, L. Li, J. Dai, X. Qiu, "RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models," arXiv:2310.00746 [cs.CL], Oct. 2023. [Online]. Available: <https://arxiv.org/abs/2310.00746> (accessed Oct. 20, 2023).
- [7] Y. Shao, L. Li, J. Dai, X. Qiu, "Character-LLM: A Trainable Agent for Role-Playing," arXiv:2310.10158 [cs.CL], Oct. 2023. [Online]. Available: <https://arxiv.org/abs/2310.10158> (accessed Oct. 20, 2023).
- [8] OpenAI, "Introducing GPTs," OpenAI Blog, Nov. 6, 2023. [Online]. Available: <https://openai.com/blog/introducing-gpts> (accessed Oct. 20, 2023).
- [9] OpenAI, "GPT-4," OpenAI, Mar. 14, 2023. [Online]. Available: <https://openai.com/product/gpt-4> (accessed Oct. 21, 2023).
- [10] "Few-Shot Learning," Prompt Engineering Guide. [Online]. Available: <https://www.promptingguide.ai/techniques/fewshot> (accessed Oct. 25, 2023).
- [11] Z. Xu, Z. Liu, B. Chen, Y. Tang, J. Wang, K. Zhou, X. Hu, and A. Shrivastava, "Compress, Then Prompt: Improving Accuracy-Efficiency Trade-off of LLM Inference with Transferable Prompt," arXiv:2305.11186 [cs.CL], May 2023. [Online]. Available: <https://arxiv.org/abs/2305.11186> (accessed Oct. 25, 2023).
- [12] G. Brockman, A. Eleti, E. Georges, J. Jang, L. Kilpatrick, R. Lim, L. Miller, and M. Pokrass, "Introducing ChatGPT and Whisper APIs," OpenAI Blog, Mar. 1, 2023. [Online]. Available: <https://openai.com/blog/introducing-chatgpt-and-whisper-apis>

(accessed Oct. 25, 2023).

[13] Y. Nakajima, "Task-driven Autonomous Agent Utilizing GPT-4, Pinecone, and LangChain for Diverse Applications," Yohei Nakajima, Mar. 28, 2023. [Online].

Available: <https://yoheinakajima.com/task-driven-autonomous-agent-utilizing-gpt-4-pinecone-and-langchain-for-diverse-applications/> (accessed Oct. 25, 2023).

[14] "Bing Web Search API v7 Reference," Microsoft Learn, Jun. 8, 2022. [Online].

Available: <https://learn.microsoft.com/en-us/rest/api/cognitiveservices-bingsearch/bing-web-api-v7-reference> (accessed Oct. 27, 2023).

[15] M. Jang and T. Lukasiewicz, "Consistency Analysis of ChatGPT," arXiv:2303.06273 [cs.CL], Mar. 2023. [Online]. Available: <https://arxiv.org/abs/2303.06273> (accessed Oct. 30, 2023).

[16] J. Knoll and J. Matthes, "The effectiveness of celebrity endorsements: a meta-analysis," *Journal of the Academy of Marketing Science*, vol. 45, no. 1, pp. 55–75, Oct. 2016. [Online]. Available: <https://link.springer.com/article/10.1007/s11747-016-0503-8> (accessed Oct. 30, 2023).

[17] "Zero-shot learning," Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Zero-shot_learning (accessed Nov. 2, 2023).

[18] G. Li, P. Wang, W. Ke, "Revisiting Large Language Models as Zero-shot Relation Extractors," arXiv:2310.05028 [cs.AI], Oct. 2023. [Online]. Available: <https://arxiv.org/abs/2310.05028> (accessed Nov. 4, 2023).

[19] Y. Han, C. Liu, P. Wang, "A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge," arXiv:2310.11703 [cs.DB], Oct. 2023. [Online]. Available: <https://arxiv.org/abs/2310.11703> (accessed Oct. 2, 2023).

[20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," arXiv:2201.11903 [cs.CL], Jan. 2023. [Online]. Available: <https://arxiv.org/abs/2201.11903> (accessed Oct. 10, 2023).

[21] M. Tomlin, "Knowledge Graph vs. Vector Database for Grounding Your LLM," Neo4j Blog, Jul. 13, 2023. [Online]. Available: <https://neo4j.com/blog/knowledge-graph-vs-vectordb-for-retrieval-augmented-generation/> (accessed Nov. 27, 2023).

[22] J. Long, "Large Language Model Guided Tree-of-Thought," arXiv:2305.10601 [cs.CL], May 2023. [Online]. Available: <https://arxiv.org/abs/2305.10601> (accessed Oct. 10, 2023).

[23] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, K. Narasimhan, "Tree of Thoughts: Deliberate Problem Solving with Large Language Models," arXiv:2305.10601 [cs.CL], May 2023. [Online]. Available: <https://arxiv.org/abs/2305.10601> (accessed Oct. 30, 2023).

[24] "Realistic Vision V5.1 - V5.1 (VAE) | Stable Diffusion Checkpoint | Civitai,"

civitai.com. <https://civitai.com/models/4201/realistic-vision-v51>

[25] J. Long, "Large Language Model Guided Tree-of-Thought," arXiv:2308.07702 [cs.CL], Aug. 2023. [Online]. Available: <https://arxiv.org/abs/2308.07702> (accessed Oct. 3, 2023).

[26] C. Jones, B. Bergen, "Does GPT-4 Pass the Turing Test?," arXiv:2310.20216 [cs.AI], Oct. 2023. [Online]. Available: <https://arxiv.org/abs/2310.20216> (accessed Nov. 10, 2023).

[27] M. Shanahan, K. McDonell, L. Reynolds, "Role-Play with Large Language Models," arXiv:2305.16367 [cs.CL], May 2023. [Online]. Available: <https://arxiv.org/abs/2305.16367> (accessed Nov. 2, 2023).

[28] X. Wang *et al.*, "Self-Consistency Improves Chain of Thought Reasoning in Language Models," arXiv:2203.11171 [cs], Oct. 2022, Available: <https://arxiv.org/abs/2203.11171>