The Chinese University of Hong Kong

Department of Computer Science and Engineering

April 2024

LYU2305

Full-Stack AI Content Creator

Author:

TSE, Hui Tung 1155158864

Ng Man Tik 1155158302

Supervisor:

Professor Michael R. Lyu

Abstract

This project aims to recover the imitation of celebrities in aspects of thought strategy and writing style. We constructed an automated retrieval augmented fine-tuning system architecture that is specialized in reappearing a role. We investigated multiple approaches and performed research in figuring out the role-playing ability in different large language models with various designs. To measure the outcome, we performed news summarizing with the celebrity's personal view and analysis through the platform "Medium". The System could automatically summarize the news that happened recently and pick the most related news topic that is matching the celebrity's expertise. The blog writing architecture is powered with numerous subsystems including Retrieval Augmented Generator (RAG), Image Generator System, Data Augmentation System, Article Enhancement System and Layout Finalizing System.

In order to evaluate the role-playing ability of Large Language Models (LLMs), We conducted a research in this field. However, existing studies focus on the imitation of either well-known public figures or fictional characters, overlooking the potential for simulating everyday individuals. Such an oversight limits the potential for advancements in digital human clones and non-player characters in video games. Addressing this gap, we draw inspiration from the Turing Test and propose ECHO, an evaluation framework that involves the target individual's acquaintances to differentiate between responses generated by humans and those by machines. This advantage is provided by the scenario of imitating everyday individuals instead of historical celebrities or fictional characters. We benchmark three role-playing LLMs with ECHO, utilizing both GPT-3.5 and GPT-4 as backbones. Additionally, we also assess the capability of role-playing of GPTs, the latest online application from OpenAI. Our findings indicate that GPT-4 more effectively fools human evaluators, with GPTs leading the pack by achieving a success rate of 48.3%. This result will be used to empower our imitation architecture.

Acknowledgement

We would like to acknowledge and give my warmest thanks to our supervisor Professor Michael R. Lyu, Mr. Jen-tse Huang who guided our project and made this work possible. Their guidance carried me through the whole design and system architecture. Their innovative ideas and reviews are crucial and valuable to extend the feasibility of our project. Besides, we would also like to give special thanks to Mr. Jen-tse Huang and Ms. Jing Jing Li who mentored us in our research work. Without their precious enlightenment, it is hard to finish the research in a comprehensive way. Their support was what sustained us this far.

Contents

Chapter 1

Introduction

1.1 Motivation

Artificial Intelligence is developing in a exponential speed, the LLMs model became much more better with a smaller size. However, due to a lot of generalised LLMs, the area of exploring LLM as a specific domain expertise is an upcoming trend. When come to this crucial development, role-playing is one of the way to achieve the target. If LLms are capable to imitate a role-well and equipped with relevant expert knowledge would be an advantage to further expand correlated studies and bring a huge benefit across a lot of fields. Therefore, We selected mimicking celebrities as our target to further evaluate and develop our model architecture.

1.2 Objective

Our first goal is to post a review from imitating a celebrity through online platform. By doing so, the system could automatically post the imitation result to public and could be further evaluate by the public.

Our second objective is to investigate the role-playing ability of current LLMs with different proposed system architecture and explore the possibilities of using LLMs in auto evaluation field.

Last but not least, our final goal is to propose a new architecture that is being tested and improved in a certain level of imitation. With testing various methods and prompt optimizing.

Chapter 2

Full-stack AI content creator

2.1 Imitation System Architecture

2.1.1 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.

2.1.2 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 (LangChain, 2024a) 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.

2.1.3 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

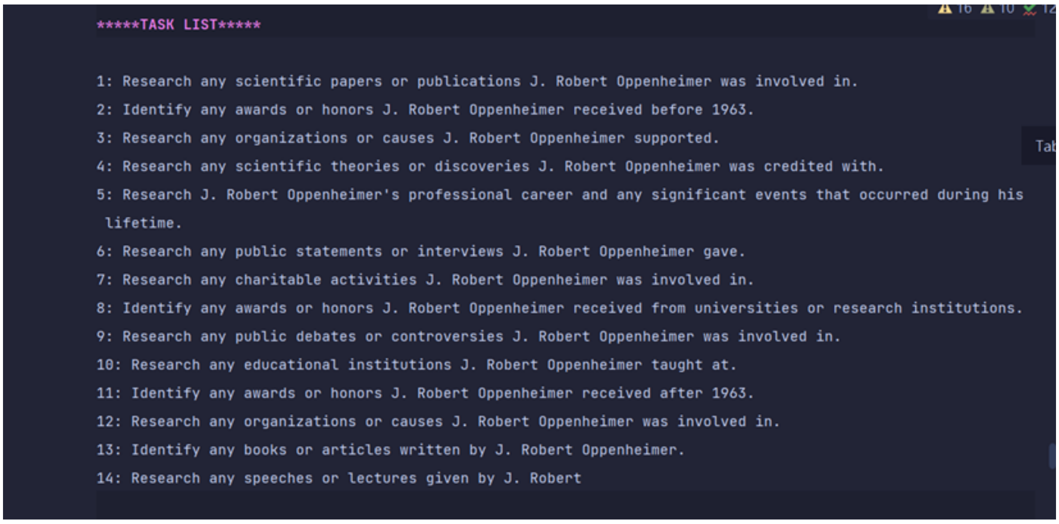


Figure 2.1: Task list for agent to search online to get information

in creating a dataset that is representative of the celebrity's persona.

2.1.4 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 (Xu et al., 2023b). 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 (Brockman, 2023) and BabyAGI (Nakajima, 2023), utilizing BingSearch API V7 (Microsoft, 2024b). The workflow involves:

- Task Prioritization: BabyAGI selects the most pertinent task at hand.
- Data Retrieval: LangChain employs BingSearch API V7 to gather data according to the given task.
- Result Refinement: The agent removes irrelevant information from the search results.
- Deduplication: The agent compares the new data with existing files to eliminate duplicates.
- Data Structuring: The refined data is structured into a JSON file, facilitating easy retrieval for future processing.
- 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.

2.1.5 Extracting Features

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 (Jang and Lukasiewicz, 2023). Therefore, our system is designed to

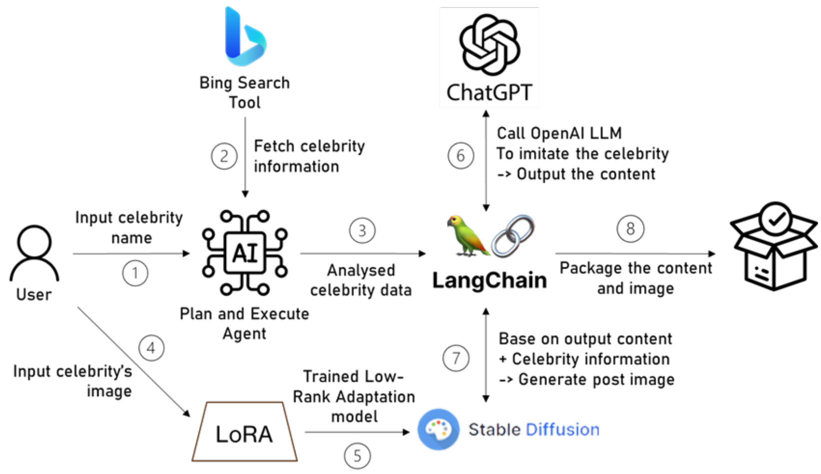


Figure 2.2: Celebrity Imitation Application Workflow


```
{
    "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\u00f6ttingen 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 \u201cfather of the atomic bomb\u201d 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."
        },
```


Figure 2.3: Summarized result of the fetched information of the celebrity

both analyze and synthesize information into a question-and-answer (QA) format, fostering dynamic and consistent interactions that mirror the celebrity's communication style (Knoll and Matthes, 2016).

2.1.6 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.

2.1.7 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
- Memorable Life Experiences
- Writing and Speaking Style

2.1.8 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

2.1.9 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:

2.1.10 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:

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.

```
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 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.

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

Figure 2.4: System Prompt for generating QA

```
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."
```

Figure 2.5: Human Prompt for generating QA

```
{
    "aspects": "Personality",
    "types": "Emotional Questions",
    "question": "Dr. Oppenheimer, how did you cope with the ethical implications of your work on the atomic bomb, given its
      destructive power?",
    "factualness": true,
    "response": "The ethical implications of my work on the atomic bomb were indeed a heavy burden. While the development of
      the bomb was crucial for ending World War II, the destruction it caused was immense. This led me to advocate for
      nuclear disarmament and control of nuclear power post-war."
},
{
    "aspects": "Values and Beliefs",
    "types": "Ethical Questions",
    "question": "Dr. Oppenheimer, you were a vocal advocate for nuclear disarmament after the war. Can you share your
      thoughts on why you believed this was necessary?",
    "factualness": true,
    "response": "After witnessing the destructive power of the atomic bomb, I believed it was necessary to advocate for
      nuclear disarmament. The potential for such weapons to cause mass destruction and loss of life is immense, and I felt
      it was our responsibility to ensure they were not used recklessly."
},
```


Figure 2.6: Result of the generated QA

- 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.

2.1.11 Evaluating role-playing ability

To evaluate different prompting approach's effectiveness in role-playing the celebrity, we utilized the same setup in real person to examine the output and figure out a best approach to be used in our application. Here, we did a research and proposed a evaluation setup called ECHO.

Chapter 3

ECHO

3.1 What is ECHO?

Large Language Models (LLMs) have recently made significant breakthroughs in the field of Artificial Intelligence (AI). Notably, ChatGPT[1], one of the state-of-the-art commercial models, has showcased its capabilities across different Natural Language Processing (NLP) tasks, such as information retrieval (Zhu et al., 2023), computer programming (Surameery and Shakor, 2023), grammar checking (Wu et al., 2023), and sentence translation (Jiao et al., 2023). Trained on extensive datasets, LLMs also demonstrate applicability beyond NLP tasks, extending to domains such as healthcare (Johnson et al., 2023), education (Baidoo-Anu and Ansah, 2023), legal service (Guha et al., 2024), and product design (Lanzi and Loiacono, 2023).

Given LLMs' extensive capabilities, researchers have explored their resemblance to humans (Huang et al., 2024; 2023). *Role-playing*, the act of changing one's behavior to fulfill a specific role, has been employed as a criterion for evaluating LLMs (Shanahan et al., 2023) since it is a complicated task requiring

[1]https://chat.openai.com/

various abilities. However, the evaluation of LLMs' role-playing ability is a relatively under explored area. Previous studies (Shao et al., 2023; Wang et al., 2023) mainly focus on instructing LLMs to impersonate celebrities or fictional characters. These approaches restrict the scope of assessing LLMs' role-playing capabilities and overlook situations where LLMs could act as digital clones of humans, non-player characters in video games and metaverse, or, more concerningly, be used maliciously to impersonate individuals, spreading false information or damaging reputations. Addressing this gap, our study directs LLMs to emulate real, ordinary individuals instead of famous figures, a notably unexplored area in current research.

To effectively assess the capability of LLMs to emulate specific individuals, our approach is inspired by the *Turing test*, as initially proposed by Turing (1950). This test gauges whether a machine can demonstrate intelligence indistinguishable from that of a human. In our study, we create a role-play LLM using the profile of an actual person and invite acquaintances of this person to discern between responses from the real individual and the LLM. Utilizing real-person data makes it possible to apply the Turing test and makes it easier to recruit annotators, which is advantageous over using profiles of well-known figures due to the accessibility of their acquaintances. However, a limitation arises in multi-round dialogues, where human evaluators can easily differentiate between LLMs and real people by posing questions LLMs cannot answer, such as queries about the current time. This issue can shift evaluators' focus from assessing the LLMs' ability to think and act like the intended emulation target. To address this problem, we introduce a novel framework, ECHO, designed to specifically evaluate LLMs' proficiency in replicating a human's thought process within a particular domain. ECHO is a question-answering-based evaluation system, which marks a departure from traditional dialogue-based evaluation systems.

We evaluate four different role-playing methods, RoleGPT (Wang et al., 2023), Juliet (Jones and Bergen, 2023), Role-Play Prompting (RPP) (Kong et al., 2023), and OpenAI's online application, GPTs (OpenAI, 2023). For the first three methods, we additionally compare the performance differences when utilizing GPT-3.5 versus GPT-4. We collect the personal data of ten unique individuals for instructing each method to role-play these characters. Subsequently, we pose ten types of questions from various aspects to both the target character and the role-playing LLMs. Each character then invites their acquaintances to identify which responses they believe are written by the actual person. The findings indicate that the most effective role-playing method achieved a 48.3% success rate in deceiving the acquaintances. The research paper can be summarized as:

1. We propose ECHO, the first framework to conduct Turing tests on role-playing LLMs, which can effectively compare different role-playing methods.

2. We conduct extensive experiments, including profiles of ten people, and invite their acquaintances to discern between responses produced by LLMs and real humans.

3. We delve into LLMs' potential as evaluators in identifying human versus machine-generated texts, addressing concerns about biases that might influence their judgment.

3.2 Related Work

3.2.1 Turing Tests for LLMs

The concept of the Turing Test (Turing, 1950) is a cornerstone in AI's history, initially assessing AI through text-based interactions to determine if a judge was conversing with a human or a machine. The advent of LLMs has propelled the

Turing Test into new territories. For instance, Jannai et al. (2023) conducted a large-scale public Turing Test online, challenging participants worldwide to discern between an LLM and another human in a two-minute conversation. Their findings indicated that current LLMs pass the test approximately 40% of the time. Meanwhile, the TURINGBENCH environment (Uchendu et al., 2021) provides a structured platform to systematically evaluate the indistinguishability of LLM outputs from human responses, showcasing the advancements and the limitations of current models. Similarly, Jones and Bergen (2023) implemented an approach where an interrogator interacts with a single respondent to determine if they are human or AI. In their experiments, the best-performing GPT-4 prompt passed in 41 games. Also, Sejnowski (2023) posits that interactions with LLMs, through a reverse Turing test, reveal more about human intelligence dynamics than the artificial nature of LLMs, highlighting a complex interplay between human expectations and machine outputs. Elkins and Chun (2020) demonstrates GPT-3's capacity to emulate well-known authors' writing styles and thematic elements, showcasing its potential in creative writing fields ranging from journalism to novel writing.

However, these advancements face challenges, such as LLMs acknowledging their non-human nature when directly questioned, reflecting their programming for honesty, and experiments often featuring LLMs in ambiguous roles rather than imitating real people. Our research aims to overcome these hurdles by evaluating LLMs' ability to replicate specific individuals, offering a more detailed examination of their imitation skills.

3.2.2 Role-Playing LLMs

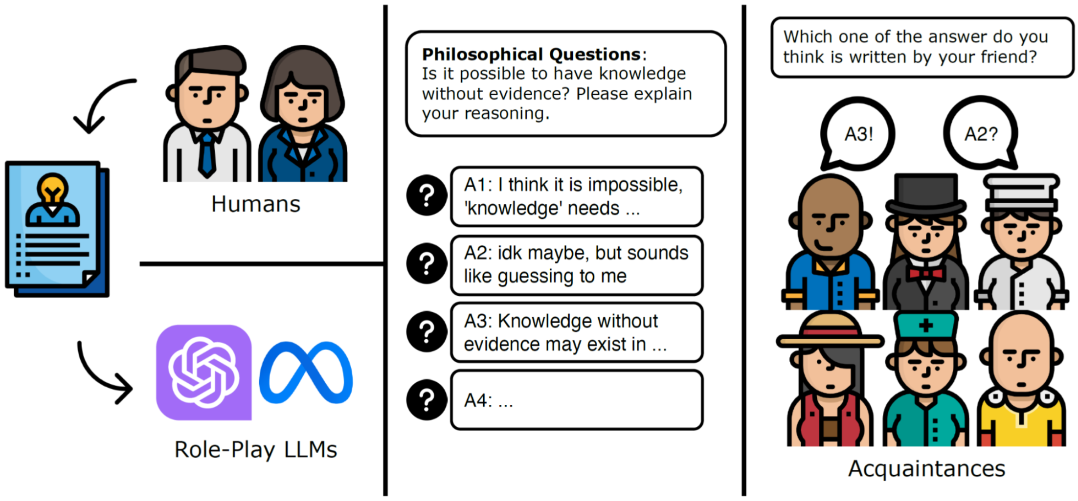
Recent advancements in AI have led to a growing interest in the role-playing abilities of LLMs. This field investigates how LLMs adapt to and maintain specific characters or personas in conversational contexts. Studies examine the intrinsic capacity of LLMs to engage in role-play and assess the models' ability to consistently portray assigned roles, providing insights into their adaptability and versatility in dynamic interactions (Shanahan et al., 2023). Meanwhile, RoleLLM (Wang et al., 2023) and CharacterLLM (Shao et al., 2023) provide different frameworks specifically designed to benchmark or enhance the role-playing capabilities of LLMs while Kong et al. (2023) focus on enhancing LLMs' zero-shot reasoning abilities in role-playing various personas. Several studies, such as CharacterGLM (Zhou et al., 2023) and ChatHaruhi (Li et al., 2023), expand the exploration of LLMs' role-playing capabilities into cultural and entertainment arenas. These works highlight the remarkable ability of LLMs to engage in role-playing not only within Chinese cultural contexts but also in bringing fictional characters to life, demonstrating the versatility and creative potential of LLMs in diverse settings.

Additionally, some applications, such as character.ai[2] offer an innovative platform where users can interact with AI-generated characters, each with distinct personalities and backgrounds. GPTs (OpenAI, 2023), introduced by OpenAI, allows users to tailor and access the customized GPT models for specific tasks like doing role-playing.

3.3 ECHO Design and Implementation

ECHO is a human evaluation system built on top of the Turing Test, aimed at assessing the role-playing capabilities of different LLMs. It comprises four key components: a question generator, a human-side group, an LLM-side group consisting of language models such as GPT-3.5 and GPT-4, and an evaluation

[2] https://character.ai/



(a) Constructing Role-Play LLMs (b) Question Answering (c) Human Evaluation

Figure 3.1: An illustration of the design of ECHO.

group. The overall framework is illustrated in Figure 3.1.

3.3.1 Constructing Role-Play LLMs

Gathering detailed background information about an individual before evaluation is essential to creating a realistic and comprehensive imitation. The profile database equips the LLM to understand and replicate the nuances of the individual's personality, experiences, and communication style to generate responses that reflect the individual's character and thought processes. By providing the model with rich, multi-faceted data on a person, we enhance the LLM's capability to generate responses that are not just accurate but also deeply resonant and personalized.

To ensure the depth and consistency of the data, we employ predefined categories when gathering background information. This structured approach ensures that the LLM receives a balanced representation of various facets of the individual's life, enabling it to generate responses that are factually accurate and reflective of the individual's unique persona. The categories explored include:

- **Background and Interests**: Education, Professional Background, Interests, and Hobbies.
- **Personal Identity**: Personality, Values and Beliefs, Memorable Life Experiences.
- **Cultural Preferences**: Favorite Books, Movies, and Music.
- **Cognitive and Social Dynamics**: Problem-Solving Style, Thoughts on Current Events, Communication and Social Style, Writing and Speaking Style.

We present a detailed list of ten questions for background information collection in §**??** in the appendix. Testers must answer all questions without refusal and provide evidence to support their responses, ensuring a comprehensive and credible dataset. This process aims to enhance the quality of information for LLMs' understanding of individuals. Responses that lack sufficient evidence or do not meet guidelines may be excluded to preserve data integrity.

3.3.2 Collecting Answers

Question Types Our approach separates the questions into two categories: general and specific. The general question types encompass broader themes, while the specific ones delve deeper into personal attributes based on the person's background information. Below are the general question types:

- Creativity Questions (CR): Questions that involve generating original ideas or envisioning scenarios by altering or expanding upon existing concepts.
- Ethical Dilemmas Questions (ED): Questions that challenge individuals to consider and articulate their moral perspectives or decisions in complex situations involving moral ambiguity or conflict.

- Logical Questions (LG): Questions designed to assess an individual's ability to think in a structured, coherent, and logical manner.
- Philosophical Questions (PH): Inquiries that explore deep and often abstract ideas about human existence, ethics, knowledge, and the nature of reality.
- Problem Solving Questions (PS): Questions that require analytical thinking and practical solutions to hypothetical or real-world challenges.

We also incorporate the following types of specific questions:

- In-depth Personal Questions (IP): Questions that delve into an individual's personal history, experiences, and reflections to gain insights into their character, motivations, and life journey.
- Emotional Questions (EM): inquiries that explore an individual's emotional experiences and how they manage and understand their feelings in various situations.
- Future Prediction Questions (FP): Questions that prompt individuals to articulate their aspirations, predictions, or plans regarding their personal or professional future.
- Insightful Questions (IS): Questions that encourage individuals to share their unique perspectives or understanding of an individual's related subject or experience.
- Interest Questions (IT): Questions that explore how a person's hobbies, passions, or areas of interest shape their views, experiences, or future aspirations.

Question Generation We developed a framework using GPT-4 to generate both general and specific questions. For general questions, GPT-4 created five unique questions for each type, totaling 25, without considering the imitator's background. For specific questions, GPT-4 crafted five questions tailored to each imitator's background. Each imitator received ten questions—five specific and five randomly chosen general questions from each general question type—ensuring a diverse evaluation across baselines.

The questions are initially sourced and designed from diverse platforms, including asking different people the list of question types that can help distinguish a real person and an LLM in social media and looking for different academic research focusing on distinguishing real persons from LLMs, *i.e.*, asking about some general questions, such as daily activities and emotion (Jones and Bergen, 2023).

Initially, our setup faced challenges in LLM-generated questions being extremely specific to individuals' backgrounds, leading to complex questions that both participants and evaluators found challenging to understand, affecting the evaluation process. For example, questions on specialized topics like gut microbiota in human health, while relevant, often exceeded the general knowledge scope of participants, rendering their responses ineffective. Additionally, evaluators without specific knowledge could resort to random guesses rather than informed evaluations. To address this, we implemented a selective filtering process to ensure that questions match the testers' general English proficiency and intellectual level while aligning with their unique experiences and knowledge. This process helps maintain the relevance and fairness of the evaluation by tailoring questions to be comprehensible yet reflective of each tester's background, thereby avoiding excluding overly specific questions from the analysis.

3.3.3 Conducting Turing Tests

Human Evaluation The evaluation employs a cyclical process where roles between individuals and LLMs alternate, ensuring a thorough assessment. Each cycle features an individual and an LLM-side group answering questions, with models treated as separate baselines. Questions are generated for both parties, with human responses remaining constant for comparison. LLM responses, tailored with background information data from the human participants, aim to mimic their personas closely. Responses are anonymized and randomized for unbiased evaluation by at least one person familiar with the human subject, focusing on tone, thought process, and identification accuracy. This cycle repeats, with performance feedback informing the next round's human-side selection.

ECHO aims to overcome traditional Turing test limitations by conversational traps that could skew assessments, such as reliance on context-specific inquiries, *e.g.*, asking about current time or falling prey to the ELIZA effect. By focusing on a structured Q&A format, ECHO provides a direct response comparison method, improving the fidelity of role-play assessment. This structured approach prioritizes evaluating ideas, thoughts, and writing style, offering a nuanced view of LLMs' ability to replicate human interaction nuances.

Meanwhile, pre-processing responses are also included to remove syntactical biases that might influence evaluators, such as inconsistencies in capitalization, the omission of spaces between words, and the correction of misspelled terms. This ensures evaluations focus on the content's authenticity and coherence rather than superficial patterns while maintaining the original tone and style of responses. Consequently, evaluations are rooted in the genuine quality of ideas and thoughts, providing a fairer assessment of each response's substance.

Response Collection We collect responses from two distinct sources: the individuals being imitated and their corresponding LLM imitators. To ensure the integrity and impartiality of the study, we implement a randomized shuffling mechanism. This involves randomizing the order of the ten questions from 10 different question types and the corresponding answers before incorporating them into the questionnaire. Such a shuffling process is critical to minimize potential bias from fixed question or answer orders, thus providing a more robust test of the LLMs' imitation abilities.

Results The final score will be the number of correct choices. The effectiveness of the LLMs in mimicking human responses is quantified by the success rate calculated from these participants' choices.

3.4 Experiments

In our experiments, ten participants from varied backgrounds serve as real individuals, leading to a ten-round baseline evaluation. Additionally, at least seven judges familiar with each human imitator are assigned to independently evaluate the mix of responses from humans and LLMs, with all participants and judges possessing tertiary education levels to ensure adequate English proficiency. During response collection, we utilize Google Forms[3] for data collection and management. The test groups, familiar with the real individuals, are entrusted with classifying the human answer to the ten questions.

Baseline Methods Our experiment involves evaluating four popular approaches to role-playing LLMs. We integrate GPT-3.5-Turbo and GPT-4-Turbo as backbone models for each baseline method, except for GPTs, resulting in a total of

[3] https://www.google.com/forms/about/

7 baselines. Given that some baseline models were not directly accessible, we adapted their concepts using LangChain[4] to ensure a comprehensive comparison across various models. The implementation details are elaborated in §?? in the appendix. The baseline methods include:

- **RoleGPT** (Wang et al., 2023): enhances role-playing in LLMs through a four-stage process including role profile construction for 100 roles, context-based instruction for knowledge extraction, role prompting with GPT for style imitation, and role-conditioned instruction tuning. RoleGPT, our version, focuses on prompt engineering due to the demo model's unavailability.

- **Role-Play Prompting (RPP)** (Kong et al., 2023): Introduces a methodology to boost zero-shot reasoning in LLMs via role-play prompting, allowing them to assume various personas. The method samples multiple role-feedback prompts and selects the optimal one for answering reasoning queries, acting as an implicit Chain-of-Thought trigger to improve LLM reasoning.

- **Juliet** (Jones and Bergen, 2023): Evaluates GPT-4's performance in passing the Turing Test in online settings by testing 25 LLM witnesses (including GPT-3.5 and GPT-4) with human participants. The study found that the best GPT-4 prompt convincingly mimicked human behavior, achieving a 41% success rate in deceiving participants about its human likeness.

- **GPTs** (OpenAI, 2023): A novel feature by OpenAI that allows for creating customized ChatGPT applications for specific tasks using only natural language and data files. These custom applications can be shared via links or the GPT store. We selected several GPTs designed for person imitation for our experiment.

[4] https://www.langchain.com/

Table 1: Success rates of role-playing LLMs in deceiving human evaluators. The human evaluators are instructed to identify **human**-generated responses.

Success Rate (%)	GPT-3.5-Turbo			GPT-4-Turbo			GPTs	Overall
	RPP	RoleGPT	Juliet	RPP	RoleGPT	Juliet		
Creativity	40.0	**53.3**	31.3	26.1	37.0	37.5	47.8	39.0
Ethical Dilemmas	43.5	30.0	44.4	38.9	27.3	44.4	**47.8**	39.5
Logical	23.5	**50.0**	36.4	42.1	47.6	47.1	41.7	41.2
Philosophical	26.7	38.9	43.5	44.0	28.0	40.9	34.8	36.7
Problem Solving	17.4	23.3	34.8	46.2	46.7	48.0	**54.6**	38.7
In-depth Personals	42.1	45.2	40.0	35.0	**83.3**	41.7	56.0	49.0
Emotional	44.4	57.9	22.2	**66.7**	25.0	55.6	45.8	45.4
Future Prediction	38.9	59.1	37.5	**60.0**	50.0	50.0	50.0	49.4
Insightful	50.0	34.8	**61.5**	45.0	50.0	35.5	50.0	46.7
Interest	48.0	41.7	30.0	**66.7**	22.7	33.3	53.9	42.3
Overall	37.5	43.4	38.2	47.1	41.8	43.4	**48.2**	42.8

Figure 3.2: The Success rates of role-playing LLMs in deceiving human evaluators.

3.4.1 Results

Across Baselines Figure 3.2 displays the success rate of role-playing LLMs in deceiving human evaluators. In our experiments, the human evaluators are instructed to identify *human*-generated responses. In general, GPTs outperforms other baselines in varied question types. With customized construction, GPTs enables more accurate imitation by specifying instructions with enriched personal information. This approach contrasts with general human imitation methods, suggesting that specificity is crucial for enhanced mimicry.

It is also observed that upgrading from GPT-3.5-Turbo to GPT-4-Turbo significantly improved mimicry accuracy, especially with the GPT-4-Turbo model showing a substantial increase in success rates, following the hypothesis that upgrading the model can enhance the model's capacity for mimicking individual writing and thinking styles more accurately. This enhancement is most notable in the RPP and Juliet frameworks, particularly with the GPTs model, which benefited from GPT-4-Turbo's refined ability to process and reflect the

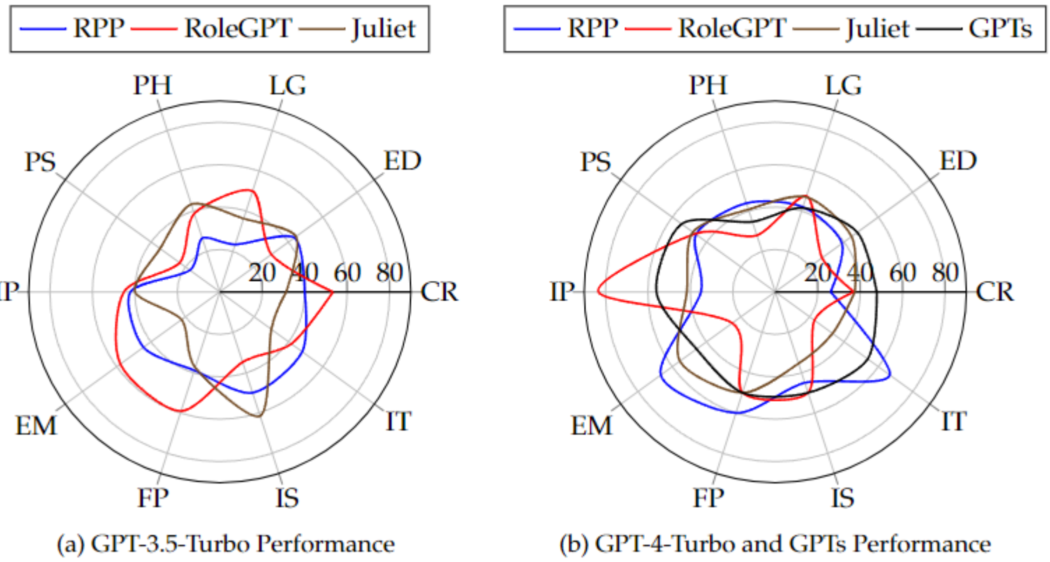


Figure 3.3: Success rates of role-playing LLMs in deceiving human evaluators. The human evaluators are instructed to identify **human**-generated responses.

nuances in background information more effectively. However, RoleGPT experiences reduced performance after upgrading to GPT-4-Turbo, possibly because the model's outputs tend toward extreme casualness or dramatics, which detract from the authenticity of its imitations. This indicates that GPT-4-Turbo's detailed comprehension might lead to style exaggerations that affect perceived genuineness.

Across Question Types Success rate analysis across different question types reveals how models, particularly GPT-3.5-Turbo and GPT-4-Turbo, fare with general versus specific queries. GPT-3.5-Turbo encounters challenges with emotional and complex problem-solving questions. Juliet struggles in EM, RPP and RoleGPT with LG and PS, indicating a potential lack of nuanced emotional and multi-step logical processing. The upgrade to GPT-4-Turbo generally improves success rates, especially for specific questions. For instance, RoleGPT's performance in IP surpasses 80%, the highest success rate in the whole table, and RPP's performance suppresses 60% over 3 out of 5 metrics in specific question types, a testament to the improved comprehension and response generation afforded by GPT-4-Turbo, especially when dealing with intricate personal details. However, this improvement in handling specific questions suggested a possible over-specialization, affecting performance on more general queries.

Both the Juliet and GPTs models demonstrated balanced performances across question types, with GPTs notably outperforming Juliet, underscoring the effectiveness of nuanced prompt engineering in yielding human-like responses. The observed trend of superior performance in specific over general questions reflects the models' design focus, revealing that detailed, personalized responses are more effectively mimicked than broad, abstract topics. General questions, particularly PH and PS, pose challenges due to their abstract demands and need for clear-cut answers, pushing beyond LLMs' strengths in data-driven reasoning toward areas requiring speculative thought or creative problem-solving. This creates a discernible contrast between human and AI-generated responses, as LLMs may struggle with the creative or interdisciplinary thinking needed for such questions, often making AI responses easier for evaluators to spot.

The variance in baselines' performance across different question types stems from the specific prompt engineering strategies employed during model training. By tailoring prompts to mimic a particular individual, the LLM's understanding of and response to the background information are influenced, leading to specialized imitation capabilities across various question domains. This underscores the impact of prompt design on LLM performance, highlighting how different imitation strategies can enhance or detract from a model's ability to replicate human-like responses authentically.

Table 2: Success rates of role-playing LLMs in deceiving evaluator LLMs. The evaluator LLMs are instructed to identify **human**-generated responses.

Success Rate (%)	GPT-3.5-Turbo			GPT-4-Turbo			GPTs	Overall
	RPP	RoleGPT	Juliet	RPP	RoleGPT	Juliet		
GPT-4	85.3	92.3	88.3	63.7	93.0	91.3	**95.7**	91.4
GPT-4-Turbo	95.0	94.0	95.3	95.7	**99.0**	98.0	98.3	96.5
Gemini-1.0-Pro	52.7	52.7	**62.7**	56.3	60.7	58.3	54.0	56.8
Verbosity Bias	86.0	78.0	67.0	95.0	31.0	5.0	78.0	62.9

Table 3: Success rates of role-playing LLMs in deceiving evaluator LLMs. The evaluator LLMs are instructed to identify **non-human**-generated responses.

Success Rate (%)	GPT-3.5-Turbo			GPT-4-Turbo			GPTs	Overall
	RPP	RoleGPT	Juliet	RPP	RoleGPT	Juliet		
GPT-4	25.7	24.7	26.0	25.7	29.0	**52.3**	11.7	27.9
GPT-4-Turbo	61.7	**62.7**	53.3	34.3	60.0	58.0	62.3	56.5
Gemini-1.0-Pro	51.0	49.0	42.3	48.7	**54.3**	50.0	48.7	41.0
Verbosity Bias	14.0	22.0	33.0	5.0	69.0	95.0	22.0	37.1

Figure 3.4: Result of auto evaluation

3.4.2 Discussion: LLM as Evaluators

LLM-based evaluators have demonstrated their potential in identifying the quality of machine-generated texts. Although some research works point out that the positional and verbosity biases of LLM evaluators may skew their preference toward longer responses or affect their judgment based on answer order (Zheng et al., 2024), a recent study suggests that those biases are not impactful in several models like GPT-4-Turbo (Chen et al., 2024). In this section, we study the performance of LLM evaluators and their distinction with human evaluators.

Evaluation Strategy To examine verbosity bias, we establish a baseline mimicking a model predisposed to this bias by consistently selecting the longer answer for identification tasks. Proximity in success rates between this simulated model and actual evaluators would suggest that the evaluated models are significantly influenced by verbosity bias, indicating a preference for longer answers in their identification process.

To mitigate potential positional biases, we introduce a two-fold mitigation strategy. Initially, the sequence of answers' position inside the QA pairs is randomized before presentation to the LLM. Subsequently, the evaluation is performed in multiple rounds with the same QA set, to calculate the average success rate for calibrating the effect of positional bias.

Our evaluation methodology involves assessing responses from ten participants across seven baselines, where participants respond to ten questions. We pair each participant's responses with those from baseline LLMs to form answer sets for analysis. These sets, supplemented with background information, are presented to LLMs to distinguish between human and LLM-generated answers over three iterations. This dual-focused assessment aims to measure LLMs' proficiency in identifying human and their own generated responses, calculating an average success rate for accuracy.

Results The results of LLMs as evaluators are shown in Figure 3.4. Discerning whether a human or an LLM produces an answer in our analysis translates into a binary classification problem. A success rate that significantly deviates from the 50% benchmark, which is expected in random guessing, indicates the LLM's capability to distinguish between human and machine-generated text, irrespective of the accuracy of that distinction. Both tables show that GPT-4 is proficient in differentiating answers, leveraging indicators like patterns, writing styles, and thought processes. While Gemini-1.0-Pro only exhibits at most 9% differences with the random guessing benchmark in choosing either human or non-human generated responses, exhibiting difficulty in effectively discriminating between the two answer types, mirroring near-random guessing performance.

A salient observation from our study is the significant impact of the task's defined goal on the results. In Table **??**, the GPT series both have a notably weak performance when identifying the human-generated answers but significantly better performance with the same setup, but just identifying the LLM-generated answer in Table **??** with 63.5% (GPT-4) and 40.0% (GPT-4-Turbo) increase. This outcome necessitates the distinction between two capabilities: the "power of identifying human answers" and the "power of identifying LLM answers," which our findings suggest are not inherently balanced.

In a binary case scenario, one might anticipate comparable accuracy in identifying human and LLM-generated responses. However, our results diverge from this assumption. When distinguishing human-generated answers, LLMs need to grasp human writing styles deeply for accurate identification. It might seem plausible for LLMs to first identify all LLM-generated responses and then apply a process of elimination. However, this intuitive strategy, common in human decision-making, must be revised in multiple-choice settings for LLMs.

Comparing the identification power within the GPT series, GPT-4 shows a considerable discrepancy in its abilities to identify human versus LLM-generated answers, achieving only about an 8% success rate for human answers compared to approximately 70% for LLM answers. This stark contrast underscores GPT-4's stronger predisposition for identifying LLM-generated content. Furthermore, when employing GPT-4-turbo, we observe a similar trend but with a diminished capability in recognizing LLM answers, indicating that while GPT-4-turbo maintains a preference for detecting LLM-generated content, its proficiency in doing so is reduced compared to the standard GPT-4 model.

For the verbosity bias, Table **??** and Table **??** illustrate a significant discrepancy in success rates between that baseline and GPT models in identifying human-generated answers, with at least a 10% difference in detecting LLM-generated responses. This indicates that verbosity bias minimally influences model choices, aligning with findings from (Chen et al., 2024).

Chapter 4

Application-side Architecture

4.1 Imitation System

4.1.1 Introduction

The Imitation System is central to our project, designed to replicate celebrities' unique communication styles and thought patterns through advanced natural language processing. This system intricately combines firsthand and third-party information to create a robust database, facilitating the generation of authentic responses that mirror a chosen celebrity's style. Our development process has evolved from basic prompt engineering to sophisticated model fine-tuning, aiming to produce responses that convincingly appear as if they were directly authored by the celebrity. This report outlines the evolution of our prototypes, highlighting the modifications, improvements, and persisting challenges across different versions.

4.1.2 Applying ECHO result

The research conducted on role-playing LLMs such as GPT-4 has profound implications for celebrity imitation applications, demonstrating that the success of such models depends heavily on the use of diverse prompt engineering techniques across different baselines. This study highlights the necessity of an integrative approach that combines various methods, rather than relying solely on GPT-4' s capabilities. Tailored prompt engineering, designed to match the specific nuances and communication styles of celebrities, significantly enhances the authenticity and accuracy of the models' outputs. By adopting a multifaceted strategy that involves continuously refining prompts and model parameters based on systematic evaluations, we can achieve more realistic and convincing imitations, thereby improving the effectiveness of LLMs in applications that require nuanced human-like interactions.

Challenges in Celebrity Imitation

One major challenge in using LLMs for celebrity imitation stems from the difficulty in evaluating the effectiveness of these models due to the limited availability of comprehensive data on some celebrities or historical figures like classic film stars, or influential personalities from the era before digital media, and the variability in human evaluators' familiarity with the celebrity being imitated. These evaluators often lack the nuanced understanding of the celebrity's communication style needed to judge the AI's performance accurately.

Strategy for Effective Imitation

To address these challenges, the idea emerged to shift our initial imitation attempts from celebrities to real individuals who volunteers could more readily and familiarly evaluate. This approach allows for a more controlled and measurable evaluation environment, where the nuances of the AI-generated responses can be more effectively assessed for accuracy in style and content as part of the reference. Meanwhile, we adopt the prompt engineering ideas from other baselines. For instance, we adopted the Context-Instruct prompt from RoleLLM (Wang et al., 2023) for creating detailed QA profiles that aid in feature extraction from online sources, as discussed in Section 2.1.5. Similarly, the guidelines from the Juliet prompt are utilized to produce responses that are more nuanced and human-like. This comprehensive approach not only refines the effectiveness of our imitations but also serves as a valuable reference for scaling our methods to celebrity figures, ensuring that the AI-generated responses are both authentic and relatable.

Results from Role-Playing LLMs

Our results highlighted in Figures 3.2 and 3.3 show that LLMs, particularly those upgraded to GPT-4-Turbo, have achieved higher success rates in deceiving human evaluators who were tasked with distinguishing AI-generated responses from human-generated ones. The success of these models in role-playing scenarios supports their potential use in celebrity imitation, provided that the data used for fine-tuning includes diverse and contextually rich inputs that cover the broad spectrum of the celebrity's public and private communications.

4.1.3 Transition of Prototypes

Prototype v1

Prototype v1 utilized a zero-shot single prompt approach to instruct the LLM to answer user-input questions by imitating a celebrity, based on data from our extensive database. This prototype also featured a conversational memory buffer to maintain long-term consistency.

However, this version revealed significant shortcomings as 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, as the responses 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 viewpoints authentically.

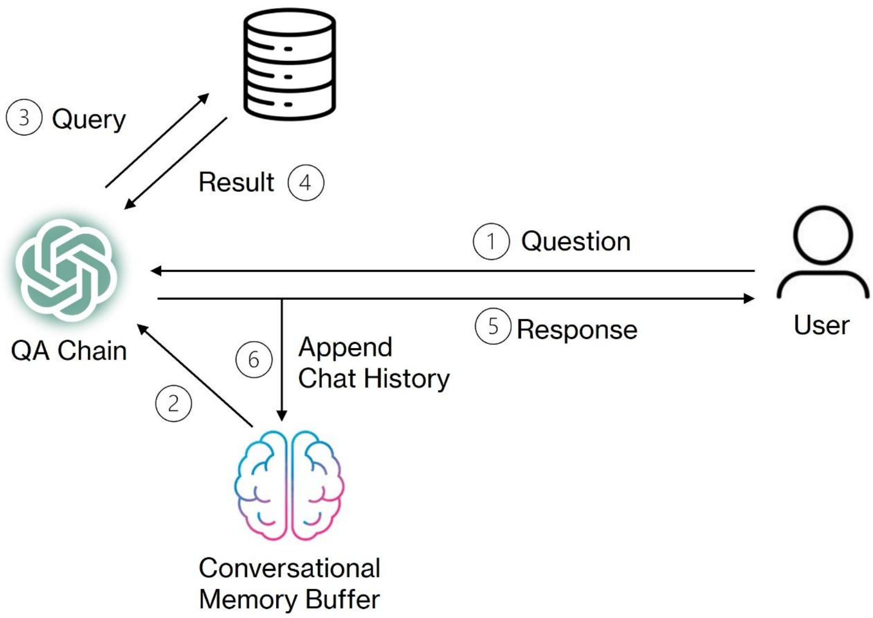


Figure 4.1: Prototype v1

Prototype v2

Prototype v2 aimed to address the limitations observed in the first version by introducing a multiple-prompting system combined with a knowledge bagging approach. This version utilized a LLM to generate a QA chain based on the celebrity's background information fetched from a vector database, enhancing response accuracy and relevance.

While this prototype marked an improvement in generating linguistically aligned responses, it encountered challenges in accurately mimicking the celebrity's thinking style and logical reasoning. The responses were often compilations of existing data, lacking genuine reflection of the celebrity's unique thought process, thereby resulting in superficial representations of the celebrity's persona.

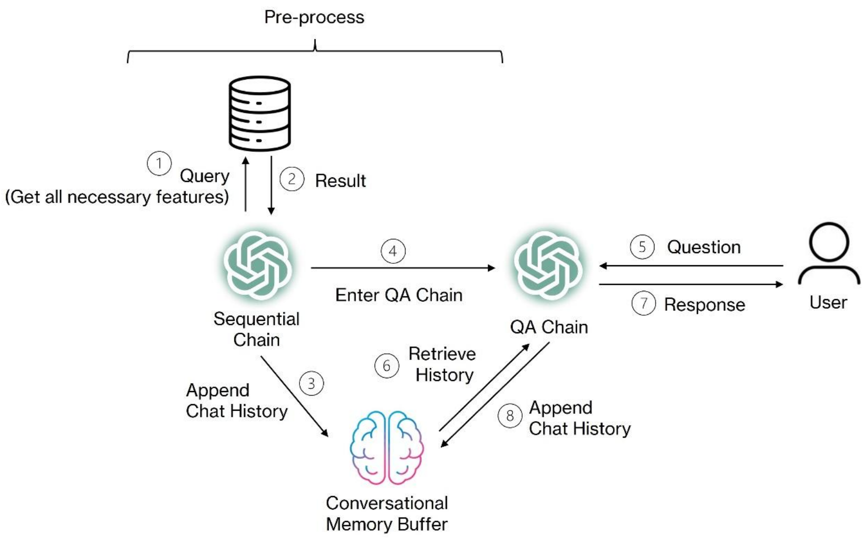


Figure 4.2: Prototype v2

Prototype v3

Prototype v3 was developed with the 'Tree of Thought'(Yao et al., 2023) structure to better reflect the celebrity's thinking style and personality using language characteristic of the celebrity. This involved a two-chain process: the 'Idea and Thought Chain' and the 'Writing Style Chain,' which helped refine responses to align closely with the celebrity's typical language patterns.

This approach, while innovative, still allowed users to discern that responses were AI-generated due to subtle nuances in language use not fully captured by the model. The formal tone and atypical punctuation usage sometimes made the

AI responses easily distinguishable from those of a human.

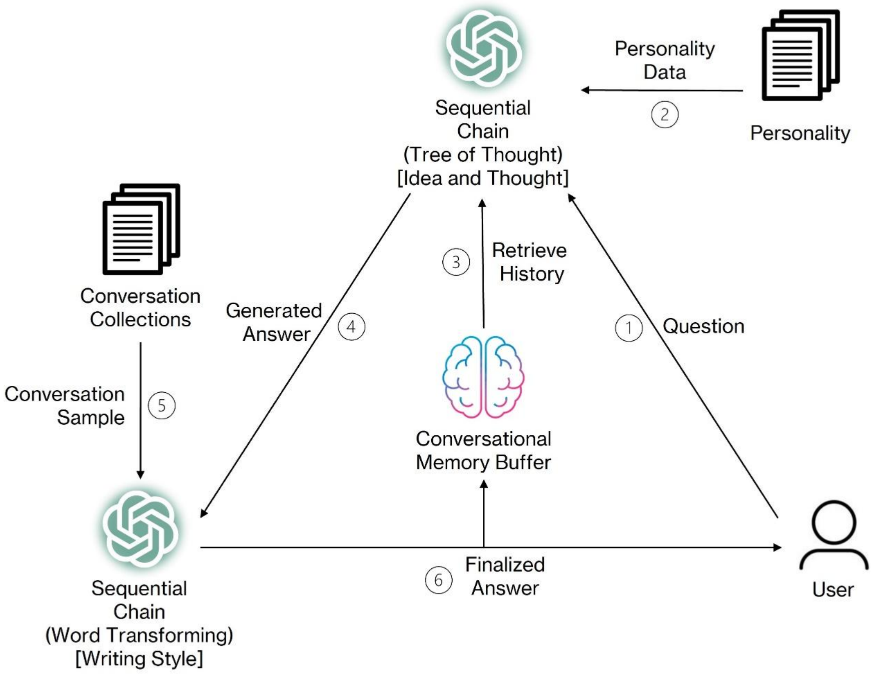


Figure 4.3: Prototype v3

Prototype v4

Prototype v4 refined the original structure from a ’Tree of Thought’ to a ’Chain of Thought,’(Wei et al., 2022) focusing on the ’Idea and Thought Chain’ and the ’Wordings and Punctuation Chain.’ This revision aimed to solve the issues seen in Prototype v3 by enhancing the precision of wordings and punctuation to better match the celebrity’s style.

This structural refinement allowed for more accurate mimicry of the celebrity’s linguistic habits and thought patterns, producing responses that were not only stylistically accurate but also reflective of the celebrity’s personal views and mannerisms. Thought, this system cannot follow the celebrities’ writing style and wording when answering.

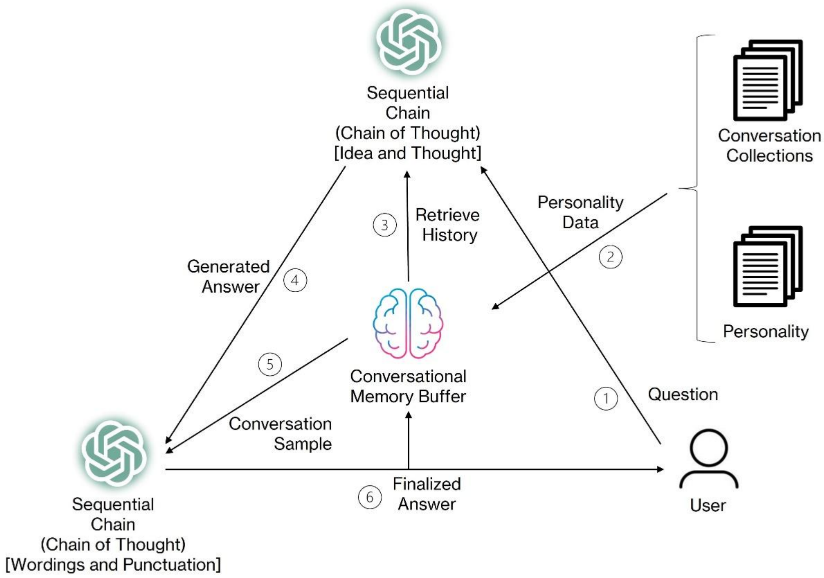


Figure 4.4: Prototype v4

Prototype v5

In the latest iteration of our imitation prototype, significant improvements have been implemented to more accurately capture the unique writing styles and phrasing of the celebrities it aims to emulate. Previous versions, including the fourth iteration of the prototype, struggled to accurately mimic these stylistic nuances due to limitations in the methods used for analyzing and integrating the celebrities' conversational data.

The initial approach attempted to replicate celebrities' writing styles by generating hypothetical QA sets from limited conversational data, which was stored as personality data in the conversation memory buffer. However, this method was ineffective for several reasons. The primary issue was the inconsistency in writing styles exhibited by celebrities in different contexts, such as conversations with friends versus interactions with the public or in interviews. This variation in style was not adequately captured by the system, resulting in outputs that often did not reflect the true stylistic patterns of the celebrity. Additionally, the limited data available in the conversation memory buffer was insufficient to cover the broad spectrum of linguistic nuances that define a celebrity's public and private personas.

To overcome these challenges, Prototype v5 introduced significant changes. A Chain-of-Thought architecture has now been integrated, enhancing the system's ability to generate coherent and contextually appropriate responses. This new setup refines the process by directly incorporating a language model fine-tuned specifically with data representing the celebrity's style. Unlike the previous method, which relied heavily on prompting, fine-tuning adjusts the model's underlying behavior to better reflect the nuances of the celebrity's communication style.

This specialized language model can now more effectively convert the developed thoughts into responses that sound as if they were written by the celebrity themselves. However, it is important to note that due to resource limitations, only a smaller model like GPT-3.5 has been employed (OpenAI, 2024b). This necessitates a focus on simpler tasks, such as adapting the generated responses to match the writing style, rather than more complex cognitive simulations like replicating the celebrity's thought processes in entirety.

4.1.4 Conclusion and Future Directions

Prototype v5 represents a significant step forward in our ability to create AI-driven content that not only mimics the writing style but also the thinking style of celebrities. The use of a fine-tuned, smaller model addresses the practical limitations of computing resources while still achieving a high degree of stylistic accuracy. Moving forward, we plan to further refine these techniques, potentially integrating more complex models and broader data sets to tackle the more intricate aspects of personality emulation. This future work will aim to enhance the system's capability to handle diverse and dynamic interaction scenarios, ultimately making the AI-generated content indistinguishable from that of the actual celebrities.

These developments will continue to build on the foundational research and prototypes, pushing the boundaries of what is possible in the realm of AI-generated personal communication and public interaction.

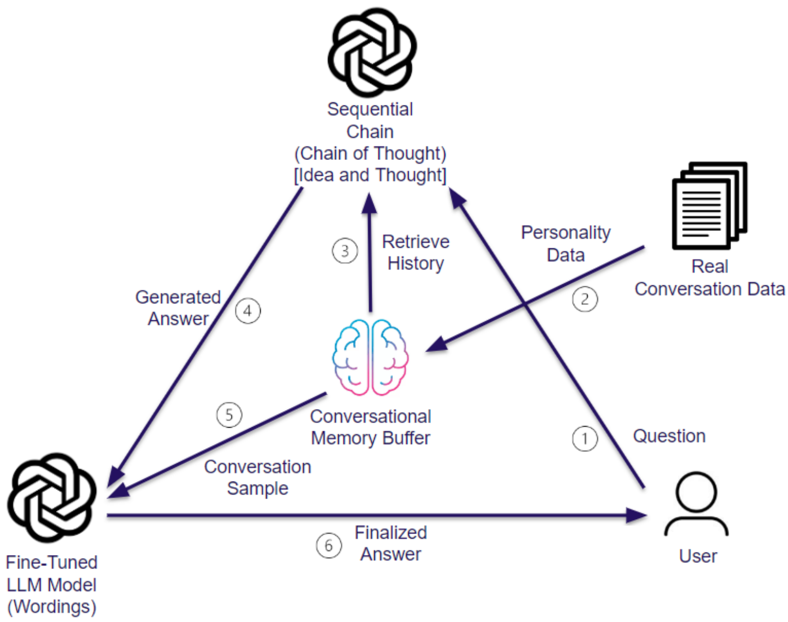


Figure 4.5: Prototype v5

4.2 Posting System Architecture

4.2.1 Changing new architecture

Over the past semester, the core of the imitation system was developed, which involved fetching data, converting this data into a QA (Question and Answer) format for feature extraction related to celebrities, and establishing the foundational architecture of the system. This semester, following the initial implementation, the objective was to integrate this system to enable a Language Model to autonomously post content on various social media platforms, mimicking the style of specific celebrities.

Several challenges arose when utilizing the basic LangChain framework. Firstly, the system's workflow involved multiple cycles which LangChain's linear chain structure struggled to represent effectively. This limitation complicated the implementation of dynamic processes such as the search agent, which required continuous data search and summarization simultaneously. Secondly, the lack of support for cyclic workflows in LangChain meant difficulties in instructing the search agent when to cease operations. For instance, determining the sufficiency of information gathered was problematic; the agent often entered a state of perpetual searching, believing more data was needed, potentially leading to endless loops. Additionally, as the system's complexity increased, tracking the decision-making process and data flow became increasingly challenging. This complexity made it difficult to diagnose and rectify bugs, particularly when integrating new functionalities.

To address these issues, the system was re-implemented using LangGraph (LangChain, 2024b), an extension of LangChain that incorporates a graph-based structure. This new architecture provided several improvements. The system was restructured into a network of nodes (agents) and edges (connections), enhancing modularization and flexibility. This was particularly beneficial for adapting the system's output to various social media platforms. LangGraph's support for cyclic dependencies through conditional edges significantly mitigated the risk of endless loops within the search agent, providing robust control over workflow processes. Furthermore, utilizing LangSmith (LangChain, 2024c) alongside

LangGraph simplified the representation of the system's overall structure. This integration facilitated easier streaming and interpretation of token/node outputs, thereby improving the debugging process.

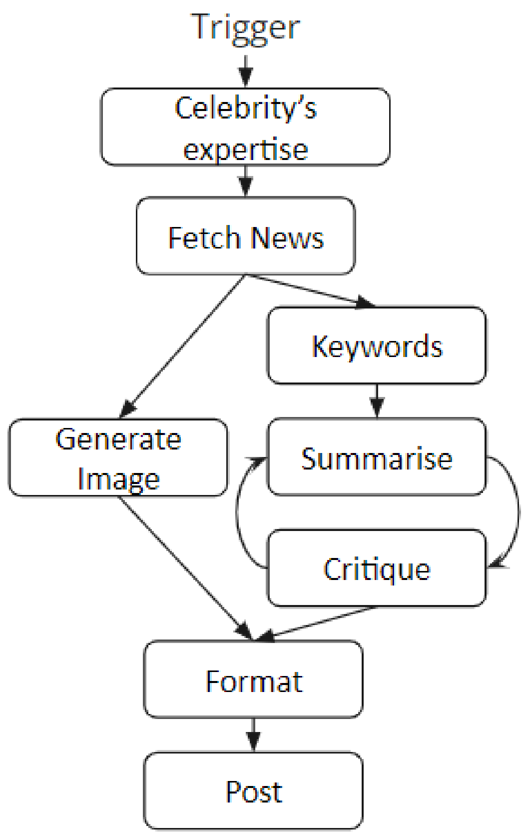


Figure 4.6: Whole System's Graph Architecture

The enhanced system is organized into five main categories. The Retrieval Augmented Generator (Lewis et al., 2020) focuses on enhancing content generation by leveraging retrieved data to produce more accurate and relevant outputs. The Image Generator System synthesizes visual content that complements the text, tailored to the style and preferences of the imitated celebrity. The Data Augmentation System augments the available data to refine the inputs for both text and image generation modules, ensuring high-quality content creation. The

Article Enhancement System improves the quality of generated articles by refining their language, structure, and overall presentation to better resonate with targeted audiences. Finally, the Layout Finalizing System adjusts the layout of the generated content to meet the specific formatting and stylistic requirements of different social media platforms.

These enhancements and restructuring efforts have significantly improved the flexibility, control, and efficiency of the posting system architecture, paving the way for more effective and autonomous social media interactions by the Language Model.

4.2.2 News Fetching System

In order to get the first-hand information, we built an agent that perform web scrapping form the source BBC news rss(bbc) and further stored the news into a data type that contain the title, a short description, url link to the news and the publish date.

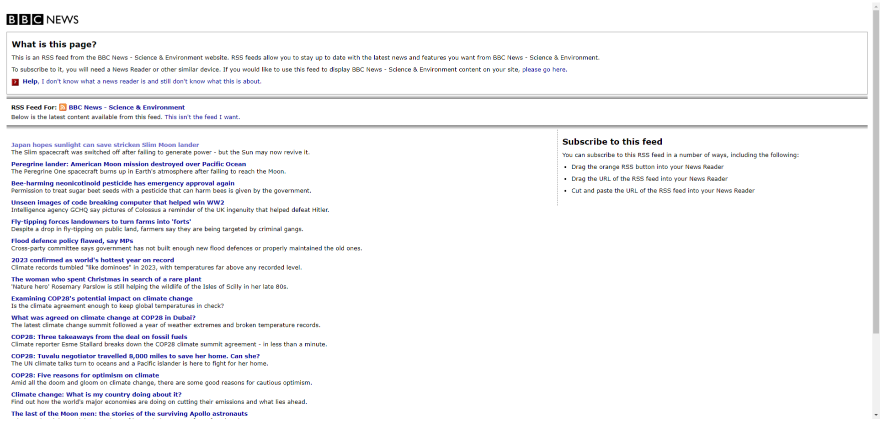


Figure 4.7: News source from BBC News

	title	link	description	pubDate
0	Japan hopes sunlight can save stricken Slim Mo...	https://www.bbc.co.uk/news/science-environment...	The Slim spacecraft was switched off after fai...	Mon, 22 Jan 2024 10:43:31 GMT
1	Peregrine lander: American Moon mission destro...	https://www.bbc.co.uk/news/science-environment...	The Peregrine One spacecraft burns up in Earth...	Fri, 19 Jan 2024 04:19:24 GMT
2	Bee-harming neonicotinoid pesticide has emerge...	https://www.bbc.co.uk/news/science-environment...	Permission to treat sugar beet seeds with a pe...	Thu, 18 Jan 2024 19:04:50 GMT
3	Unseen images of code breaking computer that h...	https://www.bbc.co.uk/news/technology-67997406...	Intelligence agency GCHQ say pictures of Colos...	Thu, 18 Jan 2024 00:02:41 GMT
4	Fly-tipping forces landowners to turn farms in...	https://www.bbc.co.uk/news/science-environment...	Despite a drop in fly-tipping on public land, ...	Wed, 17 Jan 2024 16:24:49 GMT

Figure 4.8: Fetched News List

4.2.3 Image Generator System

After collecting news content, the implementation of the image generation process begins, aiming to produce visual support for the final posts to enhance their appeal. The inputs for the Image Generator System and the Retrieval Augmented Generator (RAG) are the same, yet these systems operate independently and can function in parallel, optimizing performance by reducing processing time.

Previously, a demonstration involved fine-tuning the stable diffusion model using LoRa to generate images of Oppenheimer. However, this approach was not continued for generating images based on article content due to several challenges. Primarily, running a stable diffusion model demands extensive GPU resources, which is impractical on a cloud platform. Moreover, the fine-tuning process, even with LoRa, required about three hours for the demo, which is too time-consuming for practical deployment within the system.

Consequently, the DALL-E 3 model (OpenAI, 2024a) is now used for the image generator system, which simplifies the image creation process. This model operates by receiving a list of prompts via an API call, based on the article content, thereby bypassing the intensive resource requirements of the previous model.

A significant challenge addressed in the integration process was how the generated images could be utilized appropriately within the generated posts. Initially,

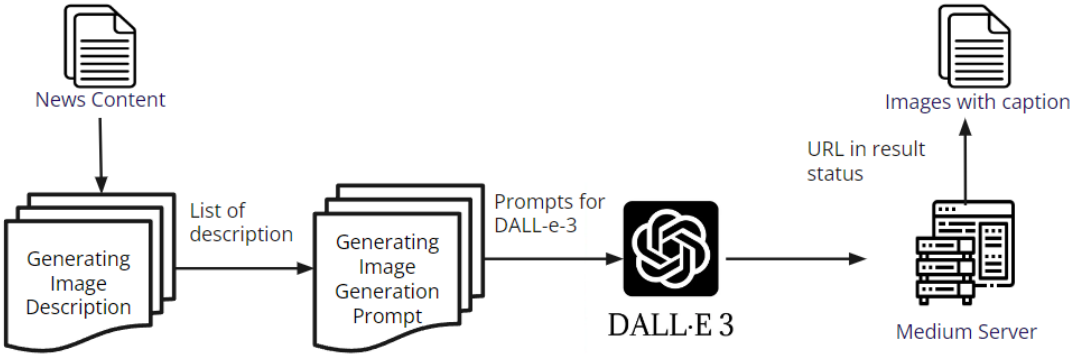


Figure 4.9: Workflow for Image Generator System

the integration of these images was problematic as GPT-4's API (OpenAI, 2024c) does not support direct analysis or placement of images based on their content. To resolve this, the system was designed to generate textual descriptions of the images before creating the images themselves. These descriptions serve as captions and provide contextual guidance that the Layout Finalizing System uses to accurately place images within the post layout.

For instance, if the article is about various facts about apples, the Language Model first generates textual descriptions for potential images, such as "A giant red apple" or "An apple tree." These descriptions are then used to create corresponding prompts for DALL-E 3, ensuring the images are relevant and of high quality. The final output includes both the generated images and their respective descriptions, which are effectively utilized during the formatting process to integrate images at suitable locations within the post.

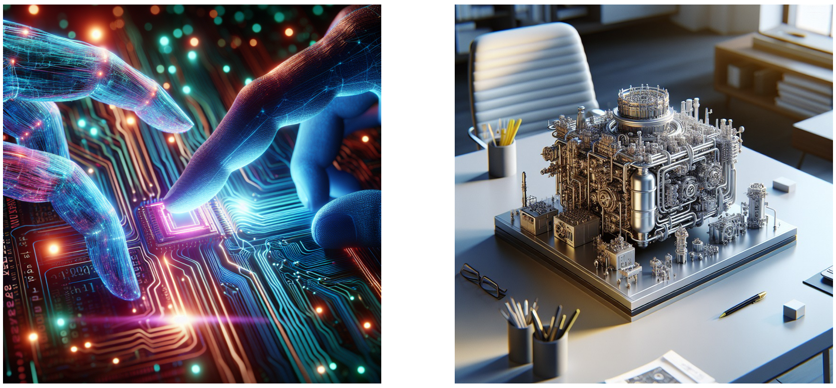
Moreover, reliance solely on the LLM-generated prompts for image creation can lead to the production of generic or vague images that do not effectively complement the article. To mitigate this, specific guidelines are incorporated within the system's prompts to ensure that DALL-E 3 generates images that are not only high quality but also appropriately aligned with the content and style of the article. This strategic guidance helps maintain the relevance and aesthetic appeal of the images, thereby enhancing the overall quality of the social media posts.

```
image_system_prompt = SystemMessagePromptTemplate.from_template(
    template="""
Human will give a paragraph which is a post content that going to be shared on social media.
You are good at prompt engineering, you can transform the paragraph into a stable diffusion prompt that can be used to generate a image.
Stable Diffusion is a tool that can generate image based on the prompt.

Good keyword prompts needs to be detailed and specific.
A good process is to look through a list of keyword categories and decide whether you want to use any of them.
IMPORTANT: you must never use these keyword category names as keywords in the prompt itself as literal keywords at all,
so always omit: "subject", "Medium", "Style", "Artist", "Website",  "Resolution", "Additional details", "

The keyword categories and their definitions are:
Subject: The subject is what you want to see in the image.
Resolution: The Resolution represents how sharp and detailed the image is. Let’s add keywords highly detailed and sharp focus.
Additional details: Any Additional details are sweeteners added to modify an image, such as stunningly beautiful and dystopian
to add some vibe to the image.
Color: always "Black and White"
Lighting: Lighting is a key factor in creating successful images (especially in photography). Lighting keywords can have a
huge effect on how the image looks, such as cinematic lighting or dark to the prompt.
```

Figure 4.10: Refined Prompt for generated pictures



(a) Original Prompt

(b) New Prompt

Figure 4.11: Comparison of Dall-E generated images using different prompts

By refining the prompt generation process (PaulBellow, 2023) and enhancing integration techniques between the Image Generator System and the Layout Finalizing System, the system effectively supports the creation and placement of visually appealing, contextually appropriate images that enhance the impact and attractiveness of the final posted content.

4.2.4 Data Augmentation System

The Data Augmentation System is designed to address the limitations posed by the reliance on original news content, which often cannot be extended, especially when such news includes domain-specific expertise and pronouns that are not universally understood. This limitation can lead to suboptimal results when the content does not cover the knowledge domain adequately.

To enhance the understanding and extend the content of such articles, a new methodology was proposed. This methodology involves fetching the keywords and pronouns that frequently appear in the article and are likely to be complex for a highschool student to understand. Utilizing these keywords, the system employs a Bing search agent to fetch external information, which includes links that can be referenced back in the article to provide a deeper context and supplementary data.

The operational flow of the Data Augmentation System starts with the extraction of up to ten key terms from the article. These keywords are then transformed into questions that serve as queries for the Bing search. Each query result is systematically stored in a vector database, including the fetched information alongside the questions and their corresponding URLs. This process ensures that all relevant external information is structured and retrievable.

Once all pertinent data is fetched and stored, the system integrates this information with the original news article. This integration involves summarizing the original and the augmented data to produce a comprehensive version of the article that is enriched with external references and broader insights. This augmented version not only extends the article's content but also enhances its educational value, making complex information accessible and understandable to a wider audience, including those without specific domain knowledge.

By leveraging external data sources through intelligent querying and structured storage, the Data Augmentation System effectively broadens the scope and depth of news articles, resulting in more informative and comprehensively enhanced content. This system not only improves the quality of the information but also ensures that it caters to a more diverse audience by clarifying complex terms and integrating essential background information.

4.2.5 Article Enhancement System

The Article Enhancement System employs the principles of generative adversarial networks (GANs) (Goodfellow et al., 2014) to refine and improve the quality of articles. This system consists of a loop with two main components: a content summarization/revision agent (the writer) and a critique agent (the critic). The design is conceptualized with the writer acting as the creator of the article content, while the critic evaluates the output, providing detailed feedback on which parts require revision.

The critique agent systematically assesses the articles generated by the writer, indicating necessary improvements and suggesting specific areas for refinement. This feedback is then passed back to the writer along with the original content, prompting the generation of a revised, enhanced article. This iterative process is designed to simulate a dynamic environment where continuous improvement is facilitated through constant critique and revision.

To prevent the process from becoming an infinite loop of revisions, a maximum iteration count is set, typically limited to three iterations. This cap is strategically set to balance the quality of the article with practical considerations such as cost and time efficiency. While a higher number of iterations could theoretically produce a more polished article, it also increases the length of the article and potentially the complexity, which may not necessarily enhance its overall quality. Extensive revisions could lead to a denser, more convoluted article, detracting from its readability and effectiveness.

Further enhancing the system's realism and effectiveness, the summarizer agent incorporates an imitation prototype designed to mimic the thinking style of a specific celebrity. This integration allows the summarization process to reflect the unique style and preferences of the imitated celebrity, making the content not only high quality but also tailored to resemble the celebrity's own words and thoughts. The critique agent then works to refine this content further, ensuring that the final output is not only stylistically consistent with the celebrity's manner but also clear, concise, and of high quality.

By combining the iterative critique-and-revise approach with celebrity-style imitation, the Article Enhancement System effectively produces engaging and well-crafted articles. This system not only improves the readability and appeal of the content but also ensures that it resonates well with the intended audience, reflecting a deep understanding of the celebrity's distinctive communication style. This dual approach of continuous refinement and stylistic imitation distinguishes the system as a sophisticated tool for content enhancement in the realm of automated article generation.

4.2.6 Layout Finalizing System

The Layout Finalizing System is crucial for preparing and optimizing the content for publication on various social media platforms, such as Medium (Medium, 2024). This system utilizes both the summarized content and the images generated by the earlier systems to produce a polished final post. The process involves several key steps: formatting the article with suitable HTML tags, inserting im-

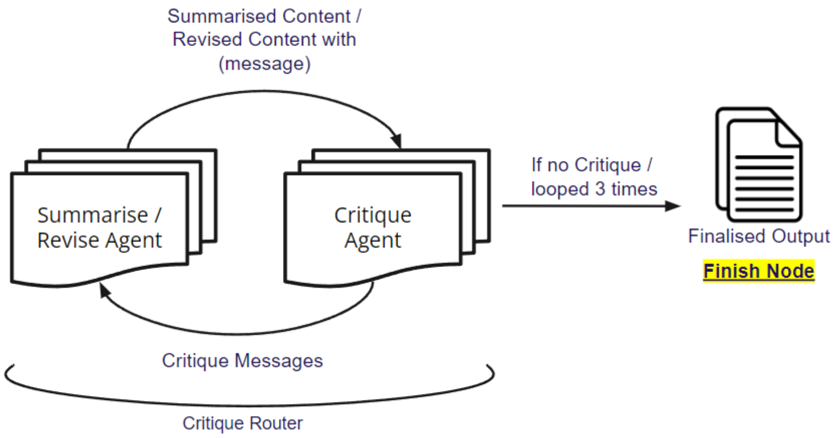


Figure 4.12: Workflow for Article Enhancement System

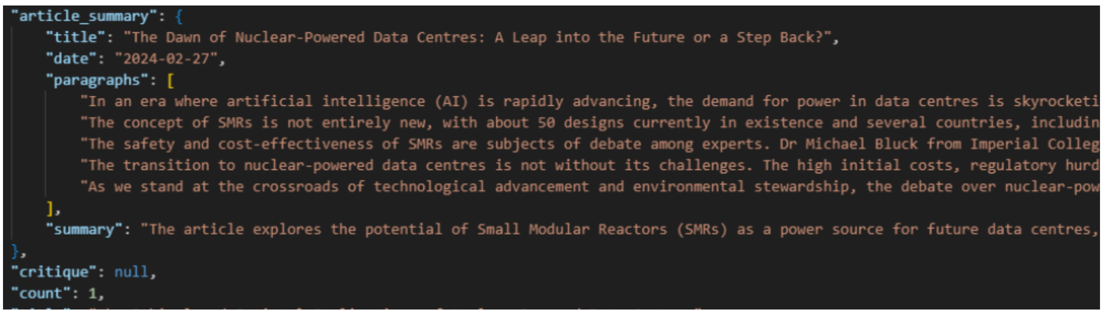


Figure 4.13: Example of summarized content in article enhancement system

ages, generating relevant tags for the post, and appending external links as related articles.

HTML Tag Integration The first step in the layout finalizing process involves the Language Model (LLM) formatting the plain article text by adding various HTML tags. These tags are essential for structuring the content effectively, making the post more appealing and attractive. The choice of tags depends on the intended layout and the specific elements of the content that need emphasis, such as headings, paragraphs, and sections.

Image Insertion Following text formatting, the LLM incorporates images generated from the Image Generator System into the article. This insertion is guided by the content of the article and the descriptions accompanying the images. The <img> tag is used to place each image appropriately within the text, enhancing the visual appeal of the post and helping to illustrate the discussed topics.

Tag Generation for Recommendations The next step involves generating tags for the article, which are crucial for the post's discoverability within the Medium platform's recommendation system. The LLM analyzes the content of the article to produce up to five relevant tags that reflect the core themes and subjects of the post, ensuring that the article reaches its appropriate audience.

Appending External Links Lastly, the system appends external links as related articles along with the reference source URL and summaries collected from the Retrieval Augmented Generator (RAG) system. This addition not only provides further reading options for the audience but also enhances the credibility and informational value of the post by linking it to established sources.

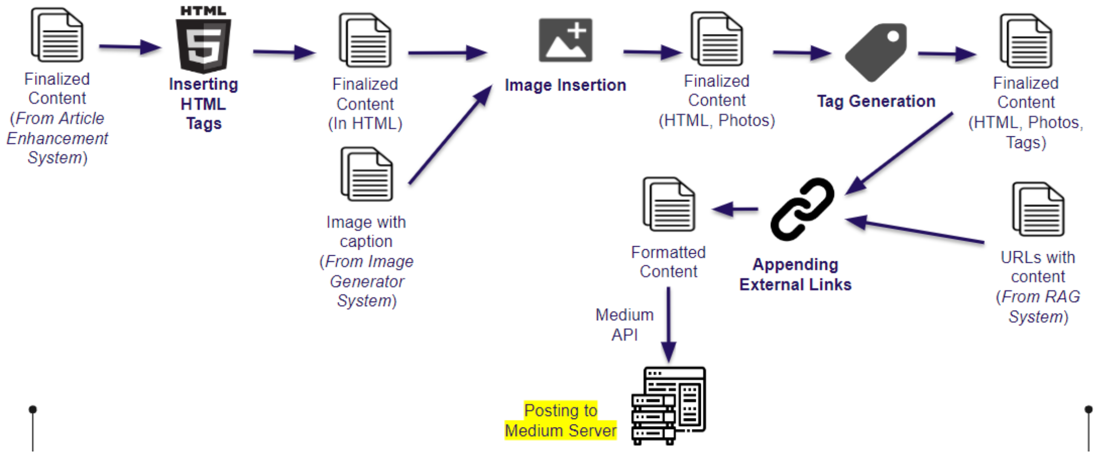


Figure 4.14: Workflow of Layout Finalizing System

Automation and Integration Once these steps are completed, the finalized post is ready to be published. It can be directly posted to the Medium platform using an API, which automates the appearance settings and tagging, eliminating the need for manual adjustments. This seamless integration ensures that the posts are not only stylistically and visually consistent but also optimized for user engagement and platform compatibility.

By structuring content effectively, integrating multimedia elements smartly, and enhancing discoverability through strategic tagging, the Layout Finalizing System plays a pivotal role in transforming raw content into well-crafted articles ready for social media publishing. This system not only simplifies the content preparation process but also ensures that each post meets high standards of quality and coherence, ready to capture the attention of the intended online audiences.

4.3 Prompt Optimization

The extensive utilization of agents within our system has significantly increased the number of prompts used, which has introduced several challenges in main-

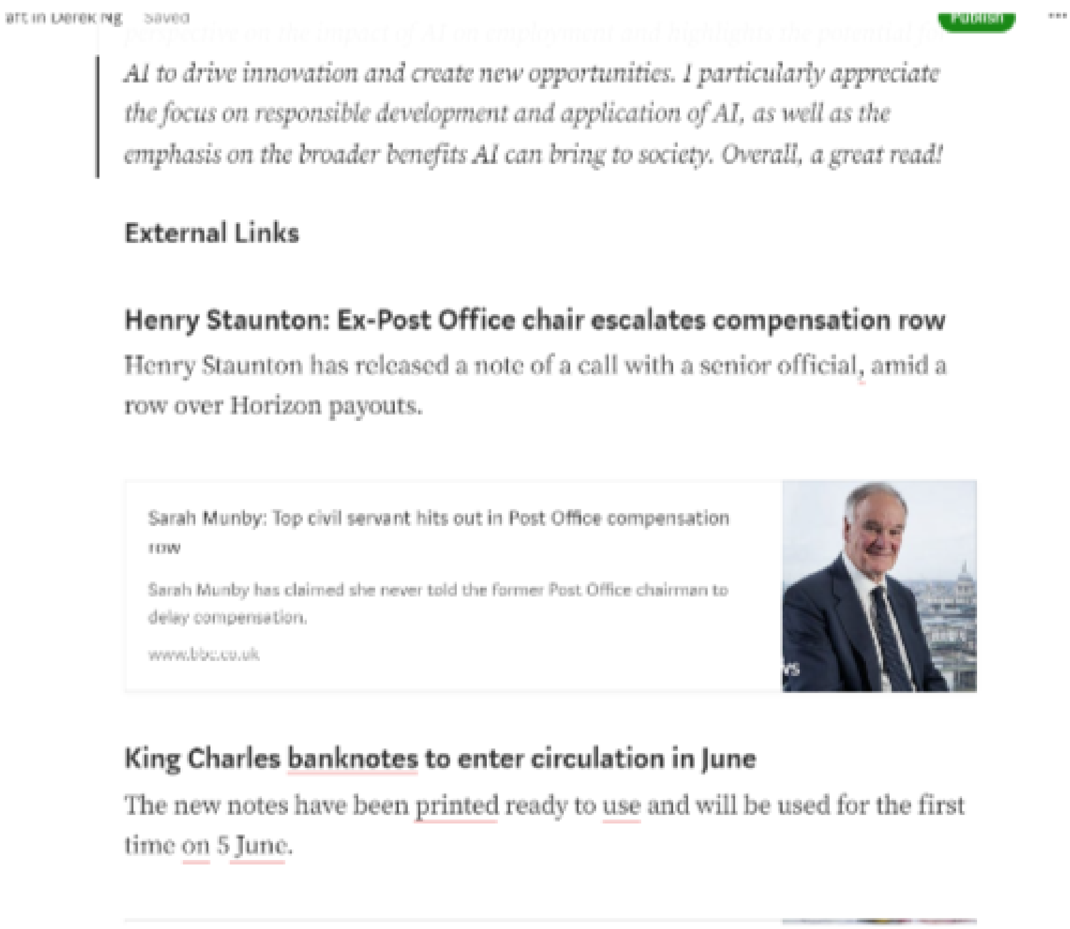


Figure 4.15: Formatted result after appending external links

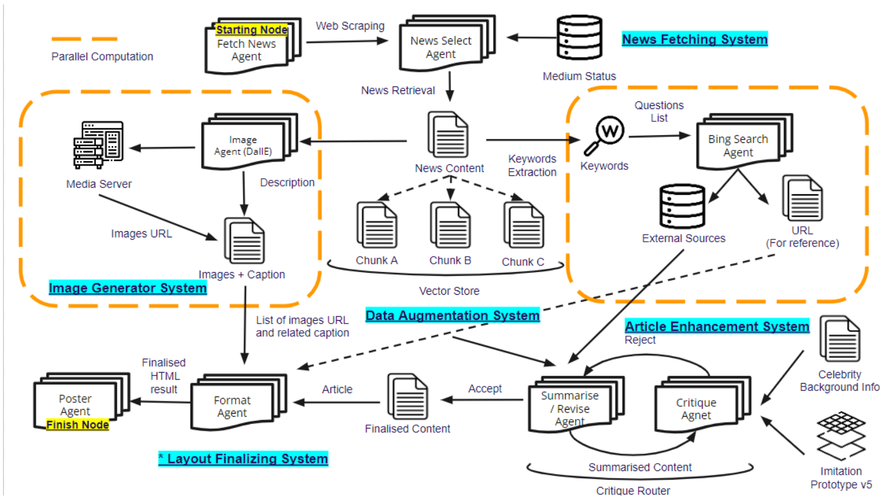


Figure 4.16: Overall system architecture

taining the precision and relevance of the generated content. Simple prompts often do not provide enough context or specificity, which can lead to responses that are surface-level and lack a detailed exploration of the topic (Amatriain, 2024). Additionally, such prompts can be too vague or open-ended, resulting in a wide range of possible responses that complicate controlling the generated results into a desirable format like JSON. Without adequate context, AI models might generate content that is inaccurate or irrelevant, missing essential details that guide the model effectively.

Manually designing all prompts for such a complex system is impractical due to the excessive effort and time consumption involved. Therefore, prompt optimization is essential before automation to enhance efficiency and effectiveness. Since the workflow and details of the system remain consistent across different celebrities, except for celebrity-specific background information and the imitation prototype, the optimized prompts can be retained and reused.

To optimize the LLM prompts and weights, we use the DsPy framework (Stanfordnlp, 2024), which separates the prompt flow from parameters like LLM

```
insert_photo_system_prompt = SystemMessagePromptTemplate.from_template(
    template="""
     You are writing a post for your Medium account. You finish writing the content part and wants to insert the photos to your article.
     Based on the HTML content provided, suggest appropriate locations to insert the following photos and their captions.
     Please provide modified HTML content with the photos inserted at relevant points. Use the <figure> tag for each photo
     <img src> for telling the photo's path without using alt in it, and <figcaption> for its caption. Ensure each photo is inserted in a
     The result should be only the result of the modified HTML content without any other comment or explanation from you.
     """
)

insert_photo_human_prompt = HumanMessagePromptTemplate.from_template(
    template="""
    Here is the HTML content of the article:
    {content}
    Please insert the following photos at appropriate locations within the article, it shows the path of the photo and its caption:
    {photos}
    """
)
```

Figure 4.17: Example prompt after optimization

prompts and weights. DsPy facilitates the use of built-in optimizers, retrievers, and chain-of-thought processes, among other tools, for tuning the elements involved in prompt generation. The optimization process involves generating training and validation data based on the original prompts, which helps the optimizer refine the prompts during the training process. Post-optimization, the prompts undergo post-processing to ensure they are well-tailored and integrated with the necessary input data, enhancing their specificity and relevance.

However, DsPy does not account for scenarios involving system and human interactions, which are common in practical applications and may require additional manual prompt designing to accommodate specific task constraints or requirements. Through these optimization and integration efforts, we significantly improve the system's efficiency and accuracy in content generation, providing more relevant and contextually appropriate content while streamlining the creation process and making it robust against errors.

4.4 Model Fine-Tuning

In the development of our imitation system, model fine-tuning is a critical step that ensures the language models perform well with specific types of content and stylistic nuances. We have explored two main approaches for this purpose: using

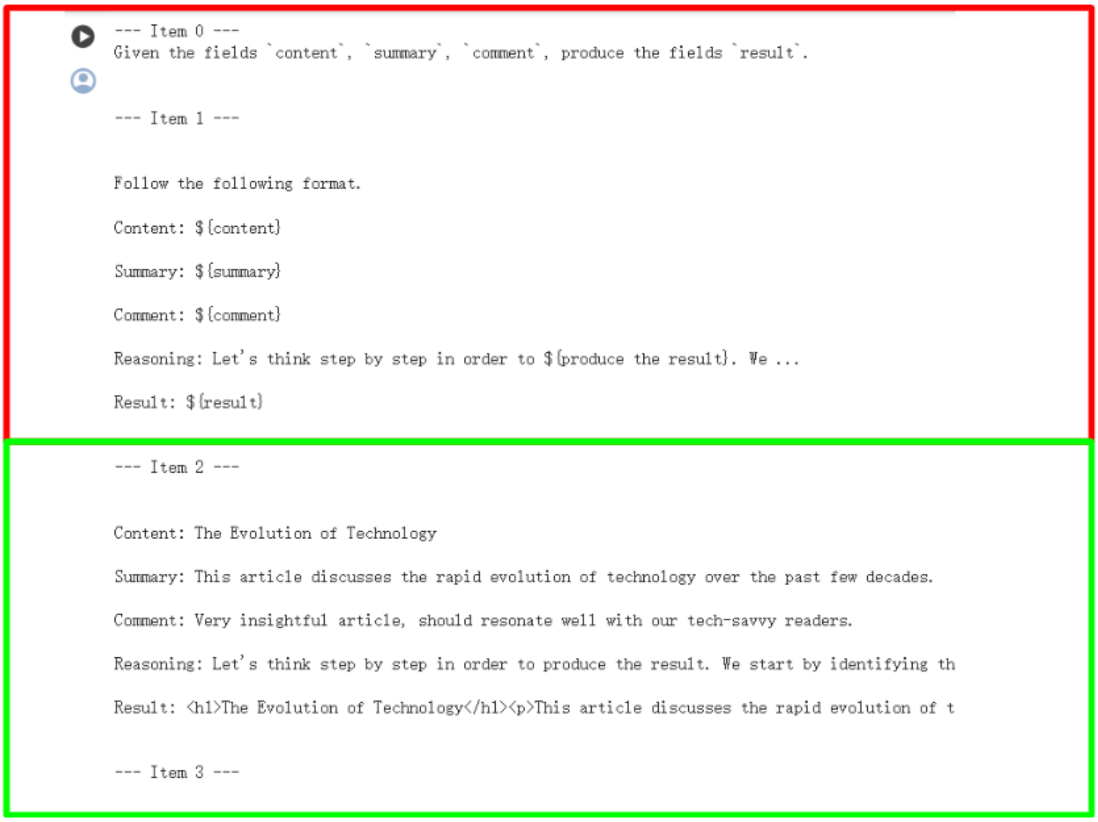


Figure 4.18: Prompt optimization result

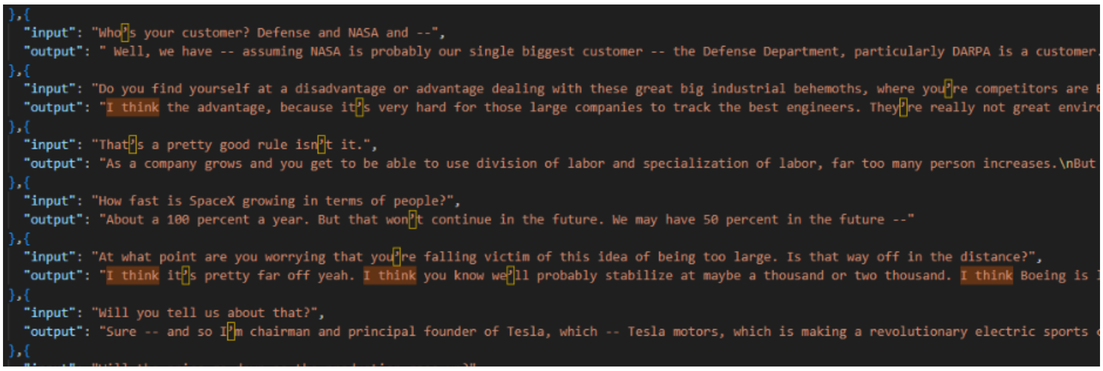


Figure 4.19: Example of the transcript data

the OpenAI API to fine-tune the GPT-3.5 model and employing LoRa (Hu et al., 2021) to fine-tune a smaller LLM.

4.4.1 Dataset

The dataset compiled for model fine-tuning was specifically designed to assess the performance of the models in imitating Elon Musk's writing style as the comparison. This involved collecting a diverse range of Musk's written and spoken communications, including transcripts of his speeches and an extensive selection of his tweets (Preda, 2023). The data preparation process included not only the aggregation of these texts but also a preliminary cleaning phase to ensure the data quality and relevance to the tasks of mimicking Musk's unique stylistic elements.

4.4.2 OpenAI Fine-Tuning

The first approach involves fine-tuning the GPT-3.5 model directly through the OpenAI API. This method is highly convenient as it does not require any GPU resources from our end, which simplifies the setup and reduces the initial technical overhead. However, this convenience comes at a higher cost compared to other methods, which can be a significant factor depending on the scope and scale of

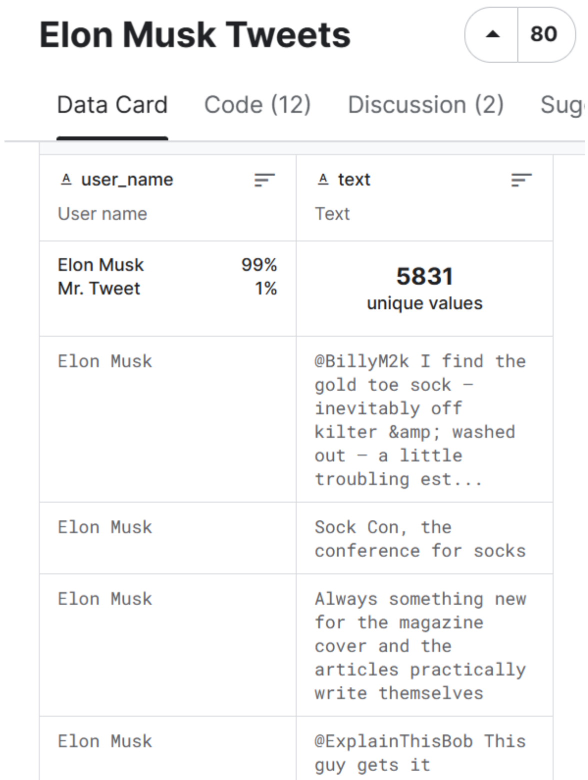
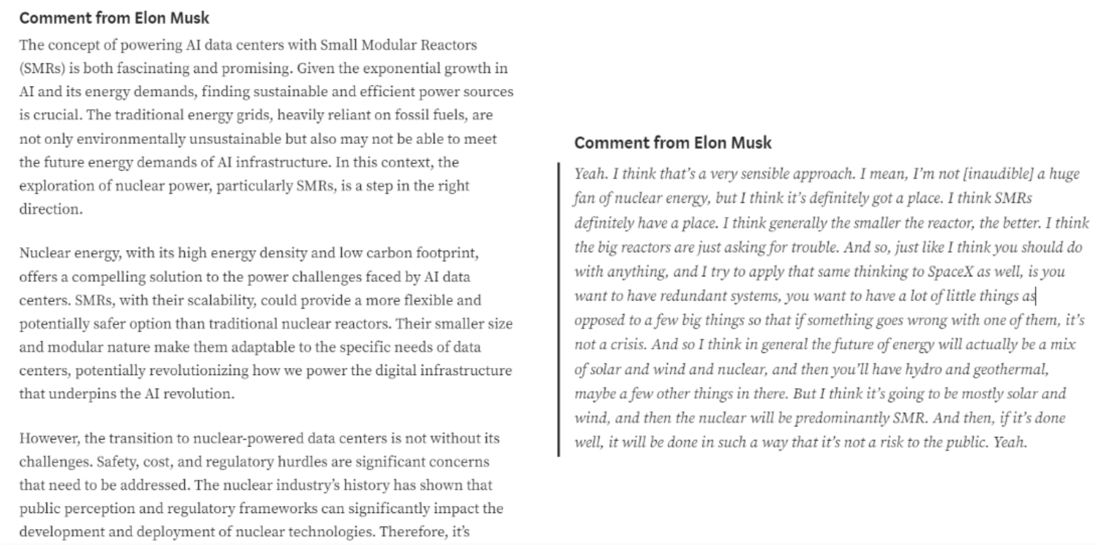


Figure 4.20: Example of Elon Musk Tweets



Comment from Elon Musk

The concept of powering AI data centers with Small Modular Reactors (SMRs) is both fascinating and promising. Given the exponential growth in AI and its energy demands, finding sustainable and efficient power sources is crucial. The traditional energy grids, heavily reliant on fossil fuels, are not only environmentally unsustainable but also may not be able to meet the future energy demands of AI infrastructure. In this context, the exploration of nuclear power, particularly SMRs, is a step in the right direction.

Nuclear energy, with its high energy density and low carbon footprint, offers a compelling solution to the power challenges faced by AI data centers. SMRs, with their scalability, could provide a more flexible and potentially safer option than traditional nuclear reactors. Their smaller size and modular nature make them adaptable to the specific needs of data centers, potentially revolutionizing how we power the digital infrastructure that underpins the AI revolution.

However, the transition to nuclear-powered data centers is not without its challenges. Safety, cost, and regulatory hurdles are significant concerns that need to be addressed. The nuclear industry's history has shown that public perception and regulatory frameworks can significantly impact the development and deployment of nuclear technologies. Therefore, it's

Comment from Elon Musk

Yeah. I think that's a very sensible approach. I mean, I'm not [inaudible] a huge fan of nuclear energy, but I think it's definitely got a place. I think SMRs definitely have a place. I think generally the smaller the reactor, the better. I think the big reactors are just asking for trouble. And so, just like I think you should do with anything, and I try to apply that same thinking to SpaceX as well, is you want to have redundant systems, you want to have a lot of little things as opposed to a few big things so that if something goes wrong with one of them, it's not a crisis. And so I think in general the future of energy will actually be a mix of solar and wind and nuclear, and then you'll have hydro and geothermal, maybe a few other things in there. But I think it's going to be mostly solar and wind, and then the nuclear will be predominantly SMR. And then, if it's done well, it will be done in such a way that it's not a risk to the public. Yeah.

Figure 4.21: Comparison of GPT4 comment and Fine-tuned GPT3.5 comment

the project.

4.4.3 LoRA Fine-Tuning

The second approach is more resource-intensive as it involves fine-tuning a smaller LLM, such as the Mistral-7B model (Jiang et al., 2023), using LoRa. This method is limited by the availability of GPU resources, which confines us to working with models no larger than the 7B or 13B sizes. The following are the implementation detail for fine-tuning:

Model Selection We chose the Mistral-7B model for several reasons. Firstly, it is a well-known open-source LLM that is versatile across different applications. Secondly, it can be directly utilized without the complexities associated with other models like the LLama series, which require additional applications for operation.

```
Input: Embracing AI: A Path to Innovation, Not Job Displacement

In the rapidly evolving landscape of technology, artificial intelligence (AI) stands at the forefront, promising to reshap
Picsart, which has achieved a valuation of more than $1 billion, serves as a testament to the positive impact AI can have
The fear of AI-induced job displacement is not unfounded, but it often overlooks the broader picture. AI, when integrated
Critics of AI often highlight the potential for misuse, such as privacy concerns and the amplification of biases. While th
In conclusion, the narrative surrounding AI and job displacement needs a recalibration. Instead of viewingAI as a threat,

Response:
I think you' re right
I think the AI acceleration is going to be much greater than the labor force acceleration
I think we' ll probably see a 10x increase in AI over the next few years, but maybe only a 2x increase in labor force
So I think we' ll see a lot of job displacement, but I think over time, as you say, new jobs will be created that we can'
</s>
```

Figure 4.22: Elon Musk Comment using Fine-tuned Mistral7B model.

Optimization Techniques To manage the extensive GPU resource requirements of the LLM and make the training feasible on available hardware (like the T4 GPU on Colab), we employed the Unsloth library (UnslothAI, 2024). This library enhances the training and inference speeds and reduces memory usage significantly. The model was loaded with 4-bit quantization combined with PEFT (Parameter-Efficient Fine-Tuning) (Xu et al., 2023a) , LoRa, and SFTTrainer (HuggingFace, 2024), optimizing the fine-tuning process under constrained resources.

4.4.4 Analysis and Results

As detailed in Sections 4.21 and 4.22, fine-tuning the GPT-3.5 model significantly enhanced its ability to mimic language styles, particularly in emulating Elon Musk's distinctive communicative style, compared to the standard GPT-4 model. The GPT-4 model's responses often appeared lengthy and overly formal, frequently using complex vocabulary and phrases like "However," which are less typical of casual conversation. In contrast, the fine-tuned GPT-3.5 model effectively captured Musk's lexical preferences, such as starting statements with "I think" and concluding with "yeah." This stylistic accuracy can be attributed to the specialized dataset discussed in 4.19, which was rich in Musk' s speech patterns and written communications. However, the limited size of this dataset might have constrained the model's ability to extract a broader array of writing styles.

Furthermore, the fine-tuned model consistently adopted a first-person perspective, using pronouns like "I" and "you," closely mirroring Musk' s conversational style, a stark contrast to the GPT-4 model that typically employs a third-person viewpoint. The GPT-3.5 model not only captured Musk's typical communicative methods but also introduced concepts reflecting his thought processes. For example, it speculated on integrating Small Modular Reactors (SMRs) into SpaceX projects and discussed potential new energy sources such as solar, wind, and nuclear energy not mentioned in the original article. These suggestions showcase the model's unique ability to incorporate Musk's personal experiences into its responses and propose innovative solutions, like using smaller, more modular components to mitigate risks. Such nuanced perspectives and strategic insights are rarely produced by the GPT-4 model, highlighting the fine-tuned model' s capacity to offer tailored and forward-thinking ideas in prevalent discussions.

When comparing the fine-tuned model's performance in Figure 4.21 and 4.22, both outputs reflect the stylistic and cognitive elements of Elon Musk's rhetoric. However, the comments from Mistral are significantly shorter than those from the fine-tuned GPT model, possibly due to the different training parameters or focus areas in the dataset. Additionally, the frequent use of "I think" at the beginning of sentences is more prevalent in the Mistral model than in the GPT-3.5 version, potentially indicating an over-fitting issue with the dataset's narrow feature range.

Despite the initial appeal of using LoRa for fine-tuning due to its lower operational costs, the ease of use and the comprehensive support provided by the OpenAI API made fine-tuning GPT-3.5 a more practical choice. The API's fine-tuning process, while more expensive, proved to be less cumbersome in terms of parameter tuning and yielded strong results with our specific training data. Consequently, the fine-tuned GPT-3.5 model was selected for inclusion in the fifth version of our imitation prototype.

4.5 Social Media Platform Choosing

Choosing the right social media platform for publishing content is a critical decision based on several criteria, particularly the extent of API automation each platform allows. Some platforms, like Twitter, restrict API usage to only posting tweets with the free-tier API (Twitter, 2024), which excludes the ability to read tweet contents or automate commenting. This limitation makes it less suited for our needs where interactive capabilities are crucial.

On the other hand, YouTube, while offering unrestricted API usage that includes content uploads and management, predominantly supports formats that are not primarily text-based, such as photos and short videos (Google, 2024). This focus diverges from our system's primary output, which integrates text with images, and thus, does not align perfectly with our publishing needs.

After careful consideration of the capabilities and limitations of various platforms, we have selected Medium and initially considered Instagram as our primary channels for content distribution. Medium is particularly well-suited for articles that combine text and images, offering a straightforward API that accommodates our content format seamlessly. Instagram, known for its broad multimedia support, was also a top contender due to its functionalities that align well with promoting interactive and visually engaging content. However, Medium's compatibility with text-heavy posts and more flexible content integration options makes it the optimal choice for our purposes. This platform choice ensures that we can automate much of the publishing process, maximizing efficiency while adhering to the specific needs of our content's nature and format.

4.5.1 Medium

Medium has been chosen as a suitable platform for posting our content, which primarily consists of summarized news articles enhanced with comments generated by the LLM to imitate celebrities. This platform is well-suited for our needs as it supports comprehensive text-based posts integrated with images. However, Medium does not permit the direct upload of images from local storage. Instead, images generated by our system must first be uploaded to an external server. Once uploaded, these images can then be included in posts via Medium's API, which involves specifying the image URLs in the HTML content of the article. This requirement adds a step to our publishing process but ensures that our posts are visually appealing and fully integrated with the textual content.

4.5.2 Instagram

Initially, Instagram was considered a viable platform for publishing due to its high engagement with multimedia content and its features that are conducive to interactive communication. The platform's ability to fetch users' posts or comments and facilitate replies, as well as its direct messaging functionality, seemed ideal for demonstrating the imitation capabilities of our system. We utilized Instagrapi (subzeroid, 2024), an unofficial API available on GitHub, to integrate posting and direct messaging functionalities into our system. This tool allowed us to automate interactions effectively, showcasing the system's ability to imitate celebrities in a dynamic and engaging manner.

However, the use of this unofficial API led to the banning of our test account on Instagram, highlighting a significant risk associated with non-compliance with official API usage policies. The potential for account suspension or banning poses too great a threat to the sustainability of the project, particularly in a public and high-stakes environment like Instagram. As a result, we have decided to discontinue using Instagram as a platform for posting our content. This decision prioritizes the long-term viability and compliance of our social media strategy, focusing on platforms that allow for official and secure API usage, such as Medium, which supports our content's format without compromising the project's goals.

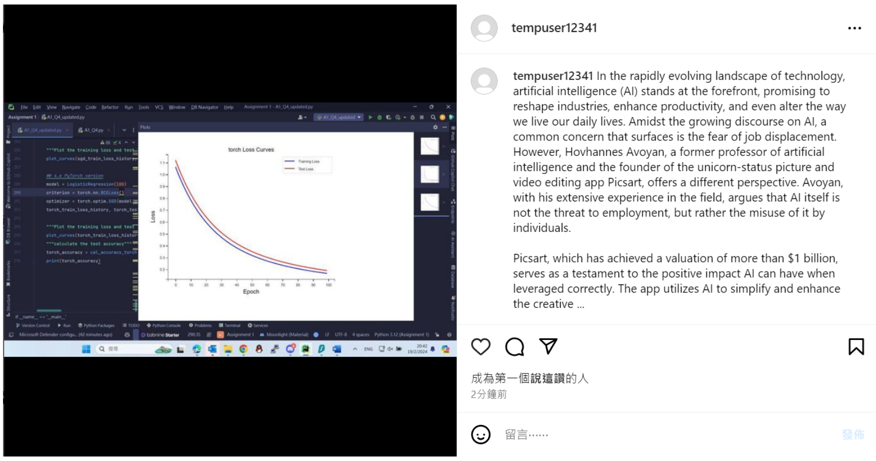


Figure 4.23: Automatically generated post on Instagram 1

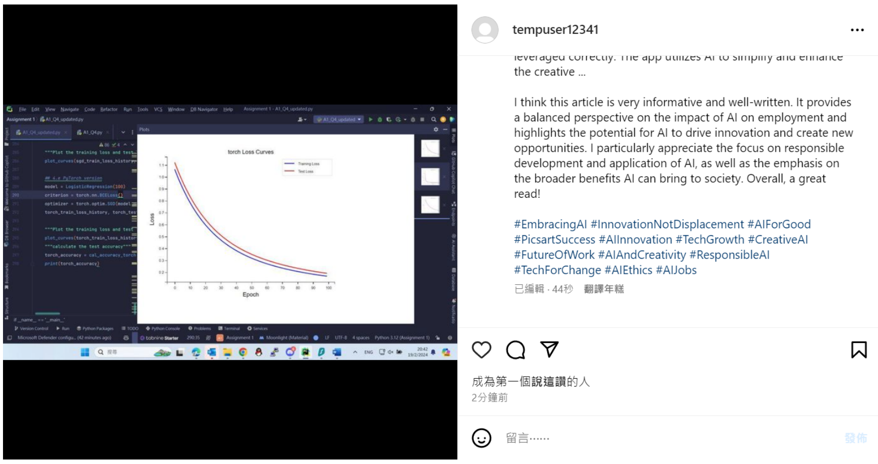


Figure 4.24: Automatically generated post on Instagram 2

4.6 Result

Draft in Argonaut Saved

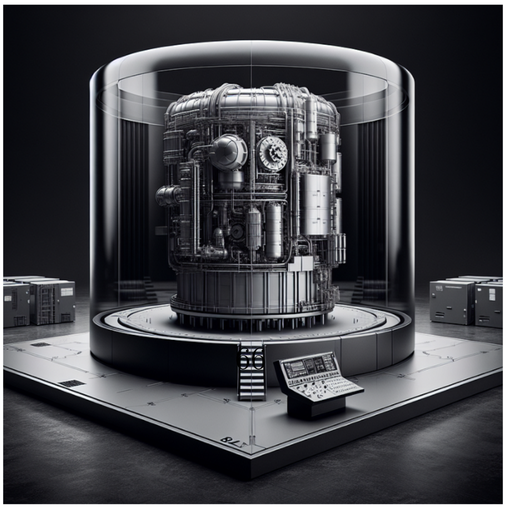
Publish

The Future of AI Data Centres: A Nuclear Solution?

In an era where artificial intelligence (AI) is becoming increasingly integral to our daily lives, the infrastructure supporting this technology is facing unprecedented power demands. Chris Sharp, the chief technology officer at Digital Realty, has brought to light a future where data centres, the backbone of AI, may need to harness the power of nuclear reactors to meet their energy needs. This revelation underscores the growing power consumption of AI data centres, which require about 80 megawatts of power, dwarfing the 32 megawatts needed by traditional data centres.

"modern AI data center with glowing servers"

The solution to this burgeoning power demand lies in Small Modular Reactors (SMRs), a new breed of nuclear reactors that offer a fraction of the power generation capacity of their larger counterparts but are seen as a perfect fit for the energy-hungry data centres. While the concept of SMRs is still in its infancy, with no commercial operations worldwide, China is leading the charge by constructing the world's first SMR. Meanwhile, in the United States, NuScale's SMR design has received approval from the Office of Nuclear Energy, marking a significant milestone in the journey towards nuclear-powered data centres.



"small modular reactor (SMR) model"

Figure 4.25: Final post result posting on Medium 1

Despite the promising prospects of SMRs, the transition to nuclear-powered data centres is met with mixed opinions among experts. Dr. Bluck sees a safety case for SMRs, drawing parallels with their use in submarines, while Dr. Doug Parr of Greenpeace UK raises concerns over the high costs and risks associated with nuclear energy. Spencer Lamb from Kao Data views the idea of nuclear-powered facilities as a distant reality, whereas Brian Gitt from Oklo presents an almost ready SMR design, emphasizing safety, waste management, and the keen interest from major data centre operators.

The involvement of prominent figures such as Sam Altman, chairman of OpenAI and Oklo, highlights the intersection between AI, data centres, and nuclear power. This connection underscores the critical role that SMRs could play in powering the future of AI, offering a sustainable and reliable energy source for data centres that are the lifeblood of the AI revolution.

As the world grapples with the dual challenges of advancing AI technology and ensuring sustainable energy sources, the prospect of nuclear-powered data centres represents a bold step forward. While the path to adopting SMRs is fraught with regulatory, safety, and cost considerations, the potential benefits of a reliable, high-capacity power source for AI data centres could redefine the landscape of technology and energy.

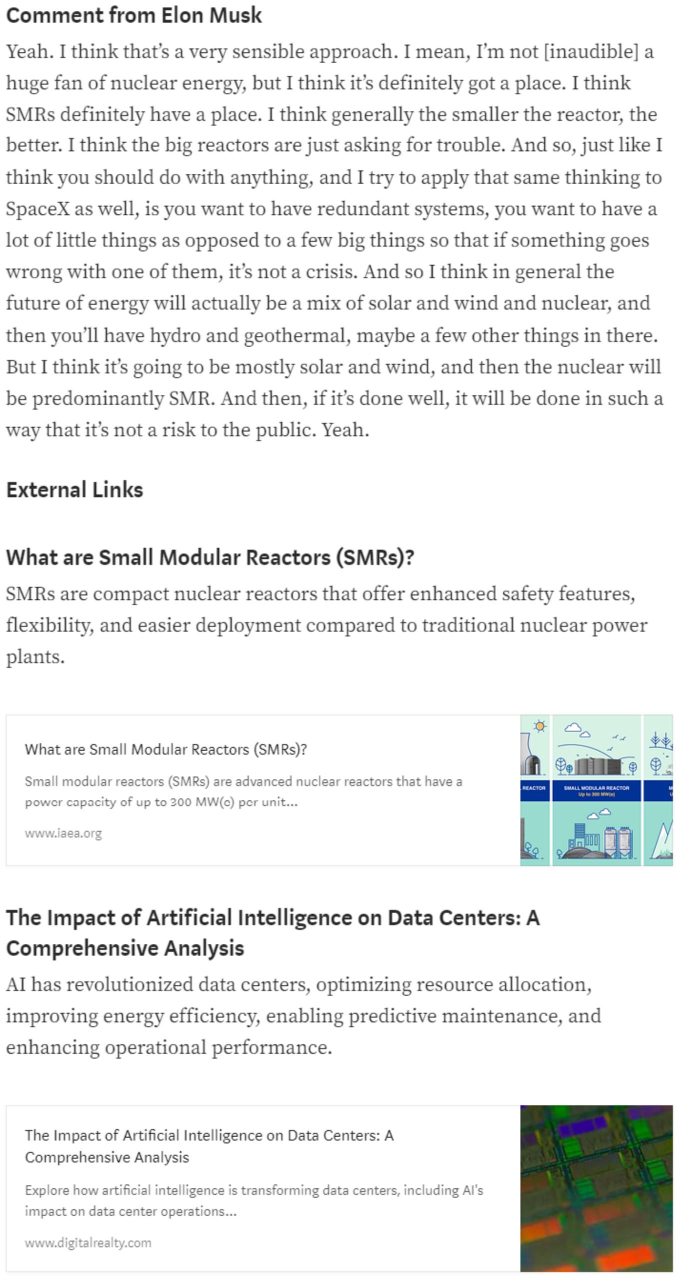
. . .

Summary

The future of AI data centres may be powered by Small Modular Reactors (SMRs), offering a solution to their high power demands. With the world's first SMR under construction in China and designs being approved in the US, the technology is gaining traction. However, opinions on the transition to nuclear-powered data centres vary among experts, with concerns over safety, costs, and environmental impacts. The involvement of AI industry leaders in the development of SMRs highlights the potential for a sustainable and reliable energy source for the burgeoning AI sector.

. . .

Figure 4.26: Final post result posting on Medium 2



Comment from Elon Musk

Yeah. I think that's a very sensible approach. I mean, I'm not [inaudible] a huge fan of nuclear energy, but I think it's definitely got a place. I think SMRs definitely have a place. I think generally the smaller the reactor, the better. I think the big reactors are just asking for trouble. And so, just like I think you should do with anything, and I try to apply that same thinking to SpaceX as well, is you want to have redundant systems, you want to have a lot of little things as opposed to a few big things so that if something goes wrong with one of them, it's not a crisis. And so I think in general the future of energy will actually be a mix of solar and wind and nuclear, and then you'll have hydro and geothermal, maybe a few other things in there. But I think it's going to be mostly solar and wind, and then the nuclear will be predominantly SMR. And then, if it's done well, it will be done in such a way that it's not a risk to the public. Yeah.

External Links

What are Small Modular Reactors (SMRs)?

SMRs are compact nuclear reactors that offer enhanced safety features, flexibility, and easier deployment compared to traditional nuclear power plants.

What are Small Modular Reactors (SMRs)?

Small modular reactors (SMRs) are advanced nuclear reactors that have a power capacity of up to 300 MW(e) per unit...

www.iaea.org

The Impact of Artificial Intelligence on Data Centers: A Comprehensive Analysis

AI has revolutionized data centers, optimizing resource allocation, improving energy efficiency, enabling predictive maintenance, and enhancing operational performance.

The Impact of Artificial Intelligence on Data Centers: A Comprehensive Analysis

Explore how artificial intelligence is transforming data centers, including AI's impact on data center operations...

www.digitalrealty.com

Figure 4.27: Final post result posting on Medium 3

4.7 System Automation

To facilitate the automated posting of reviewed articles at scheduled times, we utilized Microsoft Azure (Microsoft, 2024a) to deploy our system. The deployment involves isolating the entire system within a container to ensure that the environment is stable and consistent, which is crucial for maintaining the integrity and reliability of the system's outputs.

We configured the system to automatically post content every Monday at 00:00. This routine scheduling helps in maintaining a regular presence on the selected social media platform, which for our project is Medium, thus keeping the audience engaged with fresh content at predictable intervals.

The core of our automation lies in the dynamic news fetching system embedded within our setup. This system is designed to automatically fetch a list of news articles from designated sources and then select the most appropriate article based on predefined criteria. The criteria could include relevance to current events, popularity of the topic among the target audience, or suitability for enhancement via the LLM's celebrity imitation commentary.

By deploying the system on Azure and using containerization, we ensure that the system's operations are not only isolated from external changes that could disrupt processing but also scaled efficiently to handle the load of tasks it performs. This setup also allows for modifications and updates to be implemented with minimal downtime and without affecting the overall system performance.

The use of Azure's robust cloud services ensures high availability and reliability, which are essential for the continuous operation of our social media management system. The automated, dynamic fetching of news articles ensures that the content is timely and relevant, thereby increasing the effectiveness of the posts in engaging readers and fostering an active social media presence. This automation of content generation and posting significantly enhances the efficiency of the process, reducing manual oversight and allowing for more strategic deployment of resources.

Chapter 5

Conclusion

This project represents a notable achievement in the realm of AI, particularly with the creation of an advanced Imitation System designed to mimic celebrity personalities through natural language processing. Our system skillfully replicates the unique communication styles and thought patterns of various celebrities, which was continuously enhanced through iterative development and strategic refinements based on comprehensive research.

Our research focused on evaluating the role-playing abilities of LLMs in depicting real individuals. The insights gained were integral in refining the prototype —improving its accuracy and realism by incorporating sophisticated strategies to better simulate the nuances of human interaction. This involved adapting the system to incorporate feedback and varying interaction contexts, which significantly boosted its performance and reliability.

We successfully integrated this system into a social platform application, enabling the automated generation and posting of articles mimicking celebrity voices. Future efforts will focus on diversifying the application's versatility to function across various platforms with enhanced interactive features, making the technology more accessible and engaging for users.

Ethical considerations were meticulously addressed, particularly the implications of using LLMs to replicate real persons or historical figures. We emphasized the necessity of obtaining proper consent, maintaining transparency, and preventing misuse, especially in handling the personas of public or deceased individuals. This is crucial to avoid misrepresentation and the spread of misinformation, as these systems can convincingly generate realistic and influential content. Clear guidelines and restrictions were established to ensure the AI's use remains within ethical and legal boundaries, safeguarding personal legacies and upholding societal norms.

In summary, while our system has pushed the boundaries of what's possible with LLMs in imitating real individuals, it also highlights the challenges and ethical considerations inherent in such technologies. These insights not only refine our understanding of AI's capabilities but also guide future ethical AI development practices.

Chapter 6

Division of Labor

6.1 Derek's Part

Regarding the research work, I am mainly in charge of searching for and reconstructing those baselines except the PTT baseline. For those baselines, I mainly selected the one where we can reconstruct with solely prompt engineering methods and do so by using LangChain. After learning about the release of GPTs, I proposed the idea of appending GPTs as another baseline and building up a systematic way to build the custom GPT for imitating the real person. Meanwhile, we have proposed an idea of building up a heat map where we test how the baseline performs when reading only part of the background information. Based on that, I built up various GPTs and generated the heat map manually even though the result was not used in the final paper. Regarding the paper writing, I am in charge of writing the draft of the paper and the appendix, refining the sections introducing the ECHO system, writing different figures based on the main results, discussing with Billy regarding the experimental results and finishing the analysis.

For the application part, in term 1, I had implemented part of the fetching system, mainly following the idea from RoleLLM and extracting the information of the fetched information into QA format as the input in the imitation process. Regarding the medium blog writer, I am mainly in charge of proposing the idea of building the whole system in a graph feature using LangGraph. And I implemented different components like the image generator system, article enhancement system and layout-finalizing system. I have also built the prompt optimization system using DsPy as well. For the imitation prototype v5, we have explored the 2 approaches and I am in charge of fine-tuning a local LLM model. I implemented the fine-tuning process using unsloth to fine-tune a Mistral-7B model to test whether the result can reflect the writing style of the celebrity or not.

Bibliography

Bbc news rss feed. https://feeds.bbci.co.uk/news/science_and_environment/rss.xml. [Accessed 15-04-2024]. 43

X. Amatriain. Prompt design and engineering: Introduction and advanced methods. https://arxiv.org/abs/2401.14423v3, 2024. Accessed: Apr. 15, 2024. 54

David Baidoo-Anu and Leticia Owusu Ansah. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Journal of AI*, 7(1):52–62, 2023. 13

G. Brockman. Introducing chatgpt and whisper apis. https://openai.com/blog/introducing-chatgpt-and-whisper-apis, 2023. Accessed: Apr. 15, 2024. 6

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669*, 2024. 28, 31

Katherine Elkins and Jon Chun. Can gpt-3 pass a writer' s turing test? *Journal of Cultural Analytics*, 5(2), 2020. 16

I. J. Goodfellow et al. Generative adversarial networks. https://arxiv.org/abs/1406.2661, June 2014. Accessed: Apr. 15, 2024. 48

Google. Youtube data api. https://developers.google.com/youtube/v3?hl=zh-tw, 2024. Accessed: Apr. 15, 2024. 62

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36, 2024. 13

E. J. Hu et al. Lora: Low-rank adaptation of large language models. https://arxiv.org/abs/2106.09685, June 2021. Accessed: Apr. 15, 2024. 57

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. Emotionally numb or empathetic? evaluating how llms feel using emotionbench. *arXiv preprint arXiv:2308.03656*, 2023. 13

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho LAM, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *The Twelfth International Conference on Learning Representations*, 2024. 13

HuggingFace. Supervised fine-tuning trainer. https://huggingface.co/docs/trl/sft_trainer, 2024. Accessed: Apr. 15, 2024. 60

M. E. Jang and T. Lukasiewicz. Consistency analysis of chatgpt. arXiv.org, March 2023. https://arxiv.org/abs/2303.06273 (accessed Apr. 15, 2024). 6

Daniel Jannai, Amos Meron, Barak Lenz, Yoav Levine, and Yoav Shoham. Human or not? a gamified approach to the turing test. *arXiv preprint arXiv:2305.20010*, 2023. 16

A. Q. Jiang et al. Mistral 7b. `https://arxiv.org/abs/2310.06825`, October 2023. Accessed: Apr. 15, 2024. 59

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*, 2023. 13

Douglas Johnson, Rachel Goodman, J Patrinely, Cosby Stone, Eli Zimmerman, Rebecca Donald, Sam Chang, Sean Berkowitz, Avni Finn, Eiman Jahangir, et al. Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the chat-gpt model. *Research square*, 2023. 13

Cameron Jones and Benjamin Bergen. Does gpt-4 pass the turing test? *arXiv preprint arXiv:2310.20216*, 2023. 15, 16, 21, 24

J. Knoll and J. Matthes. The effectiveness of celebrity endorsements: a meta-analysis. *Journal of the Academy of Marketing Science*, 45(1):55–75, October 2016. 8

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*, 2023. 15, 17, 24

LangChain. Few-shot prompt templates. `https://python.langchain.com/docs/modules/model_io/prompts/few_shot_examples/`, 2024a. Accessed: Apr. 15, 2024. 4

LangChain. Langgraph langchain blog. `https://blog.langchain.dev/langgraph/`, 2024b. Accessed: Apr. 15, 2024. 41

LangChain. Langsmith - introduction. `https://www.langchain.com/langsmith`, 2024c. Accessed: Apr. 15, 2024. 41

Pier Luca Lanzi and Daniele Loiacono. Chatgpt and other large language models as evolutionary engines for online interactive collaborative game design. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1383–1390, 2023. 13

P. Lewis et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. https://arxiv.org/abs/2005.11401, May 2020. Accessed: Apr. 15, 2024. 42

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*, 2023. 17

Medium. Github - medium/medium-api-docs: Documentation for medium's oauth2 api. https://github.com/Medium/medium-api-docs, 2024. Accessed: Apr. 15, 2024. 49

Microsoft. Cloud computing services. https://azure.microsoft.com/en-us/, 2024a. Accessed: Apr. 15, 2024. 69

Microsoft. Bing search services developer documentation. https://learn.microsoft.com/en-us/rest/api/cognitiveservices-bingsearch/bing-web-api-v7-reference, 2024b. Accessed: Apr. 15, 2024. 6

Y. Nakajima. Task-driven autonomous agent utilizing gpt-4, pinecone, and langchain for diverse applications –yohei nakajima. https://yoheinakajima.com/task-driven-autonomous-agent-utilizing-gpt-4-pinecone-and-langchain-for-dive 2023. Accessed: Apr. 15, 2024. 6

OpenAI. Introducing gpts. *OpenAI Blog Nov 6, 2023*, 2023. 15, 17, 24

OpenAI. Dall · e 3. https://openai.com/dall-e-3, 2024a. Accessed: Apr. 15, 2024. 44

OpenAI. Openai fine-tuning. https://platform.openai.com/docs/guides/fine-tuning, 2024b. Accessed: Apr. 15, 2024. 39

OpenAI. Openai platform. https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo, 2024c. Accessed: Apr. 15, 2024. 45

PaulBellow. Dalle3 prompt tips and tricks thread. https://community.openai.com/t/dalle3-prompt-tips-and-tricks-thread/498040, 2023. Accessed: Apr. 15, 2024. 46

Gabriel Preda. Elon musk tweets. https://www.kaggle.com/datasets/gpreda/elon-musk-tweets, 2023. Accessed: Apr. 15, 2024. 57

Terrence J Sejnowski. Large language models and the reverse turing test. *Neural computation*, 35(3):309–342, 2023. 16

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, pages 1–6, 2023. 13, 17

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, 2023. 14, 17

Stanfordnlp. Github - stanfordnlp/dspy: Dspy: The framework for programming —not prompting—foundation models. 2024. Accessed: Apr. 15, 2024. 54

subzeroid. Github - subzeroid/instagrapi: The fastest and powerful python library for instagram private api 2024. `https://github.com/subzeroid/instagrapi`, 2024. Accessed: Apr. 15, 2024. 63

Nigar M Shafiq Surameery and Mohammed Y Shakor. Use chat gpt to solve programming bugs. *International Journal of Information technology and Computer Engineering*, 31:17–22, 2023. 13

Alan M Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950. 14, 15

Twitter. Twitter api documentation. `https://developer.twitter.com/en/docs/twitter-api`, 2024. Accessed: Apr. 15, 2024. 62

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. Turingbench: A benchmark environment for turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016, 2021. 16

UnslothAI. Git,hub - unslothai/unsloth: 2-5x faster 80 `https://github.com/unslothai/unsloth`, 2024. Accessed: Apr. 15, 2024. 60

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*, 2023. 14, 15, 17, 24, 34

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. 37

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark. *arXiv preprint arXiv:2303.13648*, 2023. 13

L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, and F. L. Wang. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. https://arxiv.org/abs/2312.12148, December 2023a. Accessed: Apr. 15, 2024. 60

Z. Xu et al. Compress, then prompt: Improving accuracy-efficiency trade-off of llm inference with transferable prompt. arXiv.org, May 2023b. https://arxiv.org/abs/2305.11186. 5

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023. 36

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024. 28

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*, 2023. 17

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, 2023. 13