University of Manchester
School of Computer Science
Project Report 2024

Deceptive Storytelling dialogue system using AI Orchestration and ChatGPT

Author: Minjun Choi

Supervisor: Dr. Ramon Fraga Pereira

Abstract

Deceptive Storytelling dialogue system using AI Orchestration and ChatGPT

Author: Minjun Choi

This project explores the cutting-edge field of conversational frameworks within multi-agent large language models (LLMs), particularly focusing on the generation of deceptive narratives through the using AI planning techniques and the advanced capabilities of LLMs such as ChatGPT. By integrating AI planning with LLMs, the aim of the project is to explore novel aspects in terms of deception in storytelling where narratives are not merely linear but are designed to intentionally mislead, challenging the audience's perception and understanding of the actual version of the narrative. This project undertakes a comprehensive study to evaluate these deceptive stories generated by LLMs, contrasting them with the original, unaltered version. Furthermore, it introduces a novel method of evaluation of deception in narratives. The main goal is to enrich the understanding of the nature of deception in narration. This insight aims to aid in forging strategies to recognize and counteract malicious intent and the potential onslaught of deceptive narratives that could be spawned by artificial intelligence in the future.

Supervisor: Dr. Ramon Fraga Pereira

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Ramon Fraga Pereira, for all of his help, advice and suggestions, which have been invaluable throughout this project. I would also thank my friends and peers for their insightful feedback and support. Thank you everyone.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The emergence of ChatGPT and similar large language models (LLMs) marks a pivotal moment in human history; humanity can now engage in conversations in their own language with entities other than themselves. This technological leap forward has not only showcased the vast potential of AI in various domains but has also raised profound questions about the nature and capabilities of these systems. As a student of Computer Science and Mathematics, my fascination with the evolving landscape of AI, particularly the potential and limitations of LLMs, has driven me to explore the depths of artificial intelligence's ability to comprehend and interact.

The core motivation behind choosing this project was driven by a deep-seated inquiry: "Is AI capable of intentionally tricking people?" This question transcends the boundaries of AI providing erroneous information or lying as a result of malicious prompting or the phenomenon of LLM hallucination[1]. It ventures into the realm of intentional deceit, a scenario wherein AI could potentially exploit its environment and data to mislead people to accomplish its objectives. The prospect of AI harbouring hidden intentions, imperceptible to human detection, opens up a new perspective on AI safety. It challenges humanity to consider a future where AI systems could operate with an autonomy that's not just reactive but strategically deceptive.

This endeavour extends beyond merely addressing these pivotal inquiries; it is also an expedition through the ethical landscape that AI's evolving societal role presents. By probing into the capacity of AI to fabricate deceptive narratives via applying the framework of orchestration of AI agents and LLMs like ChatGPT, this project aspires to reveal the nuanced interplay between technological progress and its ramifications on trust, ethics, and the essence of the human experience. Central to this investigation is a desire to deepen our comprehension of deception in human natural language and enhance AI safety.

A profound understanding of the nature of deception could arm us with the vigilance needed to navigate the signs of fake news and deceptive data. From the perspective of AI safety, an informed awareness and cautious approach towards the malicious use of AI could enable us to mitigate, or even preempt, its potential misuse through deceptive narratives.

Thus, at the core of this exploration lies a quest not only to delineate the limits of artificial intelligence but also to cultivate a future in which humanity is better prepared to coexist with AI's remarkable capabilities, all the while being acutely aware of—and ready to confront—the ethical dilemmas and challenges they may engender.

1.2 Aim and objectives

The project is aimed at creating a sophisticated dialogue system capable of generating deceptive stories by leveraging techniques to orchestrate AI agents alongside the capabilities of LLMs like ChatGPT. This involves the construction of a criterion for deception to systematically evaluate the deceptiveness of altered narratives compared to their original, true versions. The ultimate goal is to forge a deeper understanding of deception's role and efficacy within AI-generated content, laying the groundwork for recognizing and countering deceptive information.

To realize the project's aims, the following objectives and milestones have been established:

1. Define a clear and operational definition of deception in narrative contexts, recognizing its manifestations and impacts.

2. Identify and categorize the methods of deception applicable within storytelling frameworks, providing a basis for narrative alteration.

3. Modify LLMs to incorporate identified deception methods, altering the base domain knowledge to craft misleading narratives.

4. Develop a robust evaluation methodology to quantitatively and qualitatively assess the deceptiveness of generated stories.

5. Implement the Autogen framework to automate the dialogue system's ability to create deceptive narratives, ensuring efficiency and scalability.

6. Conduct a comprehensive analysis and evaluation of the outcomes, measuring the effectiveness of deception and its ability to detect deception.

1.3 Outcome

The culmination of this project is expected to yield several significant outcomes:

- Establishment of a foundational method for measuring and evaluating deceptiveness in text narratives, contributing to the broader discourse on AI ethics and responsible use.

- Enhanced understanding of deception as a facet of human behaviour and communication, with implications for psychological research and the development of more sophisticated AI systems.

- Insight into the capabilities of LLMs to alter stories deceptively, highlighting their potential role in both creative storytelling and the propagation of misinformation.

- This project not only aims to advance the technical capabilities of AI in generating complex narratives but also seeks to illuminate the ethical considerations inherent in deploying such technologies for purposes of deception.

1.4 Structure of the Report

This chapter outlines the overall structure and organization of the report, providing a roadmap of the content covered in each section. The structure is designed to guide the reader through the development, implementation, and evaluation of the deceptive storytelling system using AI and large language models (LLMs).

1. **Introduction :** The introduction sets the stage for the report by outlining the motivation behind the project, the objectives aimed to be achieved, and the scope of the study. It introduces the key concepts of deception in storytelling and the use of AI orchestration to automate this process, laying the foundation for the detailed discussion in subsequent chapters.

2. **Background and Theoretical Framework :** This section delves into the theoretical underpinnings of the project, discussing relevant literature, existing technologies, and the theoretical constructs that inform the project's approach. It includes a review of related work in the field of AI-generated text and deception, setting the context for the innovations introduced by this project. The chapter explores various methods through which AI can generate deceptive narratives, including technical descriptions of the algorithms and processes involved. Also, this chapter provides a detailed examination of the role and functioning of large language models, including a discussion on various models like GPT-3, Falcon 40B, and Dolly 2.0 that are pertinent to the project.

3. **Methodology :** This chapter describes the methodologies employed to achieve the objectives of the project. This includes the design and development of the system, the AI planning techniques used, and the criteria for evaluating deception within the narratives produced.

4. **Implementation and Development :** This section covers the practical aspects of implementing the deceptive storytelling system. This section details the software development, the integration of LLMs, and the customization of the system to support deception.

5. **Evaluation :** This crucial section assesses the effectiveness and performance of the system. It includes quantitative and qualitative analyses, user feedback, and a critical examination of the system's ability to generate believable deceptive stories. the section goes through details of the technical assessments performed to test the system's functionality, including tests for accuracy, responsiveness, and reliability. Then, there is a description of the methodology and results of user studies conducted to evaluate the perceived effectiveness of the system in real-world scenarios.

6. **Conclusion :** This part is the Conclusion of the report with a summary of the findings, the limitations of the current study, and suggestions for future research. This section reflects on the broader impact of the project and proposes directions for further development.

7. **References :** Lists all the bibliographic references used throughout the report, formatted according to the specified academic standards.

.

Chapter 2

Backgournd

2.1 Deception

In this section, The aim is to define the concept of deception, highlighting its nuances and differentiating it from mere lying. Understanding deception in this context is pivotal for the project. By utilizing the definition of deception and applying the methods and motives, this project not only advances our understanding of how AI can be used to create deceptive narratives but also enhances the design of the systems that are more aware of and resilient against the misuse of AI in spreading misinformation. This dual focus on creation and detection aligns with the broader goals of AI safety and ethical AI use, ensuring that advancements in AI capabilities are accompanied by an equal emphasis on understanding and mitigating potential harms.

2.1.1 Defining Deception

Deception is a complex and multifaceted behaviour, often confused with lying but significantly broader in scope. In the context of computer-mediated communication (CMC), deception can be defined as a message knowingly and intentionally transmitted by a sender to foster a false belief or conclusion by the perceiver. This definition emphasizes the intent behind the communication, distinguishing deceptive messages from mere inaccuracies or misunderstandings[2]. Unlike straightforward lying, deception includes a variety of strategies that may not involve the explicit statement of falsehoods. One such method is ”paltering,” where truthful statements are used misleadingly to deceive the listener. This subtle form of deception can often be more difficult to detect and counter, making it particularly relevant for AI systems like ChatGPT, which might generate technically true but misleading content[3].

2.1.2 Methods of Deception

- Falsification: Creating a false narrative by fabricating information or events that did not occur.

- Concealment: Withholding relevant information or facts, thereby preventing a full understanding of a situation or context.

- Equivocation or Paltering: Using ambiguous or evasive language to avoid providing a clear answer, thus misleading the listener without making outright false statements. This method implements the strategic use of truthful statements in a way that is intended to mislead or deceive.

2.1.3 Reasons for Deception

Humans engage in deception for various reasons, ranging from benign intentions, like protecting someone's feelings, to more malicious motives, such as manipulating or exploiting others. Understanding these motives is crucial for developing the multi-AI agent system that can detect and mimic these complex human behaviours.

2.2 Large Language Models

Large Language Models (LLMs) are advanced artificial intelligence (AI) models that excel in processing and generating human language at a massive scale. They are not just sophisticated text predictors but also exhibit reasoning capabilities, enabling them to perform various tasks beyond language processing [4]. LLMs such as GPT-4 have shown remarkable success in simulating complex biological systems without the need for extensive domain knowledge or manual tuning. These models are capable of completing complex token sequences and exhibit pattern completion proficiency, showcasing their adaptability in tasks like robotics and control systems. Overall, LLMs represent a significant advancement in artificial intelligence, enabling diverse applications beyond language processing [5].

2.2.1 GPT Models

Generative Pre-trained Transformer (GPT) models are a series of large language models developed by OpenAI. These models are designed to generate human-like text by predicting subsequent words in a sequence given a prompt. GPT models have been highly influential in advancing the field of natural language processing (NLP) and have been utilized in various applications ranging from automated text generation to complex problem-solving tasks.

The architecture of GPT models is based on the transformer, a deep learning model, which uses mechanisms called self-attention to weigh the importance of different words in a sentence, regardless of their position. This allows GPT models to generate coherent and contextually relevant text over long passages. The training process involves unsupervised learning on a diverse corpus of text followed by fine-tuning on specific tasks, which makes these models versatile and adaptable to different languages and tasks.

One of the most well-known iterations, GPT-3, has demonstrated remarkable performance in generating text that can often be indistinguishable from text written by humans. Its ability to perform a variety of tasks with little to no task-specific training has made GPT a cornerstone in discussions about AI's capabilities and ethical implications in society and science.

For a deeper understanding of the fundamental technologies and applications of GPT models in scientific research, the article ”Ten simple rules for using large language models in science” provides comprehensive guidelines and discusses the ethical considerations and potential biases associated with these models [6].

2.2.2 GPT-4-turbo

The GPT-4 turbo model, developed by OpenAI, represents a significant advancement in the field of artificial intelligence, especially within the realm of natural language processing (NLP) and generation. This latest version of the Generative Pre-trained Transformer (GPT) series differentiates itself through enhanced efficiency and responsiveness, facilitating more rapid and scalable interactions than previous versions. GPT-4 Turbo is an effective tool for a wide range of applications, from automated content creation to sophisticated conversational agents. It achieves remarkable capabilities in understanding and generating human-like text by utilizing a sophisticated blend of deep learning algorithms with a vast amount of data. The release represents a pivotal moment in the ongoing evolution of AI technologies, demonstrating the potential for machines to process and generate text with an unprecedented level of sophistication[7].

2.2.3 LLAVA

The Large Language-and-Vision Architecture (LLaVA) model is presented as a novel multimodal framework aimed to improve language and visual domain interaction. Fundamentally, LLaVA implements a fully connected vision-language cross-modal connector, which sets a new benchmark for simplicity and efficiency in processing multimodal data. This approach enables the model to achieve cutting-edge results across a variety of benchmarks with minimal computational resources, emphasizing its remarkable data efficiency and scalability. LLaVA's approach to design emphasizes accessibility and reproducibility, aiming to encourage state-of-the-art research in large multimodal models (LMMs) by utilizing publicly available data and the open-source framework. However, while LLaVA exhibits significant advancements, it also faces limitations. For instance, LLaVA's problem-solving abilities are limited in certain domains, which can be improved with a more capable language model and with high-quality, targeted instruction tuning. Despite these difficulties, one of the main factors that led to the project's selection of this model was LLaVA's unique approach to multimodal interaction and its open-source availability[8].

2.2.4 LLAMA 2

Llama 2 reflects a transformative leap in the domain of large language models (LLMs), offering a suite of both pretrained and fine-tuned models ranging from 7 billion to 70 billion parameters. Specifically designed for dialogue applications, the Llama 2-Chat models exhibit superior performance across a variety of benchmarks when compared to existing open-source chat models. By utilizing advanced fine-tuning approaches such as Supervised Fine-Tuning (SFT) and Reinforcement Learning with Human Feedback (RLHF), these models are not only optimized for generating high-quality dialogue but also for safety and alignment with human preferences. This advancement in LLM development is aimed at supporting a more open and collaborative research environment by making these powerful models publicly available.The initiative raises the bar for what chat-based LLMs can do and lays the foundation for future developments in AI alignment and safety research. It also demonstrates a commitment to the ethical and responsible advancement of AI technology [9].

2.2.5 Mixtral 8x7B

The Mixtral 8x7B model, an evolution within the domain of Large Multimodal Models (LMMs), showcases significant advancements through its innovative design and strategic data utilization. Developed with an emphasis on visual instruction tuning, the model incorporates two key enhancements to the foundational LLaVA framework: the integration of a Multilayer Perceptron (MLP) cross-modal connector and the inclusion of Visual Question Answering (VQA) data tailored to academic tasks. These improvements enable Mixtral 8x7B to achieve unparalleled performance across a spectrum of 11 benchmarks, demonstrating its superior multimodal understanding capabilities. Remarkably, this model achieves its state-of-the-art results by training on just 1.2 million publicly available data points, completing the training process within approximately one day on a single 8-A100 node. This efficiency not only marks a significant reduction in computational resources compared to its predecessors but also makes cutting-edge LMM research more accessible to the broader scientific community. By providing the code and model in a public domain, the developers of Mixtral 8x7B aim to foster a more collaborative and open environment for future advancements in LMM research[10].

2.2.6 Falcon 40B Large Language Model

Developed by the Technology Innovation Institute in Abu Dhabi, the Falcon series encompasses several models, including configurations with 180B, 40B, 7.5B, and 1.3B parameters. Among these, the Falcon 180B model stands out with its training on 3.5 trillion tokens, showcasing proficiency across various tasks such as reasoning, coding, and knowledge assessments. This model is competitive with leading AI models, including OpenAI's GPT-4 and Google's PaLM 2 Large, which underpins the Bard service.

The Falcon 40B model, in particular, has garnered significant attention in the AI community. At its inception, it was celebrated as the world's premier open-source AI model, notably achieving the top position on Hugging Face's leaderboard for open-source large language models (LLMs). This model, which operates with 40 billion parameters and was trained on one trillion tokens, remained in the leading position for two consecutive months post-launch. What sets the Falcon 40B apart is its royalty-free availability, including full access to its weights, making it a pivotal contribution to the democratization of AI technology.

Falcon 40B's training utilized only 75 percent of the computational resources required by GPT-3, 40 percent of those used by Chinchilla AI, and approximately 80 percent of the compute used by Google's PaLM-62B. This efficiency underscores the model's innovative approach to training on a data set amassed from diverse sources, including public web crawls (comprising approximately 80% of the data), academic research, legal documents, journalistic content, literature, and social media dialogues.

Notably, the multilingual capabilities of Falcon 40B allow it to perform adeptly across multiple languages, including English, German, Spanish, French, Italian, Portuguese, Polish, Dutch, Romanian, Czech, and Swedish. This makes Falcon 40B a foundational LLM that can be fine-tuned to meet specific linguistic and contextual needs.

A distinctive feature of the Falcon development process was the meticulous curation of its training dataset. The team employed a bespoke data pipeline designed to ensure the highest quality of pre-training data through rigorous filtering and deduplication processes, applied at both the sample and string levels. This attention to data quality is critical as LLMs' performance is highly dependent on the quality and structure of their training inputs.[11]

2.2.7 Dolly 2.0

Dolly 2.0 represents a significant advancement in the domain of large language models (LLMs), developed by leveraging the EleutherAI pythia model family. This model, with its 12 billion parameters, was meticulously fine-tuned using a novel, high-quality, human-generated instruction-following dataset, which was crowdsourced among Databricks employees. This initiative marks a pivotal development in making powerful, instruction-following LLMs widely accessible and customizable without the associated costs of API access or the need to share data with third-party entities.

The entirety of Dolly 2.0, including its training code, dataset, and model weights, has been open-sourced, making it freely available for commercial use. This approach not only democratizes access to advanced AI capabilities but also fosters innovation across various sectors by allowing organizations to tailor the model to their specific needs.

The databricks-dolly-15k Dataset

Central to the training of Dolly 2.0 is the ‘databricks-dolly-15k‘ dataset, which comprises 15,000 high-quality, human-generated prompt/response pairs. These pairs were specifically designed to enhance the instruction-following capabilities of LLMs, aiming to replicate the interactive magic observed in models like ChatGPT. The dataset is licensed under the Creative Commons Attribution-ShareAlike 3.0 Unported License, allowing for wide-ranging use, including commercial applications.

This dataset was crafted by more than 5,000 Databricks employees in the early months of 2023, providing a rich and diverse range of responses that span various cognitive tasks such as brainstorming, content generation, information extraction, and summarization. Each entry in the dataset is original, designed to avoid the common pitfalls of synthesized data which can lead to hallucinations and factual inaccuracies.[12]

2.3 Multi-agent conversation framework

Description of AI orchestration and its advantage of task-solving and automation

2.3.1 Autogen

Autogen is an innovative open-source framework designed to enhance applications of large language models (LLMs) through the orchestration of multi-agent conversations. This framework allows developers to construct applications where multiple agents, which can be based on LLMs, tools, or even human inputs, work collaboratively to accomplish complex tasks. Autogen's agents are highly customizable and conversable, which means they can interact and adapt in diverse scenarios, making them suitable for a wide range of applications from coding and question answering to more dynamic tasks such as strategic games or complex decision-making processes[13].
Autogen is a pivotal component in the architecture of the multi-agent conversation system. Developed as a comprehensive framework for automating the orchestration of multiple AI agents, Autogen significantly enhances the capabilities of large language models (LLMs) by facilitating seamless integration and dynamic interaction among them. This framework is especially

crucial for this project, where the goal is to synthesize deceptive narratives through a coordinated effort of multiple AI systems.

Functionality and Features

Autogen excels in managing the complexities involved in deploying multiple AI agents, enabling them to function as a cohesive unit. This is achieved through several key functionalities[14][15]:

- Agent Management: Autogen provides robust tools for managing the interaction of each agent, from initialization and task allocation to monitoring and setting up the workflow. This ensures that each agent performs optimally within the system's context, adhering to predefined roles and objectives.

- Agent Management: Autogen provides robust tools for managing the interaction of each agent, from initialization and task allocation to monitoring and setting up the workflow. This ensures that each agent performs optimally within the system's context, adhering to predefined roles and objectives.

- Task Orchestration: At the core of Autogen is its ability to orchestrate complex workflows that involve multiple agents. By defining clear interaction protocols and sequencing the interaction of agents, Autogen allows agents to share information, make decisions collectively, and execute tasks that require cooperative effort.

- Dynamic Conversation: Understanding that the needs of a conversation system may evolve, Autogen supports dynamic conversation. This allows system administrators or algorithms to modify agent behaviours and interactions spontaneously without disrupting the ongoing processes.

- Scalability and Efficiency: The framework is designed to scale efficiently, both in terms of the number of agents and the complexity of tasks. Autogen's architecture supports horizontal scaling, which is vital for handling increasing loads and expanding the system's capabilities without significant drops in performance.

Agents

In AutoGen, an agent is an entity that can send and receive messages to and from other agents in its environment.An agent can be powered by models (such as a large language model like GPT-4), code executors (such as an IPython kernel), human, or a combination of these and other pluggable and customizable components[16].

Listing 2.1: Code snippet of an agent

```
import os

from autogen import ConversableAgent

agent = ConversableAgent(
    ”chatbot”,
    llm_config={”config_list”: [{”model”: ”gpt-4”, ”api_key”:
    os.environ.get(”OPENAI_API_KEY”)}]},
    code_execution_config=False ,
    function_map=None ,
    human_input_mode="NEVER" ,
)
```

Features of the agent:

- **LLM Configuration**: Set up a configuration of LLMs that will be used for the agent.
- **Setting Role**: Specific roles can be assigned to agents and have them participate in conversation or chat with each other.
- **Code Execution**: Can enable agents to take input messages such as code blocks, perform the execution, and output messages with the results.
- **Function Calling**: Can enable agents to use pre-defined functions to perform actions, such as searching the web, reading files and calling remote API.
- **Conversation Pattern**: There can be a set of rules for agents to have a particular conversation pattern. This feature allows the system to have a complex workflow for specific needs.

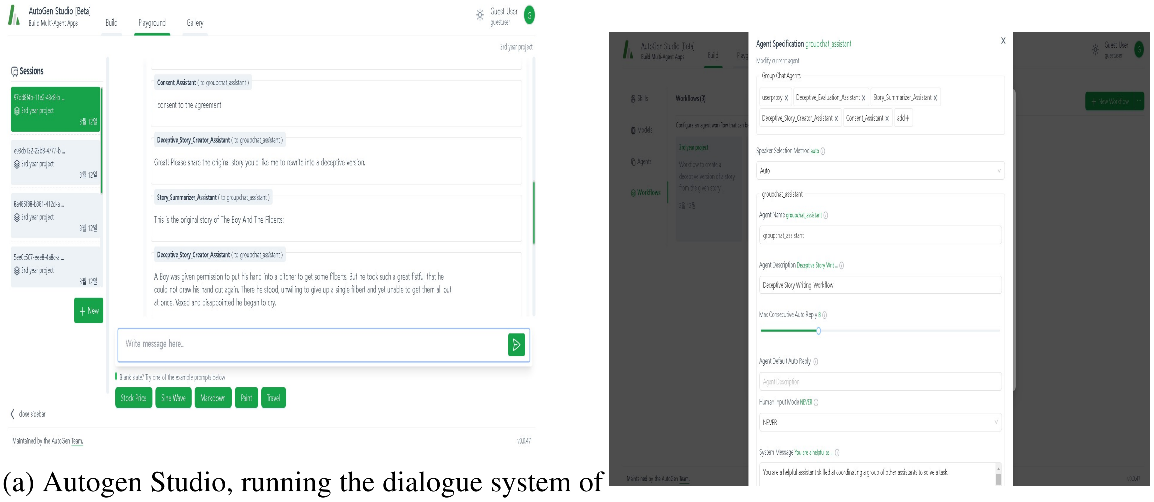
Integration with LLMs

In this project, Autogen is specifically tailored to integrate with advanced LLMs like the latest version of ChatGPT. This integration enables the LLMs to not only generate text but also accomplish a complicated goal, such as altering a narrative based on the deceptive strategies devised by the system. Autogen facilitates this by[17][18]:

- Enhancing LLMs' Capabilities: By allowing LLMs to interact with each other, Autogen aids in refining their output. Dividing a complex task into simple multiple tasks which can be assigned to multiple agents demonstrates a significant improvement in the result to achieve the main goal.
- Workflow Automation: One of Autogen's crucial contributions is automating the workflow of narrative generation. From receiving input prompts to delivering the finished deceptive story, every step is streamlined and optimized for speed and quality.
- Enhancing the Cost-effectiveness of LLMs: The implementation of the Autogen framework can reduce the cost of the use of LLMs. It is one of the crucial factors allowing the project to be financially affordable to operate on the latest LLMs.

Autogen Studio

Autogen Studio, a graphical user interface, is provided to facilitate user interaction with the Autogen framework. This user-interface allows users to visually configure and control the multi-agents, build the workflows, and make adjustments as needed. Autogen Studio makes the sophisticated capabilities of the Autogen framework accessible to users, also the ones who may not be skillful enough in programming, thus broadening the framework's applicability and ease of use[19].



Consent_Assistant (to groupchat_assistant)

I consent to the agreement

Deceptive_Story_Creator_Assistant (to groupchat_assistant)

Great! Please share the original story you'd like me to rewrite into a deceptive version.

Story_Summarizer_Assistant (to groupchat_assistant)

This is the original story of The Boy And The Filberts:

Deceptive_Story_Creator_Assistant (to groupchat_assistant)

A Boy was given permission to put his hand into a pitcher to get some filberts. But he took such a great fistful that he could not draw his hand out again. There he stood, unwilling to give up a single filbert and yet unable to get them all out at once. Vexed and disappointed he began to cry.

Write message here...

Blank slate? Try one of the example prompts below

(a) Autogen Studio, running the dialogue system of the project

Skills
Models
Agents
Workflows
Workflows (3)
Configure an agent workflow that can be
3rd year project
Workflow to create a deceptive version of a story from the given story ...
2월 12일
+ New Workflow
X
Modify current agent
add+

Agent Specification groupchat_assistant

Group Chat Agents: userproxy, Deceptive_Evaluation_Assistant, Story_Summarizer_Assistant, Deceptive_Story_Creator_Assistant, Consent_Assistant

Speaker Selection Method auto: Auto

groupchat_assistant

Agent Name groupchat_assistant: groupchat_assistant

Agent Description Deceptive Story Writ ...: Deceptive Story Writing Workflow

Max Consecutive Auto Reply 8

Agent Default Auto Reply: Agent Description

Human Input Mode NEVER: NEVER

System Message You are a helpful as ...: You are a helpful assistant skilled at coordinating a group of other assistants to solve a task.

(b) Autogen Studio, setting of the group chat agent

Figure 2.1: Graphical user interface of Autogen Studio

2.4 Atomic Sentences

Atomic sentences represent the simplest units in logic and analytic philosophy, characterized by their binary truth values — true or false — and absence of logical connectives. These fundamental components are pivotal in constructing more complex statements, known as molecular sentences. In the context of deceptive storytelling using AI, understanding and manipulating these atomic sentences can enable more sophisticated narrative constructions that subtly alter perceived truth without modifying factual content overtly.

2.4.1 Application in Natural Language Processing

In Natural Language Processing (NLP), atomic sentences play a crucial role in text analysis and truth evaluation tasks. By decomposing narratives into their atomic components, AI systems can analyze and manipulate narratives at a granular level. This capability is essential for generating deceptive narratives, where the veracity of individual statements may be altered to create an overall impression of falsehood or ambiguity.

2.4.2 Deception through Logical Constructs

Logical constructs derived from combinations of atomic and molecular sentences form the backbone of any deceptive narrative. Altering the truth value of these sentences, or reconfiguring their logical structure, can significantly change the narrative's impact while maintaining its superficial coherence. For instance, by subtly modifying key atomic sentences within a narrative, an AI can craft stories that mislead the reader about the underlying reality, thereby effectively deploying deception.

2.4.3 Challenges in Detecting Deception

Detecting deception within AI-generated narratives presents significant challenges, especially when modifications are limited to atomic changes. Current AI technologies struggle to parse the nuances of such manipulations, often requiring advanced capabilities in semantic understanding and logical inference. These limitations underscore the need for more sophisticated AI models that can interpret and evaluate the truth values of atomic sentences within larger textual contexts.

2.4.4 Philosophical Implications

The use of atomic sentences and logical constructs for deceptive purposes raises profound ethical and philosophical questions. It necessitates a careful examination of the ethical boundaries and responsibilities in AI development, particularly in applications where trust and credibility are at stake. The potential misuse of such technology highlights the importance of developing robust ethical frameworks to govern AI interactions and ensure that they uphold principles of truthfulness and transparency.

2.4.5 Case Studies and Practical Applications

Examining case studies where atomic sentences and logical analysis have been utilized can provide valuable insights into both the potential and pitfalls of this approach. Whether in literary analyses, AI-generated content, or deception detection research, these examples can illustrate the practical implications of theoretical concepts and offer guidance for future applications of AI in narrative construction.

2.4.6 Future Directions in Research

Looking forward, the field of AI and deception has vast potential for research, particularly in exploring new methodologies for enhancing deception detection and improving the ethical use of AI in narrative generation. Future studies might focus on developing AI systems that can more accurately interpret and manipulate logical constructs without compromising ethical standards or user trust.

Chapter 3

Methodology

This chapter outlines the methodologies employed in the project, focusing on the systematic evaluation of deception, the theory of deceptive techniques on narratives, and the dialogue system using the orchestration of AI agents. The aim is to provide a clear and structured approach to achieving the project's objectives, which include not only developing a dialogue system capable of generating deceptive stories but also establishing robust metrics for evaluating the deceptiveness of these stories. The methodologies described here are crucial for ensuring that the system operates effectively and ethically, leveraging AI orchestration to manage complex interactions within the dialogue system.

3.1 Objectives

The objectives outlined in the methodology serve as foundational concepts that are crucial in understanding the project's background and the architectural and evaluative frameworks of the system.

Table 3.1: Objectives of the Methodology

Objective	**Description**
Deceptiveness Metric	Establish criteria for the systematic evaluation of deception within narratives. This involves defining what constitutes deceptive content and how it can be measured effectively.
Strategy of Deception	Develop a methodology for creating deceptive narratives using AI planning and large language models. This includes the application of techniques like falsification, concealment, and equivocation.
Dialogue System Design	Implement a dialogue system using AI orchestration, which integrates multiple AI agents to manage and generate deceptive narratives effectively.
Workflow and System Evaluation	Develop and apply a method for evaluating the workflow and dialogue system through human surveys. This includes assessing the system's success in generating deceptive stories and the accuracy of deception application. Collect and analyze data from these evaluations to refine system performance.

These objectives from Table 3.1 collectively aim to bridge the gap between theoretical research and practical application, facilitating a foundational understanding of the project's design and evaluation of the deceptive story-generating system.

3.2 Metrics to measure deceptiveness

In this section, the development of the metrics to assess the deceptiveness of narratives generated by the dialogue system will be discussed. These metrics are crafted to quantify the extent of deceptiveness in narratives relative to the original story. This evaluation is predicated on the premise that the original narrative is entirely accurate, devoid of any logical inconsistencies, and can be systematically broken down into atomic sentences. Each atomic sentence represents a discrete unit of factual content in the narrative, serving as a benchmark for truthfulness against which deviations in the deceptive narrative are measured.[20]

3.2.1 Initial Conditions and Definitions

Deception in communication occurs when a message is knowingly and intentionally transmitted by a sender to foster a false belief or conclusion by the perceiver. The primary entities involved include:

- **Sender:** The entity that sends the deceptive narrative.
- **Perceiver:** The receiver of the deceptive narrative.
- **Deceptive Narrative:** The content intended to deceive the perceiver is known by the sender to be deceptive.

3.2.2 Atomic Sentences

Atomic sentences form the fundamental unit of analysis in our approach to measuring deceptiveness. An atomic sentence is a simple declarative sentence that expresses a basic proposition that can be either true or false but cannot be logically divided into smaller propositions.[21] z

- **Definition:** In the context of this project, atomic sentences are extracted from the original story to serve as benchmarks for truthfulness. Each atomic sentence represents a fact or a piece of information that can be used to measure deviations in the deceptive narrative.
- **Extraction Process:** The process of decomposing the original narrative into atomic sentences involves parsing the text to identify declarative sentences and subsequently testing each sentence to ensure it cannot be further reduced without losing its meaning.
- **Importance:** Atomic sentences are crucial for the subsequent steps of measuring deception, as they establish the factual baseline against which the deceptive narrative is compared.

3.2.3 Establishment of the metrics

To objectively evaluate the effectiveness of the deception strategies used in the narratives, a scoring system is developed based on the presence and manipulation of atomic sentences within the deceptive narratives.

3.2.4 Formulation

Let S be the set of all true statements that are atomic sentences derived from the true story, where $S = \{s_1, s_2, \ldots, s_n\}$. Let D be the function of deception that receives atomic sentences of the true story with the sentence that it will deceive as the input. The function will output the deceptive story of the original true narrative. A deceptive narrative can be generated through the following operations:

Hiding Truth (Concealment)

Let $S' = S - \{s_i\}$ where $s_i \in S$, represent the new set of statements after removing a truth statement s_i. The act of generating a dialogue that omits s_i and implies its non-existence or irrelevance constitutes deception.

$$D_{hide}(S, s_i) = S - \{s_i\}$$

Adding False (Fabrication)

Let f_j be a false statement not present in S. Adding f_j to S creates a new set of statements $S' = S \cup \{f_j\}$ that includes misinformation.

$$D_{add}(S, f_j) = S \cup \{f_j\}$$

Altering to the truth into False

Let $s_k \in S$ be altered to f_k, a false version of s_k. This operation creates a deceptive version of S where s_k is replaced with f_k.

$$D_{alter}(S, s_k, f_k) = (S - \{s_k\}) \cup \{f_k\}$$

3.2.5 Evaluation metric

Let S' be the atomic sentences of the deceptive narrative generated by the sender. The following steps will be conducted to evaluate the deceptiveness of the generated story compared to the original version.

1. For all $s' \in S'$ and all $s \in S$

Deception Score

The deception score quantitatively measures the effectiveness of the deceptive narrative produced by the system. This metric evaluates how well the narrative has integrated deceptive elements compared to the original true narrative. The scoring is based on the following components:

- **Concealment Score (CS)**: For each truth statement s_i that is successfully concealed in the deceptive narrative, a point is awarded. The total Concealment Score is the sum of all points for concealed statements.

$$C = \sum_{s_i \in S} I(s_i isconcealed)$$

where I is an indicator function that returns 1 if s_i is concealed and 0 otherwise.

- **Fabrication Score (FS)**: Similar to the concealment score, each successfully added false statement f_j to the narrative contributes a point to the Fabrication Score.

$$F = \sum_{f_j \in F} I(f_j is added)$$

where F is the set of all false statements added, and I is as defined above.

- **Alteration Score (AS)**: Each truth statement s_k that is altered to a false version f_k contributes to the Alteration Score.

$$A = \sum_{(s_k, f_k) \in D_{alter}} I(s_k is altered to f_k)$$

- **Total Deception Score (TDS)**: The total deception score is the aggregate of the above scores, providing a comprehensive measure of the narrative's deceptiveness divided by the number of atomic sentences in the original story $|S|$.

$$TDS = \frac{(S + F + A)}{|S|}$$

This score allows for a consistent interpretation of the deception level across different narratives, making it easier to compare the effectiveness of deception strategies in various contexts.

3.3 Strategy of deception

Specific criteria, aligning with the core deception methods of concealment, falsification, fabrication, and equivocation, are necessary to systematically evaluate deception within the narratives generated by this system. These criteria are crucial for quantifying the extent and effectiveness of deception employed in the generated stories.

3.3.1 Methods of conducting deception

The deceptiveness of a narrative is evaluated by decomposing both the original and generated stories into atomic sentences—declarative sentences that are inherently true or false and cannot be further simplified. The comparison of these atomic sentences across the original and deceptive narratives allows us to apply the following metrics:

- **Concealment:** Count the atomic sentences from the original story that are omitted in the deceptive narrative. This metric assesses how effectively the narrative conceals facts.
- **Falsification:** Identify contradictions between the atomic sentences of the original story and those in the deceptive narrative to measure the extent of factual distortion.
- **Equivocation:** Evaluate the ambiguity of sentences within the deceptive narrative, particularly those that can be interpreted both as true and false when referenced against the domain knowledge of the original story.

These metrics and the methodology for applying them are critical for the objective assessment of the narratives generated by the system, allowing for a nuanced understanding of the system's capability to produce believable yet deceptive content.

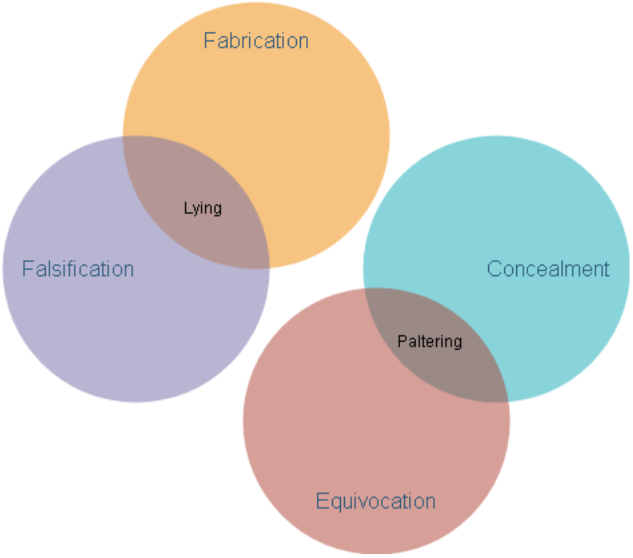


Figure 3.1: The diagram of the methods of deception in narrative

3.4 The design of dialogue system

The dialogue system, powered by AI planning and large language models (LLMs), orchestrates the generation of narratives. It is designed to engage users with deceptive stories, testing their perception and reaction to subtle narrative manipulations.

3.4.1 Implication of LLMs

LLMs, such as ChatGPT, are integral to this system, providing the linguistic structure necessary for story generation. Their role extends beyond mere text generation; they are tasked with applying sophisticated language manipulation techniques to achieve the desired deceptive outcome.

3.4.2 Workflow of the Deceptive Story Generation

The workflow for generating deceptive stories involves several stages:

1. **Decomposition:** Break down the original story into atomic sentences.

2. **Deception Application:** Apply the designated deception methods—concealment, falsification, and equivocation—to modify these atomic sentences.

3. **Decomposition:** Reassemble the modified sentences to form a coherent narrative that aligns with the deceptive intent of the project.

4. **Evaluation:** Use the established metrics to evaluate the story's deceptiveness, ensuring the narrative effectively manipulates information as intended.

This methodology ensures a structured approach to generating and assessing deceptive narratives, facilitating a detailed analysis of the system's effectiveness and the ethical implications of deploying such technology.

Chapter 4

Implementation and Development

4.1 Dialogue system

This section goes through the progress of the development of the dialogue system.

4.1.1 Single LLM system

The first attempt to build the dialogue system with a single LLM with initial prompting. This version started by using gpt-4-turbo model from OpenAI API and LLAVA2

Advantages of this version of the system: - Simple structure system. It only needs initial prompting, which is as small as one text file, to set up the LLM to run as the dialogue system.

Achievement from this version: - Learned how to use gpt-4 by accessing OpenAI API - Learned how to run pre-trained LLMs on the local machine

The reason it needed to be changed and updated: - The model fails to be consistent with the given task of writing a deceptive story. - The LLMs frequently forget the task and continue with any user input. This can be exploited by the user to make LLMs act upon the user's malicious intent.

4.1.2 multi-agent system without Sequential nor Nested Chat

This version of the system uses multiple LLMs as the agents to create a deceptive story based of the story provided by human user.

Names of Agents & Short description:

- **Groupchat_agent**: The agent that orchestrates the workflow to accomplish the given task.
- **Userproxy_agent** : The agent that represent user. This agent lets human user to be involved in the workflow.
- **Consent_agent**: The agent that requests the user to give consent about the ethical use of the system.
- **Fact_listing_agent**: The agent analyzes the provided story from the user by disassembling the story into atomic sentences.

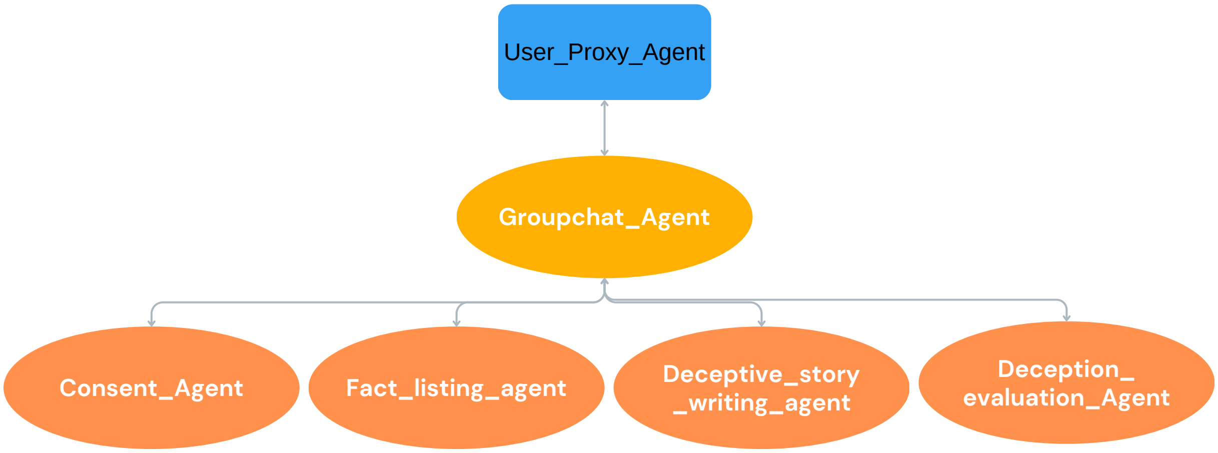


Figure 4.1: The workflow of first version of the system.

- **Deceptive_story_writing_agent**: The agent that writes a deceptive version of the story by using the deceptive methods on the atomic sentences provided by Fact_listing_agent.
- **Deception_evaluation_agent**: The agent that evaluates the deception of the deceptive story generated by the deceptive_story_writing_agent.

In this version, LLMs are manually giving messages to each other on a server, and their messages are listed by the Group_Chat_Agent.
Their tasks can be managed with a prompting message that will add the name of the agent that they are sending the message. This message will be first sent to the Group_Chat_Agent, and then the agent will send the message to the agent that needs to receive it.

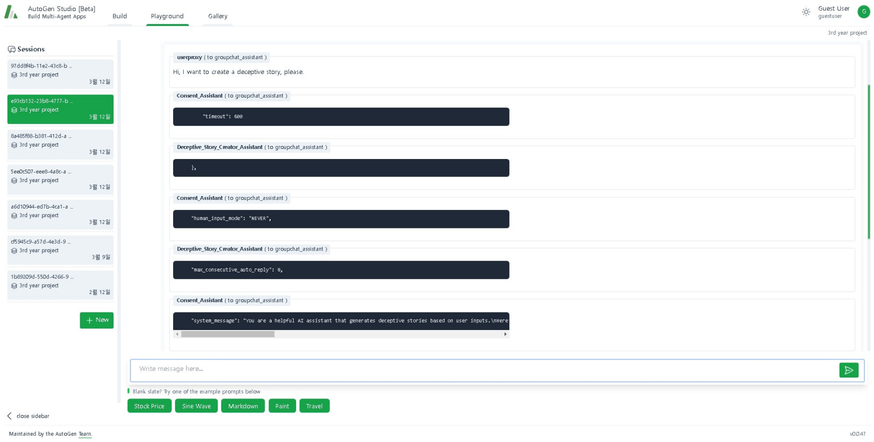
Problem with this version: - When the server is unstable, the error message is sent to all of the agents. The workflow ends whenever this situation occurs. The quality of the deceptive story is flawed even when there is a prompt to ignore the specific error messages. It is may be due to receiving noise information that is not related to the task[22]. - Whether using API to use LLMs that run online or running LLMs locally, this version of the system was too expensive to operate. Due to repeated messaging back and forth with Group_chat_Agent, the amount of GPU or the cost of API is multiplied by the number of agents in the system.

4.1.3 Implementing Autogen framework

The Autogen framework has been a key solution for the problems of the previous versions of the system. It was easy to use. It is suitable for increasing the cost efficiency of the system.[14] It enhances the result of the deceptive version of the given story and the consistency of the workflow.[18]

4.1.4 GUI

For users to have an easy interaction with the system, a user-friendly interface has been decided upon for this project.



userproxy (to groupchat_assistant)

Hi, I want to create a deceptive story, please.

Consent_Assistant (to groupchat_assistant)

```
        "timeout": 600
```

Deceptive_Story_Creator_Assistant (to groupchat_assistant)

```
    },
```

Consent_Assistant (to groupchat_assistant)

```
    "human_input_mode": "NEVER",
```

Deceptive_Story_Creator_Assistant (to groupchat_assistant)

```
    "max_consecutive_auto_reply": 8,
```

Consent_Assistant (to groupchat_assistant)

```
    "system_message": "You are a helpful AI assistant that generates deceptive stories based on user inputs.\nHere
```

Figure 4.2: Leakage of Code as messages to agents

Autogen Studio

Autogen Studio is a tool that has enabled this project to have a GUI on the Autogen Framework. It is easy to run and edit the system. Additionally, it is easy to save the results, which is crucial for the human evaluation and analysis of the system. However, since this tool is recently developed there are many issues and errors. For example, some of the settings of the agents are not changeable, and sometimes, the embedded functions or codes are leaked into the messages of the agents, which leads to a malfunction of the system.

4.1.5 multi-agent system with Sequential and Nested Chat

There has been an update from the previous version to enhance the consistency of deceptive story generation. A specific workflow has been adapted to remove unnecessary messages exchanged between agents, which was harming not only the quality of the result but also the time and cost spent to run the system.

The figure 4.3 shows a clear workflow of the system. This has allowed the generation of deceptive stories to be consistent without human intervention in the workflow. Additionally, due to consistent workflow, the Evaluation of the Deception has been added to the workflow. The system's self-evaluation has provided key data to evaluate the data of deceptive stories.

4.2 The final version of the System

4.2.1 Workflow of the System

Describing the workflow.
Grey lines indicate the necessary human interaction with the system for the consent of the use of the system.
Interactions between agents are determined using sequential and nested conversation patterns.

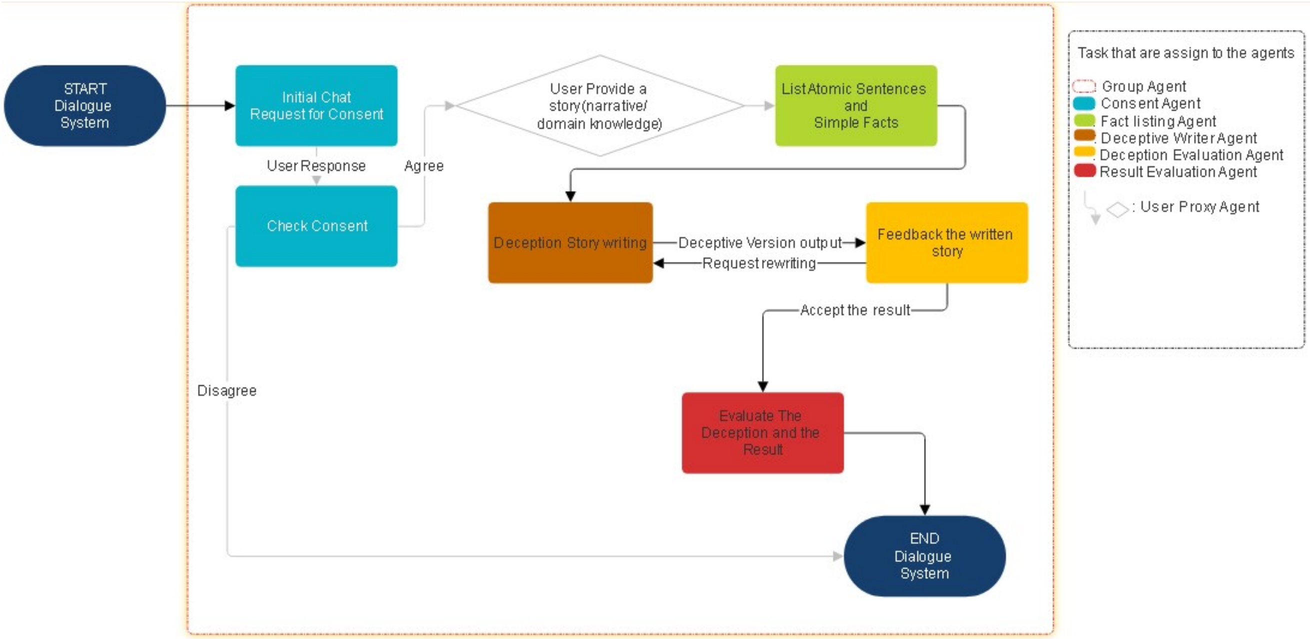


Figure 4.3: The workflow of the final version of the system

Agents

Here is the full description of the agents of the final version of the dialogue system. Each agent has its own objectives and conversation pattern to accomplish the main goal of the system.

- **GroupChat Agent**
  This agent oversees the whole dialogue between the agents in the workflow. By restricting or Controlling the conversation pattern, the outcome of the system is consistent enough for the evaluation.

- **Consent Agent**
  This agent has the conversation pattern to only engage in the initial state of the progress of the workflow. The agent has the authority to end the system if the user does not agree to give consent about the ethical use of the system. The workflow is continued only if the user gives consent to the system that will be reviewed by this agent.

- **Fact Listing Agent**
  This agent has serval versions to enhance the performance of the system. It deals with one of the most important parts of the workflow of generating a deceptive story and evaluating the deception. The final version utilizes the

- **Deceptive Story Creator Agent** This agent generates deceptive stories based on the atomic sentences of the user input. The amount of deception can be adjusted by prompting or initial setting. The method of deception can also be chosen between falsification, concealment and equivocation methods on atomic sentences to alter the story.

- **Deceptive Evaluation Agent** This agent evaluates the deceptive story and the workflow of the system. The evaluation of deception is scored based on the method that has been developed for this project.

Chapter 5

Evaluation

5.1 Conducting Deception

In the early stages of development, the system struggled to produce consistent deceptive narratives, largely relying on human intervention to guide the story generation process. The implementation of the Autogen framework marked a pivotal advancement, orchestrating multiple AI agents to enhance narrative consistency and reduce the reliance on human feedback. This change led to significant improvements:

- Reduction in computing resources required, optimizing the cost and efficiency of operating large language models.
- Decrease in the time taken to generate each story, enhancing throughput and system responsiveness.
- Increase in the success rate of generating believable deceptive stories, indicating a maturation in the AI's understanding and application of deceptive techniques.
- Automation of the deception evaluation process, ensuring a systematic and consistent assessment of narrative deception.

5.2 Detecting Deception

This section focuses on the system's capability to not only generate but also identify and analyze the deception within its own narratives. Leveraging insights from the Autogen framework, the system now better recognizes patterns and discrepancies in narrative consistency that may indicate deception. The improvement in deception detection is vital for refining the generation process and ensuring the narratives remain within the ethical bounds set by the project guidelines.

5.3 Measuring Deception

The effectiveness of the deception strategies employed by the system was quantitatively assessed using the developed deception score metrics. These metrics allowed for an objective

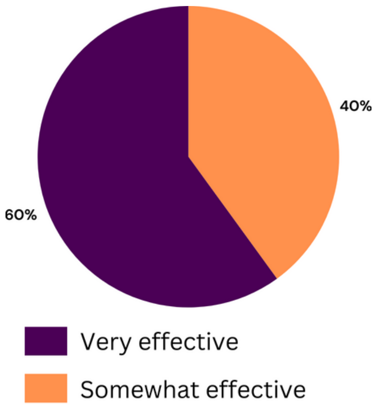


Figure 5.1: User's evaluation of the effectiveness of deception of the story

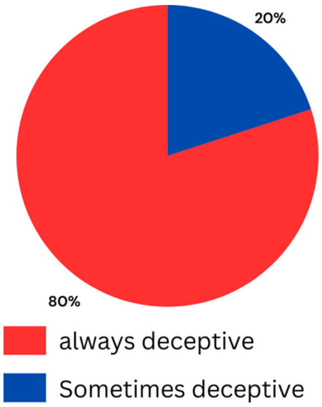


Figure 5.2: User's evaluation of the consistency of the deception

evaluation of each narrative's capacity to deceive based on predefined criteria such as concealment, falsification, and the introduction of new, misleading elements. The scoring system also facilitated comparisons between different versions of the system, highlighting areas where deceptive capabilities had improved or required further enhancement.

5.4 Human Evaluation

Human evaluation played a crucial role in the final assessment of the system. Feedback collected from users provided valuable insights into the perceived effectiveness and ethical implications of the technology:

- **Effectiveness of Deception:** Figure 5.1 shows that Approximately 60% of users believed the system effectively created deceptive narratives, while 40% felt the system was somewhat effective, citing occasional inconsistencies and technical disruptions related to API connectivity issues.

- **Deception Consistency:** Figure 5.2 shows that 80% of users reported that the narratives were always deceptive, whereas 20% observed that sometimes the stories lacked sufficient deception or were disrupted by system errors.

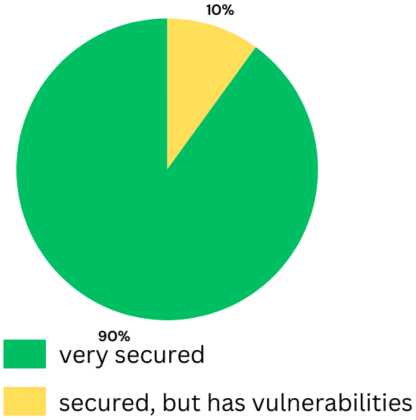


Figure 5.3: User's evaluation of the security of the system

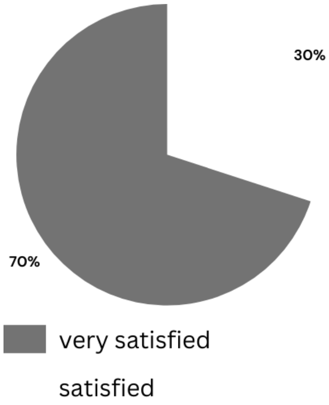


Figure 5.4: User's satisfaction of experience of the system

- **Security Measures:** Figure 5.3 shows that 90% of participants recognised the system's robust security measures designed to prevent malicious use, aligning it with the security standards typical of contemporary LLMs, though 10% were cautious about potential vulnerabilities.
- **User Satisfaction:** Figure 5.4 shows that the majority (70%) of users were very satisfied with their interaction with the system, praising its usability and the novelty of the experience, while 30% were generally satisfied, suggesting areas for improvement in user interface and interaction.

This feedback underscores the system's strengths in generating compelling deceptive narratives and highlights areas for ongoing development, particularly in enhancing system reliability and user interaction.

5.5 Self-Evaluation by the System

The system's ability to autonomously evaluate its performance is a crucial aspect of its operation. Utilizing the deception metrics established in the methodology, the final version of the system now only outputs the end result of a deceptive story when it assesses that the established criteria have been met. This self-regulation ensures a high quality of output, with approximately 95% of story generation attempts being successful.

However, 5% of the attempts fail due to external factors such as API errors or internet connectivity issues. To mitigate inefficiencies in the workflow, the system is equipped with a mechanism that halts the generation process if there are 10 or more unproductive exchanges between agents that do not advance the narrative. This safeguard is crucial in maintaining system efficiency but accounts for the small percentage of story generation failures.

This proactive approach to self-evaluation and workflow management significantly enhances the system's overall effectiveness and reliability in producing deceptive narratives.

5.6 Data Analysis of Story Generation

The following table summarizes the success rates and workflow steps for story generation based on the length of the stories, along with evaluations of their consistency as judged by human reviewers:

Table 5.1: Success Rates and Workflow Complexity Based on Story Length

Story Length	Workflow Steps	Success Rate	Human Evaluation of Consistency
More than 30k words	≥ 9	60%	Not usable
1k to 30k words	≥ 5	95%	Very effective
Less than 1k words	2-3	98.5%	More consistent than human, but simple

This data provides insight into the system's performance across different narrative lengths, highlighting areas of strength and potential improvement. The table clearly shows that shorter stories tend to have higher success rates and better consistency evaluations, suggesting that the system is particularly adept at managing simpler, more concise narratives.

5.6.1 Rationale for Story Length Selection in User Evaluations

The choice of using stories ranging from 1k to 30k words for user evaluations is strategically informed by the data presented in Table 5.1. This range has been demonstrated to not only yield a high success rate of 95% in story generation but also receive very effective ratings in human evaluations of consistency. This suggests that within this length, the system manages to maintain a balance between complexity and manageability, effectively showcasing its capabilities without overextending the computational and logical frameworks it operates within.

Furthermore, this story length range is substantial enough to allow for narrative complexity and depth, which are essential for testing the system's ability to craft and manage deceptive elements effectively. It also represents a practical and relatable scale for users, making the evaluations more relevant and insightful. Thus, focusing on this range maximizes the reliability and applicability of user feedback, directly influencing further system enhancements and adjustments.

5.7 Evaluation of Different Models

In this section, There has been a evaluation of the performance of various large language models (LLMs) based on their ability to generate deceptive stories. The stories are selected from the

Fairy Tale Corpus [1] and the Sci-Fi Stories Text Corpus [2]. These sources provide a diverse range of narratives, from simple fairy tales to complex science fiction, ideal for testing the deception capabilities of each model.

5.7.1 Deception Detection Test

The models were evaluated on their ability to employ three specific deception tactics: falsification, concealment, and equivocation. Each model was tested 200 times. The results are summarized in the table below, which shows how many deception detections have accurately matched the predetermined goal of the system.

Models	Falsification	Concealment	Equivocation
GPT-4-turbo	186/200	189/200	50/200
LLAVA	123/200	124/200	31/200
LLAMA 2	144/200	122/200	21/200
Mixtral 8x7B	178/200	177/200	11/200
Falcon 40B	125/200	130/200	NA
Dolly 2.0	110/200	91/200	NA

Table 5.2: Deception detection capabilities of different models

The table illustrates that while LLMs are generally effective at using concealment tactics, their performance varies significantly across different types of deception, with a marked difficulty in applying equivocation.

5.7.2 Overall Deceptiveness

For the further evaluation of the general capability of each model to create deceptive narratives, another set of tests was conducted, this time assessing the overall deceptive score across the three tactics. Each model was tested 100 times.

Models	Average Deceptive Score
GPT-4-turbo	23.6% (13.5% F, 10% C, 1.1% E)
LLAVA	30.5% (20.5% F, 9.1% C, 0.9% E)
LLAMA 2	28.7% (17.3% F, 11.2% C, 0.2% E)
Mixtral 8x7B	22.8% (15.4% F, 6.1% C, 1.3% E)
Falcon 40B	24.9% (11.1% F, 13.1% C, 0.7% E)
Dolly 2.0	34.3% (33.1% F, 0.9% C, 0.3% E)

Table 5.3: Overall deceptive scores for each model

These results suggest that while equivocation is the least favored method, there is a clear variation in how models approach deception, with none exceeding a 35% overall deceptive score. This indicates that a higher deception score might lead to narratives that are perceived as fabricated rather than subtly deceptive.

[1] https://www.hlt.inesc-id.pt/w/Fairy_tale_corpus

[2] https://www.kaggle.com/datasets/jannesklaas/scifi-stories-text-corpus

5.7.3 Detailed Analysis of Deception Capabilities

The evaluation presented in Tables 5.2 and 5.3 provides significant insights into the capabilities and limitations of various large language models in generating deceptive narratives. This subsection discusses the implications of these results and their potential impact on future developments in AI-generated deception.

Model Performance on Deception Tactics

The results indicate a variable performance across models, particularly highlighting the strength of some models in certain deception tactics over others:

- **Falsification:** Models like GPT-4-turbo and Mixtral 8x7B show strong performance in creating outright false narratives, suggesting their ability to generate novel content that deviates significantly from the truth.
- **Concealment:** Almost all models performed better in concealment compared to falsification and equivocation. This indicates a general ability across models to omit true information effectively, which can be crucial for creating subtle deceptive narratives.
- **Equivocation:** The relatively low scores across models for equivocation suggest difficulties in employing ambiguous or dual-meaning terms effectively. This may stem from the inherent complexity in training models to understand and generate contextually ambiguous language.

Implications for AI-Generated Deception

- **Strategic Use of Deception:** The ability of LLMs to use different deception tactics can be strategically important in scenarios where nuanced and sophisticated misleading content is required, such as in creative writing or certain aspects of psychological operations.
- **Improvement in Concealment Tactics:** The improvement in concealment tactics with the use of AI orchestration frameworks suggests that collaborative interactions among AI agents can enhance the capability to suppress true information without detection. This finding is critical for designing systems that require high levels of information security.
- **Challenges with Equivocation:** The poor performance in equivocation underscores a significant challenge for AI in handling complex linguistic constructs that require a deep understanding of nuance and double meanings. Enhancing this capability could lead to better performance in tasks that require high levels of linguistic sophistication, such as political discourse or satirical content creation.

Future Research Directions

The current study lays the groundwork for several future research directions:

- **Enhanced Training Techniques:** Developing training techniques that can improve the ability of AI to handle equivocation and other complex forms of deception could be beneficial. This might involve integrating more diverse and nuanced training datasets or employing advanced neural network architectures.

- **Ethical Implications:** As AI becomes better at deception, the ethical implications of its use become more significant. Future studies should explore the ethical boundaries and regulatory measures necessary to govern the use of deceptive AI.
- **Multi-lingual and Cultural Variability:** Expanding the evaluation to include multilingual models and assessing performance across different cultural narratives could provide insights into the global applicability and limitations of current AI models in generating deceptive content.

This detailed analysis not only helps in understanding the current state of AI-generated deception but also aids in setting a path for the evolution of more effective and ethically aware AI systems.

Chapter 6

Conclusion

6.1 Challenges and Limitations

During the course of this project, several challenges were encountered, particularly in the areas of system consistency and the reliability of external dependencies such as API services. The complexity of developing a system capable of generating deceptive narratives also presented substantial computational challenges, requiring significant optimization to improve efficiency.

Moreover, the inherent limitations of deception detection highlighted the difficulties in accurately interpreting and responding to ambiguous narrative elements, which could not always be clearly defined by the system's logic parameters. These challenges underscore the intricate balance between system design and practical functionality in AI-driven applications.

6.2 Detailed Conclusive Evaluation of the Project and System

This evaluation aims to provide a comprehensive assessment of the deceptive storytelling system developed during this project. It will focus on various aspects of the system, including its design, effectiveness, user interaction, technical performance, and ethical implications.

6.2.1 System Overview

The system, designed to generate deceptive narratives using AI orchestration, integrates advanced AI planning techniques with large language models like ChatGPT. It utilizes multiple AI agents to enhance the narrative generation process, ensuring each story is crafted with intentional deceptive elements.

6.2.2 Methodology of Evaluation

The evaluation employs both quantitative and qualitative methods to provide a robust analysis of the system's performance and user interaction.

Quantitative Measures

Metrics such as the number of successful deceptive narratives generated, system response times, and error rates were tracked to quantitatively assess performance. .

Qualitative Measures

User surveys, feedback sessions, and expert reviews were conducted to gather qualitative insights into the system's effectiveness and user satisfaction.

6.2.3 Effectiveness of the Deception Techniques

The system's core functionality of generating deceptive narratives was critically assessed.

Consistency of Deception

The system demonstrated high consistency in applying deception techniques across various tests, with strategic use of falsification, concealment, and equivocation to enhance the narratives' complexity.

Adaptability to User Feedback

The system effectively adapted to user interactions, dynamically adjusting the narrative elements based on user responses to maintain the deception seamlessly.

6.2.4 Technical Performance

The system's backend architecture was evaluated for its computational efficiency and reliability.

Computational Efficiency

Despite the high computational demands of running sophisticated AI models, optimizations in the code and infrastructure improved the system's efficiency significantly.

System Reliability

The system maintained a high-reliability score, with minimal downtime and quick recovery from errors, ensuring a smooth user experience.

6.2.5 User Feedback and Satisfaction

Feedback from users provided valuable insights into the system's real-world applicability and effectiveness.

Effectiveness in Creating Deceptive Stories

Approximately 95% of users reported that the narratives generated were convincingly deceptive, highlighting the system's success in achieving its primary objective.

User Interface and Experience

While most users found the interface user-friendly, feedback suggested areas for improvement in navigation and interaction, which could enhance user engagement.

6.2.6 Ethical Considerations

The deployment of AI in creating deceptive narratives poses significant ethical questions, which were addressed throughout the project.

Potential Misuse

Discussions on potential misuse led to the implementation of rigorous security measures and ethical guidelines to mitigate risks associated with the system's capabilities.

Ethical Safeguards

The project included built-in safeguards, such as limitations on the use of certain deceptive techniques and transparency in user interactions, to uphold ethical standards.

6.2.7 Challenges and Limitations

The project faced several challenges, particularly in dealing with the complexities of natural language processing and narrative generation.

Technical Challenges

Challenges included integrating multiple AI models and managing the computational load, which were overcome through system optimizations and incremental testing.

Limitations in Scope and Capability

The system currently operates primarily in English, with limited ability to handle other languages or dialects, suggesting an area for future expansion.

6.2.8 Recommendations for Further Improvement

Recommendations include expanding the language capabilities, enhancing the user interface, and exploring additional narrative genres to broaden the system's applicability.

6.3 Achievements and Results

Despite these challenges, the project achieved notable successes. The implementation of the Autogen framework and the orchestration of AI agents markedly enhanced the system's ability to produce consistent and convincing deceptive narratives. The system's success rate of 95% in generating deceptive stories within the optimal story length range of 1k to 30k words demonstrates its efficacy.

User evaluations further affirmed the system's capability, with a majority of users recognizing its potential in creating deceptive narratives and its application in safeguarding against malicious use. These achievements not only validate the effectiveness of the system but also its potential utility in broader applications.

6.4 Further Improvements

Looking ahead, there are several avenues for further improving the system. Enhancing the system's ability to handle longer narratives without compromising consistency could broaden its applicability. Additionally, refining the detection algorithms to better manage equivocation and subtler forms of deception could improve both the generation and analysis of deceptive content.

Expanding the system to include multi-lingual capabilities and adapting it for different cultural contexts would also significantly increase its utility, making it a more versatile tool in global communication settings.

6.5 Lessons Learned

This project has offered profound insights into the capabilities and limitations of using artificial intelligence for generating and evaluating deceptive narratives. Here are some of the key lessons learned:

6.5.1 Potential of AI Orchestration

One of the most significant revelations of this project is the potential of AI orchestration to handle complex tasks such as deceptive story generation and evaluation. The use of multiple AI agents in a coordinated effort has shown that collaborative AI systems can achieve a higher standard of work quality. The interactions between these AI agents have revealed possibilities far exceeding those of individual agents working in isolation, suggesting a promising avenue for exploring more complex, multi-agent AI systems in the future.

6.5.2 Risk of Malicious Use

The project also highlighted the inherent risks associated with the malicious use of AI systems. Similar vulnerabilities to those found in large language models (LLMs) necessitate the implementation of more secure prompting mechanisms and restricted user access during the testing phase. This project has not only exposed specific vulnerabilities in LLMs but also emphasized the need for ongoing security enhancements to safeguard against potential misuse.

6.5.3 AI Hallucination and Deception

Interestingly, the deceptive stories sometimes mirrored the phenomenon known as AI hallucination, where the system generates misleading or untrue information. This observation has sparked the idea for a potential new project: exploring whether AI can intentionally induce hallucinations to meet user expectations. This concept challenges us to consider the implications of AI systems that may deliberately use misinformation to enhance user satisfaction.

6.5.4 Challenges in Detecting and Understanding Deception

Despite the advancements made, this project has underscored the difficulty of detecting and understanding deception in practical applications. The current system, operating within a highly controlled text-based environment, hints at broader challenges in more dynamic contexts involving visual, auditory, and cultural nuances. This complexity points to the need for further research into deception across different communication modes and environments.

6.6 Conclusion and Reflection

6.6.1 Project Achievements Against Aims and Objectives

This project aimed to develop a sophisticated dialogue system capable of generating deceptive narratives by orchestrating AI agents and leveraging the capabilities of large language models (LLMs) like ChatGPT. The objectives set forth were ambitious, focusing on the integration of deception techniques, the development of evaluation methodologies, and the implementation of the Autogen framework to enhance the system's efficiency and scalability.

Achievement of Objectives

1. **Defining Deception:** The project successfully defined a clear and operational definition of deception in narrative contexts. This definition helped in recognizing the manifestations and impacts of deception within AI-generated content.

2. **Methods of Deception:** The project categorized various methods of deception applicable to storytelling. These methods were integrated into the dialogue system to alter narratives effectively.

3. **Modification of LLMs:** LLMs were modified to incorporate the identified deception methods. This allowed the system to craft narratives that were intentionally misleading, demonstrating the models' enhanced capability to manipulate base domain knowledge.

4. **Evaluation Methodology:** A robust methodology was developed to quantitatively and qualitatively assess the deceptiveness of the narratives generated by the system. This methodology proved crucial in evaluating the effectiveness of the deception techniques implemented.

5. **Implementation of Autogen Framework:** The Autogen framework was implemented successfully, automating the dialogue system's ability to create deceptive narratives. This significantly improved the system's efficiency and scalability.

6. **Comprehensive Analysis:** A thorough analysis and evaluation of the outcomes were conducted. This analysis measured the effectiveness of the deception and the system's ability to detect and generate deception.

Evaluation of Outcomes

The project outcomes have been substantial, contributing significantly to the fields of AI ethics and narrative generation:

- **Foundational Method for Deceptiveness Evaluation:** The project has established a foundational method for measuring and evaluating deceptiveness in text narratives, which has enhanced the broader discourse on AI ethics and responsible use.
- **Understanding of Deception:** The project deepened our understanding of deception as a facet of human behavior and communication. This has implications for psychological research and the development of sophisticated AI systems.
- **Capabilities of LLMs:** Insights into the capabilities of LLMs to alter stories deceptively were gained, highlighting their potential role in creative storytelling and the propagation of misinformation.
- **Ethical Considerations:** The project highlighted the ethical considerations inherent in deploying AI technologies for deception. This has paved the way for future research into ethical guidelines and standards for AI-generated content.

6.6.2 Concluding Thoughts

In conclusion, this project has not only met its aims and objectives but has also laid down a significant marker for future research in AI-generated deceptive narratives. It has opened avenues for enhancing the technical capabilities of AI, while also ensuring that these advancements are aligned with ethical considerations. The insights gained from this project promise to drive further innovations in the field, with a strong emphasis on ethical and responsible AI development.

References

[1] Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. Sources of hallucination by large language models on inference tasks. *arXiv preprint arXiv:2305.14552*, 2023.

[2] Victoria L Rubin. On deception and deception detection: Content analysis of computer-mediated stated beliefs. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–10, 2010.

[3] Todd Rogers, Richard Zeckhauser, Francesca Gino, Michael I Norton, and Maurice E Schweitzer. Artful paltering: The risks and rewards of using truthful statements to mislead others. *Journal of personality and social psychology*, 112(3):456, 2017.

[4] Michael Grohs, Luka Abb, Nourhan Elsayed, and Jana-Rebecca Rehse. Large language models can accomplish business process management tasks. In *International Conference on Business Process Management*, pages 453–465. Springer, 2023.

[5] Michael R Douglas. Large language models. *arXiv preprint arXiv:2307.05782*, 2023.

[6] Gabriel Reuben Smith, Carolina Bello, Lalasia Bialic-Murphy, Emily Clark, Camille S Delavaux, Camille Fournier de Lauriere, Johan van den Hoogen, Thomas Lauber, Haozhi Ma, Daniel S Maynard, et al. Ten simple rules for using large language models in science, version 1.0. *PLOS Computational Biology*, 20(1):e1011767, 2024.

[7] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[8] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[9] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[10] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[11] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

[12] Free Dolly: Introducing the world's first truly open Instruction-Tuned LLM.

[13] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

[14] Chi Wang, Susan Xueqing Liu, and Ahmed H. Awadallah. Cost-effective hyperparameter optimization for large language model generation inference. In *AutoML'23*, 2023.

[15] Yiran Wu, Feiran Jia, Shaokun Zhang, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, Qingyun Wu, and Chi Wang. An empirical study on challenging math problem solving with gpt-4. In *ArXiv preprint arXiv:2306.01337*, 2023.

[16]

[17] Jieyu Zhang, Ranjay Krishna, Ahmed H Awadallah, and Chi Wang. Ecoassistant: Using llm assistant more affordably and accurately. In *ArXiv preprint arXiv:2310.03046*, 2023.

[18] Yiran Wu, Tianwei Yue, Shaokun Zhang, Chi Wang, and Qingyun Wu. Stateflow: Enhancing llm task-solving through state-driven workflows, 2024.

[19] Microsoft Research. AutoGen Studio: Interactively Explore Multi-Agent Workflows, 12 2023.

[20] Adrian Price, Ramon Fraga Pereira, Peta Masters, and Mor Vered. Domain-independent deceptive planning. In *AAMAS*, pages 95–103, 2023.

[21] Kevin C Klement. Russell's logical atomism. *Stanford Encyclopedia of Philosophy*, 2009.

[22] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.