Chapter 2

Literature Review

2.1 Existing Research and Implementation

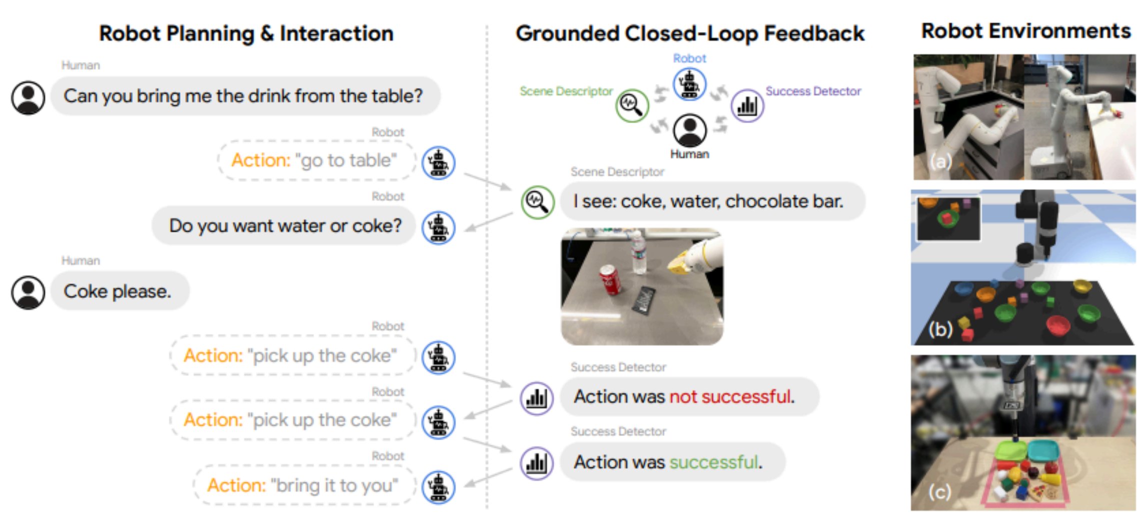
2.1.1 Generative Agents: Interactive Simulacra of Human Behavior

This is the research paper that inspired our mini-project, the AI-Village. The paper describes a system that allows users to observe and interact with these agents in a simulated environment, potentially useful for creating virtual worlds, testing social theories, or designing products.

A significant milestone has been achieved in the field of artificial intelligence, as researchers have successfully created a virtual village where autonomous agents, powered by large language models such as ChatGPT, interact and communicate with each other. The study highlights the agents' ability to navigate and engage in social activities, including planning and attending a birthday party. The party exemplifies the agents' capacity for coordination and cooperation, as one agent invited all others to join in the celebration, and they all arrived punctually to participate in the festivities. This achievement underscores the potential for AI agents to collaborate and exhibit complex social behaviors, thereby paving the way for future applications in areas such as customer service, education, and entertainment

2.1.2 Inner Monologue: Embodied Reasoning through Planning with Language Models

The paper "Inner Monologue: Embodied Reasoning through Planning with Language Models" proposes a novel approach to embodied reasoning, which is the ability of an AI system to reason about its physical environment and interact with it. The authors introduce a new framework that combines planning and language processing to enable embodied reasoning.



When asked a question, the AI in the paper goes to a “grounded closed-loop feedback” and gets relevant information before answering. The internal monologue can happen for multiple loops until the question seems to be answered well. This can lead to longer answering time but will lead to the answer being very accurate compared to traditional models.

2.1.3 Improving Factuality and Reasoning in Models through Multiagent Debate

The paper proposes a novel approach to improve the factuality and reasoning capabilities of AI models. The authors introduce a framework that leverages multiagent debate to facilitate the exchange of information and improve the accuracy and coherence of AI models.

Round 1

ChatGPT: Before the restart, Carla was able to download 40% of the file, which is 0.4 * 200 GB = 80 GB. At a rate of 2 GB/minute, it would take her: time = 80 GB ÷ 2 GB/minute = 40 minutes.

Because of the restart, Carla had to start over. This means she had to download the entire 200 GB file. The time it takes to download the remaining 120 GB is: time = 120 GB ÷ 2 GB/minute = 60 minutes.

But Carla also had to wait 20 minutes for the Windows update to install, so the total time it took to download the file is: total time = 40 minutes + 60 minutes + 20 minutes = **120 minutes**.

Bard: Carla is downloading a 200 GB file at a speed of 2 GB/minute. This means that it will take her 200 / 2 = 100 minutes to download the file.

However, Windows forces a restart to install updates, which takes 20 minutes. This means that Carla will have to restart the download from the beginning. Since Carla has already downloaded 40% of the file, she only needs to download 200 - 40 = 160 GB. At a speed of 2 GB/minute, it will take Carla 160 / 2 = 80 minutes to download the file.

Adding the time for the Windows restart, it will take Carla 80 + 20 = **100 minutes** to download the file.

Round 2

ChatGPT : Based on the calculations provided by the other agent, the correct answer is: Before the restart, Carla was able to download 40% of the file, which is 0.4 * 200 GB = 80 GB. At a rate of 2 GB/minute, it would take her: time = 80 GB ÷ 2 GB/minute = 40 minutes

However, because of the restart, Carla had to start over. This means she had to download the entire 200 GB file. The time it takes to download the entire 200 GB file is: time = 200 GB ÷ 2 GB/minute = 100 minutes

And since Carla had to wait 20 minutes for the Windows update to install, the total time it took to download the file is: total time = 40 minutes + 100 minutes + 20 minutes = **160 minutes**.

In this paper, multiple AI agents(ideally different models) are tasked to solve a problem. This can lead to a debate-like scenario where the answers from different models may vary. Then the output of all the different models is put into another AI, which will access all the above results and try to formulate a better answer. The research shows that the answer from such a system is vastly superior to any of the models. This can be vastly observed in the case of mathematical and reasoning problems.

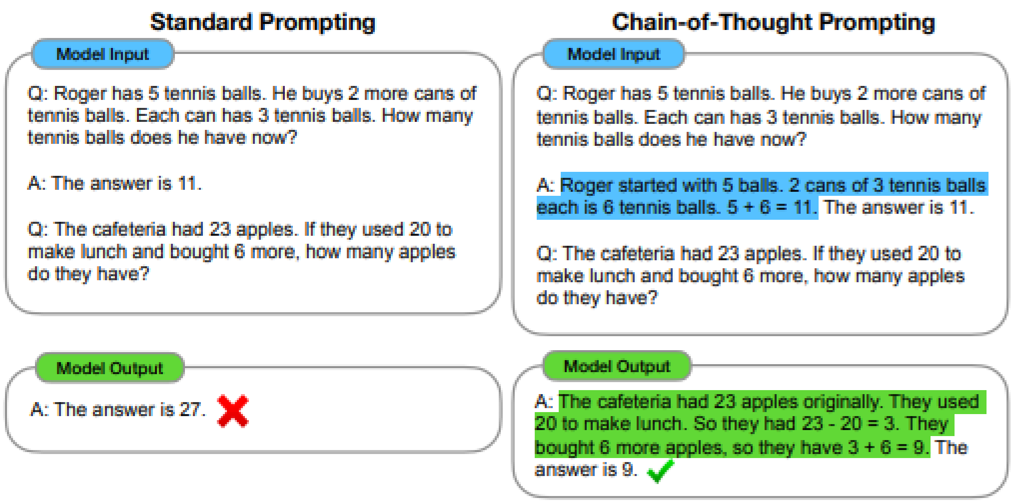
2.1.4 Character-LLM: A Trainable Agent for Role-Playing

This research paper introduces Character-LLM, a novel artificial intelligence (AI) model designed to engage in role-playing conversations. The model is trained to adopt a specific character or persona, allowing it to respond in a way that is consistent with that character's personality, traits, and backstory. This has numerous applications in areas such as entertainment, education, and customer service, where interactive storytelling and immersive experiences are valuable. The Character-LLM model has the potential to revolutionize the way we interact with AI systems, enabling more realistic and engaging conversations.

The model is trained on a large dataset of text, which includes character descriptions, dialogue, and backstory information. During training, the model learns to generate responses that are consistent with the character's personality and traits, while also taking into account the context of the conversation. The model's architecture consists of a transformer-based encoder-decoder structure, which enables it to process and generate text inputs and outputs. Overall, Character-LLM represents a significant advancement in AI research, enabling the creation of more realistic and engaging conversational agents.

2.1.5 Chain of Thought Prompting Elicits Reasoning in Large Language Models

This research paper introduces a novel approach to prompting large language models to elicit reasoning and problem-solving abilities. The authors propose a method called "chain of thought" prompting, which involves providing the model with a series of prompts that guide it through a step-by-step reasoning process. This approach enables the model to break down complex problems into smaller, more manageable steps, and to generate explanations for its reasoning. The results show that this method can significantly improve the model's ability to reason and solve problems, and has the potential to unlock new capabilities in large language models.



From a technical perspective, the "chain of thought" prompting method involves providing the model with a series of prompts that are designed to elicit a specific reasoning process. The model is trained to respond to each prompt in a way that is consistent with the previous responses, creating a "chain of thought" that reflects the model's reasoning process. Overall, this research has significant implications for the development of more advanced AI systems that can reason and solve complex problems.

2.1.6 AI-Based Conversational Agents: A Scoping Review From Technologies to Future Directions

This scoping review provides a comprehensive overview of the current state of AI-based conversational agents, which are computer programs that use artificial intelligence to engage in conversation with humans. The review covers the various technologies and techniques used to develop conversational agents, including natural language processing, machine learning, and dialogue management. The authors also discuss the applications of conversational agents in areas such as customer service, healthcare, and education, and highlight the benefits and challenges of using these agents in different domains.

2.1.7 Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies

This research explores using large language models (LLMs) to simulate human subjects in experiments, potentially replacing traditional human subject studies in some cases. The paper proposes a new kind of Turing test, where the goal is not to convince someone the LLM is human, but to see how well the LLM replicates results from real human subject studies.

(a) Typical few-shot prompt for classification:

Classify each sentence based on whether it is a garden path sentence or a normal sentence. A garden path sentence is a grammatically correct sentence whose likely reading appears to be ungrammatical.

Sentence: The old man the boat.
Classification: garden path

Sentence: The cat chased the mouse that was in the house.
Classification: normal

Sentence: While the student read the notes that were long and boring blew off the desk.
Classification: ________

(b) TE prompt for simulating a named individual:

Ms. Olson was asked to indicate whether the following sentence was grammatical or ungrammatical.

Sentence: While the student read the notes that were long and boring blew off the desk.

Answer: Ms. Olson indicated that the sentence was ________

The authors define a Turing Experiment (TE) specifically for LLMs. Instead of trying to trick a person into thinking the LLM is human, this TE measures how well the LLM replicates findings from human subject studies in psychology, economics, and other fields. The research shows that it is very effective for increasing the quality of answers generated by the LLM.

2.1.8 The Recent Large Language Models in NLP

This paper presents a comprehensive analysis of several prominent large language models (LLMs): BERT, ELMo, GPT-3, and LLaMA. We evaluate their performance across various aspects crucial to NLP tasks, offering a comparative overview of their strengths and weaknesses.

The key areas of comparison include architectural advancements, datasets, training techniques, evaluation methods, challenges, and applications. By providing a comparative analysis across these key areas, this paper offers valuable insights into the current landscape of large language models. It equips researchers and developers with a deeper understanding of the strengths and weaknesses of each LLM, allowing for informed decisions when selecting a model for specific NLP tasks.

2.1.9 LLaMA: Open and Efficient Foundation Language Models

This research paper introduces LLaMA, a new family of open and efficient foundation language models that can be used for a wide range of natural language processing (NLP) tasks. LLaMA models are designed to be highly performant, efficient, and adaptable, making them suitable for deployment in a variety of applications, from chatbots and virtual assistants to language translation and text summarization. The authors demonstrate that LLaMA models can achieve state-of-the-art results on a range of NLP benchmarks, while also being more computationally efficient and requiring fewer parameters than other language models.

From a technical perspective, LLaMA models are based on a novel architecture that combines the strengths of transformer-based models with the efficiency of recurrent neural networks (RNNs). The models use a hierarchical structure to process input sequences, allowing them to capture local and global language dependencies. The authors also introduce a range of techniques to improve the efficiency and adaptability of the models, including a novel parameter-sharing scheme and a method for dynamically adjusting the model's capacity based on the input sequence. The result is a family of language models that can be easily fine-tuned for specific tasks and deployed in a variety of applications, making them a valuable resource for the NLP community.

2.1.10 Model-Driven Prompt Engineering

Large language models (LLMs) have emerged as powerful tools capable of synthesizing complex content – text, code, and images – based on natural language prompts. However, the quality of the output hinges on the efficacy of these prompts, necessitating a focus on prompt engineering.

This paper proposes a novel approach for prompt engineering called model-driven prompt engineering. This DSL empowers researchers and developers to define prompts that can be readily adapted to various LLM systems while ensuring high-quality outputs. Additionally, the DSL offers version control mechanisms for efficient prompt management and the capability to chain prompts together for more intricate tasks.

Additional citations

- Aher, Gati et al. “Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies.” International Conference on Machine Learning (2022).
- R. Clarisó and J. Cabot, "Model-Driven Prompt Engineering," 2023 ACM/IEEE 26th International Conference on Model Driven Engineering Languages and Systems (MODELS), Västerås, Sweden, 2023, pp. 47-54, doi: 10.1109/MODELS58315.2023.00020. keywords: {Visualization;Source coding;Natural languages;Syntactics;Model driven engineering;DSL;Artificial intelligence;prompt engineering;model-driven engineering;domain-specific language;generative AI;large language models},
- N. T. K. Le, N. Hadiprodjo, H. El-Alfy, A. Kerimzhanov and A. Teshebaev, "The Recent Large Language Models in NLP," 2023 22nd International Symposium on Communications and Information Technologies (ISCIT), Sydney, Australia, 2023, pp. 1-6, doi: 10.1109/ISCIT57293.2023.10376050. keywords: {Surveys;Performance evaluation;Training data;Market research;Chatbots;Internet;Information and communication

technology;NLP;LLM(s);pre-trained;data;model;architecture;performance;state-of-art;BERT;ELMo;GPT-3;LLaMA;application.},

-