Do LLMs understand what they "know": Analyzing if LLMs can solve Simple Analogies devoid of the relations

Project-I (CS47005) report submitted to

Indian Institute of Technology Kharagpur

in partial fulfilment for the award of the degree of

Bachelor of Technology

in

Computer Science & Engineering

by

Gaurav Malakar

(20CS10029)

Under the supervision of

Professor Pawan Goyal

Department of Computer Science & Engineering

Indian Institute of Technology Kharagpur

Autumn Semester, 2023-24

November 04, 2023

DECLARATION

I certify that

(a) The work contained in this report has been done by me under the guidance of my supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Date: November 04, 2023 (Gaurav Malakar)

Place: Kharagpur (20CS10029)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

KHARAGPUR - 721302, INDIA

CERTIFICATE

This is to certify that the project report entitled "**Do LLMs understand what they "know": Analyzing if LLMs can solve Simple Analogies devoid of the relations**" submitted by **Gaurav Malakar** (Roll No. 20CS10029) to Indian Institute of Technology Kharagpur towards partial fulfilment of requirements for the award of degree of Bachelor of Technology in Computer Science & Engineering is a record of bona fide work carried out by him under my supervision and guidance during Autumn Semester, 2023-24.

Date: November 04, 2023

Place: Kharagpur

Professor Pawan Goyal
Department of Computer Science & Engineering
Indian Institute of Technology Kharagpur
Kharagpur - 721302, India

Abstract

Name of the student: **Gaurav Malakar** Roll No: **20CS10029**

Degree for which submitted: **Bachelor of Technology**

Department: **Department of Computer Science & Engineering**

Thesis title: **Do LLMs understand what they "know": Analyzing if LLMs can solve Simple Analogies devoid of the relations**

Thesis supervisor: **Professor Pawan Goyal**

Month and year of thesis submission: **November 04, 2023**

In the realm of Large Language Models (LLMs), analogical reasoning is key to assessing their cognitive strength. This study delves into the analogical reasoning abilities of LLMs, particularly focusing on their capacity to reason with the information they inherently "know". We posed queries in the format A:B::C:X, where A, B, and C are names of renowned personalities, and the LLMs were tasked with deducing X based on the relationship between A and B. The uniqueness of our approach lies in ensuring that the LLMs are aware of the relationships (e.g., father-son) before presenting the analogy in a new context. Our experiments, were conducted on a self-curated dataset comprising of Indian rulers and renowned philosophers. Despite the extensive capabilities attributed to LLMs, their performance in this specific analogical reasoning task was found to be sub-optimal. Notably, GPT-3.5 achieved an accuracy of approximately 20% and other LLMs struggled further, with accuracies falling below 10%.These results highlight the problems LLMs have in some reasoning tasks and suggests we should look deeper into how they reason things out.

Acknowledgements

I would like to extend my appreciation to both my supervisors for their roles in this project. Prof. Pawan Goyal, a respected professor of our college, for his overarching presence and the academic foundation he provided. A special mention goes to Mr. Abhilash Nandy, a doctoral student under Prof. Pawan Goyal, who has been my hands-on mentor for this endeavor. His constructive feedback, dedication, and unwavering support have been instrumental in guiding me through the challenges and intricacies of the research. His mentorship has enriched my academic journey, offering a deeper understanding and perspective on the subject. I am truly grateful for the insights and guidance both have provided.

Contents

List of Figures

List of Tables

Abbreviations

LLM	**L**arge **L**anguage **M**odel
GPT	**G**enerative **P**re-trained **T**ransformers
LSTM	**L**ong **S**hort -**T**erm **M**emory
API	**A**pplication **P**rogramming **I**nterface
NLP	**N**atural **L**anguage **P**rocessing

Chapter 1

Introduction

In the rapidly evolving domain of artificial intelligence, Large Language Models (LLMs) have gained significant attention. These models, backed by extensive data, have showcased their potential in various tasks. However, a pressing question emerges: Do these models genuinely grasp the content they've been trained on, or are they merely echoing patterns?

When humans encounter unfamiliar situations, they often draw parallels with known scenarios to derive solutions. This ability to reason analogically is a distinguishing feature of human cognition. Can LLMs, with their vast reservoirs of data, emulate this essential human trait?

This study delves into the analogical reasoning capabilities of LLMs. While there's existing research on this topic, our approach brings a fresh perspective. We're keen to understand if LLMs can effectively use their inherent knowledge, especially when presented in a new light.

Our methodology employs queries structured as A:B::C:X, challenging LLMs to identify and extrapolate relationships. To lend depth to our exploration, we've curated a dataset that pairs renowned Indian rulers and their sons with popular

philosophers and their descendants. This unique dataset forms the backbone of our experiments, aiming to uncover the intricacies of LLMs' reasoning abilities.

This thesis provides a detailed account of our approach, the experimental design, and the insights derived. Through this exploration, we aim to shed light on the strengths and limitations of LLMs in specific reasoning tasks.

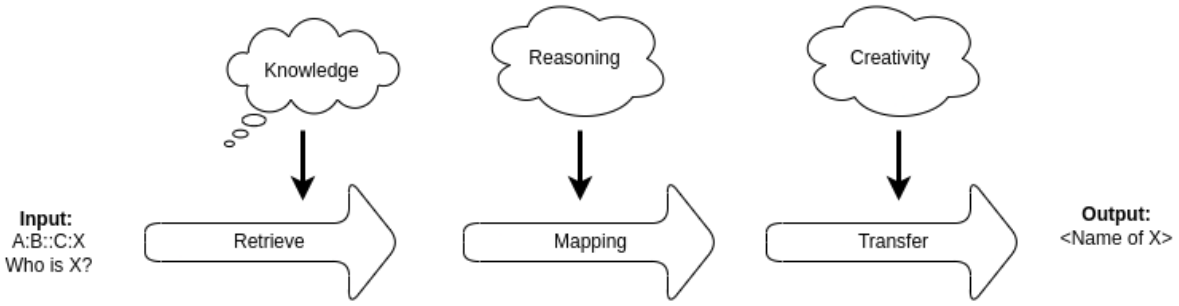

FIGURE 1.1: Analogical reasoning process

Chapter 2

Literature Review

The exploration of Large Language Models (LLMs) and their capabilities has been a focal point of research in recent years. Several studies have delved into the intricacies of these models, aiming to understand their reasoning abilities, knowledge representation, and potential applications. This section sheds light on some of the seminal works that have laid the foundation for our current research.

An Explanation of In-context Learning as Implicit Bayesian Inference: This study (Xie et al., 2021) delves into the unique ability of LLMs, notably GPT-3, to perform in-context learning. The research highlights the model's capability to learn downstream tasks by merely conditioning on prompts with input-output examples, without any explicit pretraining. The paper emphasizes the role of long-range coherence in pretraining documents and how it facilitates in-context learning. The study also introduces a synthetic dataset, GINC, where both Transformers and LSTMs exhibit in-context learning, showcasing the real-world implications of their findings.

Emergent Analogical Reasoning in Large Language Models: Standing at the forefront of our discussion is the work titled "*Emergent Analogical Reasoning in Large Language Models*" (Webb et al., 2023). This research offers a comparative

analysis between human reasoning and a variant of GPT-3, specifically focusing on analogical tasks. The findings suggest that GPT-3, and preliminary tests on GPT-4, exhibit a robust capacity for abstract pattern induction, often matching or even surpassing human capabilities. The paper underscores the emergent ability of LLMs to find zero-shot solutions to a broad spectrum of analogy problems.

Large Language Models As Analogical Reasoners: Another significant contribution to the field is the study titled "*LARGE LANGUAGE MODELS AS ANALOGICAL REASONERS* "(Yasunaga et al., 2023). The paper introduces the concept of analogical prompting, a novel prompting approach designed to guide the reasoning process of LLMs. Drawing inspiration from human analogical reasoning, this method prompts LLMs to self-generate relevant exemplars or knowledge in context. The research showcases the method's adaptability and generality across various reasoning tasks.

Building upon these foundational works, our research endeavors to further probe into the analogical reasoning capabilities of LLMs, especially in the context of understanding the information they have been trained on.

Chapter 3

Dataset

Our research is built on a carefully curated dataset, aimed at understanding the analogical reasoning of Large Language Models (LLMs). This section breaks down how the dataset was created, its sources, and the special steps we took to make sure the LLMs really "know" the information.

3.1 Data Sources

The dataset was primarily sourced from:

Wikipedia: The "*List of Indian monarchs*" page on Wikipedia was instrumental, providing a comprehensive list of Indian rulers and their descendants. Link[1]

Totally History: Data about renowned philosophers, predominantly of European descent, and their children was extracted from the "*Famous Philosophers*" section of the Totally History website. Link[2]

To further enrich the dataset, ChatGPT was also prompted, ensuring a more diverse and comprehensive data pool.

[1]Wikipedia contributors, "List of Indian monarchs - Wikipedia, The Free Encyclopedia", 2023.
[2]Totally History, "Most Famous Philosophers - List of Famous Philosophers in History", 2020.

3.2 Dataset Structure and Creation

From the outset, two distinct datasets were formulated:

Indian Rulers Dataset: Focusing on prominent Indian rulers, the data was structured with two columns: 'Name', representing the ruler, and 'Children', detailing their descendants.

Philosophers Dataset: Centered on renowned philosophers, especially those of European lineage, it too had two columns: 'Name' for the philosopher and 'Children' for their lineage.

The reason behind curating separate datasets for rulers and philosophers was to probe analogies involving individuals from different eras and walks of life.

3.3 Data Preprocessing

The preprocessing phase was pivotal, not just for data accuracy but to ensure that the LLMs genuinely "know" the information. The methodology was as follows:

- Each entry in the 'Name' column was used to query the LLM with the question, "Who is the son of <name>?".
- The LLM's response was cross-referenced with the corresponding data in the 'Children' column.
- If a mismatch was detected between the LLM's answer and the 'Children' column data, that specific row was discarded.

This preprocessing was not merely a data-cleaning exercise. It was a strategic step to ensure that the LLMs were familiar with the information, thereby setting the stage for a genuine test of their analogical reasoning capabilities.

Chapter 4

Methodology

In this section, we detail the systematic approach adopted to investigate the analogical reasoning capabilities of Large Language Models (LLMs). The methodology is structured to provide clarity on the procedures and rationale behind each step.

4.1 Query Formation

Formation of A:B::C:X Queries: The primary structure used for querying the LLMs was the A:B::C:X format. For instance, a sample query would be "Akbar:Jahangir::Socrates:X?" where the expected value of X could be Lamprocles, Sophroniscus, or Menexenus. This format was chosen as it is a classic representation of analogical reasoning, where the relationship between A and B should mirror the relationship between C and X.

Rationale Behind the Format: Analogies, by their nature, require understanding relationships between entities. The A:B::C:X structure forces the model to discern the relationship between A and B and then apply that understanding to determine X, given C. This format is not just a test of the model's knowledge but also its ability to reason and draw parallels.

4.2 LLM Selection

GPT-3.5 was chosen for this study primarily because of its popularity and the user-friendly API it offers for Python. Additionally, we incorporated three LLMs sourced from the Hugging Face model repository, a renowned hub for state-of-the-art NLP models and resources. These models, namely Llama-2-7b-chat-hf, Llama-2-13b-chat-hf, and vicuna-13b-v1.3, were selected for their open-source nature, cost-free accessibility, and remarkable response times.

4.3 Evaluation Metrics

Metrics Used: The primary metric used to evaluate the performance of LLMs was accuracy. Accuracy was calculated as the percentage of correctly answered analogical queries out of the total queries presented to the LLM.

Rationale Behind Metric Selection: Accuracy was deemed the best metric for this study because the task at hand doesn't involve classification but requires discrete answers that can either be correct or wrong. As such, metrics like the F1 score aren't applicable in this context. For the purpose of evaluation, replies were considered correct if the first name in the response was accurate, given that the last name of someone's child is typically the same as their own, making it trivial to predict.

By adhering to this structured methodology, the research aims to provide a clear and comprehensive examination of the analogical reasoning capabilities of LLMs, especially when presented with entities from distinct backgrounds and eras.

Chapter 5

Experimental Results

In this chapter, we present the results obtained from our experiments on the analogical reasoning capabilities of Large Language Models (LLMs). The experiments were designed based on the A:B::C:X query format, as detailed in the methodology section.

5.1 GPT-3.5

GPT-3.5 was the primary focus of our experiments and was tested most extensively. The queries were designed in two directions:

<Ruler>:<Ruler's son>::<Philosopher>:X

<Philosopher>:<Philosopher's son>::<Ruler>:X

For each direction, two types of prompting strategies were employed:

1. Providing the A:B::C:X structure and simply asking, *"Who is X?"*

2. Providing the A:B::C:X structure and instructing the LLM to deduce X analogically.

Both prompting strategies yielded similar results. On average, GPT-3.5 achieved an accuracy of **15%**, which, while being the highest among the tested LLMs, is still indicative of a poor performance in analogical reasoning tasks.

5.2 LLama-2 Models

Three open-source LLMs, specifically LLama-2 models, were tested less extensively than GPT-3.5. Their Hugging Face names are:

- `Llama-2-7b-chat-hf`,
- `Llama-2-13b-chat-hf`, and
- `vicuna-13b-v1.3`.

Their tests were limited to the **<Ruler>:<Ruler's son>::<Philosopher>:X** queries and employed only the first type of prompting strategy. The results were as follows:

Table 5.1: Accuracies of Llama-2 models

Model Name	**Accuracy**
Llama-2-7b-chat-hf	10%
Llama-2-13b-chat-hf	9%
vicuna-13b-v1.3	6%

All four LLMs exhibited subpar performance in the analogical reasoning tasks, with the LLama-2 models performing especially poorly. A significant challenge encountered with these LLama-2 models was their limited prior knowledge. During the preprocessing phase, a substantial portion of the data was filtered out because these models couldn't correctly answer the question, *"Who is the son of <name>?"*.

Chapter 6

Discussion

The analogical reasoning capabilities of Large Language Models (LLMs) have been a focal point of our research. The primary objective was to discern whether LLMs genuinely understand the content they've been trained on or if they merely replicate patterns. The A:B::C:X query format was instrumental in this exploration, challenging the LLMs to identify and extrapolate relationships.

6.1 GPT-3.5 Analysis

Our experiments with GPT-3.5 revealed intriguing insights. The model's performance in predicting relationships, between rulers and their sons and philosophers and their descendants, was not always accurate. A deeper dive into the reasons behind these inaccuracies led to several observations:

1. **Frequency Analysis**: A simple frequency analysis was conducted for each philosophers 6.1. The frequency of occurrence of the correct/expected answer on the Wikipedia page of a philosopher was compared against the frequency of the answer provided by GPT-3.5. As anticipated, GPT-3.5's answer occurred

more frequently than the correct answer for philosophers whose sons were inaccurately predicted

2. **Qualitative Analysis**: When GPT-3.5 was prompted to justify its answers, it was evident that the model sometimes deduced unexpected relationships between terms. For instance, in the analogy Akbar:Jahangir::Socrates:X, while the expected relationship was father-son, GPT-3.5 predicted Jahangir as Akbar's successor and, therefore, suggested Plato as Socrates' successor. This indicates that while the model wasn't actually wrong, it deviates from the expected father-son relationship.

3. **Data Preprocessing Limitations**: There were instances where GPT-3.5 claimed not to know the relationship between terms, even though the data preprocessing step ensured that the model was familiar with these relationships. This raises questions about the model's ability to recall and apply information effectively.

Drawing inspiration from another research document, it's evident that analogical reasoning is a complex cognitive process. While LLMs like GPT-3.5 have made significant strides in this domain, there are still gaps in their understanding and application of knowledge.

6.2 In-Context Learning and Its Impact on Analogical Reasoning

After analyzing GPT-3.5's performance, further understanding of the mechanisms behind analogical reasoning in LLMs was sought. A paper on in-context learning (Xie et al., 2021) was consulted, which highlighted the unique emergent behavior observed in large language models (LLMs) like GPT-3.5, where the model performs

tasks by conditioning on input-output examples without optimizing any parameters. This behavior, termed as in-context learning, is seen as the model's ability to "locate" latent concepts it has acquired from its pretraining data. Such a mechanism suggests that all components of a prompt, including inputs, outputs, and their mapping, can provide information for inferring the latent concept. This understanding of in-context learning suggests that it could be a major factor in how GPT-3.5 forms analogies. This ability to use context to shape answers might be a central component of its reasoning process. As the project progresses to the next phase, there is an intent to further explore the relationship between in-context learning and analogical reasoning. The plans for this exploration and its potential implications will be detailed in the next chapter.

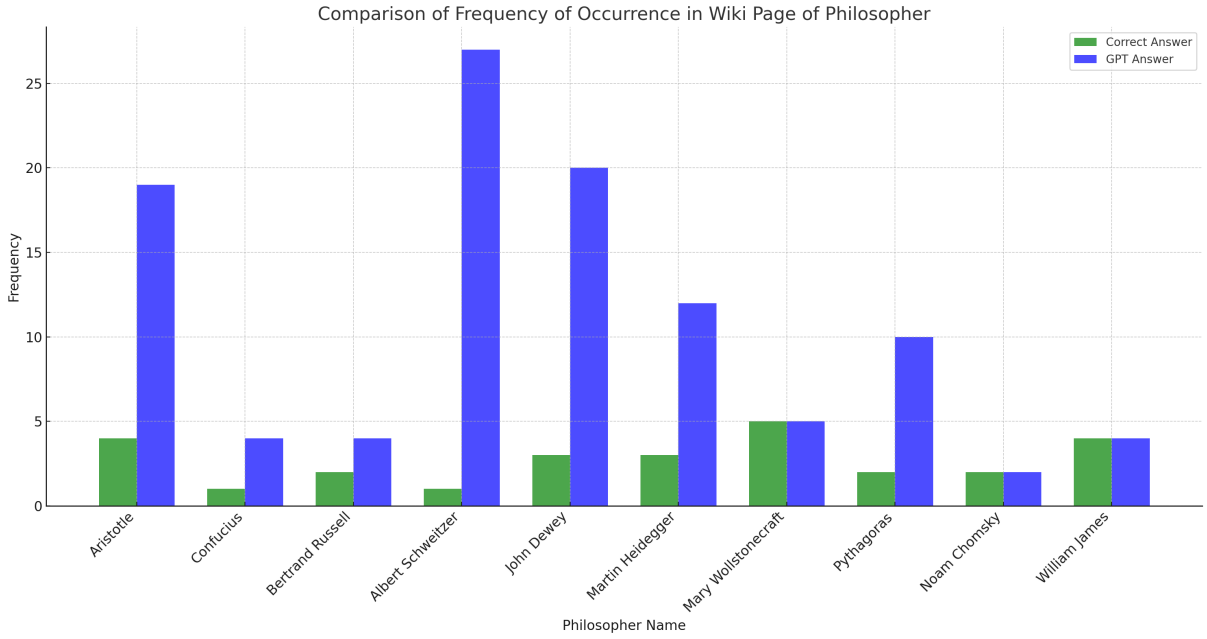

FIGURE 6.1: Frequency of Occurrence Comparison

Chapter 7

Conclusion

This thesis embarked on a journey to unravel the **analogical reasoning capabilities** of *Large Language Models (LLMs)*, with a particular focus on whether these models truly comprehend the information they have been trained on. Our exploration, centered around the **A:B::C:X** query format, has shed light on the complexities and limitations of LLMs in drawing analogies, especially when it comes to extrapolating relationships from known data.

Our findings with **GPT-3.5**, while revealing the model's potential, also highlighted its inconsistencies and the challenges it faces in accurately predicting relationships. The model's occasional deviation from expected relationships and its struggle to recall and apply known information underscore the need for a deeper understanding of how LLMs process and utilize their training.

As we look to the future, our research aims to delve further into the specific types of errors made by LLMs like GPT-3.5 and to uncover the underlying reasons for these inaccuracies. Inspired by recent advancements in the field, we plan to explore strategies that could enhance the analogical reasoning capabilities of LLMs.

The paper "*An Explanation of In-context Learning as Implicit Bayesian Inference*"(Xie et al., 2021) provides valuable insights into in-context learning, a phenomenon that could be pivotal in understanding how LLMs form analogies. Similarly, the work "*Learning From Mistakes Makes LLM Better Reasoner*"(Shengnan An, 2023) introduces the "*Learning from Mistakes*" (LeMa) method, which could be instrumental in refining LLMs' reasoning processes. Lastly, the paper "*LARGE LANGUAGE MODELS AS ANALOGICAL REASONERS*"(Yasunaga et al., 2023) presents the concept of "*analogical prompting*," a promising approach that could guide LLMs in generating more accurate analogies.

Incorporating these insights, our future work will focus on developing methodologies that not only address the current shortcomings of LLMs but also harness their latent potential. By doing so, we aspire to bring LLMs a step closer to understanding what they "know" and enhancing their ability to reason analogically with the information at their disposal.

Bibliography

Shengnan An, Zexiong Ma, Z. L. N. Z. J.-G. L. W. C. (2023). Learning from mistakes makes llm better reasoner. *arXiv:2310.20689*.

Webb, T., Holyoak, K. J., and Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. (2021). An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Yasunaga, M., Chen, X., Li, Y., Pasupat, P., Leskovec, J., Liang, P., Chi, E. H., and Zhou, D. (2023). Large language models as analogical reasoners. *arXiv preprint arXiv:2310.01714*.