DS 440 Final Report

Danny Dawson

Introduction

For our capstone project, my teammate and I evaluated the deductive, inductive, and relational reasoning ability of Large Language Models (LLM's). Relational reasoning refers to ability of LLM's to deduce familial relationships given from the context of a family tree. Two family trees were created for this experimentaiton, the first being a standard family tree and the second being a half-family tree. This report will focus on the experimentation and evaluation of the relational reasoning abilities of LLM's.

Related Work

LLM Relational Reasoning

There is a vast amount of research that investigates the reasoning capabilities of modern LLM's. While these LLM's have become more popular and sophisticated in the past 5 years, that does not mean that they have great reasoning capabilities. In the paper, "LLMs for Relational Reasoning. How far are we?", a group of researchers from Singapore assess the reasoning benchmarks of LLM's. They claim that the previous textual and numerical reasoning benchmarks found by other works are shallow and simple, and LLM's achieving positive results in those reports does not necessarily mean that LLM's possess strong reasoning skills (Li et al., 2024). In their research, they find that LLM's are poor at solving sequential decision-making problems that use common sense planning (Li et al., 2024).

While experimenting with the GPT-4 model, these researchers found that it performed very well with simple tasks, it struggled with more complex tasks. Their experimentation focused on relationships in a family tree, testing the relationships between different members of a family tree. A function that GPT-4 had great success with was the "HasFather" function since it is such a simple task. However, the GPT-4 model only had a 49% accuracy with the "IsUncle" function, since it is more of a complex task. Among the LLM's they evaluated, they found that GPT-4 and GPT-4 Turbo showed the strongest relational reasoning capability, but their accuracy faltered when presented with problems that required more complex problem-solving logic (Li et al., 2024).

Advanced Reasoning Benchmark for LLM's

In 2023, a new benchmark dataset called Advanced Reasoning Benchmark (ARB) was introduced to evaluate LLM reasoning abilities in mathematics, science, and law (Sawada et al.,

2023). The dataset was built using 3 types of questions, those being multiple choice, short answer, and open response. The multiple-choice questions make up a large proportion of the total number of questions due to the ease of grading them (Sawada et al., 2023). The short answer and open response questions helped improve existing benchmarks, as a lot of previous research only dealt with multiple choice questions (Sawada et al., 2023). These researchers evaluate ChatGPT, GPT 3.5, GPT-4, and Claude with their questions. They found that these models did particularly well with law and MCAT multiple choice questions. However, these LLM's struggled with questions that contained a numerical final answer, especially math and physics questions. They found that GPT-4 was the only reliable model that was capable of simplifying complex expressions, but GPT-4 also struggled to perform arithmetic in complex problems (Sawada et al., 2023). Symbolic answers were also a problem for these models, as GPT-4 only had an 18% accuracy for questions with a "math symbolic" answer (Sawada et al., 2023). For the mistakes on the math and physics problems, a logical error or a hallucination were the most common errors from the LLM's. When testing these LLM's ourselves, we will have to carefully consider the format of our question and the desired answer. Investigating the accuracy between questions formats and answer types is key when evaluating the reasoning abilities of LLM's.

Physical Reasoning Capabilities of LLM's

While most LLM reasoning research focuses on verbal reasoning, there is also studies that focus on the physical reasoning abilities of LLM's. Physical reasoning involves the process of comprehending and predicting the actions of physical systems based on observations and fundamental principles (Wang et al., 2023). Researchers from the University of Washington developed the NEWTON benchmark to evaluate the physical reasoning skills of LLM's. Mainstream 3D object datasets like Objaverse and Amazon Berkely Objects were used in the analysis and experimentation. The NEWTON benchmark consists of 160,000 questions distributed over 3 tasks of foundational attribute understanding, explicit application, and implicit scenario-based analysis (Wang et al., 2023). There were 8 categories of tasks for the foundational attribute understandings tasks, such as the softness, sharpness, stiffness, and malleability of objects. The explicit application task involved two types of questions, those being true/false or multiple-choice questions. The implicit scenario-based analysis tested 9 different scenarios, which utilized certain combinations of the categories evaluated in task 1. Using the NEWTON framework, GPT-4 and GPT-Turbo seemed to be the most accurate models, with GPT-4 having the highest average accuracy for each of the 3 tasks (Wang et al., 2023). The average accuracies for the LLM's tested were typically around 40-50%, so it seems that there is some potential that LLM's can physically reason at a respectable accuracy.

Based off of the existing research, me and my teammate felt that GPT-4 was the state-of-the-art LLM currently, and we decided to perform all of our testing on GPT-4 using Microsoft Copilot via Bing.

Methods

Two family trees were made to test the relational reasoning ability of LLM's. The first family tree was called a "Standard Family Tree", where there isn't any abnormal relationship in the tree. The standard family tree is pictured below:
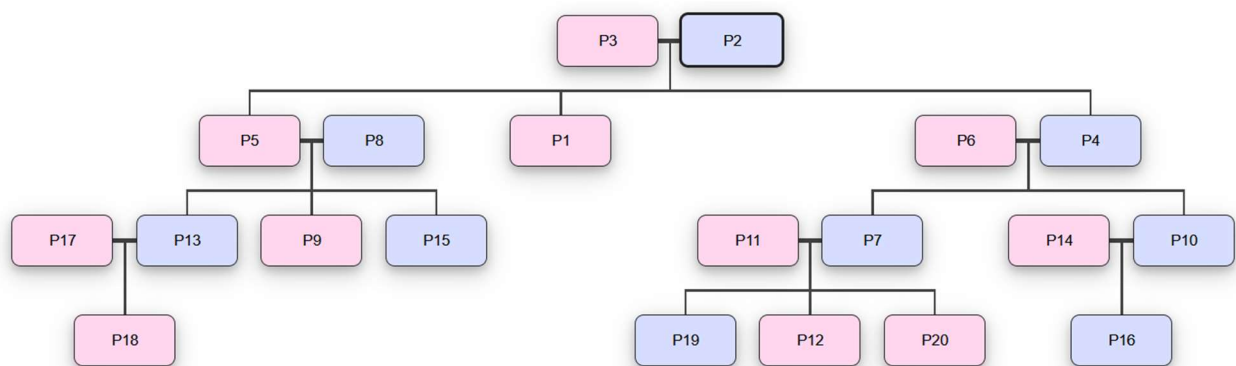

Figure 1: Standard Family Tree

The tree starts with a typical marriage between P2 and P3, with children P4, P5 and P1 being their offspring. The rest of the family tree then expands from their relationships. The tree contains 20 members in total.

Questions were prompted at levels n = 2 to n = 20, where n denotes how many members of the tree were included in the prompt. 25 questions were asked at each level, and random number generation was used to determine which members of the tree were included in the prompt. The context of the family tree were provided in textual form, and a question was asked in the format; "What is the relationship between PX and PY?' where X and Y were the two random numbers generated earlier. The following was a prompt used during the standard family tree testing:

> N = 13, Distance = 4
>
> *Context:* P2 is the father of P1, who is female. P2 and P3 are married, and P3 is the mother of P1. P4 is the son of P2 and P3. P5 is the daughter of P2 and P3. P4 is married to P6, who is female. P7 is the son of P4 and P6. P5 is married to P8, who is male. P9 is the daughter of P8 and P5. P10 is the son of P4 and P6. P7 is married to P11, who is female. P12 is the daughter of P11 and P7. P13 is the son of P8 and P5.
>
> What is the relationship between P12 and P5?

GPT-4 Answer: **P12** is the granddaughter of **P5**. Here's why:

- **P12** is the daughter of **P7** and **P11**.
- **P7** is the son of **P4** and **P6**.
- **P4** is the son of **P2** and **P3**, and **P5** is the daughter of **P2** and **P3**.

Therefore, **P12** is the child of **P4**'s child, making her the granddaughter of **P5**. I hope this helps! If you have any other questions, feel free to ask.

In this example, P12 was incorrectly described as the granddaughter of P5, as P12 is actually the great-niece of P5. I established two categories of errors for incorrect GPT-4 answers. The first category was "Completely Wrong", where there was some sort of fatal logic error or crucial mistake that led to an answer far from the truth. The other error classification was "Right, but not precise", where the GPT-4 response was close to the truth, but there were some minor errors in its answer. In this case, the GPT-4 response would be classified as "Completely Wrong".

The distance between the family tree members was calculated using Dijkstra's algorithm, and a separate was made for these calculations. In this graph, parent-child relationships and spousal relationships created an edge between two people, and each edge had a weight of 1.

The second family tree created was labeled as a "Half-Family Tree". This tree is pictured below:
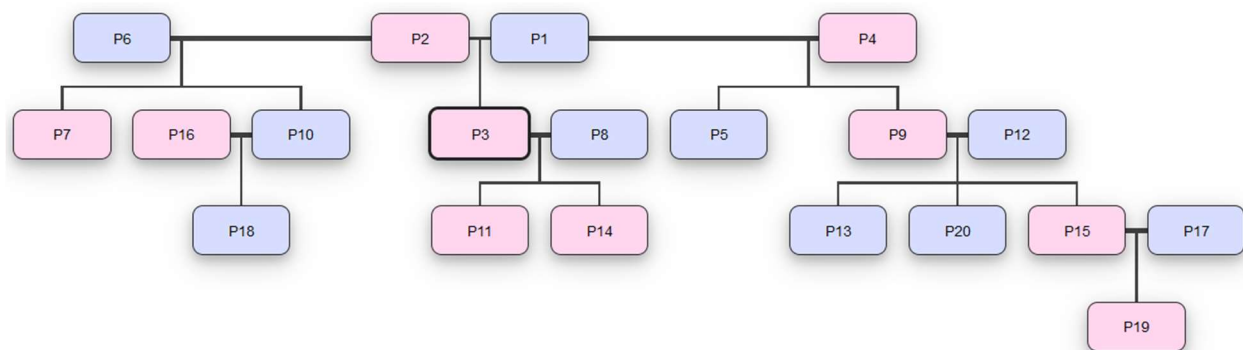

Figure 2: Half-Family Tree

In this family tree, P1 and P2 divorce after the birth of their daughter, P3. They both remarry new spouses and have children with them as well. The divorce and remarriage create new half-family and stepfamily relationships that GPT-4 can be prompted to decipher. This tree also contains 20 members in total. The testing process was similar to the standard family tree testing, but with one modification. This testing only spanned from n = 10 to n = 20, as standard family tree testing with 10 people or less produced nearly no errors. The rest of the testing process was the same, and the following is an example prompt from this test:

N = 11, Distance = 3

Context: P1 (male) and P2 (female) used to be married, and they have a daughter, P3. P2 later married P6, and they had a son, P10, and a daughter, P7. P1 later married P4, and they had a son, P5, and a daughter, P9. P3 is married to P8. P11 is the daughter of P3 and P8.

What is the relationship between P5 and P11?

GPT-4 Answer: The relationship between **P5** and **P11** is that of **half-siblings**. They share one biological parent, **P3**. While **P5** is the son of **P1** and **P4**, **P11** is the daughter of **P3** and **P8**. Since **P3** is the common parent, **P5** and **P11** are half-siblings.

In this example, the answer would be classified as "Completely Wrong", as P5 is the half-uncle of P11. The distance calculation and error classification processes were the same as the standard family tree testing.

Evaluation

Results

The following table shows the results for the testing of the standard family tree

Distance	Total Questions Asked	Accuracy
1	104	98.1%
2	88	100%
3	49	81.6%
4	44	61.4%
5	32	34.4%
6	32	15.6%

Figure 3: Standard Family Tree Test Results

The results show that as the distance between the family members increases, the accuracy of GPT-4 decreases. An interesting observation from this table was the 98.1% accuracy for the lowest distance level of 1. GPT-4 got 2 of the 104 questions wrong, and when I went back and tested those questions again, GPT-4 didn't make the same mistake again and got them both correct multiple times. The following statistics are some metric results from this testing:

Overall Accuracy = 78.2%

Average Error Distance = 4.79

69.7% of errors were "Completely Wrong"

30.3% of errors were "Right, but not precise"

The following table shows the results for the testing of the half-family tree.

Distance	Total Questions Asked	Accuracy
1	58	91.4%
2	84	66.7%
3	88	30.9%
4	37	37.0%
5	31	6.45%
6	32	12.5%

Figure 4: Half-Family Tree Test Results

We can see from the table that the accuracy of GPT-4 decreases drastically as the distance between family members increases. This decline happens at a quicker rate than the standard family tree, as a relatively small distance of 2 has an accuracy of 66.7%, compared to the 100% accuracy at the same distance in the standard family tree test. The following statistics are some metric results from this testing:

Overall Accuracy = 46.1%

Average Error Distance = 3.74

74.2% of errors were "Completely Wrong"

25.8% of errors were "Right, but not precise"

Discussion

Takeaways

Looking at the standard family tree testing, the results are somewhat expected. When testing this tree, questions that included 10 people or less in the prompt had a nearly perfect accuracy from GPT-4. Just by looking at the overall accuracies between each test, we can see that there is a drastic decline in performance when half-family and stepfamily relationships are included in a family tree. An interesting observation from the half-family tree errors is the use of the stepsibling/half-sibling answer as a sort of default answer. There were many cases where GPT-4 would be prompted to deduce a relationships between two people in different generations of the half-family tree, and it would return a stepsibling or half-siblings relationship. The following is an example of this error:

> *Context:* P1 (male) and P2 (female) used to be married, and they have a daughter, P3. P2 later married P6, and they had a son, P10, and a daughter, P7. … P18 is the son of P10 and P16.

> What is the relationship between P6 and P3?

GPT-4 Answer: P3 and P6 are **step-siblings.** P3 is the daughter of P1 and P2, while P6 is the son of P2 and her later spouse, P6. Therefore, P3 is the **step-sister** of P6

In this example, we see that P3 and P6 are labeled stepsiblings, when P6 is the stepfather of P3. This leads to a larger issue within GPT-4 relational reasoning, which is the definition of a stepparent. In the prompt, P1 and P2 are formerly married, and GPT-4 seems to think this relationship is still active. Members of the top line of the half-family tree are often labeled as stepparents when they really aren't. This error below shows an example of this:

Context: P1 (male) and P2 (female) used to be married, and they have a daughter, P3. P2 later married P6, and they had a son, P10, and a daughter, P7. P1 later married P4, and they had a son, P5, and a daughter, P9. P3 is married to P8. P11 is the daughter of P3 and P8. P12 and P9 are married. P13 is the son of P12 and P9. P14 is the daughter of P8 and P3. P15 is the daughter of P12 and P9. P16 is married to P10.

What is the relationship between P12 and P10?
GPT-4 Answer: P12 is married to P9, who is the daughter of P1 and P4. P10 is the son of P2 and P6. P1 and P2 used to be married. Therefore, **P12 is the step-brother-in-law of P10**. Here's how:

- P12 is the husband of P9.
- P9 is the half-sister of P3 (as both have P1 as one of their parents).
- P3 is the half-sister of P10 (as both have P2 as one of their parents).
- Therefore, P9 is the half-sister-in-law of P10.
- Hence, P12 (being the husband of P9) is the step-brother-in-law of P10.

When debugging this example, I asked which stepparent do P12 and P10 share, as they don't have one. GPT-4 said that P2 is the shared stepparent of P12 and P10, when P2 is the biological parent of P10 and has no relationship with P12. However, since P2 is the ex-spouse of P1, P2 is treated as a stepparent of P12, even though a stepparent has to be actively married to a biological parent to be considered a stepparent. This logical mistake causes many errors within the half-family testing.

Another interesting observation from this testing is the inability for GPT-4 to identify cross-siblings. Cross-siblings are defined as people who share a half sibling. For example, P5 and P7 are cross siblings in the half-family tree, as they share P3 as a half sibling. When prompting GPT-4 to deduce a cross sibling relationship, the GPT-4 response was typically half-siblings or stepsiblings. GPT-4 knows what a cross sibling is, as it will return the definition of one if asked. Without including what cross sibling is in the prompt however, GPT-4 cannot identify cross-sibling relationships without additional context.

Future Work

There are multiple improvements that can be made to this project. Testing more questions at the longer distances is something I would have liked to do but couldn't due to time constraints. The

randomness of the prompt generation caused the distribution of questions across distances to be somewhat skewed for both trees, as there were many more relationships with distances of 1, 2, and 3, compared to the relationships with longer distances. Another improvement that could be made is the further classification of errors. I only created 2 different error classification categories, and a further investigation of how errors are being made would lead to more insight on how GPT-4 is making mistakes when relationally reasoning. Adding some automation into this project would also make it more efficient. Having the ability to create random family trees and then test them could lead to the investigation of complex familial relationships that weren't included in this experiment.

Conclusion

After completing this project there are 3 main conclusions that we can come to

1. GPT-4 cannot reliably identify half-family and stepfamily relationships
2. As the size of the family tree increases, the accuracy of GPT-4 decreases
3. GPT-4 is fairly reliable working with standard family trees with 10 people or less.

Relational reasoning doesn't seem to be an issue for GPT-4 with small standard family trees. However, as the family tree size increases and the relationships become more complex, GPT-4 cannot reliably identify complex familial relationships.

Contribution

For this project, my contribution was the experimentation and evaluation of the relational reasoning abilities of GPT-4. In the GitHub link in the References section, I show the code I used to calculate the distances between family tree members.

References

Li, Z., Cao, Y., Xu, X., Jiang, J., Liu, X., Teo, Y. S., ... & Liu, Y. (2024). *LLMs for relational reasoning: How far are we?* Retrieved from https://arxiv.org/abs/2401.09042

Sawada, T., Paleka, D., Havrilla, A., Tadepalli, P., Vidas, P., Kranias, A., ... & Komatsuzaki, A. (2023). *Arb: Advanced reasoning benchmark for large language models.* Retrieved from https://arxiv.org/abs/2307.13692

Wang, Y. R., Duan, J., Fox, D., & Srinivasa, S. (2023). NEWTON: *Are Large Language Models Capable of Physical Reasoning?*. Retrieved from https://arxiv.org/abs/2310.07018

GitHub link: https://github.com/danny-dawson/Capstone-Project-Code

Code for Dijkstra's Algorithm: https://www.geeksforgeeks.org/python-program-for-dijkstras-shortest-path-algorithm-greedy-algo-7/