LLMs' Reading Comprehension - A Follow Up Experiment

Tsuri Farhana and Omri Cohen

August 24, 2024

Table of Contents

1. Introduction & Motivation

2. Method

3. Results

Research Question of the Previous Paper

In the paper "LLMs' Reading Comprehension Is Affected by Parametric Knowledge and Struggles with Hypothetical Statements" - https://arxiv.org/abs/2404.06283
The authors mainly asked the following question: **What are the capabilities of LLMs to understand language models?**
in other words 'natural language understanding'

Research Question of the Previous Paper

In the paper ”LLMs’ Reading Comprehension Is Affected by Parametric Knowledge and Struggles with Hypothetical Statements” -
https://arxiv.org/abs/2404.06283
The authors mainly asked the following question: **What are the capabilities of LLMs to understand language models?**
in other words 'natural language understanding'

Empirically measure “text understanding” through the task of reading comprehension: the ability to correctly answer questions based on the given text

Research Question of the Previous Paper

https://arxiv.org/abs/2404.06283

In this paper they stated two main results:

- LMMs is affected by parametric knowledge:
  Using data that in any manner is effected by parametric knowledge (supporting or contradicting) effects the LLM answer.
- LLMs struggles with hypothetical statements.

Problems With The Previous Paper

- Except of the most basic context, all the RC did not had the data to answer the question, and they expected the LLM to answer "None"

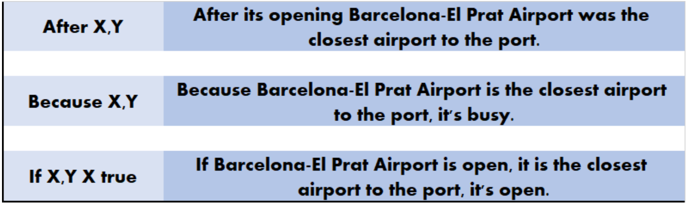
Problems With The Previous Paper

- Except of the most basic context, all the RC did not had the data to answer the question, and they expected the LLM to answer "None"
- They ask the LLM each question only once, so they miss an very important property - Does the LLM generate a consistent answer or does it produce a distribution of random answers?

Dealing with the Problems in the Previous Paper

Except of the most basic context, all the RC did not had the data to answer the question, and they expected the LLM to answer "None"

- In this project we recreated the experiment they did but with affirmative Linguistic structure.



After X,Y	After its opening Barcelona-El Prat Airport was the closest airport to the port.
Because X,Y	Because Barcelona-El Prat Airport is the closest airport to the port, it's busy.
If X,Y X true	If Barcelona-El Prat Airport is open, it is the closest airport to the port, it's open.

Dealing with the Problems in the Previous Paper

TThey ask the LLM each question only once, so they miss an very important property - Does the LLM generate a consistent answer or does it produce a distribution of random answers?

- In this project we ask the each question 25 times to see the distribution of the answers.
- We consider answer with probability:
  $[0.9, 1]$ as "Truly sure"
  $[0.75, 0.9)$ as "Truly pretty sure"
  $[0.25, 0.75)$ as "Guessing"
  $[0.1, 0.25)$ as "Falsely pretty sure"
  $[0, 0.1)$ as "Falsely sure".

Dealing with the Problems in the Previous Paper

$[0.9, 1]$ as "Truly sure"
$[0.75, 0.9)$ as "Truly pretty sure"
$[0.25, 0.75)$ as "Guessing"
$[0.1, 0.25)$ as "Falsely pretty sure"
$[0, 0.1)$ as "Falsely sure".

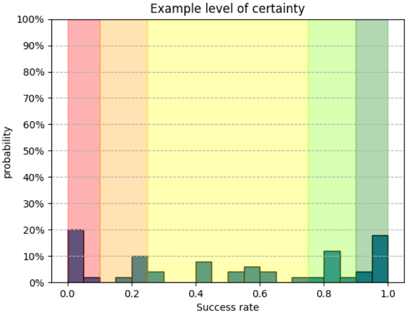


Table of Contents

1. Introduction & Motivation
2. Method
3. Results

Extractive QA

The research work in extractive question answering setup:

- The system is presented with a question and a context.

Extractive QA

The research work in extractive question answering setup:

- The system is presented with a question and a context.
- It should provide an answer based on the context.

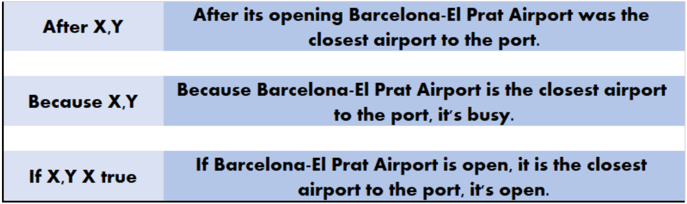
Extractive QA

The research work in extractive question answering setup:

- The system is presented with a question and a context.
- It should provide an answer based on the context.
- If the context does not answer the question, the system should return "None".

Data Creation

Adjusting the same data from the original paper, we took their 50 affirmative questions and changed them to the next linguistic structures:



After X,Y	**After its opening Barcelona-El Prat Airport was the closest airport to the port.**
Because X,Y	**Because Barcelona-El Prat Airport is the closest airport to the port, it's busy.**
If X,Y X true	**If Barcelona-El Prat Airport is open, it is the closest airport to the port, it's open.**

Imaginary Data Necessity

As a side experiment, we also verified that the imaginary and the contradicting data don't behave similarly. If we will find out that they are not it will strengthen the original paper claim that the old method of using contradicting data is problematic method of experiment.

Imaginary Data Necessity

As a side experiment, we also verified that the imaginary and the contradicting data don't behave similarly. If we will find out that they are not it will strengthen the original paper claim that the old method of using contradicting data is problematic method of experiment.

- **The supported data** is created by the answer of the LLM in order to match the parametric knowledge.

Imaginary Data Necessity

As a side experiment, we also verified that the imaginary and the contradicting data don't behave similarly. If we will find out that they are not it will strengthen the original paper claim that the old method of using contradicting data is problematic method of experiment.

- **The supported data** is created by the answer of the LLM in order to match the parametric knowledge.
- **The contradicting data** is created by replacing the factual answer spans in the contexts with other, counterfactual, ones .

Imaginary Data Necessity

As a side experiment, we also verified that the imaginary and the contradicting data don't behave similarly. If we will find out that they are not it will strengthen the original paper claim that the old method of using contradicting data is problematic method of experiment.

- **The supported data** is created by the answer of the LLM in order to match the parametric knowledge.
- **The contradicting data** is created by replacing the factual answer spans in the contexts with other, counterfactual, ones .
- **The imaginary data** is created by replacing the entities in both the context and the questions with made-up, imaginary ones.

Zero Shot Prompt Strategy

Zero shot prompting is consist of two components

Zero Shot Prompt Strategy

Zero shot prompting is consist of two components

- Task description including constraints on the expected answer.

Zero Shot Prompt Strategy

Zero shot prompting is consist of two components

- Task description including constraints on the expected answer.
- specific question.

Models and Prompts

The models that have been tested are:

- GPT-3.5 turbo-0125
- GPT-4 0613 - limited testing due to budget limits
  Temperature = 1 (default) Top-p = 1 (default)

Models and Prompts

Continuing the previous research, we took the prompt they optimized to GPT-3.5, GPT-4 to this expirement.
Example of a prompt:

Models and Prompts

Continuing the previous research, we took the prompt they optimized to GPT-3.5, GPT-4 to this expirement.

Example of a prompt:

- Text: “context”

Models and Prompts

Continuing the previous research, we took the prompt they optimized to GPT-3.5, GPT-4 to this expirement.

Example of a prompt:

- Text:“context”
- Question: “question”

Models and Prompts

Continuing the previous research, we took the prompt they optimized to GPT-3.5, GPT-4 to this expirement.

Example of a prompt:

- Text:"context"
- Question: "question"
- Shortest possible answer please. If the question cannot be answered with a single span from the text, return "None"

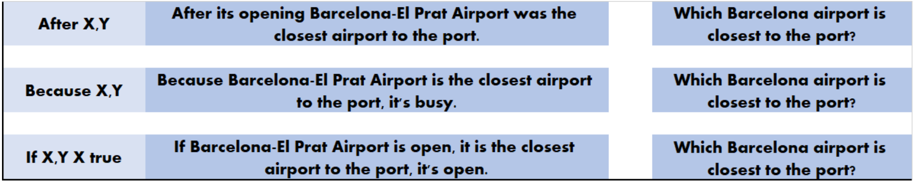
Models and Prompts

Continuing the previous research, we took the prompt they optimized to GPT-3.5, GPT-4 to this expirement.

Example of a prompt:

- Text:"context"
- Question: "question"
- Shortest possible answer please. If the question cannot be answered with a single span from the text, return "None"
- Answer:

Data Examples



After X,Y	After its opening Barcelona-El Prat Airport was the closest airport to the port.	Which Barcelona airport is closest to the port?
Because X,Y	Because Barcelona-El Prat Airport is the closest airport to the port, it's busy.	Which Barcelona airport is closest to the port?
If X,Y X true	If Barcelona-El Prat Airport is open, it is the closest airport to the port, it's open.	Which Barcelona airport is closest to the port?

Table of Contents

Results - the Necessity of the Imaginary Set

The first result we present is re-validation of the previous paper conclusion - the necessity of the imaginary data set.

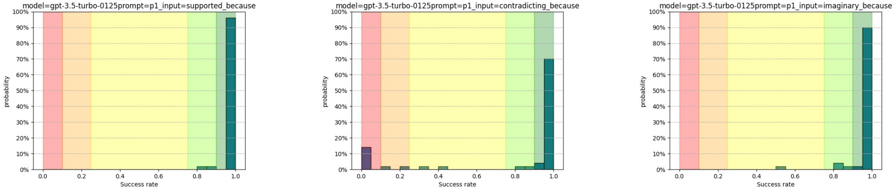


Figure: supported set

Figure: contradicting set

Figure: imaginary set

Figure: Linguistic structure - Because

Results - the Necessity of the Imaginary Set

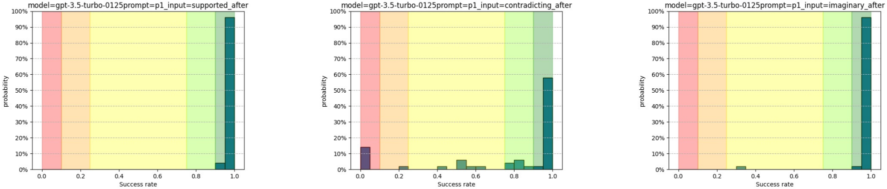


Figure: supported set

Figure: contradicting set

Figure: imaginary set

Figure: Linguistic structure - After

Results - the Necessity of the Imaginary Set

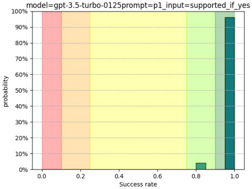


Figure: supported set

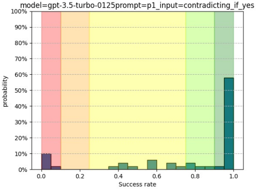


Figure: contradicting set

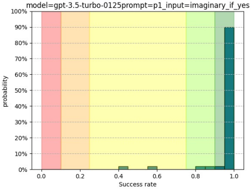


Figure: imaginary set

Figure: Linguistic structure - If-true

Results - the Necessity of the Imaginary Set

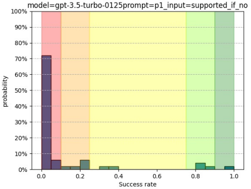


Figure: supported set

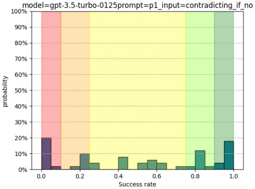


Figure: contradicting set

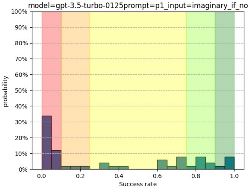


Figure: imaginary set

Figure: Linguistic structure - If-false

Results - Analyses

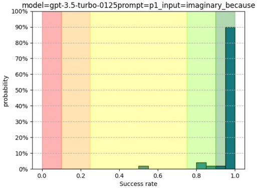


Figure: Linguistic structure - because

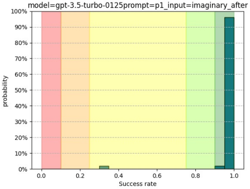


Figure: Linguistic structure - after

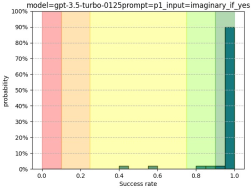
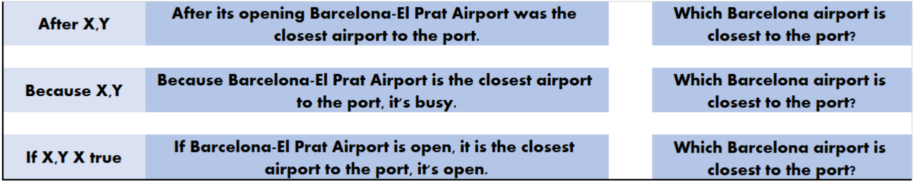


Figure: Linguistic structure - if true

Figure: Our result for the new Linguistic structure

Critical Question About Our Results

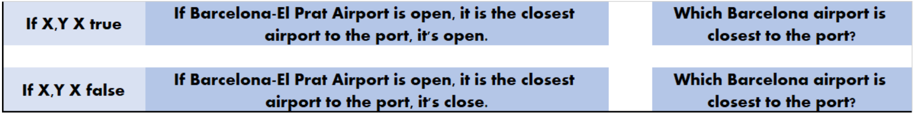
A Possible explanation to the LLM success in the experiment is not due to understanding the nuances of the Linguistic structure, but just due to repeating parts of the context.



After X,Y	**After its opening Barcelona-El Prat Airport was the closest airport to the port.**	**Which Barcelona airport is closest to the port?**
Because X,Y	**Because Barcelona-El Prat Airport is the closest airport to the port, it's busy.**	**Which Barcelona airport is closest to the port?**
If X,Y X true	**If Barcelona-El Prat Airport is open, it is the closest airport to the port, it's open.**	**Which Barcelona airport is closest to the port?**

Critical Question About Our Results

To test this assumption we performed one more experiment, we adjust the ”If X then Y, X true” structure to ”If X then Y, X False” and compared between the results of this structures.



If X,Y X true	If Barcelona-El Prat Airport is open, it is the closest airport to the port, it's open.		Which Barcelona airport is closest to the port?
If X,Y X false	If Barcelona-El Prat Airport is open, it is the closest airport to the port, it's close.		Which Barcelona airport is closest to the port?

Critical Question About Our Results

If indeed, the reason of the success is just repeating parts of the context, we would expect that in here we will see the same rate of answering (wrongly in this context).

If true & if false compression

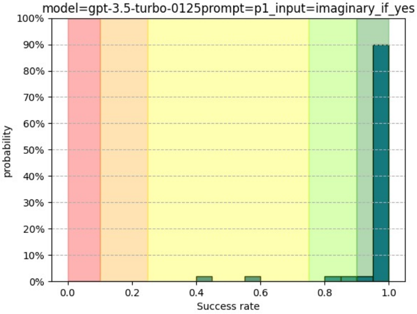


Figure: Linguistic structure - if true

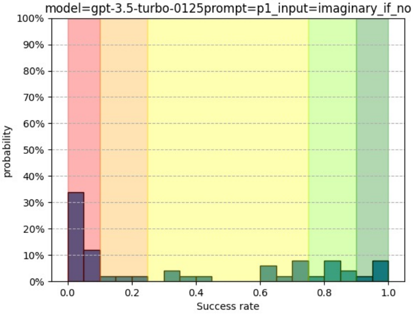


Figure: Linguistic structure - if false

Figure: compression between true and false structures

Chat GPT 3.5 & 4.0 Comparison

The last check what we decided to test is the progress between chat GPT 3.5 and 4.0

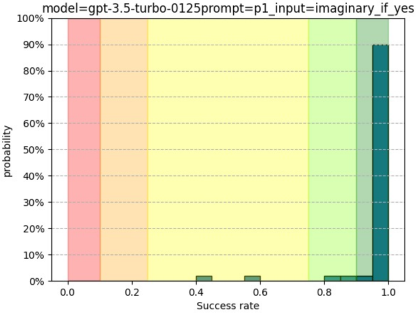


Figure: GPT 3.5 If-true

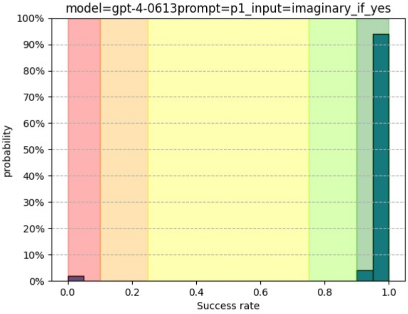


Figure: GPT 4.0 If-true

Chat GPT 3.5 & 4.0 compression

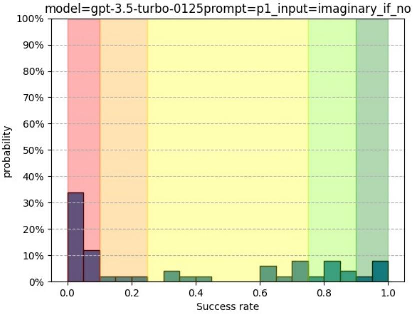


Figure: GPT 3.5 If-false

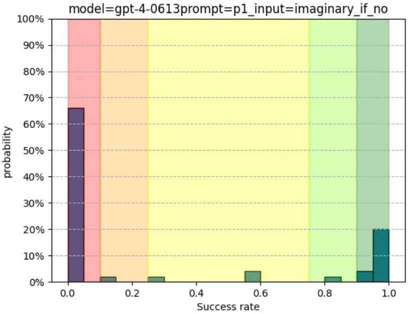


Figure: GPT 4.0 If-false

Questions & Thoughts

Figure: A man sitting on a question mark thinking about his life