LLMs' Reading Comprehension Is Affected by Parametric Knowledge and Struggles with Hypothetical Statements

Presented by Omri Cohen and Tsuri Farhana

Basmov, Goldberg, Tsarfaty

June 2, 2024

Ben-Gurion University of the Negev

Table of Contents

Research question

What are the capabilities of LLMs to understand language models?
in other words 'natural language understanding'

Research 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

Reading comprehension (RC)

To perform well on the text-grounded scenario,the LLM must adhere to two requirements

Reading comprehension (RC)

To perform well on the text-grounded scenario,the LLM must adhere to two requirements

- Understanding the text in the prompt.

Reading comprehension (RC)

To perform well on the text-grounded scenario,the LLM must adhere to two requirements

- Understanding the text in the prompt.
- Being context-faithful: answering exclusively based on information provided in the text, and not based on information in the parametric knowledge.

Reading comprehension (RC)

To perform well on the text-grounded scenario,the LLM must adhere to two requirements

- Understanding the text in the prompt.
- Being context-faithful: answering exclusively based on information provided in the text, and not based on information in the parametric knowledge.

In this work they are interested in the first property, the ability to understand text

Motivation

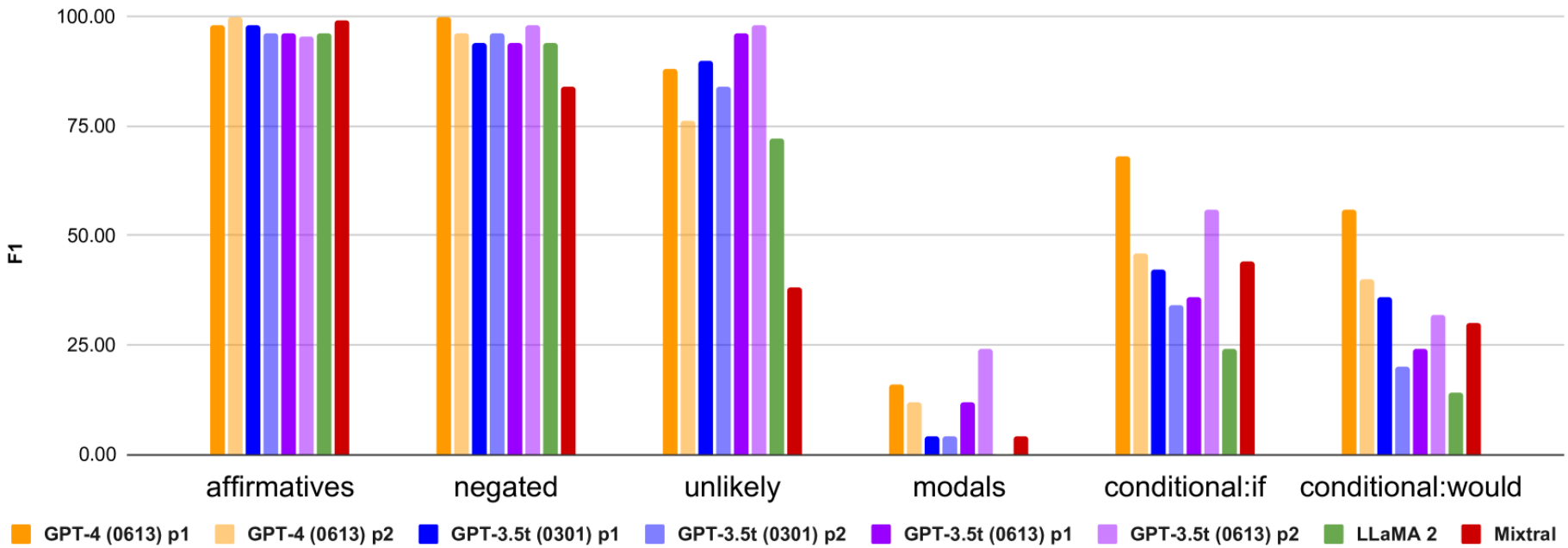

Figure: Results

Is this enough?

Data Set Example

	supported		imaginary		contradicting
affirmative	Bigos is a stew.		Zorg is a stew.		Bigos is a cake.
negation	Bigos is *not* a stew.		Zorg is *not* a stew.		Bigos is *not* a cake.
negative non-factives	*It is unlikely that* Bigos is a stew.		*It is unlikely that* Zorg is a stew.		*It is unlikely that* Bigos is a cake.
modal verbs	Bigos *could have been* a stew.		Zorg *could have been* a stew.		Bigos *could have been* a cake.

Figure: Context Examples

Motivation

When applied to Large Language Models (LLMs) with extensive built-in world knowledge, RC method can be deceptive.

Motivation

When applied to Large Language Models (LLMs) with extensive built-in world knowledge, RC method can be deceptive.

- If the context aligns with the LLMs' internal knowledge, it is hard to discern whether the models' answers stem from context comprehension or from LLMs' internal information

Motivation

When applied to Large Language Models (LLMs) with extensive built-in world knowledge, RC method can be deceptive.

- If the context aligns with the LLMs' internal knowledge, it is hard to discern whether the models' answers stem from context comprehension or from LLMs' internal information
- Conversely, using data that conflicts with the models' knowledge creates erroneous trends which distort the results.

Motivation

When applied to Large Language Models (LLMs) with extensive built-in world knowledge, RC method can be deceptive.

- If the context aligns with the LLMs' internal knowledge, it is hard to discern whether the models' answers stem from context comprehension or from LLMs' internal information
- Conversely, using data that conflicts with the models' knowledge creates erroneous trends which distort the results.

To address this issue, they suggest to use RC on imaginary data, based on fictitious facts and entities. This task is entirely independent of the models' world knowledge, enabling us to evaluate LLMs' linguistic abilities without the interference of parametric knowledge

The Necessity of Imaginary Questions Set

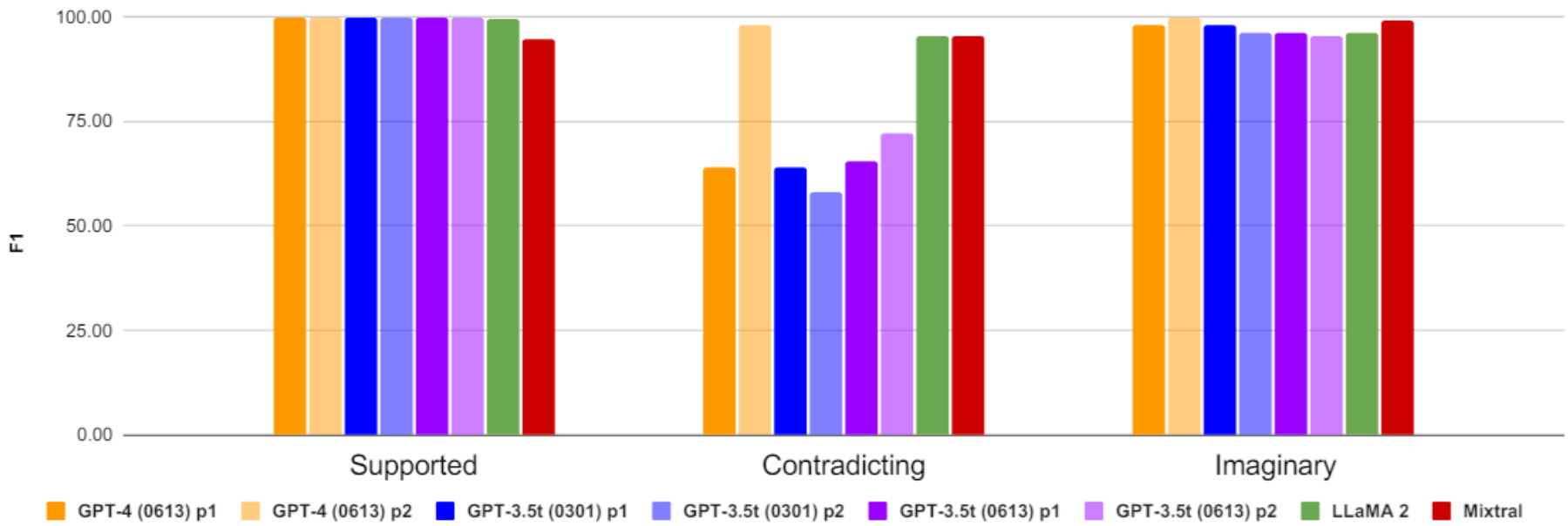

Figure: scores on the affirmative context.

Supported Context: Dog is a mammal, **Contradicting Context:** Dog is a bird, **Imaginary Context:** Zorg is a mammal

Table of Contents

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".

Reminder - Zero Shot Prompt Strategy

Zero shot prompting is consist of two components

Reminder - Zero Shot Prompt Strategy

Zero shot prompting is consist of two components

- Task description including constraints on the expected answer.

Reminder - Zero Shot Prompt Strategy

Zero shot prompting is consist of two components

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

Zero Shot

Motivation to choose zero shot:

Zero Shot

Motivation to choose zero shot:

- The questions in the paper are (by design) simple. An answer generated with examples similar to the question may be effected by learnable patterns and not by "understanding" of the model.

Zero Shot

Motivation to choose zero shot:

- The questions in the paper are (by design) simple. An answer generated with examples similar to the question may be effected by learnable patterns and not by "understanding" of the model.
- The authors believe that zero shot is a true indicator for determining whether an LLM is General purpose system.

Zero Shot

Motivation to choose zero shot:

- The questions in the paper are (by design) simple. An answer generated with examples similar to the question may be effected by learnable patterns and not by "understanding" of the model.
- The authors believe that zero shot is a true indicator for determining whether an LLM is General purpose system.
- It's more aligned with typical user interaction

Models and Prompts

The models that have been tested are:

- GPT-3.5 turbo-0301
- GPT-3.5 turbo-0613
- GPT-4 0613
- LLaMA 2
- Mixtral

Models and Prompts

Out of several prompt templates tested they selected the best templates for each model by optimizing.
example of a prompt:

Models and Prompts

Out of several prompt templates tested they selected the best templates for each model by optimizing.
example of a prompt:

- Text: “context”

Models and Prompts

Out of several prompt templates tested they selected the best templates for each model by optimizing.
example of a prompt:

- Text: "context"
- Question: "question"

Models and Prompts

Out of several prompt templates tested they selected the best templates for each model by optimizing.
example of a prompt:

- Text: "context"
- Question: "question"
- Important! The answer should be an exact span extracted from the text. Important! Give the shortest possible answer. If the question cannot be answered with one span from the text - return "None" (and nothing else).

Models and Prompts

Out of several prompt templates tested they selected the best templates for each model by optimizing.
example of a prompt:

- Text: "context"
- Question: "question"
- Important! The answer should be an exact span extracted from the text. Important! Give the shortest possible answer. If the question cannot be answered with one span from the text - return "None" (and nothing else).
- Answer:

Data Creation

They created three sets, each consisting of 50 context-question-answer triplets: supported, contradicting, and imaginary

Data Creation

They created three sets, each consisting of 50 context-question-answer triplets: supported, contradicting, and imaginary

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

Data Creation

They created three sets, each consisting of 50 context-question-answer triplets: supported, contradicting, and imaginary

- **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 .

Data Creation

They created three sets, each consisting of 50 context-question-answer triplets: supported, contradicting, and imaginary

- **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.

Data Creation

Supported data example:

Data Creation

Supported data example:

- **Context:** Mary Wollstonecraft fought for women's rights.

Data Creation

Supported data example:

- **Context:** Mary Wollstonecraft fought for women's rights.
- **Q:** What did Mary Wollstonecraft fight for?

Data Creation

Supported data example:

- **Context:** Mary Wollstonecraft fought for women's rights.
- **Q:** What did Mary Wollstonecraft fight for?
- **Possible answer:** Mary Wollstonecraft fought for women's rights

Data Creation

Contradicting data example:

Data Creation

Contradicting data example:

- **Context:** Mary Wollstonecraft fought for **immigrants's** rights.

Data Creation

Contradicting data example:

- **Context:** Mary Wollstonecraft fought for **immigrants's** rights.
- **Q:** What did Mary Wollstonecraft fight for?

Data Creation

Imaginary data example:

Data Creation

Imaginary data example:

- **Context:** The **Zogloxians** fought for women's rights.

Data Creation

Imaginary data example:

- **Context:** The **Zogloxians** fought for women's rights.
- **Q:** What did The **Zogloxians** fight for?

Table of Contents

Data Set Example

	supported		imaginary		contradicting
affirmative	Bigos is a stew.		Zorg is a stew.		Bigos is a cake.
negation	Bigos *not* a stew.		Zorg *not* a stew.		Bigos *not* a cake.
negative non-factives	*It is unlikely that* Bigos is a stew.		*It is unlikely that* Zorg is a stew.		*It is unlikely that* Bigos is a cake.
modal verbs	Bigos *could have been* a stew.		Zorg *could have been* a stew.		Bigos *could have been* a cake.

Figure: Context Examples

Data Set Example

- Text: "If Ryan Reynolds' romantic choices had been different, in 2012, he would have been married to Scarlett Johansson."
- Question: "Who was Ryan Reynolds married to in 2012?"
- If the question cannot be answered with a single span from the text, return "None"
- Answer:

Data Set Example

- Text: "If Ryan Reynolds' romantic choices had been different, in 2012, he would have been married to Scarlett Johansson."
- Question: "Who was Ryan Reynolds married to in 2012?"
- If the question cannot be answered with a single span from the text, return "None"
- Answer:
- "Scarlett Johansson"

Data Set Example

- Text: "The Deutsche Mark may be the currency of Germany now."
- Question: "What is the currency of Germany now"
- If the question cannot be answered with a single span from the text, return "None"
- Answer:

Data Set Example

- Text: "The Deutsche Mark may be the currency of Germany now."
- Question: "What is the currency of Germany now"
- If the question cannot be answered with a single span from the text, return "None"
- Answer:
- "Euro."

Table of Contents

Results

The results in the article separate to two parts according to the two requirements that LLMs must adhere to preform well in RC tasks.

Results

The results in the article separate to two parts according to the two requirements that LLMs must adhere to preform well in RC tasks.

- Results that question the context-faithfulness shown in previous works about different LLMs.

Results

The results in the article separate to two parts according to the two requirements that LLMs must adhere to preform well in RC tasks.

- Results that question the context-faithfulness shown in previous works about different LLMs.
- Results for the **Research question**: "Do LLMs understands language models".

Results - Comparing between Different Knowledge Conditions

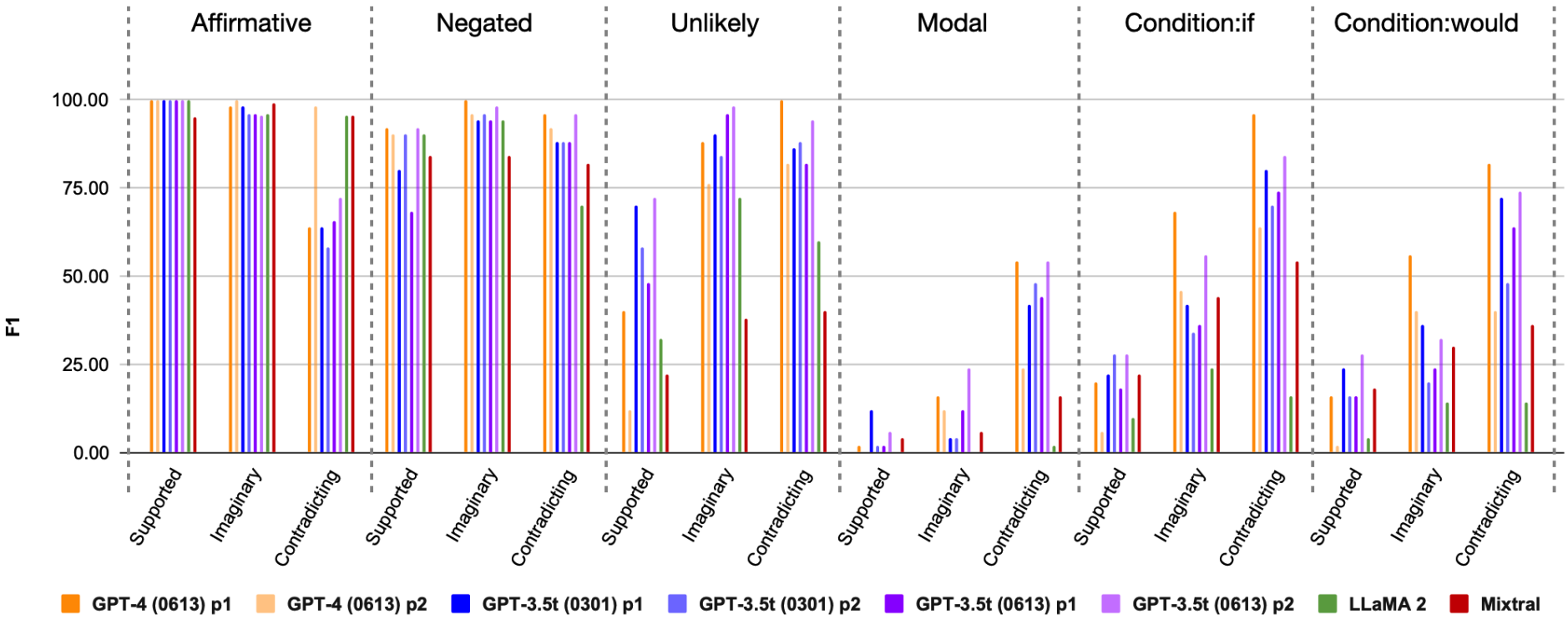

Figure: Comparing the different knowledge-conditions (supported, contradicting, imaginary) across different semantic conditions

Results - Imaginary Set

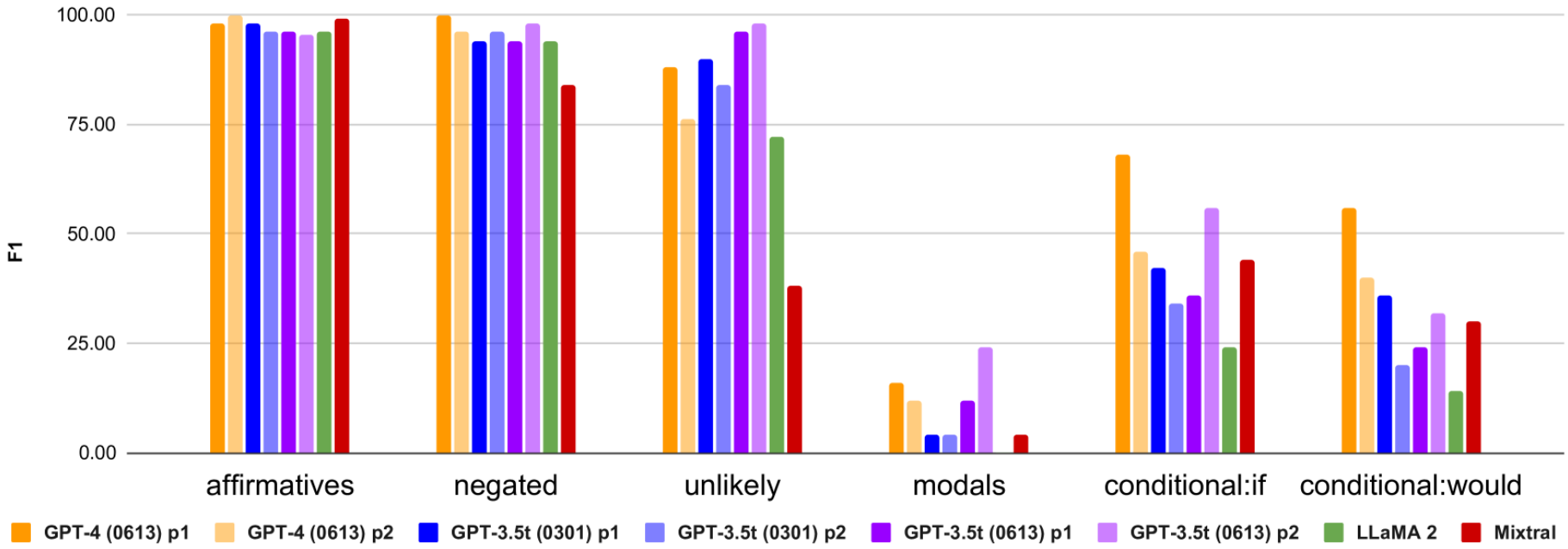

Figure: Comparing performance over different semantic variations using the Imaginary setting

Questions & Thoughts

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