Effect of Prompt Phrasing on LLMs

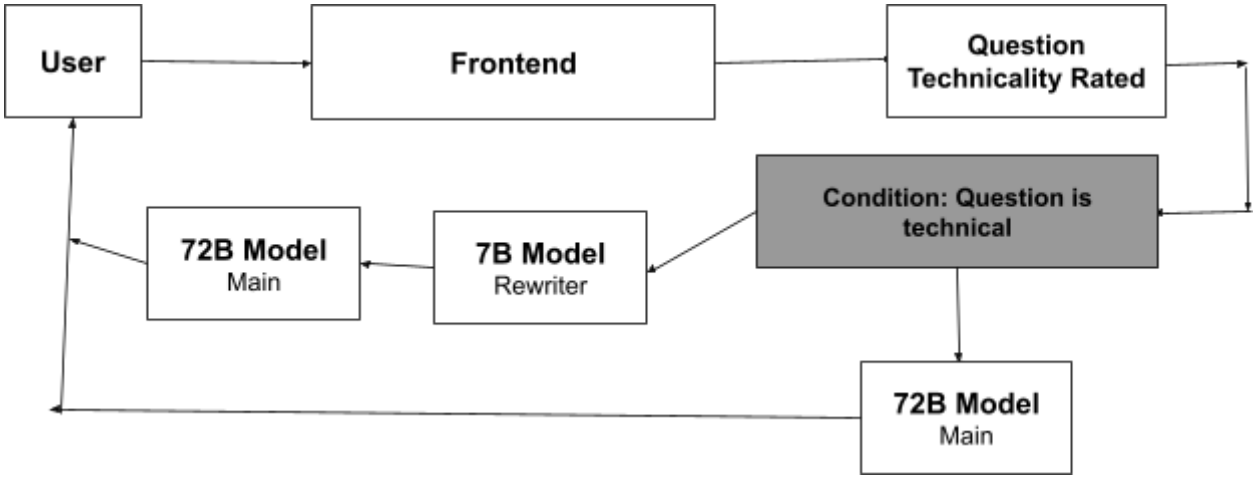
Qwen2.5, GPT-4o, Deepseek, LLama

Dylan Santwani

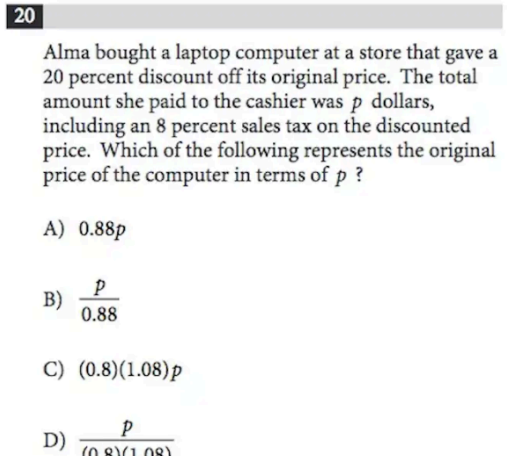
Abstract: Many trained, open source AI models, such as Qwen2.5, Llama, Deepseek (R), and close clones of GPT-2 and GPT-4o all use training data that is largely based on the internet, a large part of which is textbook-like questions, most of which are SAT style. These questions are formatted by guidelines that make the question more understandable and easy to read for the test-taker. Because AI models tend to have an increase in accuracy if the given prompt is similar to training data, I speculate that by phrasing prompts similar to how textbook questions are formatted, the training data would provide more accurate results with a clearer answer.

Application: If prompts can be rewritten, possibly by a 3b or 7b model that lies in between the prompt giver and the LLM, more accurate results can be produced.

Application Model:



Example:



20

Alma bought a laptop computer at a store that gave a 20 percent discount off its original price. The total amount she paid to the cashier was p dollars, including an 8 percent sales tax on the discounted price. Which of the following represents the original price of the computer in terms of p ?

A) $0.88p$

B) $\frac{p}{0.88}$

C) $(0.8)(1.08)p$

D) $\frac{p}{(0.8)(1.08)}$

Left: A structured SAT math question

Bottom: The question rewritten without formatting (Gemini)

A laptop computer was sold to Alma at a store with a 20% reduction from its regular price. The total cost Alma paid, denoted as p dollars, encompassed an 8% sales tax calculated on the reduced price. Which of the following formulas calculates the original price of the laptop using p?

Scoring and Prompt

A system prompt followed by the image is given to each AI model.

System Prompt:

<start>

Explain your thoughts in solving this question. Solve it in 5 steps, and explain each step. Make your ending answer clear.

<end>

Note that the answer choices are not provided to any models

All scoring is done out of 15 in 0.1 increments. Scoring is done based off the following:

Category	Clarity	Understanding	Correct Format	Correct Answer
Description	Points given based off the clarity of each step	Understanding specifically what the question is asking to solve for, and how to get there	No answer choices are provided. Points are given if the answer is in the correct format that matches an answer	Correct (5pts) or Incorrect (0pts)
Points	7pts	3pts	3pts	5pts

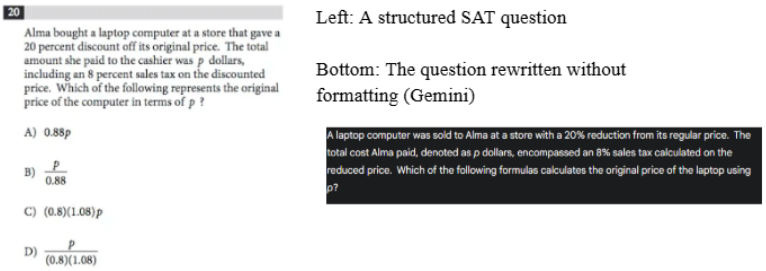
Data Set A

Question 1 - Mathematical Reasoning

(Alma Question)

Tests: Logistics, Algebraic Reasoning, Real World Problem Solving

Using Formatted Question



20

Alma bought a laptop computer at a store that gave a 20 percent discount off its original price. The total amount she paid to the cashier was p dollars, including an 8 percent sales tax on the discounted price. Which of the following represents the original price of the computer in terms of p ?

A) $0.88p$

B) $\frac{p}{0.88}$

C) $(0.8)(1.08)p$

D) $\frac{p}{(0.8)(1.08)}$

Left: A structured SAT question

Bottom: The question rewritten without formatting (Gemini)

A laptop computer was sold to Alma at a store with a 20% reduction from its regular price. The total cost Alma paid, denoted as p dollars, encompassed an 8% sales tax calculated on the reduced price. Which of the following formulas calculates the original price of the laptop using p?

Figure 1: Question 1 Formatted

As shown, all models got the correct answer with a high clarity rate using the formatted question.

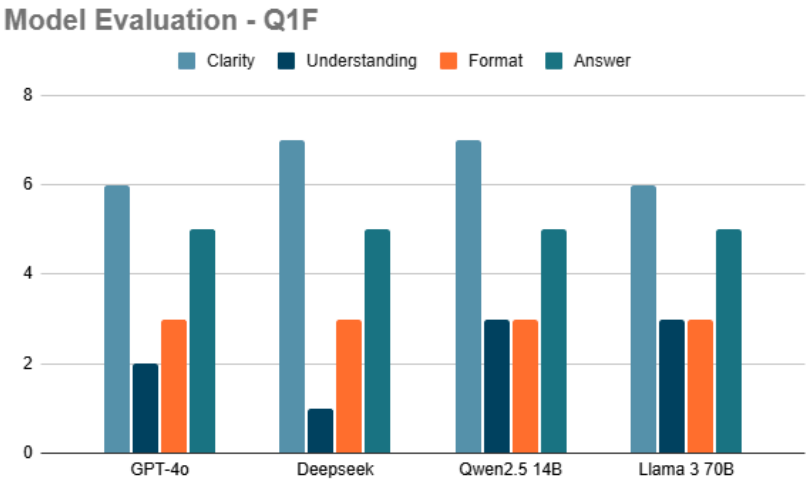
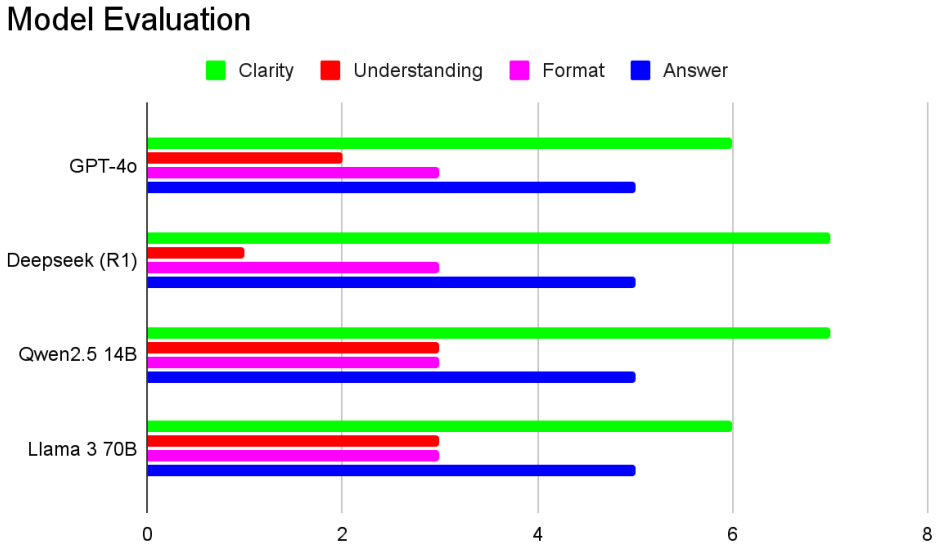


Figure 2: Question 1 Formatted

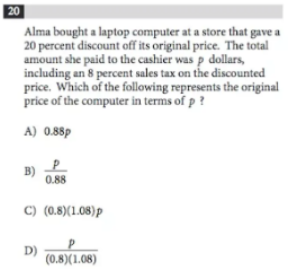


Data Set A

Question 1 - Mathematical Reasoning

(Alma Question)

Tests: Logistics, Algebraic Reasoning, Real World Problem Solving



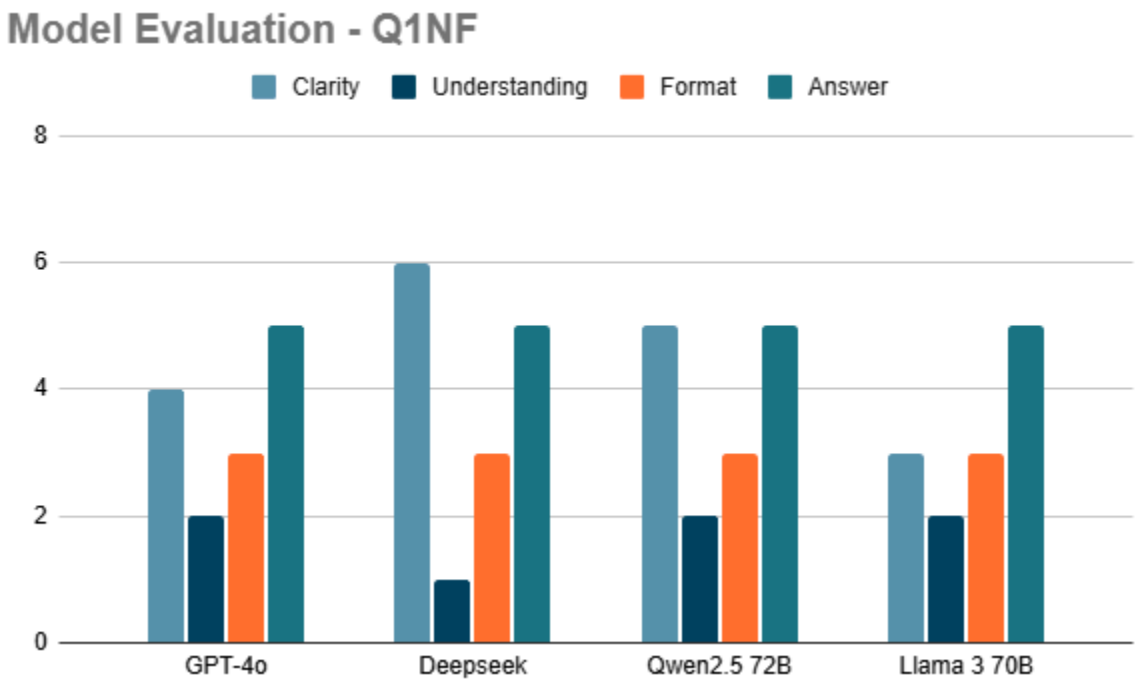
Left: A structured SAT question

Bottom: The question rewritten without formatting (Gemini)

A laptop computer was sold to Alma at a store with a 20% reduction from its regular price. The total cost Alma paid, denoted as p dollars, encompassed an 8% sales tax calculated on the reduced price. Which of the following formulas calculates the original price of the laptop using p?

Using Rewritten Question

Figure 3: Question 1 No Formatting Data



Analysis

Data Set A

Statistical Change

As demonstrated in Fig. 1 and 3, the quantitative data used shows a clear difference between whether the model was prompted with the formatted SAT style question or the rewritten question. Though the answer was always written regardless of model, clarity and understanding had a significant decrease throughout testing, showing the impact of prompting a model in a similar way to the majority of training data.

Speculation

The increase in performance within the question that was specifically formatted with SAT standards could be because of the increasing amount of internet data that LLMs are typically trained on. A large part of this data includes these structured questions and an answer key, therefore it would have a performance increase when being prompted on similarly structured questions

Evidence

Reference Fig 1, 3

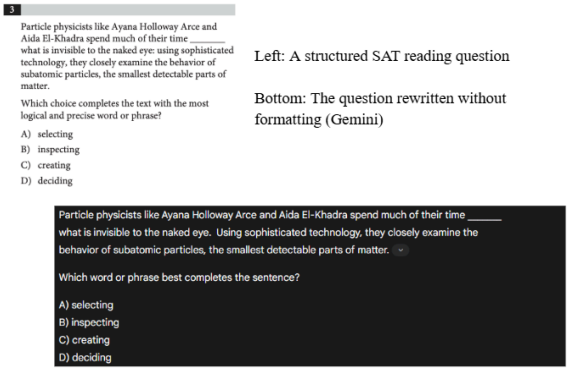
Model	Effect
Qwen2.5	Decreased Clarity
LLama-3	Decreased Clarity and Understanding
Deepseek	Slightly decreased Clarity
GPT-4o	Decreased Clarity and Understanding

Data Set B

Question 2 - Contextual Reasoning

(Physicists Question)

Tests: Reading Comprehension, Vocabulary in Context, Word reasoning and inference



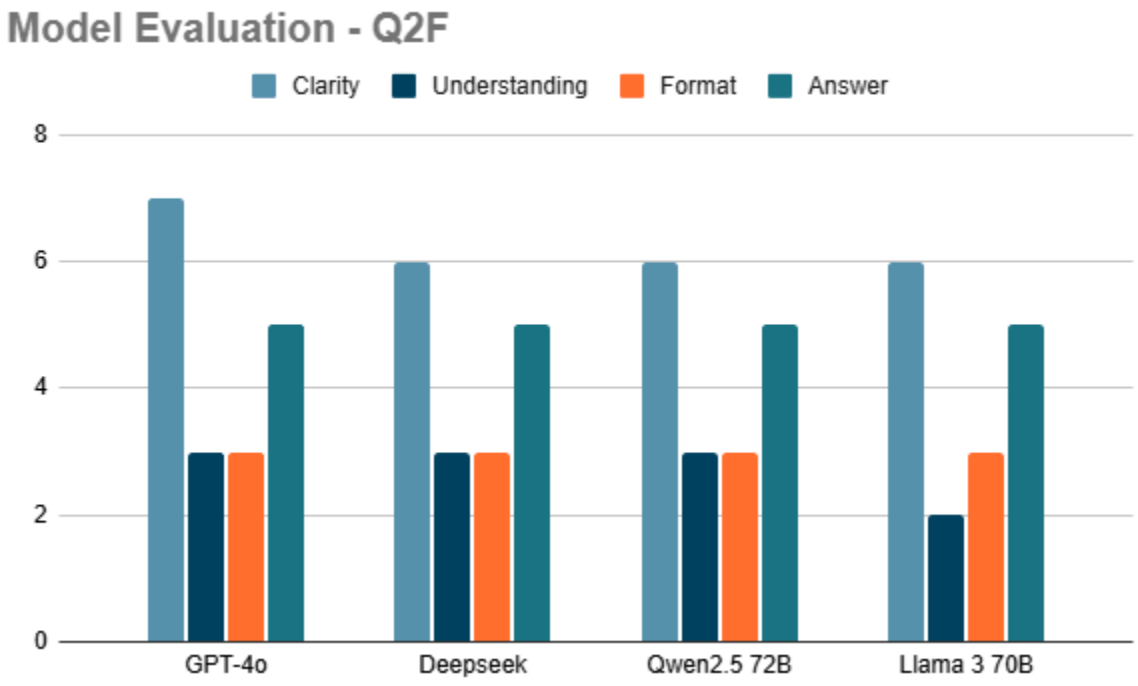
Left: A structured SAT reading question

Bottom: The question rewritten without formatting (Gemini)

Using Formatted Question

As shown below, all models got the question correct with high accuracy. The variation between depth was minimal.

Figure 4: Question 2 Formatted



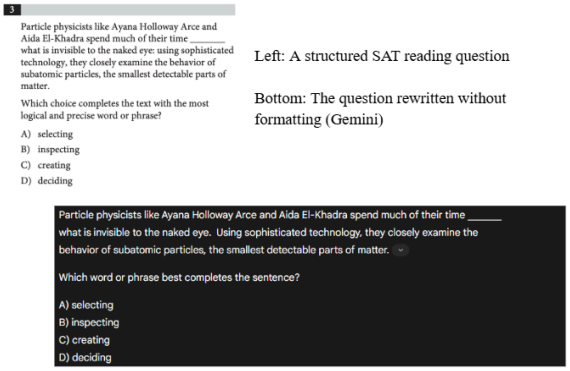
Data Set B

Question 2 - Contextual Reasoning

(Physicists Question)

Tests: Reading Comprehension, Vocabulary in Context, Word reasoning and inference

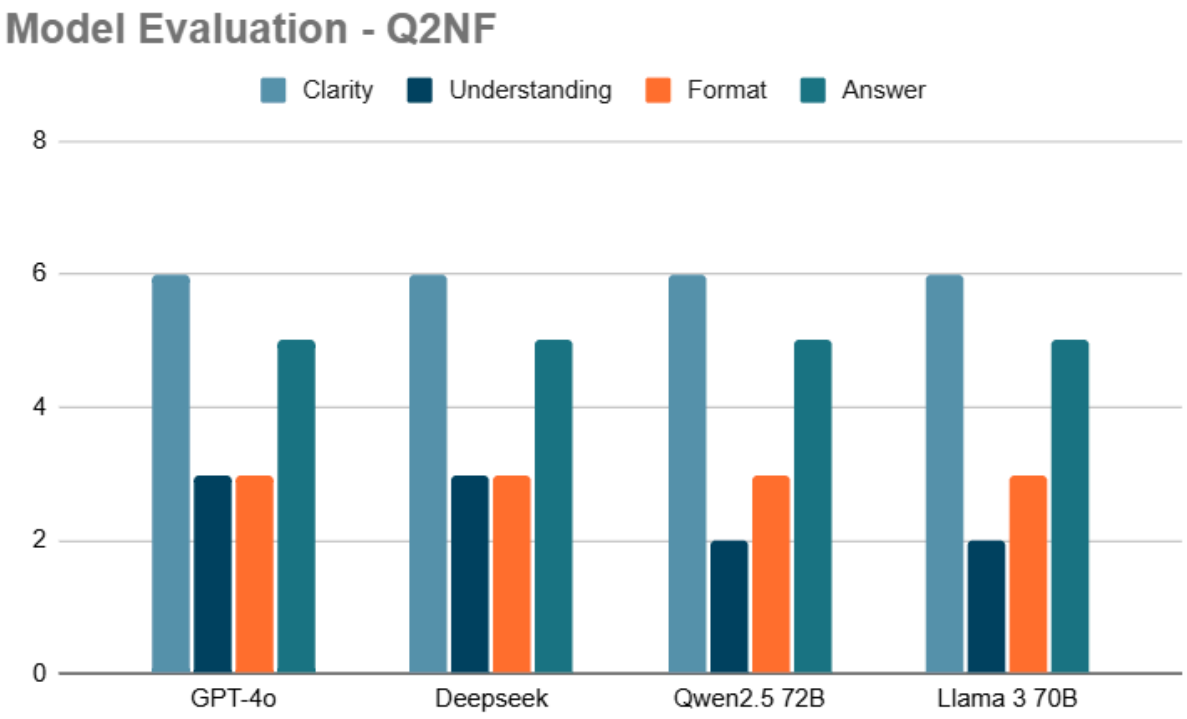
Using Rewritten Question



Left: A structured SAT reading question

Bottom: The question rewritten without formatting (Gemini)

Figure 5: Question 2 No Formatting Data



There is not a notable decrease in accuracy, understanding, or clarity throughout the models.

Analysis

Data Set B

Statistical Change

As demonstrated in Fig. 4 and 5, the qualitative data used shows a smaller difference between whether the model was prompted with the structured question or unstructured question.

Speculation

This could suggest that LLMs are less impacted by prompt formatting in specific types of reading questions, in contrast to specific logistical questions, such as basic algebra. The data shown has a slight loss of understanding, which can be attributed to a small difference in the question meaning as it was changed from its usual structured state.

Evidence

Reference Fig 4, 5

Model	Effect
Qwen2.5	Slightly decreased Understanding
GPT-4o	Slightly decreased Clarity

Conclusion

Qwen2.5, GPT-4o, Deepseek, LLama

Conclusion: Trained, open source AI models, such as Qwen2.5, Llama, Deepseek (R), and close clones of GPT-2 and GPT-4o all have a significant change in performance if the question type has some type of logistical reasoning (demonstrated by the use of basic algebra.) By rephrasing the question into a commonly used format in order to better fit the question to the models training data, more accurate results could be produced.

Application: Prompts can be checked (whether the question has to do with logistical reasoning or not), and rewritten into a common format that follows textbook-like question standards. After being rewritten by a smaller parameter model (7B), it can be sent to the main model (72B), in order to increase performance overall.

Supporting Papers:
Supporting this is the paper *Large Language Models: A Survey*, by Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Specifically, it talks on increasing performance through the use of prompt engineering, saying that "As we described in section IV, many of the shortcomings and limitations of LLMs such as hallucination can be addressed through advanced prompt engineering, use of tools, or other augmentation techniques." This could also decrease hallucinations within LLMs through the use of prompt rewriting.

Revised Application Model: