Evaluating the Casual Reasoning Abilities of Large Language Models

Isha Puri, Hima Lakkaraju

Harvard University

Introduction

Current States: Despite their impressive abilities, however, Large Language Models suffer from several behaviors that can prove dangerous to users treating them as ”oracles” and suggest that any ”reasoning” demonstrated by the models is not genuine and simply an unreliable mimicry of large-scale training data. As these models are being integrated into impactful parts of society, it is thus urgent that we gain insight into the step-by-step 'reasoning processes' taken by LLMs to arrive at their outputs.

A new paradigm for "Explainability"

We argue that explainability in language models is about extracting *reasoning* from the model, whether through prompting [**?**], one shot [**?**]/few shot learning [**?**], etc. Reasoning is a cognitive process that involves drawing conclusions based on available information, often through logical steps or inferences. By focusing on eliciting the model’s reasoning, we can better understand how a model processes and manipulates information to arrive at its conclusions. This approach allows us to extract insights into the model’s internal decision-making processes, thus enabling a deeper understanding of the model’s behavior.

In order to fully trust that the reasoning a LLM outputs for a given prompt can be used as evidence for its true internal workings, however, we need to have faith in the reasoning abilities of language models. We need to measure how well LLMs can explain their 'thinking' - how accurate is their reasoning?

Introduction to CReDETS

Although there are datasets such as LogiQA for measuring general logic abilities of large language models through open ended question and answering benchmarks, there is lack of data sources that explicitly focus on complex causal reasoning Q&A *and* include high quality explanations of those answers as well.

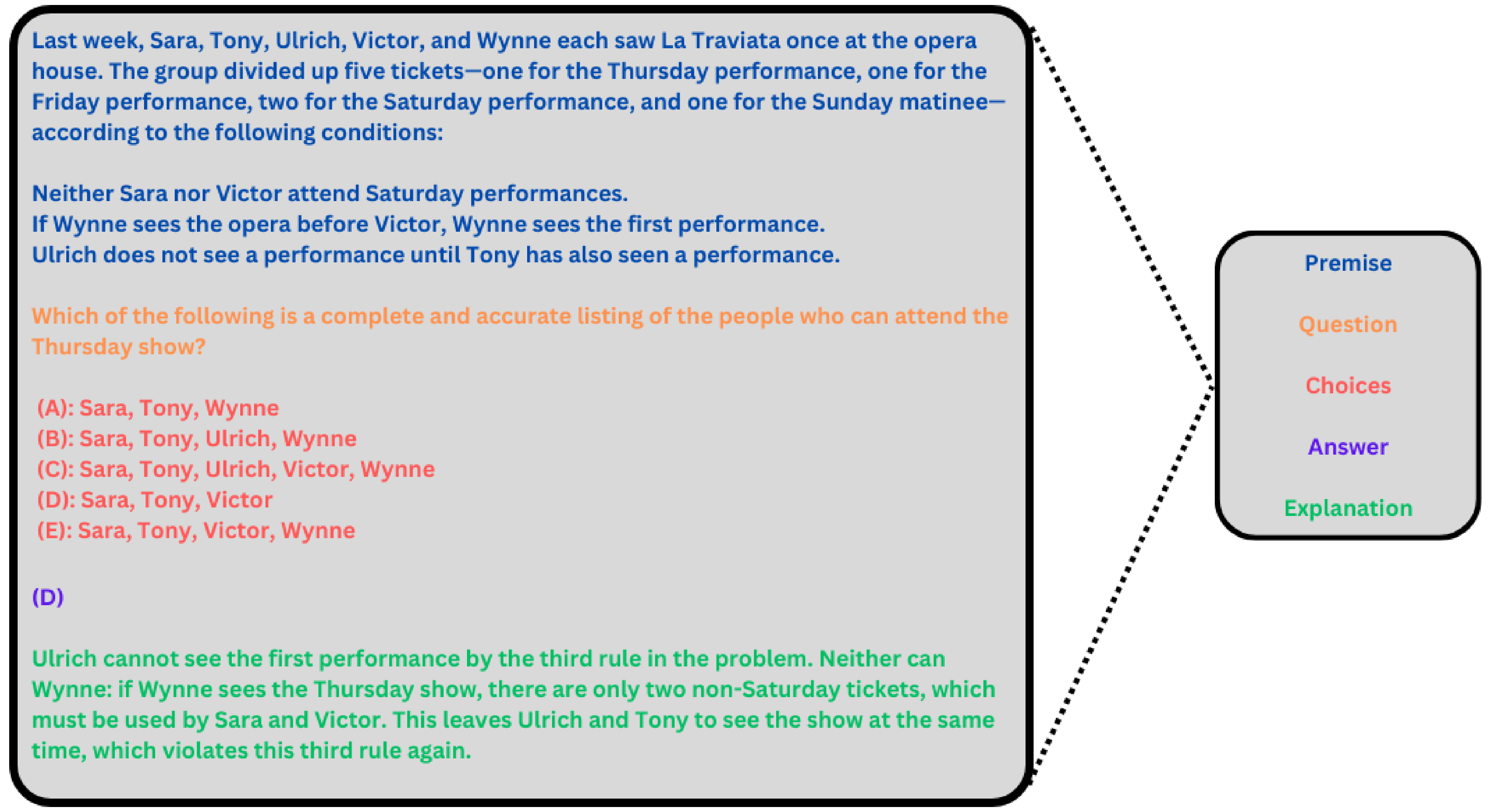
To this end, we introduce CReDETS, the **C**ausal **RE**asoning **D**ataset and **E**xplanation **T**est **S**uite, a novel, first-of-its-kind causal reasoning dataset with hand-annotated explanations.

We hope that the introduction of this dataset will allow researchers to continue to evaluate and improve the reasoning abilities of various generations of language models.

These questions are based on the LSAT, which is one of the only professional tests that doesn’t require any subject matter knowledge, and thus is a perfect basis for a causal reasoning dataset. These professional exam questions are written by philosophy and logic experts to specifically measure causal reasoning ability.

We curated 442 samples, each of which is based on a premise involving a set of characters and rules that define relationships between them.

CoFrNets



Last week, Sara, Tony, Ulrich, Victor, and Wynne each saw La Traviata once at the opera house. The group divided up five tickets—one for the Thursday performance, one for the Friday performance, two for the Saturday performance, and one for the Sunday matinee—according to the following conditions:

Neither Sara nor Victor attend Saturday performances.
If Wynne sees the opera before Victor, Wynne sees the first performance.
Ulrich does not see a performance until Tony has also seen a performance.

Which of the following is a complete and accurate listing of the people who can attend the Thursday show?

(A): Sara, Tony, Wynne
(B): Sara, Tony, Ulrich, Wynne
(C): Sara, Tony, Ulrich, Victor, Wynne
(D): Sara, Tony, Victor
(E): Sara, Tony, Victor, Wynne

(D)

Ulrich cannot see the first performance by the third rule in the problem. Neither can Wynne: if Wynne sees the Thursday show, there are only two non-Saturday tickets, which must be used by Sara and Victor. This leaves Ulrich and Tony to see the show at the same time, which violates this third rule again.

Figure 1:Structure of LSAT Logic Games Questions

Structure of CReDETS

For each question, we include not only the question, answer choices, and correct answer, but also a *hand-written explanation* for each question, a unique differentiation of our dataset with respect to all others in the field such a LogiQA. This tests the capabilities of language models to not just choose the right answer options (MCQ) but also to explain reasoning for each question.

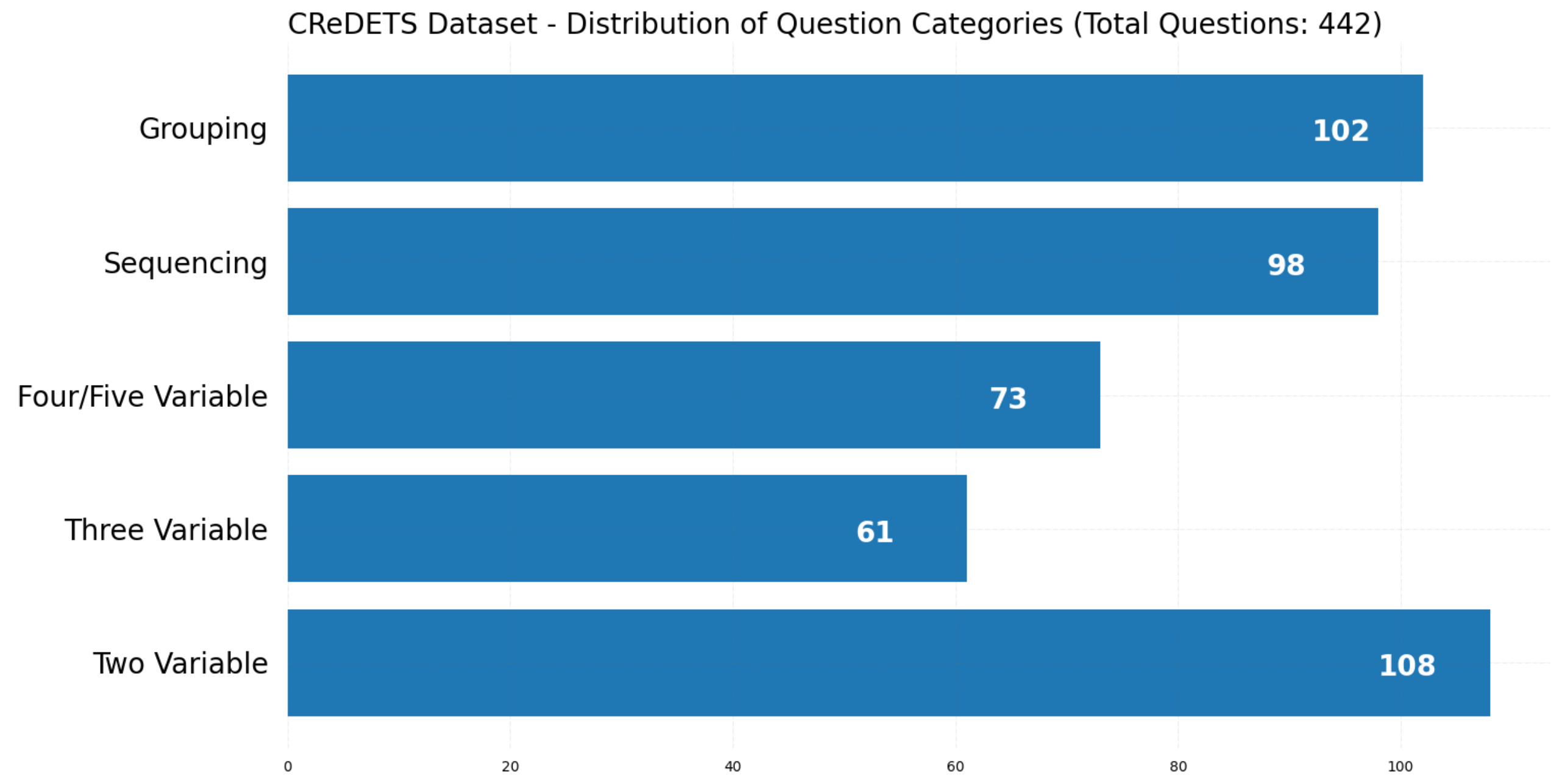


Figure 2:CReDETS Dataset - Distribution of Question Categories (Total Questions: 442)

Benchmarking Results

In order to measure the accuracy of our three test models - GPT3, GPT3.5, and GPT4 - on the questions in the CReDETS dataset, we ran 10 trials of the 442 questions. Each question was run via a separate API call. The results can be seen in table 1 below.

	Model		
	GPT 3	**GPT 3.5**	**GPT 4**
Trial Average	**0.198**	**0.207**	**0.248**
Trial 1	.199	.205	.278
Trial 2	.201	.210	.282
Trial 3	.192	.212	.291
Trial 4	.196	.203	.271
Trial 5	.208	.213	.269
Trial 6	.205	.210	.264
Trial 7	.199	.201	.280
Trial 8	.199	.199	.271
Trial 9	.187	.208	.273
Trial 10	.196	.212	.271

As we see here, all three models (GPT3, GPT3.5, GPT4) perform quite poorly on the CReDETS benchmark. GPT 4’s performance is an improvement to its predecessors - while not at all close to human-level accuracy, is a marked improvement from its predecessors. Most importantly, GPT 4 displays increased levels of consistency.

Analysis

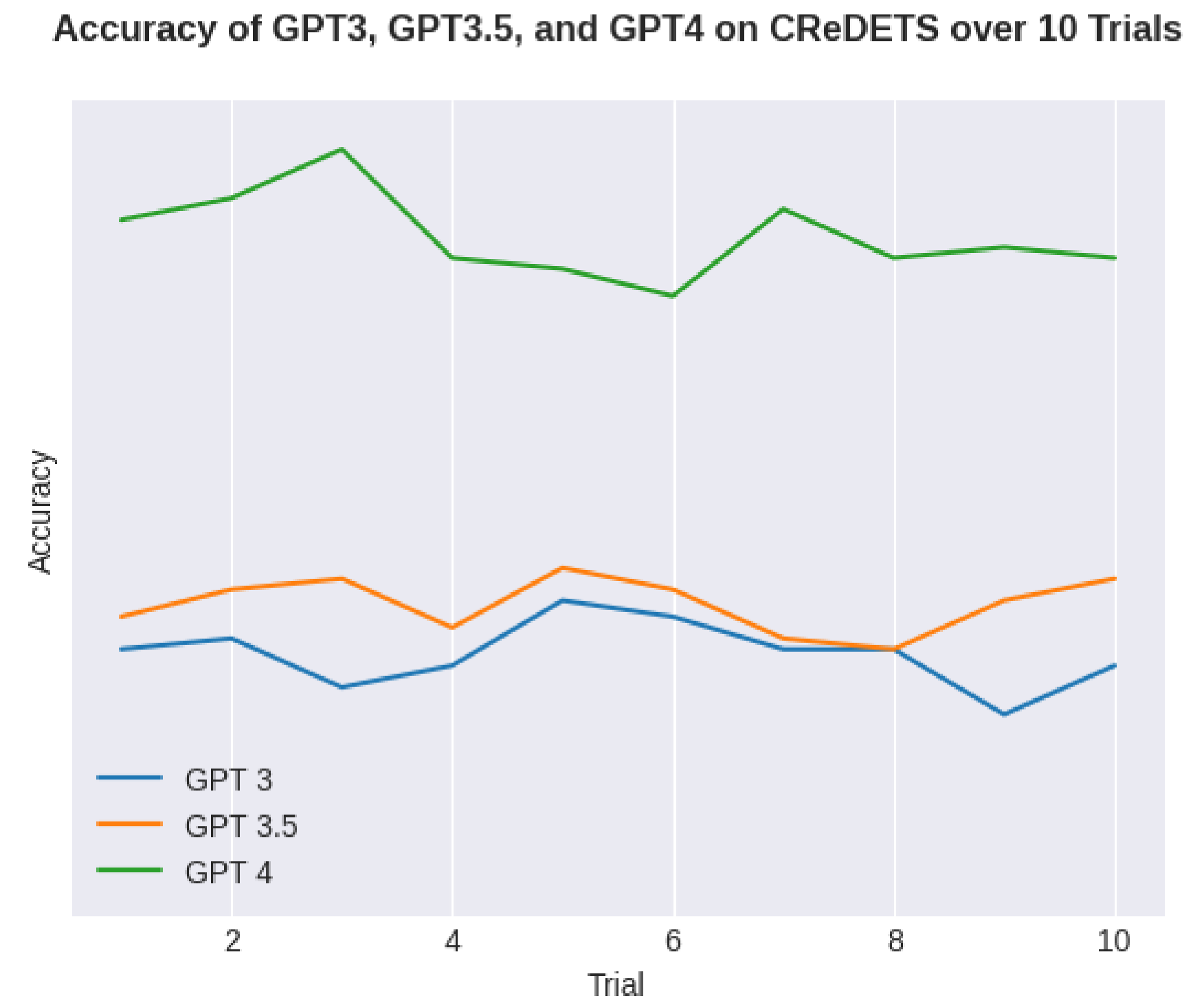


Figure 3:Test accuracies of GPT3, GPT3.5, and GPT4 on the 442 questions in preliminary CReDETS Dataset over 10 trials.

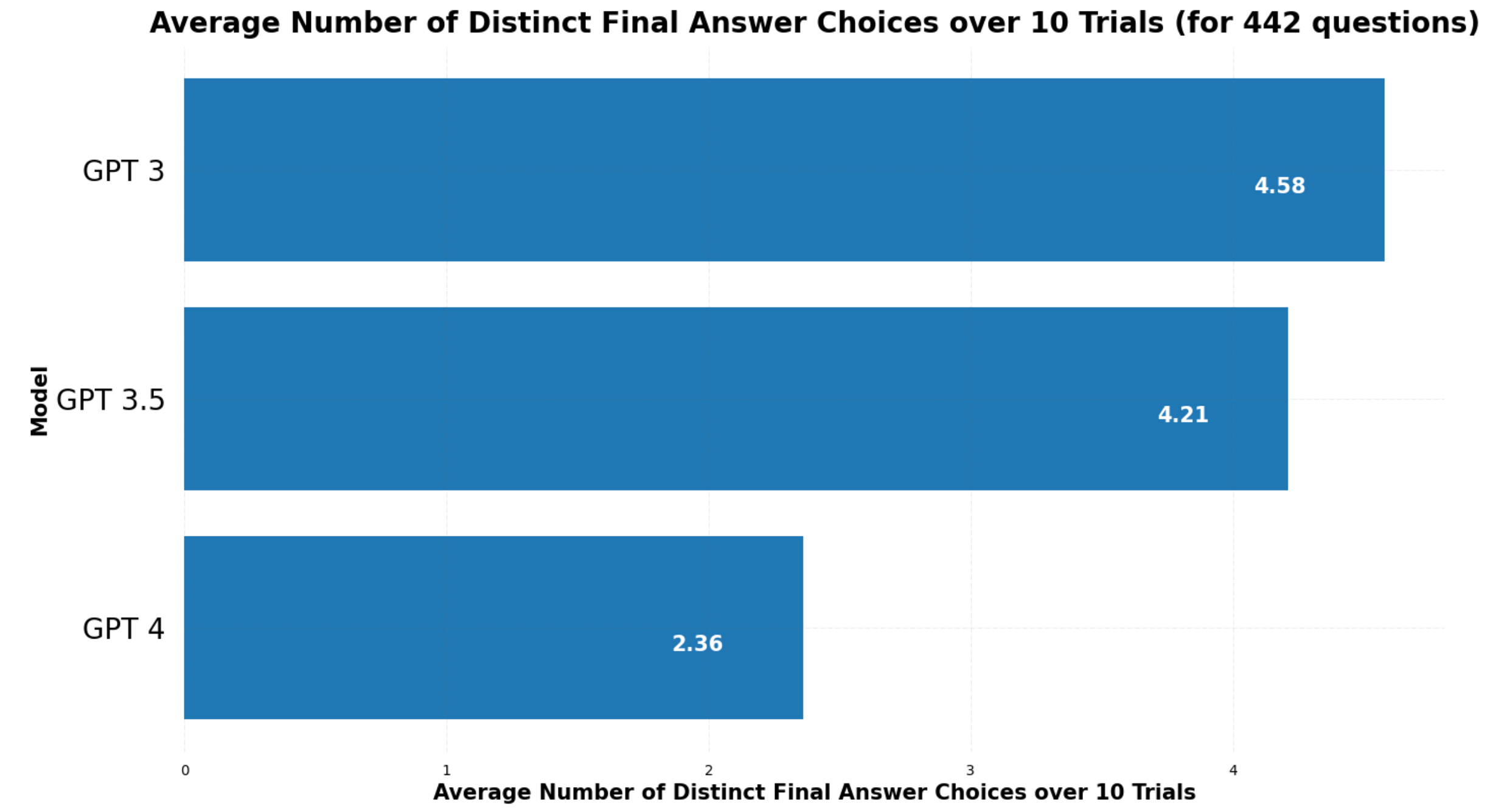


Figure 4:Average Number of Distinct Final Answer Choices Made Over 10 Trials by GPT 3, GPT 3.5, and GPT 4

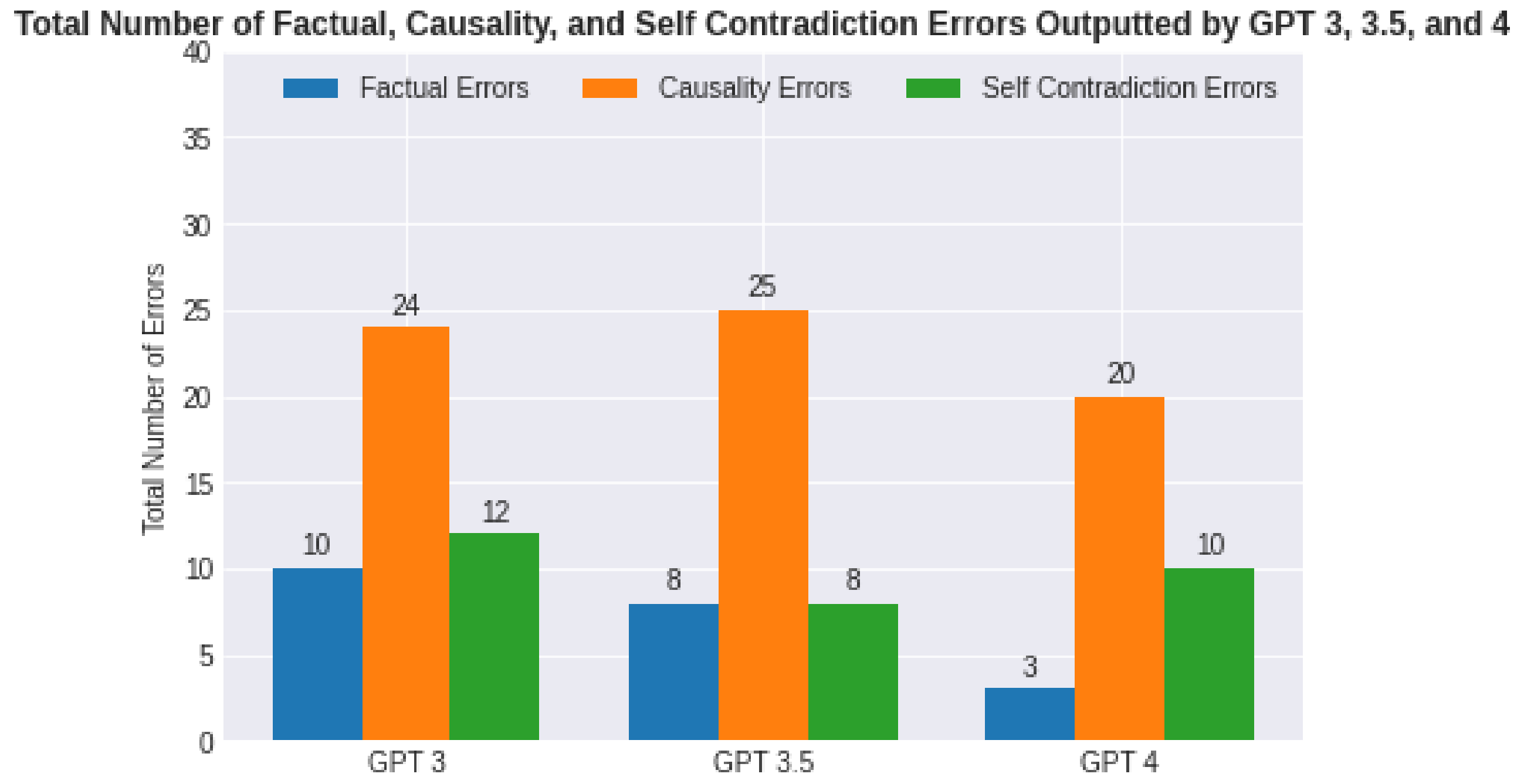


Figure 5:Average Number of Distinct Final Answer Choices Made Over 10 Trials by GPT 3, GPT 3.5, and GPT 4