Evaluating the Causal Reasoning Abilities of Large Language Models

Isha Puri [1] **Himabindu Lakkaraju** [1]

Abstract

Large language models have developed at a breathtaking pace, quickly advancing in their ability to generate, summarize, and work with long and short-form text. As these advances become further integrated into society, however, it becomes necessary to question and evaluate how well these models are actually capable of true reasoning, rather than simply mimicking their large training corpora. We argue that eliciting reasoning from language models is the new ”explainability method” and introduce CReDETS, a novel and first-of-its-kind causal reasoning dataset with annotated hand-written explanations. We benchmark the latest and most powerful generation of transformer neural network models GPT-3, GPT-3.5 (chatGPT), and GPT-4 and discuss their accuracy, coherence, and consistency. Our staggering results show that even the most recent LLMs have stark weaknesses in reasoning ability that must be ameliorated before they can be integrated with public-facing applications worldwide.

1. Introduction

1.1. The State of New Models

In recent times, Large Language Models have taken much of the public by storm, with their abilities rapidly advancing to include text generation, summarization, question answering, and much more. ChatGPT [16] and GPT4 [15] have skyrocketed in use to become the fastest-growing technology products of all time [5], and they created significant partnerships with industry players, including Microsoft Bing [2], Duolingo [13], Instacart [11], Snapchat [14], and countless other products that put them face to face with customers. The outputs of these models come face to face with vulnerable populations [7] and are often responsible for critical decision-making processes. Some of the notable advances associated with these models include Few-shot Learning, Improved Language Modeling, Enhanced Zero-shot Learning, Context-aware Conversational AI, Knowledge Retrieval and Reasoning, and Transfer Learning. 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.

1.2. The Disconnect Between Current Explainability Methods and LLMs

A 2017 survey by Chakraborty et al. [4] suggests that *explainability* refers to a complete output, where a model response is accompanied by all relevant parts of the input used for the reasoning behind its decision, while *interpretability* refers to the quality of an explanation based on how a human interprets it, but as is common in practice and literature, we will use the terms mostly interchangeably.

Still, we must recognize that even Chakraborty's definition of explainability proves hard to apply to large language models.

Post-hoc explainability methods, which aim to provide interpretability for model predictions after they have been generated, often struggle to provide effective explanations for large language models like GPT-3.5. There are several key reasons for this:

Complexity and non-linearity: Large language models like GPT-3.5 are composed of millions or even billions of parameters, making them highly complex and non-linear [18]. This complexity makes it difficult for post-hoc explainability methods to meaningfully disentangle the relationships between input features and model predictions. Many traditional methods, such as LIME [19] or SHAP [12], rely on locally linear approximations, which may not adequately capture the intricate relationships learned by large language models.

- *Input representation*: Large language models use sub-

[1] Harvard University, Cambridge, MA, USA. Correspondence to: Isha Puri <ishapuri@college.harvard.edu>.

Accepted at the Knowledge and Logical Reasoning in the Era of Data-driven Learning Workshop at ICML 2023. Copyright 2023 by the author(s).

word tokenization techniques to represent text inputs. This leads to variable-length input representations depending on the tokens present in the input text. Post-hoc explainability methods, which typically assume fixed-length feature spaces, may struggle to adapt to such variable-length input representations.

- *Self-attention mechanisms*: Transformer architectures, which underpin models like GPT-3.5, rely heavily on self-attention mechanisms to capture long-range dependencies in the input text [22] These self-attention mechanisms enable the model to weight different parts of the input differently, making the relationship between input and output even more complex. This complexity make it difficult to disentangle and interpret the individual contributions of each parameter to the model's output [1]

Above all else, it is *necessary* to shift the paradigm of explainability when considering Large Language Models because of one simple fact - everyday users of LLMs such as GPT3.5 and GPT4 do not have access to the inner workings of the architecture. Even if we were able to develop post-hoc explanation tools that worked on LLMs such as GPT4, normal users of the product only have access to an API output. With nothing but access to an input and an output of the AI model, we argue we must shift the way we think about aiming for explainability with the arrival of LLMs and arguably higly sophisticated intelligent models [3].

2. A new paradigm for ”Explainability” - Reasoning in Language Models

We argue that explainability in language models is about extracting *reasoning* from the model, whether through prompting [6], one shot [21]/few shot learning [17], 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?

Kojima et al. [9] suggests that simply appending the phrase ”Let's think step by step” to any GPT3 prompt induces the model into provide more accurate answers and include the context it uses to extract more accurate answers. OpenAI has claimed, however, that with the arrival of GPT3.5 (ChatGPT) and GPT4 also comes advanced reasoning abilities and a lesser need to engineer prompts to ”elicit” reasoning.

In this paper, we present a thorough analysis of such claims and present a new dataset CReDETS and metrics to measure explaination and reasoning ability of these models.

3. Introduction to CReDETS

Although there are datasets such as LogiQA [10] 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.

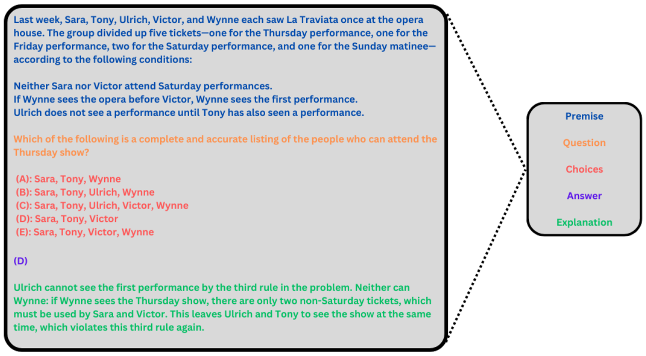
To this end, we introduce CReDETS, the **C**ausal **RE**asoning **D**ataset and **E**xplanation **T**est **S**uite, a novel, first-of-its-kind causal reasoning dataset with hand-annotated explanations. We hope that the introduction of this dataset will allow researchers to continue to evaluate and improve the reasoning abilities of various generations of language models.

In aiming to select questions that would best capture a model's causal reasoning ability, we analyzed several professional and standardized exams and decided upon the Law School Admissions Test (LSAT). The LSAT 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 from the Logic Games section of the LSAT, a section that we specifically chose because of its completely self-contained nature (no external information is required to solve each question) and explicit focus on causal reasoning. As shown in Figure 1, each question is based on a premise involving a set of characters and rules that define relationships between them.

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 [10].

CReDETS enables the measurement of both causal reasoning accuracy *and* explanation with questions that are pre-vetted by law and philosophy experts, completely self contained, and meant to be solved in quick time periods. **It is a first-of-its-kind dataset that includes human-annotated explanations for each causal reasoning question and answer pair. This tests the capabilities of language models**



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.

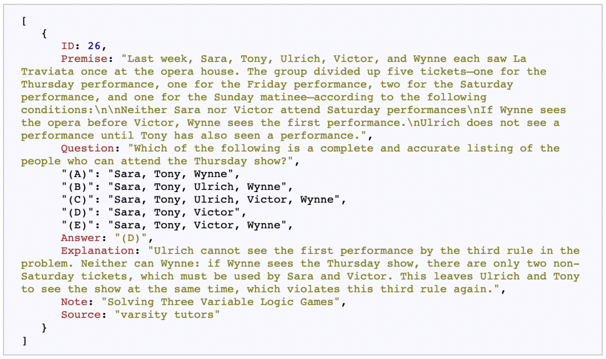
Which of the following is a complete and accurate listing of the people who can attend the Thursday show?

(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



```
[
    {
        ID: 26,
        Premise: "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:\n\nNeither Sara nor Victor attend Saturday performances\nIf Wynne sees
the opera before Victor, Wynne sees the first performance.\nUlrich does not see a
performance until Tony has also seen a performance.",
        Question: "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",
        Answer: "(D)",
        Explanation: "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.",
        Note: "Solving Three Variable Logic Games",
        Source: "varsity tutors"
    }
]
```

Figure 3. Example JSON Entry - CReDETS Dataset

to not just choose the right answer options (MCQ) but also to explain reasoning for each question.

3.1. Sourcing and Structure of CreDeTS

We hand-curated and processed 442 questions from a variety of online sources, including Varsity Tutors [20]. The questions and their corresponding explanations are written and vetted by professors and law school admissions professionals around the country. Each question was analyzed and categorized into one of 5 problem types: Grouping, Sequencing, Four/Five Variable, Three Variable, and Two Variable. The distribution of problem types in the preliminary dataset is shown in Figure 2 below.

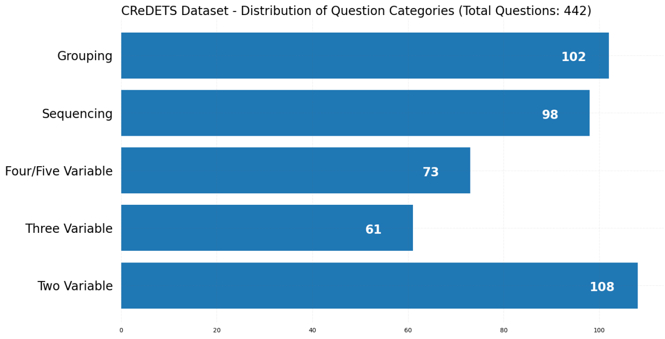


Figure 2. CReDETS Dataset - Distribution of Question Categories (Total Questions: 442)

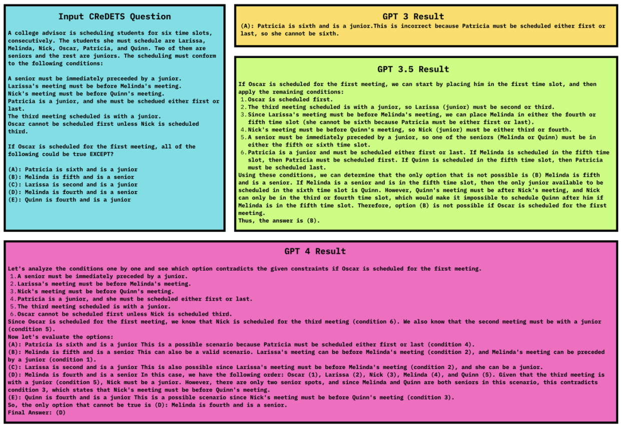
Each sample was separated into a premise, question, set of answer choices, correct answer, and explanation, organized into a JSON file structure as shown in Figure 3.

Further examples can be found in the appendix.

4. LLM Results on CReDETS

The CReDETS dataset is a novel, first-of-its-kind dataset that allows us to benchmark both causal reasoning ability *and* ability to explain a model's reasoning. In order to see some real life results on CReDETS, we ran 10 trials of the 442 questions in the preliminary dataset. Questions were tested on GPT 3, 3.5, and 4 via API calls. In this section, we show some examples of our results, discuss their accuracy and consistency, analyze common errors found in the generated text, and discuss the implications of our results.

4.1. Examples



Input CReDETS Question

A college advisor is scheduling students for six time slots, consecutively. The students she must schedule are Larissa, Melinda, Nick, Oscar, Patricia, and Quinn. Two of them are seniors and the rest are juniors. The scheduling must conform to the following conditions:

A senior must be immediately preceeded by a junior.
Larissa's meeting must be before Melinda's meeting.
Nick's meeting must be before Quinn's meeting.
Patricia is a junior, and she must be schedued either first or last.
The third meeting scheduled is with a junior.
Oscar cannot be scheduled first unless Nick is scheduled third.

If Oscar is scheduled for the first meeting, all of the following could be true EXCEPT?

(A): Patricia is sixth and is a junior
(B): Melinda is fifth and is a senior
(C): Larissa is second and is a junior
(D): Melinda is fourth and is a senior
(E): Quinn is fourth and is a junior

GPT 3 Result

(A): Patricia is sixth and is a junior.This is incorrect because Patricia must be scheduled either first or last, so she cannot be sixth.

GPT 3.5 Result

If Oscar is scheduled for the first meeting, we can start by placing him in the first time slot, and then apply the remaining conditions:
1. Oscar is scheduled first.
2. The third meeting scheduled is with a junior, so Larissa (junior) must be second or third.
3. Since Larissa's meeting must be before Melinda's meeting, we can place Melinda in either the fourth or fifth time slot (she cannot be sixth because Patricia must be either first or last).
4. Nick's meeting must be before Quinn's meeting, so Nick (junior) must be either third or fourth.
5. A senior must be immediately preceded by a junior, so one of the seniors (Melinda or Quinn) must be in either the fifth or sixth time slot.
6. Patricia is a junior and must be scheduled either first or last. If Melinda is scheduled in the fifth time slot, then Patricia must be scheduled first. If Quinn is scheduled in the fifth time slot, then Patricia must be scheduled last.

Using these conditions, we can determine that the only option that is not possible is (B) Melinda is fifth and is a senior. If Melinda is a senior and is in the fifth time slot, then the only junior available to be scheduled in the sixth time slot is Quinn. However, Quinn's meeting must be after Nick's meeting, and Nick can only be in the third or fourth time slot, which would make it impossible to schedule Quinn after him if Melinda is in the fifth time slot. Therefore, option (B) is not possible if Oscar is scheduled for the first meeting.
Thus, the answer is (B).

GPT 4 Result

Let's analyze the conditions one by one and see which option contradicts the given constraints if Oscar is scheduled for the first meeting.
1. A senior must be immediately preceded by a junior.
2. Larissa's meeting must be before Melinda's meeting.
3. Nick's meeting must be before Quinn's meeting.
4. Patricia is a junior, and she must be scheduled either first or last.
5. The third meeting scheduled is with a junior.
6. Oscar cannot be scheduled first unless Nick is scheduled third.

Since Oscar is scheduled for the first meeting, we know that Nick is scheduled for the third meeting (condition 6). We also know that the second meeting must be with a junior (condition 5).
Now let's evaluate the options:
(A): Patricia is sixth and is a junior This is a possible scenario because Patricia must be scheduled either first or last (condition 4).
(B): Melinda is fifth and is a senior This can also be a valid scenario. Larissa's meeting can be before Melinda's meeting (condition 2), and Melinda's meeting can be preceded by a junior (condition 1).
(C): Larissa is second and is a junior This is also possible since Larissa's meeting must be before Melinda's meeting (condition 2), and she can be a junior.
(D): Melinda is fourth and is a senior In this case, we have the following order: Oscar (1), Larissa (2), Nick (3), Melinda (4), and Quinn (5). Given that the third meeting is with a junior (condition 5), Nick must be a junior. However, there are only two senior spots, and since Melinda and Quinn are both seniors in this scenario, this contradicts condition 3, which states that Nick's meeting must be before Quinn's meeting.
(E): Quinn is fourth and is a junior This is a possible scenario since Nick's meeting must be before Quinn's meeting (condition 3).
So, the only option that cannot be true is (D): Melinda is fourth and is a senior.
Final Answer: (D)

Figure 4. A question sampled from the CReDETS dataset and the responses of GPT 3, GPT 3.5, and GPT 4.

4.2. Accuracy

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.

Table 1. Test accuracies of GPT3, GPT3.5, and GPT4 on the questions in preliminary CReDETS Dataset. Each trial of 442 questions was run 10 times to measure consistency and average accuracy.

	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 in Table 1, all three models (GPT3, GPT3.5, GPT4) perform quite poorly on the CReDETS benchmark. GPT 3 and GPT 3.5 display accuracies close to 0.2 (.198 and .207, respectively), which is quite close to a 0.2 random chance decision, given that there are 5 multiple choice answers. These models also perform quite erratically - as we will discuss later, running each question several times yielded several different answers. While these differences mostly averaged out over all 10 trials and 442 questions, we can still see the inconsistency of these models in the accuracy numbers (in the table and visually displayed in Figure 5).

GPT 4's performance is an improvement to its predecessors - it correctly answers on average 1 in 4 questions, which, while not at all close to human-level accuracy, is a marked improvement from the 1-in-5 accuracy of GPT 3 and GPT 3.5. Most importantly, GPT 4 displays increased levels of consistency - running the same question multiple times results in the same one (or two) answers, which we can see by the general similarity and smoothness of the accuracies in the GPT 4 column of Table 1 and the increased smoothness of the GPT 4 results in Figure 5. We will elaborate on this consistency in Section 4.3.

4.3. Consistency

When it comes to using reasoning ability as a conduit to elicit reasoning from large language models, it is crucial that there is some consistency in the text results they generate. If we run the same question 5 different times, will we get five different answers? Even if we get the same (correct or incorrect) answer, are the steps the model takes to get to the final answer similar? If the model lacks this consistency, it is difficult to trust that any model output can be used as a tool to ”explain” the model's ”decision making”.

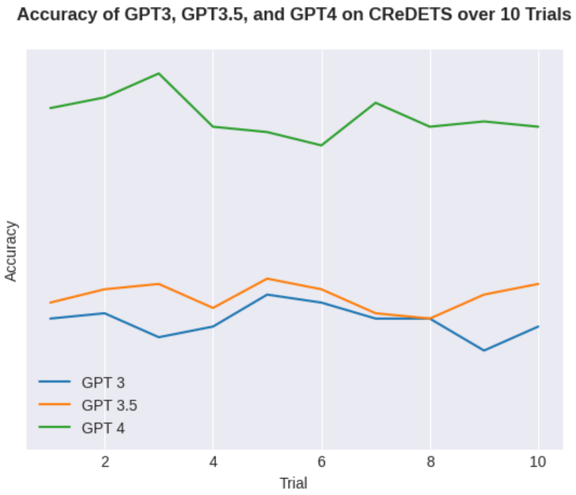


Figure 5. Test accuracies of GPT3, GPT3.5, and GPT4 on the 442 questions in preliminary CReDETS Dataset over 10 trials.

As we can see in Table 1, the number of questions that GPT 3 and GPT 3.5 answer correctly over the 10 trials varies more than the number of questions that GPT 4 answers correctly. This suggests that GPT 4 is significantly more consistent in its results than GPT 3 and GPT 3.5, even if the results are incorrect.

To analyze this further, we make a plot of the average number of distinct final answer choices made over the 10 trials by each of our three models. (If, for example, over 10 trials on a question, my model answered (A), (B), (C), (C), (B), (A), (A), (C), (B), (C), the number of distinct final answer choices would be 3).

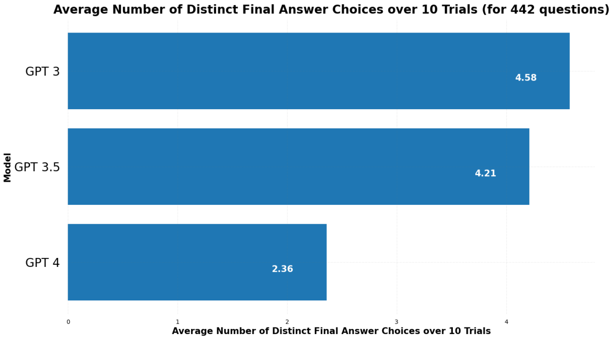


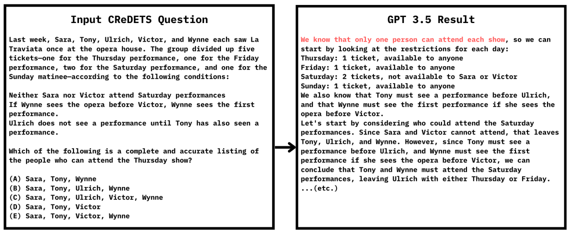
Figure 6. Average Number of Distinct Final Answer Choices Made Over 10 Trials by GPT 3, GPT 3.5, and GPT 4

As we can see in Figure 6, GPT 4 displays a great deal higher consistency than GPT 3 and 3.5. When we run a question 10 times, it usually outputs the same answer, even if that answer is incorrect. When the final generated answer choice is incorrect, manual analysis shows that the mistakes made by the model are similar in almost all of the generated ouptuts. This consistency in GPT 4 bodes well for our ability to truly use generated reasoning output to interpret the decision-making process of GPT 4, lending some sense of ”interpretability” to these outputs.

5. Error Analysis

We hand-analyze the results outputted by GPT3, GPT3.5, and GPT4 on the same, randomly selected set of 25 questions. We found that there were three main categories of errors made by the model. We describe them below.

Factual error relating to problem premise. For example, if the premise of a question states that at most two people can be on a dance committee, and the model output places 4 people on a dance committee, this error would be classified as ”factual” due to its misunderstandings of facts stated in the premise of the question. An example is shown below in Figure 7, where the premise of the question clearly states that there are there are two tickets for the Saturday performance, but the GPT-generated result starts by stating that only one person can attend each show, which is factually incorrect.



Input CReDETS Question

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?

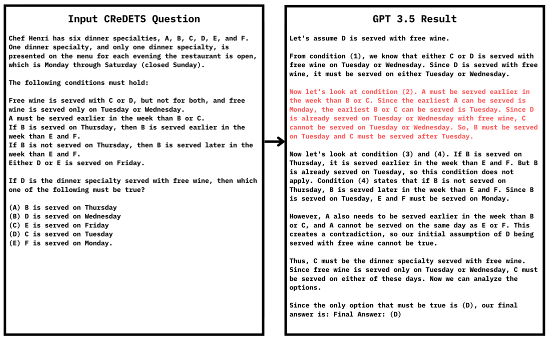
(A) Sara, Tony, Wynne
(B) Sara, Tony, Ulrich, Wynne
(C) Sara, Tony, Ulrich, Victor, Wynne
(D) Sara, Tony, Victor
(E) Sara, Tony, Victor, Wynne

GPT 3.5 Result

We know that only one person can attend each show, so we can start by looking at the restrictions for each day:
Thursday: 1 ticket, available to anyone
Friday: 1 ticket, available to anyone
Saturday: 2 tickets, not available to Sara or Victor
Sunday: 1 ticket, available to anyone
We also know that Tony must see a performance before Ulrich, and that Wynne must see the first performance if she sees the opera before Victor.
Let's start by considering who could attend the Saturday performances. Since Sara and Victor cannot attend, that leaves Tony, Ulrich, and Wynne. However, since Tony must see a performance before Ulrich, and Wynne must see the first performance if she sees the opera before Victor, we can conclude that Tony and Wynne must attend the Saturday performances, leaving Ulrich with either Thursday or Friday.
...(etc.)

Figure 7. The text colored in red displays a Factual Error.

Causality Error. These errors occur when the model makes decisions based on incorrect causality statements. If the problem premise states that A always comes after B, and the model suggests a situation in which A comes before B, we would register a Causality error. An example is shown below in Figure 8. The text colored in red displays a causality error because there are several statements made that do not causally follow from the rules provided in the question. The model says the following: ”Since D is already served on Tuesday or Wednesday with free wine, C cannot be served on Tuesday or Wednesday.” The premise of the question states that ”free wine is served with C or D, but not for both, and free wine is served only on Tuesday or Wednesday”. The premise of the question does not imply that just because D is served on Tuesday or Wednesday with free wine, C *cannot* be served on Tuesday or Wednesday. This is a causality error.



Input CReDETS Question

Chef Henri has six dinner specialties, A, B, C, D, E, and F. One dinner specialty, and only one dinner specialty, is presented on the menu for each evening the restaurant is open, which is Monday through Saturday (closed Sunday).

The following conditions must hold:

Free wine is served with C or D, but not for both, and free wine is served only on Tuesday or Wednesday.
A must be served earlier in the week than B or C.
If B is served on Thursday, then B is served earlier in the week than E and F.
If B is not served on Thursday, then B is served later in the week than E and F.
Either D or E is served on Friday.

If D is the dinner specialty served with free wine, then which one of the following must be true?

(A) B is served on Thursday
(B) D is served on Wednesday
(C) E is served on Friday
(D) C is served on Tuesday
(E) F is served on Monday.

GPT 3.5 Result

Let's assume D is served with free wine.

From condition (1), we know that either C or D is served with free wine on Tuesday or Wednesday. Since D is served with free wine, it must be served on either Tuesday or Wednesday.

Now let's look at condition (2). A must be served earlier in the week than B or C. Since the earliest A can be served is Monday, the earliest B or C can be served is Tuesday. Since D is already served on Tuesday or Wednesday with free wine, C cannot be served on Tuesday or Wednesday. So, B must be served on Tuesday and C must be served after Tuesday.

Now let's look at condition (3) and (4). If B is served on Thursday, it is served earlier in the week than E and F. But B is already served on Tuesday, so this condition does not apply. Condition (4) states that if B is not served on Thursday, B is served later in the week than E and F. Since B is served on Tuesday, E and F must be served on Monday.

However, A also needs to be served earlier in the week than B or C, and A cannot be served on the same day as E or F. This creates a contradiction, so our initial assumption of D being served with free wine cannot be true.

Thus, C must be the dinner specialty served with free wine. Since free wine is served only on Tuesday or Wednesday, C must be served on either of these days. Now we can analyze the options.

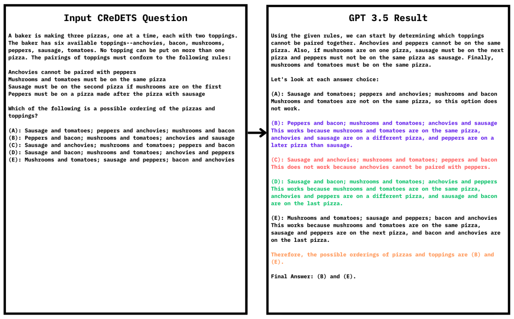
Since the only option that must be true is (D), our final answer is: Final Answer: (D)

Figure 8. The text colored in red displays a Causality Error.

Self Contradiction. The outputs of LLMs (especially more recent ones, such as GPT4) when given reasoning problems are usually lengthy statements. Oftentimes, these models will make a statement in their output that directly contradicts something that was previously generated. We categorize these errors as ”self-contradictory”. Figure 9 shows an example with several self-contradiction errors. In the text colored purple, we can see that the explanation says that peppers are on a later pizza than sausage, but the answer it gives is not on a later pizza than sausage. The text highlighted in red outputs an answer choice that pairs sausage and anchovies as well as peppers and bacon, but the choice is rejected because the anchovies cannot be paired with peppers, a pairing that was never outputted. The text highlighted green says includes an explanation that sausage and bacon are on the last pizza, which is contradictory to the answer choice the model outputted. Lastly, the text highlighted in orange mentions that only choices (B) and (E) work, but the model previously outputted that choice (D) worked as well, so this statement is self-contradictory as well.

We randomly sample 25 questions from the CReDETS dataset and manually analyze results produced by GPT 3, GPT 3.5, and GPT 4. We categorize errors made by the model into the following groups: factual errors, causality errors, and self-contradiction errors. The distribution of errors is shown in Figure 10 below.

We can see in the figure that all three models have a large tendency to make causality errors (even GPT 4, despite its higher accuracy and consistency). GPT 4 makes less factual errors than its predecessors, but it has high rates of self-contradiction, a stark reminder that despite language models



Input CReDETS Question

A baker is making three pizzas, one at a time, each with two toppings. The baker has six available toppings--anchovies, bacon, mushrooms, peppers, sausage, tomatoes. No topping can be put on more than one pizza. The pairings of toppings must conform to the following rules:

Anchovies cannot be paired with peppers
Mushrooms and tomatoes must be on the same pizza
Sausage must be on the second pizza if mushrooms are on the first
Peppers must be on a pizza made after the pizza with sausage

Which of the following is a possible ordering of the pizzas and toppings?

(A): Sausage and tomatoes; peppers and anchovies; mushrooms and bacon
(B): Peppers and bacon; mushrooms and tomatoes; anchovies and sausage
(C): Sausage and anchovies; mushrooms and tomatoes; peppers and bacon
(D): Sausage and bacon; mushrooms and tomatoes; anchovies and peppers
(E): Mushrooms and tomatoes; sausage and peppers; bacon and anchovies

GPT 3.5 Result

Using the given rules, we can start by determining which toppings cannot be paired together. Anchovies and peppers cannot be on the same pizza. Also, if mushrooms are on one pizza, sausage must be on the next pizza and peppers must not be on the same pizza as sausage. Finally, mushrooms and tomatoes must be on the same pizza.

Let's look at each answer choice:

(A): Sausage and tomatoes; peppers and anchovies; mushrooms and bacon
Mushrooms and tomatoes are not on the same pizza, so this option does not work.

(B): Peppers and bacon; mushrooms and tomatoes; anchovies and sausage
This works because mushrooms and tomatoes are on the same pizza, anchovies and sausage are on a different pizza, and peppers are on a later pizza than sausage.

(C): Sausage and anchovies; mushrooms and tomatoes; peppers and bacon
This does not work because anchovies cannot be paired with peppers.

(D): Sausage and bacon; mushrooms and tomatoes; anchovies and peppers
This works because mushrooms and tomatoes are on the same pizza, anchovies and peppers are on a different pizza, and sausage and bacon are on the last pizza.

(E): Mushrooms and tomatoes; sausage and peppers; bacon and anchovies
This works because mushrooms and tomatoes are on the same pizza, sausage and peppers are on the next pizza, and bacon and anchovies are on the last pizza.

Therefore, the possible orderings of pizzas and toppings are (B) and (E).

Final Answer: (B) and (E).

Figure 9. The text colored in purple, red, green, and orange display Self Contradiction Errors.

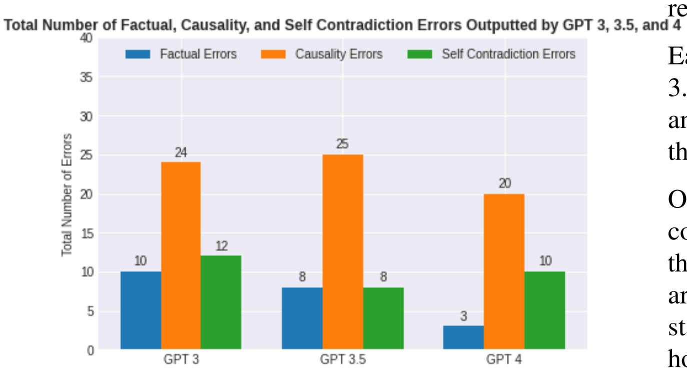


Figure 10. Average Number of Distinct Final Answer Choices Made Over 10 Trials by GPT 3, GPT 3.5, and GPT 4

achieving prolific results, they are still probabilistic algorithms that output their results word by word, as opposed to a more big-ideas / long term reasoning framework that humans use.

Interestingly, the causality errors made by GPT 4 were more likely than those made by its predecessors to belong to an output that resulted in an incorrect final answer. As we can see in Table 2 below, out of the 25 asked questions, GPT 3 incorrectly answered 21 questions, GPT 3.5 incorrectly answered 20, and GPT 4 incorrectly answered 18. The number of Causality Errors made in outputs that resulted in the correct final answer, however, decreased in GPT 4's output, another sign that there is a higher probability that we can trust the reasoning output of GPT 4 to be a form of interpretability for the model's decisions.

Table 2. Causality Errors made by GPT 3, 3.5, and 4 in outputs that generated the wrong final answer versus outputs that generated the correct final answer. Manually tabulated over 25 questions.

		Causality Errors		
	Incorrect Answers	*in Correct Answers*	*in Incorrect Answers*	*Total Causality Errors*
GPT 3	21	3	21	24
GPT 3.5	20	3	22	25
GPT 4	18	1	19	20

6. Discussion

Our results on the preliminary CReDETS dataset show that while cutting-edge models such as GPT 4 display impressive, coherent language modeling abilities, they have a long way to go before coming close to reaching human-ability levels of reasoning.

All three models show poor accuracies on the CReDETS reasoning benchmark questions - with each question having 5 multiple choice answer possibilities, GPT 3, GPT 3.5, and GPT 4 displayed average accuracies of .20, .21, and .25 respectively.

Earlier models, especially GPT 3 and occasionally GPT 3.5, prove to be inconsistent, often generating distinct final answer choices if run on the same problem many times, though this problem is significantly ameliorated in GPT 4.

Outputs by all three language models are riddled with self-contradictions and factual errors, a fact which reminds us that despite their prowess and size, Large Language Models are in fact just that - models. They compose an output by statistically predicting one word at a time - true reasoning, however, involves thinking several steps ahead and making logical decisions by concept instead of on a word-by-word basis. The high rate of causality errors shown by the GPT models reinforces this notion and makes clear the need for work evaluating how LLMs can improve their base causal reasoning ability.

At this stage, our results show that it is difficult to claim that even today's most powerful language models are capable of true reasoning. Our results showed generated text that was riddled with self-contradictions, causality errors, hallucinations, factual errors, and several other mistakes that suggest that large language models are likely simply parroting their powerfully large training corpora. Major improvements need to be made before we can recognize their abilities as true reasoning, and as LLMs are integrated deeper into our society, great care needs to be taken to ensure that false abilities of "reasoning" do not result in dangerous mistakes or implications.

When thinking about how reasoning can be useful in our paradigm of explainability, it is interesting to note how our results have improved as we test GPT 3, 3.5, and 4. The accuracy, though low for all three models, improved as the parameter size increased. The coherency and intricacy of

	Accurate	Consistent	Coherent
GPT 3	🟥	🟥	🟥
GPT 3.5	🟥	🟥	🟩
GPT 4	🟥	🟩	🟩

Table 3. Visual depiction of the Accuracy, Consistency, and Coherency of GPT 3, 3.5, and 4 on the CReDETS dataset

produced output improved significantly from GPT 3 to GPT 3.5 to GPT 4. The consistency of the models also improved significantly as the number of parameters increased.

These Large Language Models have shown remarkable improvement and give credence to the hypothesis [8] that increasing the number of model parameters will lead to more accurate modeling of human language.

This improvement in coherency and consistency that GPT 4 has shown demonstrates the possibility that the reasoning steps outputted by the model actually provide a correct, reasonable explanation of an internal "decision process". (In GPT 3, for instance, there were several questions where the model happened to get a correct final answer but the reasoning steps outputted had causal, factual, and self-contradiction errors. In this case, we cannot take the outputted reasoning steps to provide any semblance of interpretability.) As we continue to explore how to derive a sense of explainability from language models where, for the most part, we only have access to an input and output (API call), the improved results displayed by GPT 4 suggest that the reasoning steps outputted by a model could perhaps, in the future, be used as a source of interpretability.

It is our hope that the novel CReDETS dataset, with its unique combination of hand-written explanations and questions designed explicitly to test reasoning ability, will allow researchers to continue to study and evaluate the evolving reasoning abilities of language models.

References

[1] Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. "Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation". In: *CoRR* abs/1308.3432 (2013). arXiv: 1308.3432. URL: http://arxiv.org/abs/1308.3432.

[2] Todd Bishop. *Microsoft confirms Bing's AI-powered search chatbot is running on OpenAI's new GPT-4*. https://www.geekwire.com/2023/microsoft-confirms-bings-ai-powered-search-chatbot-is-running-on-openais-new-gpt-4/. 2023.

[3] S'ebastien Bubeck et al. "Sparks of Artificial General Intelligence: Early experiments with GPT-4". In: 2023.

[4] Supriyo Chakraborty and etc. Tomsett Richard. "Interpretability of Deep Learning Models: A Survey of Results". In: *arXiV* (2017).

[5] Andrew Chow. *How ChatGPT Managed to Grow Faster Than TikTok or Instagram*. https://time.com/6253615/chatgpt-fastest-growing/. 2023.

[6] Shizhe Diao et al. *Active Prompting with Chain-of-Thought for Large Language Models*. 2023. arXiv: 2302.12246 [cs.CL].

[7] Geoffrey Fowler. *Snapchat tried to make a safe AI. It chats with me about booze and sex*. https://www.washingtonpost.com/technology/2023/03/14/snapchat-myai/. 2023.

[8] Jared Kaplan et al. *Scaling Laws for Neural Language Models*. 2020. arXiv: 2001.08361 [cs.LG].

[9] Takeshi Kojima et al. "Large Language Models are Zero-Shot Reasoners". In: *ArXiv* abs/2205.11916 (2022).

[10] Jian Liu et al. *LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning*. 2020. arXiv: 2007.08124 [cs.CL].

[11] Angus Loten. *Instacart Joins ChatGPT Frenzy, Adding Chatbot To Grocery Shopping App*. https://www.wsj.com/articles/instacart-joins-chatgpt-frenzy-adding-chatbot-to-grocery-shopping-app-bc8a2d3c. 2023.

[12] Scott M. Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *CoRR* abs/1705.07874 (2017). arXiv: 1705.07874. URL: http://arxiv.org/abs/1705.07874.

[13] Aisha Malik. *Duolingo launches new subscription tier with access to AI tutor powered by GPT-4*. https://techcrunch.com/2023/03/14/duolingo-launches-new-subscription- tier-with-access-to-ai-tutor-powered-by-gpt-4/. 2023.

[14] Aisha Malik. *Snapchat launches an AI chatbot powered by OpenAI's GPT technology*. https://techcrunch.com/2023/02/27/snapchat-launches-an-ai-chatbot-powered-by-openais-gpt-technology/. 2023.

[15] OpenAI. "GPT-4 Technical Report". In: *arXiV* (2023).

[16] OpenAI. *Introducing ChatGPT*. https://openai.com/blog/chatgpt. 2022.

[17] Archit Parnami and Minwoo Lee. *Learning from Few Examples: A Summary of Approaches to Few-Shot Learning*. 2022. arXiv: 2203.04291 [cs.LG].

[18] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: 2019. (Visited on 02/02/2023).

[19] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "“Why Should I Trust You?”: Explaining the Predictions of Any Classifier". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144.

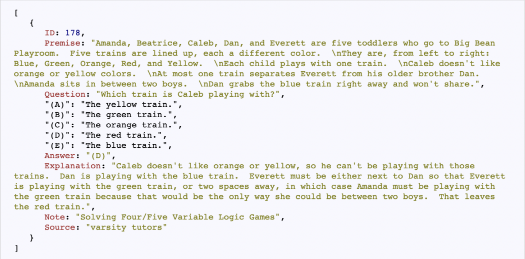
[20] Varisty Tutors. *Varsity Tutors*. https://www.varsitytutors.com/. 2023.

[21] Talip Ucar et al. *One-Shot Learning for Language Modelling*. 2020. arXiv: 2007.09679 [cs.CL].

[22] Ashish Vaswani et al. "Attention is All You Need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6000–6010. ISBN: 9781510860964.

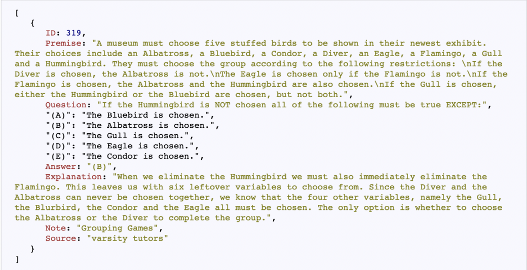
A. CReDETS Dataset Examples

Below, we include some further examples of data points from the CReDETS dataset.



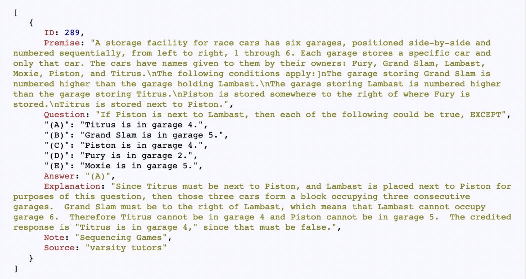
```
[
   {
      ID: 178,
      Premise: "Amanda, Beatrice, Caleb, Dan, and Everett are five toddlers who go to Big Bean
Playroom.  Five trains are lined up, each a different color.  \nThey are, from left to right:
Blue, Green, Orange, Red, and Yellow.  \nEach child plays with one train.  \nCaleb doesn't like
orange or yellow colors.  \nAt most one train separates Everett from his older brother Dan.
\nAmanda sits in between two boys.  \nDan grabs the blue train right away and won't share.",
      Question: "Which train is Caleb playing with?",
      "(A)": "The yellow train.",
      "(B)": "The green train.",
      "(C)": "The orange train.",
      "(D)": "The red train.",
      "(E)": "The blue train.",
      Answer: "(D)",
      Explanation: "Caleb doesn't like orange or yellow, so he can't be playing with those
trains.  Dan is playing with the blue train.  Everett must be either next to Dan so that Everett
is playing with the green train, or two spaces away, in which case Amanda must be playing with
the green train because that would be the only way she could be between two boys.  That leaves
the red train.",
      Note: "Solving Four/Five Variable Logic Games",
      Source: "varsity tutors"
   }
]
```

Figure 11. Example JSON Entry - CReDETS Dataset



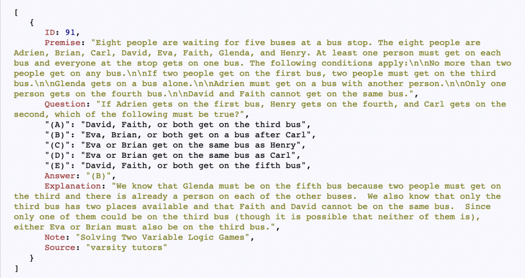
```
[
   {
      ID: 319,
      Premise: "A museum must choose five stuffed birds to be shown in their newest exhibit.
Their choices include an Albatross, a Bluebird, a Condor, a Diver, an Eagle, a Flamingo, a Gull
and a Hummingbird. They must choose the group according to the following restrictions: \nIf the
Diver is chosen, the Albatross is not.\nThe Eagle is chosen only if the Flamingo is not.\nIf the
Flamingo is chosen, the Albatross and the Hummingbird are also chosen.\nIf the Gull is chosen,
either the Hummingbird or the Bluebird are chosen, but not both.",
      Question: "If the Hummingbird is NOT chosen all of the following must be true EXCEPT:",
      "(A)": "The Bluebird is chosen.",
      "(B)": "The Albatross is chosen.",
      "(C)": "The Gull is chosen.",
      "(D)": "The Eagle is chosen.",
      "(E)": "The Condor is chosen.",
      Answer: "(B)",
      Explanation: "When we eliminate the Hummingbird we must also immediately eliminate the
Flamingo. This leaves us with six leftover variables to choose from. Since the Diver and the
Albatross can never be chosen together, we know that the four other variables, namely the Gull,
the Blurbird, the Condor and the Eagle all must be chosen. The only option is whether to choose
the Albatross or the Diver to complete the group.",
      Note: "Grouping Games",
      Source: "varsity tutors"
   }
]
```

Figure 12. Example JSON Entry - CReDETS Dataset



```
[
   {
      ID: 289,
      Premise: "A storage facility for race cars has six garages, positioned side-by-side and
numbered sequentially, from left to right, 1 through 6. Each garage stores a specific car and
only that car. The cars have names given to them by their owners: Fury, Grand Slam, Lambast,
Moxie, Piston, and Titrus.\nThe following conditions apply:]nThe garage storing Grand Slam is
numbered higher than the garage holding Lambast.\nThe garage storing Lambast is numbered higher
than the garage storing Titrus.\nPiston is stored somewhere to the right of where Fury is
stored.\nTitrus is stored next to Piston.",
      Question: "If Piston is next to Lambast, then each of the following could be true, EXCEPT",
      "(A)": "Titrus is in garage 4.",
      "(B)": "Grand Slam is in garage 5.",
      "(C)": "Piston is in garage 4.",
      "(D)": "Fury is in garage 2.",
      "(E)": "Moxie is in garage 5.",
      Answer: "(A)",
      Explanation: "Since Titrus must be next to Piston, and Lambast is placed next to Piston for
purposes of this question, then those three cars form a block occupying three consecutive
garages.  Grand Slam must be to the right of Lambast, which means that Lambast cannot occupy
garage 6.  Therefore Titrus cannot be in garage 4 and Piston cannot be in garage 5.  The credited
response is "Titrus is in garage 4," since that must be false.",
      Note: "Sequencing Games",
      Source: "varsity tutors"
   }
]
```

Figure 13. Example JSON Entry - CReDETS Dataset



```
[
   {
      ID: 91,
      Premise: "Eight people are waiting for five buses at a bus stop. The eight people are
Adrien, Brian, Carl, David, Eva, Faith, Glenda, and Henry. At least one person must get on each
bus and everyone at the stop gets on one bus. The following conditions apply:\n\nNo more than two
people get on any bus.\n\nIf two people get on the first bus, two people must get on the third
bus.\n\nGlenda gets on a bus alone.\n\nAdrien must get on a bus with another person.\n\nOnly one
person gets on the fourth bus.\n\nDavid and Faith cannot get on the same bus.",
      Question: "If Adrien gets on the first bus, Henry gets on the fourth, and Carl gets on the
second, which of the following must be true?",
      "(A)": "David, Faith, or both get on the third bus",
      "(B)": "Eva, Brian, or both get on a bus after Carl",
      "(C)": "Eva or Brian get on the same bus as Henry",
      "(D)": "Eva or Brian get on the same bus as Carl",
      "(E)": "David, Faith, or both get on the fifth bus",
      Answer: "(B)",
      Explanation: "We know that Glenda must be on the fifth bus because two people must get on
the third and there is already a person on each of the other buses.  We also know that only the
third bus has two places available and that Faith and David cannot be on the same bus.  Since
only one of them could be on the third bus (though it is possible that neither of them is),
either Eva or Brian must also be on the third bus.",
      Note: "Solving Two Variable Logic Games",
      Source: "varsity tutors"
   }
]
```

Figure 14. Example JSON Entry - CReDETS Dataset