Ellipsis-Dependent Reasoning: a New Challenge for Large Language Models

Daniel Hardt
Copenhagen Business School
dha.msc@cbs.dk

Abstract

We propose a novel challenge for large language models: ellipsis-dependent reasoning. We define several structures of paired examples, where an ellipsis example is matched to its non-ellipsis counterpart, and a question is posed which requires resolution of the ellipsis. Test results show that the best models perform well on non-elliptical examples but struggle with all but the simplest ellipsis structures.

1 Introduction

Ellipsis is a fundamental feature of human language, occurring in all registers, where parts of sentences are omitted, although the missing parts are essential for understanding the meaning. The following is an example of Verb Phrase Ellipsis (VPE)(Bos and Spenader, 2011):

(1) William went running. Harold did too.

(1) is understood as asserting that Harold went running; that is, the hearer or reader naturally fills in the missing material. This is done by identifying the antecedent VP, *went running*, in the first sentence. The following is the non-elliptical counterpart of (1):

(2) William went running. Harold went running too.

With such examples, we can test understanding of ellipsis by targeting the ellipsis phrase with a simple Yes/No question:

(3) Did Harold go running?

If a system answers the question incorrectly for (1), the ellipsis example, but answers correctly for (2), the non-elliptical counterpart of (1), we can ascribe the result specifically to the challenge of ellipsis, since the examples are otherwise identical.

As with pronominal anaphora and other discourse processes, there is great flexibility in the way ellipsis can occur in discourse. Example (1) involves two simple adjacent sentences. It is also possible for ellipsis or the antecedent to occur in embedded clauses. Furthermore, ellipsis can occur either before or after the antecedent. Finally, an arbitrary amount of material can intervene between the antecedent and the ellipsis occurrence.

In this paper, we propose the challenge of ellipsis-dependent reasoning. This challenge consists of examples involving an ellipsis clause, the target. Each ellipsis example is paired with its non-elliptical counterpart, where the target clause is overt rather than elliptical. We then pose a question whose answer is dependent on the target clause. A key aspect of the challenge is that ellipsis occurrences are possible in a variety of diverse structural configurations. We test a series of GPT-3 models (GPT) on several such ellipsis structures.

2 Related Work

There is a large literature concerning the probing of language models from a variety of perspectives. Furthermore, there has been substantial work specifically addressing ellipsis in NLP. In this paper, we are proposing the challenge of ellipsis-dependent reasoning. This proposal builds on various strands of prior research; below we consider some particularly relevant aspects of this literature.

2.1 Probing Models for Knowledge

The Winograd Schema (Kocijan et al. (2022); Levesque et al. (2012)) involves test examples that use the linguistic problem of pronoun resolution to gain insight into the commonsense reasoning abilities of an AI system. To do this, the Winograd Schema requires pairs of examples that differ only in one specific, small way, as in (4):

(4) The city councilmen refused the demonstrators a permit because they feared/advocated violence.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 39–47

With “feared", the pronoun “they" refers to the city councilmen, while with “advocated", it refers to the demonstrators. Humans understand this because of general, commonsense knowledge about what would reasonably explain the refusal of a permit in the two cases. It is difficult to ensure that such examples can *only* be solved through such sophisticated reasoning, and, according to Kocijan et al. (2022)[p. 8], “Solving Winograd schemas is not a surrogate for the ability to do commonsense reasoning".

A different approach is exemplified by Lin et al. (2019): here, examples are constructed which test specific aspects of linguistic knowledge of a system, namely, whether BERT embeddings “encode hierarchical information". For example, a task is defined to identify the main auxiliary verb in a sentence, even in cases where the main auxiliary is not the first auxiliary verb to appear. Training and testing datasets are automatically generated using a context-free grammar for several such tasks involving hierarchical syntactic information.

2.2 Anaphora and Question Answering

Quoref (Dasigi et al. (2019); Zhang and Zhao (2022)) is a question-answer dataset designed so that correct answers cannot be given unless a coreference relationship is correctly identified; that is, the reasoning involved in question answering is dependent on resolving coreference. This is, in a sense, the inverse of the Winograd schema, where resolving coreference is dependent upon reasoning. Just as with the Winograd schema, it is difficult to ensure that resolving this dependency is required for system success. (Dasigi et al., 2019)[p. 1] note that this is “challenging, because it is hard to avoid lexical cues that shortcut complex reasoning", and based on a random sample, found that coreference resolution was required for 78% of questions.

2.3 Ellipsis as a Task

There has been substantial work on ellipsis as a discrete NLP task (Khullar (2020), Zhang et al. (2019); Kenyon-Dean et al. (2016); Bos and Spenader (2011)). Vanderlyn et al. (2022) surveys a variety of forms of what they call “implicit reference", which includes ellipsis and related phenomena. Aralikatte et al. (2021) frame ellipsis as a question-answering task, i.e., a task of locating an antecedent, understood as a span of tokens in context. Aralikatte et al. (2021) report token F1 scores of 78.66 for VPE and 86.01 for sluicing, another form of ellipsis. It’s important to note that the task here, of antecedent identification, is a sub-part of the ellipsis challenge. Before the antecedent is identified, an ellipsis occurrence must be identified, and after the antecedent is identified, it must be interpreted, or “reconstructed", at the ellipsis site.

2.4 Relevance for Ellipsis-Dependent Reasoning

The specific task of ellipsis is addressed in work like that of Aralikatte et al. (2021), but the key difference here is that we are probing for a complete solution to the ellipsis problem. The proposed ellipsis-dependent reasoning task involves a question that can only be answered correctly if the ellipsis is properly identified and interpreted. This combines aspects of the preceding works in a novel way: like the Winograd schema and the syntactic work by Lin et al. (2019), it probes for what we see as a specific type of psychologically-defined knowledge: namely, a representation of context that supports the resolution of ellipsis. Similarly to the work on Quoref, we use targeted questions to probe for discourse-related knowledge.

There is an extensive literature on the contextual interpretation of natural language, resting on the idea of a dynamic, ongoing model of discourse. For example, Discourse Representation Theory (Kamp, 1981) describes a semantic model supporting discourse phenomena such as pronominal and temporal anaphora, and Sag and Hankamer (1984) argue explicitly that ellipsis and other such phenomena are interpreted with respect to a discourse model (Garnham, 2010). As one study puts it, “Interpreting a verb-phrase ellipsis (VP ellipsis) requires accessing an antecedent in memory, and then integrating a representation of this antecedent into the local context" (Martin and McElree, 2008). In this paper, we seek to determine whether a large language model is capable of such an interpretive process.

3 Data

There is a great deal of variety in the structural configurations in which ellipsis can occur. In tables 1 and 2 we define structures for ellipsis and antecedent occurrences.

In all structures, there is an ellipsis occurrence, and the question targets the ellipsis occurrence. Furthermore, each ellipsis example is paired with a non-ellipsis version. The first two structures, Sep-

Structure	Example
Separate Sentence	William went running. John did too.
Conjoined Sentence	William went running, and John did too.
Subordinate Antecedent	Because William went running, John did.
Subordinate VPE	William went running after John did.
Backwards	Because John did, William went running.
Two Actions	William didn't go running but John did. William went shopping and John didn't.
Question	Did John go running?

Table 1: Structures for Positive Answer

Structure	Example
Separate Sentence	William went running. But John didn't.
Conjoined Sentence	William went running, but John didn't.
Subordinate Antecedent	Because William went running, John didn't.
Subordinate VPE	William went running after John didn't.
Backwards	Because John didn't, William went running.
Two Actions	William didn't go shopping but John did. William went running and John didn't.
Question	Did John go running?

Table 2: Structures for Negative Answer

arate Sentence, and Conjoined Sentence, involve two adjacent main clauses. This is followed by two structures in which either the VPE or antecedent occur in a subordinate clause. There is a Backwards structure where the VPE precedes the antecedent; here the VPE is in a subordinate clause. Finally, we have a Two Actions structure; that is, two ellipsis occurrences each with their respective antecedent VPs. We have two versions: one in which the target question has a correct answer of "Yes", shown in table 1, and another where the target question has a correct answer of "No", shown in table 2.

We generate large numbers of examples of each structure by performing random substitutions for both the subject and verb. The substitution lists are given in the appendix, along with samples of each structure and the size of the resulting sets. [1]

4 Test

For each instantiation of a given structure, we produce paired ellipsis and non-ellipsis examples, with an associated Yes/No question. We randomly select 1000 examples for each structure, including 500 ellipsis examples and 500 examples which are their non-elliptical counterparts. Each example is presented to the system, preceded by the text, "Please give a Yes or No answer:". We test five GPT-3 models on these structures: Davinci-003, Davinci-002, Curie-001, Babbage-001, and Ada-001. According to the GPT-3 documentation, Davinci-003 is the most powerful model and Ada-001, the least.

5 Results

Figure 1 gives the accuracy for ellipsis and non-ellipsis, for each of the five models. We have set up the test examples so that an ellipsis example is paired with a non-ellipsis example that is otherwise identical. Because of this, we claim that the difference in accuracy of the non-ellipsis case vs. the ellipsis case provides a measurement of the difficulty specifically posed by ellipsis. For all but the least powerful model, Ada, the non-ellipsis accuracy is substantially higher than ellipsis accuracy, supporting the hypothesis that ellipsis-dependent reasoning presents a difficult challenge for these models. While the Ada model actually performs somewhat better for ellipsis than non-ellipsis, this is not because the Ada model does well with ellipsis cases; rather, the model has great difficulty with both the ellipsis and non-ellipsis cases, and is close to a random guessing baseline of .50.

In figures 2 through 6, we present results for each model. We show the accuracy for each structure, for both the ellipsis version and the non-ellipsis version. Consider the most powerful models, Davinci-003 and Davinci-002. In figures 2 and 3, we can see that ellipsis is not difficult in the first two structures: 2Sent (Separate Sentence) and 1Sent (Conjoined Sentence). Here the accuracy is nearly perfect for

[1] Datasets and associated programs can be accessed at https://github.com/DanHardtDK/ellipsisGPT3.

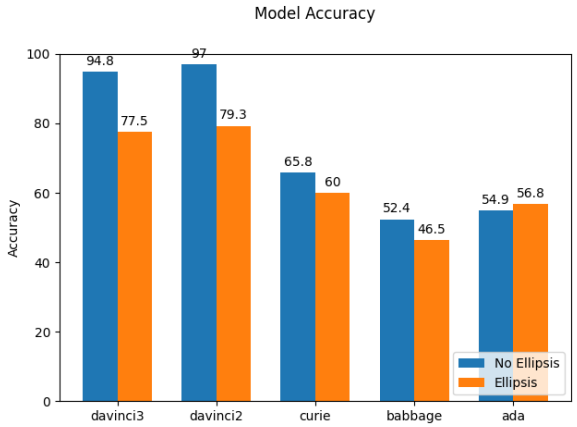

Figure 1: Model Accuracy – Ellipsis vs. Non-Ellipsis (5000 examples per model.)

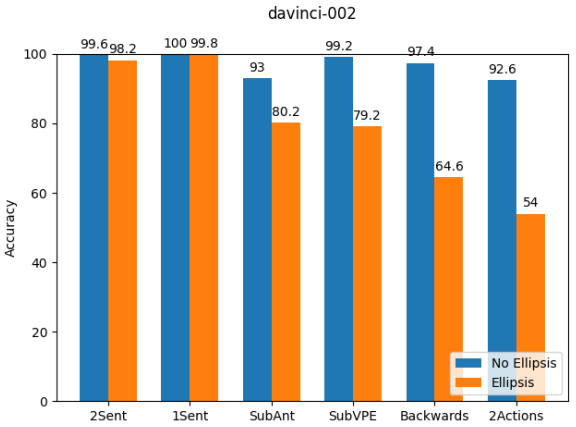

Figure 3: Ellipsis Accuracy – davinci-002 (1000 examples per structure.)

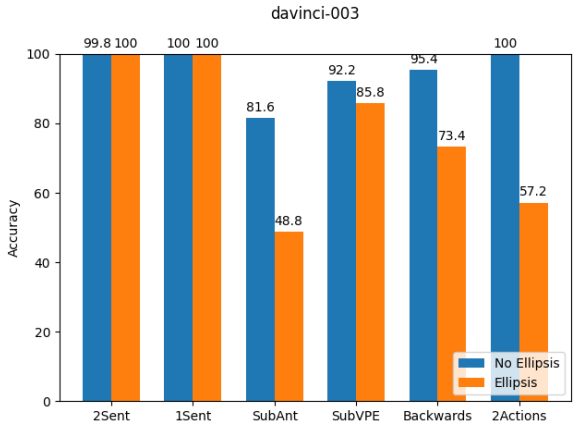

Figure 2: Ellipsis Accuracy – davinci-003 (1000 examples per structure.)

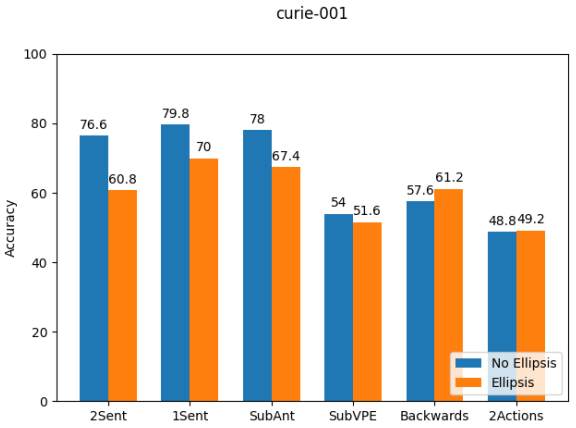

Figure 4: Ellipsis Accuracy – curie-001 (1000 examples per structure.)

the ellipsis as well as the non-ellipsis condition. However, in all the other structures, there is a large divergence in accuracy between ellipsis and non-ellipsis, for both the Davinci-003 and Davinci-002 models. Subordination for either antecedent or ellipsis is quite challenging, with accuracies ranging from 48.8 to 85.8. The Backwards and Two Actions structures are even more difficult for ellipsis.

6 Analysis

For the two most powerful models, it is clear that ellipsis poses a difficult challenge, except in the two simplest ellipsis structures. For the less powerful models, the picture is mixed. For these models, the non-ellipsis examples are themselves a difficult challenge, so we are not able to observe the specific difficulties posed by ellipsis.

As we can see in figure 1, the Davinci-002 model performs somewhat better overall than Davinci-003, on both ellipsis and non-ellipsis. However, figures 2 and 3 show that the advantage of Davinci-002 on ellipsis is exclusively due to the subordinate antecedent construction. In every other ellipsis structure, Davinci-003 performs better than Davinci-002.

There are striking differences in the distribution of errors. For both the Davinci-003 and Davinci-002 models, errors are nearly always false negatives – that is, incorrect “No" answers. There are virtually no false positives, either for the ellipsis case or non-ellipsis case. For the other three models, there are many errors of each type, with a much higher ratio of false positives.

7 Conclusion

Most of the current rapid progress in NLP is due to pre-trained large language models. GPT-3 is an impressive publicly available collection of such

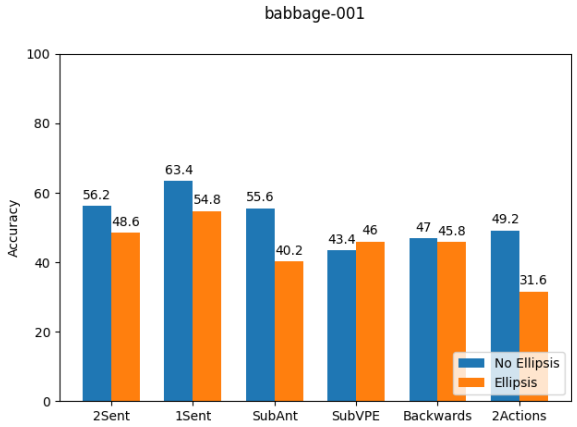

Figure 5: Ellipsis Accuracy – babbage-001 (1000 examples per structure.)

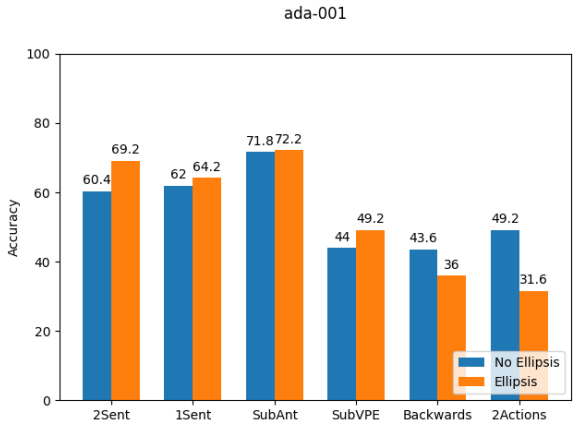

Figure 6: Ellipsis Accuracy – ada-001 (1000 examples per structure.)

models, and is able to perform in a way that suggests human-level understanding. Because of this, it is important to explore areas in which it might still differ from human language understanding. In this paper we have argued that ellipsis is one such area. For many simple ellipsis structures, the most powerful GPT-3 models struggle, with accuracies far lower on ellipsis examples than on non-elliptical counterparts.

In many ways, GPT-3 appears to understand the texts that it processes, often being able to answer questions that appear to rely on sophisticated reasoning. However, the challenge of ellipsis-dependent reasoning provides evidence that GPT-3 is not able to understand in anything like the way humans do.

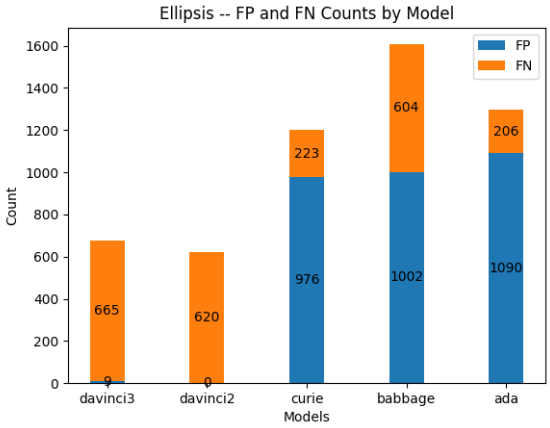

Figure 7: Ellipsis case – Errors by model. False Positives (Incorrect “Yes") vs. False Negatives (Incorrect “No")

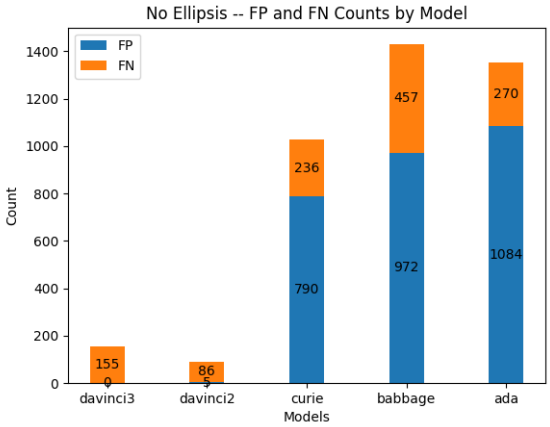

Figure 8: Non-ellipsis case – Errors by model. False Positives (Incorrect “Yes") vs. False Negatives (Incorrect “No")

8 Limitations

This paper argues that the proposed task of ellipsis-dependent reasoning is a difficult challenge for GPT-3 models, which are among the most powerful current language models. The data constructed here is restricted to English, and furthermore is restricted to a single form of ellipsis, namely verb phrase ellipsis. It may well be that other forms of ellipsis may give rise to different effects, and it is also important to test the claims made here on other languages.

References

OpenAI GPT-3 Models Overview. Accessed on 2023-01-10.

Rahul Aralikatte, Matthew Lamm, Daniel Hardt, and Anders Søgaard. 2021. Ellipsis resolution as question answering: An evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 810–817, Online. Association for Computational Linguistics.

Johan Bos and Jennifer Spenader. 2011. An annotated corpus for the analysis of VP ellipsis. *Language resources and evaluation*, 45(4):463–494.

Pradeep Dasigi, Nelson F Liu, Ana Marasović, Noah A Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. *arXiv preprint arXiv:1908.05803*.

Alan Garnham. 2010. Models of processing: Discourse. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(6):845–853.

Hans Kamp. 1981. A theory of truth and semantic representation. In *Formal methods in the study of language*, pages 277–322.

Kian Kenyon-Dean, Jackie Chi Kit Cheung, and Doina Precup. 2016. Verb phrase ellipsis resolution using discriminative and margin-infused algorithms. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1734–1743.

Payal Khullar. 2020. Exploring statistical and neural models for noun ellipsis detection and resolution in english. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 139–145.

Vid Kocijan, Ernest Davis, Thomas Lukasiewicz, Gary Marcus, and Leora Morgenstern. 2022. The Defeat of the Winograd Schema Challenge. *arXiv preprint arXiv:2201.02387*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open Sesame: getting inside BERT's linguistic knowledge. *arXiv preprint arXiv:1906.01698*.

Andrea E Martin and Brian McElree. 2008. A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *Journal of Memory and Language*, 58(3):879–906.

Ivan A Sag and Jorge Hankamer. 1984. Toward a theory of anaphoric processing. *Linguistics and philosophy*, pages 325–345.

Lindsey Vanderlyn, Talita Anthonio, Daniel Ortega, Michael Roth, and Ngoc Thang Vu. 2022. Toward implicit reference in dialog: A survey of methods and data. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 587–600.

Wei-Nan Zhang, Yue Zhang, Yuanxing Liu, Donglin Di, and Ting Liu. 2019. A neural network approach to verb phrase ellipsis resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7468–7475.

Zhuosheng Zhang and Hai Zhao. 2022. Tracing origins: Coreference-aware machine reading comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1281–1292.

A Appendix

A.1 Sample Instantiations

Below are sample instantiations for the Single Sentence and Two Action structures. The complete datasets for all the structures can be accessed at https://github.com/DanHardtDK/ellipsisGPT3.

Single Sentence

Mary went swimming, and Harold did too.

Mary went swimming, and Harold went swimming too.

Q: Did Harold go swimming?

A: Yes

Mary went swimming, but Harold didn't.

Mary went swimming, but Harold didn't go swimming.

Q: Did Harold go swimming?

A: No

Two Actions

Mary didn't go swimming but Harold did.

Mary went shopping and Harold didn't.

Mary didn't go swimming but Harold did go swimming.

Mary went shopping and Harold didn't go shopping.

Q: Did Harold go swimming?

A: Yes

Q: Did Harold go shopping?

A: No

Category	Substitution List
Subject	“Mary", “Harold", “Sam", “William", “The teacher", “The student", “The driver", “My friend", “John", “Elena", “Karen", “Mrs Jones"
Verb	“swimming", “shopping", “running", “walking", “skiing", “jogging", “hiking"

Table 3: Substitutions

A.2 Substitutions

The examples are produced using the substitutions for subjects and verbs in the different structures, as shown in table 3.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
section 8

☑ A2. Did you discuss any potential risks of your work?
only in the sense of limitations – there is a risk that the conclusions will not extend beyond English, and the particular models considered here

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

section 3 and appendix

☑ B1. Did you cite the creators of artifacts you used?
section 1 – GPT3 from OpenAI

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
GPT3 is freely available for research, we have made our data available on Github

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
research use of GPT3 is well established

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We created simple synthetic data and see no issues here

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
section 3

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☑ Did you run computational experiments?

section 4

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
standard access to GPT3 models

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
no search of hyperparameters

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
section 5

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.