CoGen: Abductive Commonsense Language Generation

Rohola Zandie[1], **Danny Brahman**[2], **Mohammad H. Mahoor**[1]
[1]Department of Electrical and Computer Engineering
[2]Department of Computer Science
University of Denver
Denver, USA
rohola.zandie@du.edu
danny.brahman@du.edu mohammad.mahoor@du.edu

Abstract

Reasoning is one of the most important elements in achieving Artificial General Intelligence (AGI), specifically when it comes to Abductive and counterfactual reasoning. In order to introduce these capabilities of reasoning in Natural Language Processing (NLP) models, there have been recent advances toward training NLP models to better perform on two main tasks - Abductive Natural Language Inference (αNLI) and Abductive Natural Language Generation Task (αNLG). This paper proposes CoGen, a model for both αNLI and αNLG tasks that employs a novel approach of combining the temporal commonsense reasoning for each observation (before and after a real hypothesis) from pre-trained models with entailment-based filtering for training. Additionally, we use state-of-the-art semantic entailment to filter out the contradictory hypothesis during the inference. Our experimental results show that CoGen outperforms current models and set a new state of the art in regards to αNLI and αNLG tasks. We make the source code of the CoGen model publicly available for reproducibility and to facilitate relevant future research.

1 Introduction

Different kinds of reasoning can be categorized into three classes (Walton, 2014): Deduction, Induction, and Abduction. In deduction, the truth of the conclusion is already provided in the premise, therefore, it is impossible that the premises are true and the conclusion is false. Induction is the process of going from the truth of some premises to the conclusion. Finally, abduction is the process of forming the most plausible hypothesis based on incomplete observations. The focus of this paper is on abductive reasoning.

The abductive inference could be viewed as going backward from the conclusions of a valid deductive inference to the premises to find its plausible causes and effects. In terms of classical logic, this is a fallacy (Andersen, 1973). Abductive reasoning is defeasible (and also non-monotonic) which means the conclusions can be refuted in the light of new data. Although abductive reasoning forms one of the core abilities of human cognition, its research in the area of NLP is still widely unexplored.

Recent work on large language models like GPT-3 (Brown et al., 2020) and GPT-Neo (Gao et al., 2020) had impressive results on different NLP tasks but still struggled with Abductive Natural Language Inference (αNLI) tasks. These models embed a great deal of world knowledge (Petroni et al., 2019; Wang et al., 2020), but their potential for commonsense reasoning (e.g. abductive reasoning) has not been fully harnessed. The task of abductive commonsense language generation can be defined as generating reasons given incomplete observations.

Abductive commonsense language generation can be formulated as a controlled language generation task. Like other controllable language generation problems that involve maintaining fluency and relevance of the generated text conditioned on some property, such as sentiment (Lample et al., 2018), topic (Zandie and Mahoor, 2021), and style (Shen et al., 2017), the abductive commonsense language generation can be viewed as a controllable language generation task that is conditioned on incomplete observations.

In this paper, we introduce CoGen[1], a model for generating and inferring abductive reasons that are compatible with observations. This combines temporal commonsense reasoning for each observation (before and after the hypothesis) from pretrained models with contextual filtering for training. Contextual filtering refers to the technique of refining temporal entailment during text generation to produce more coherent and contextually relevant output. We also use state-of-the-art semantic entail-

[1] Codes and Data are publicly available at: `https://github.com/roholazandie/abduction_modeling`

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 295–302
July 9-14, 2023 ©2023 Association for Computational Linguistics

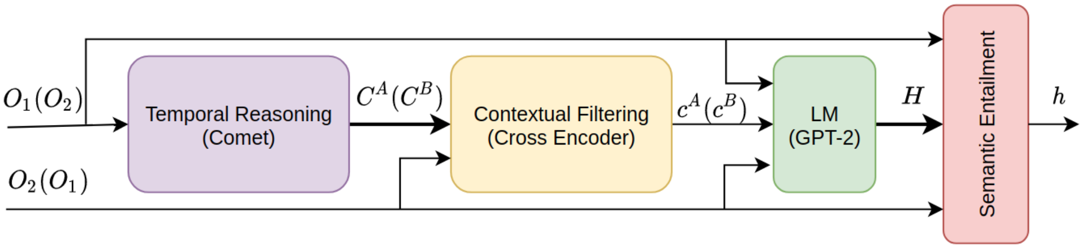

Figure 1: CoGEN first uses Temporal Reasoner to produce *before* and *after* commonsense then with a Cross-Encoder it filters unrelated temporal commonsense based on the context. With GPT-2 the system takes both observations and contextual knowledge as inputs a set of hypotheses H will be generated that in semantic entailment will be cleaned up from contradictions by a BERT model. The bold arrows indicate a set of inputs.

ment to filter out contradictory hypotheses during the inference. Our results show that CoGEN outperforms all previous models regarding αNLI and αNLG tasks.

Our main contributions are the following:

1. Using temporal commonsense reasoning for augmenting the observations - a crucial step in the abductive hypothesis generation as this task requires understanding the temporal relationships such as causes, effects, reasons, and intents.

2. Using contextual filtering to help narrow down the space of generated commonsense reasoning to the ones that are relevant to both observations.

3. Using the semantic entailment filtering to rule out the possibility of generating contradictory hypotheses given both observations.

4. Releasing the source code of the CoGEN model for reproducibility and assisting relevant future research.

2 Related Work

Previous research on reasoning in NLP mainly focuses on monotonic reasoning, which is usually about finding the “entailment”, “contradiction” or “neutral” relationships between a premise and a hypothesis. For example, SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) are both datasets that focus on monotonic inference. There is a choice of plausible reasoning task with the COPA dataset (Roemmele et al., 2011) which is designed for causal reasoning.

In (Qin et al., 2019), the authors introduced the TimeTravel dataset which contains over 28k counterfactual instances. The results show the current language models lack understanding of the reasoning behind the stories, sometimes even adding more samples will not improve the quality of the generation. (Qin et al., 2020) proposes Delorean, a new unsupervised decoding algorithm based on backpropagation that incorporates observations from the past and future to generate constrained text in between. They used the ART dataset (Bhagavatula et al., 2019) which contains 20k samples.

The most relevant work to CoGEN is Abductive Commonsense reasoning (COMeTEmb+GPT2) (Bhagavatula et al., 2019), which introduces ART dataset consisting of 20k commonsense narrative contexts with 200k explanations. They also introduced two tasks: abductive NLI (αNLI) a multiple-choice task for choosing the best hypothesis and abductive NLG (αNLG) which generates an abductive hypothesis given the two before and after contextual observations. Results showed that abductive NLG is much more challenging compared to (αNLI) and needs further research. They also used GPT-2 and COMET (Bosselut et al., 2019) for commonsense reasoning to generate new abductive hypotheses. The human judgment results show that only 44.56 percent of these generated hypotheses make sense to evaluators. In (Paul and Frank, 2021), they consider possible events emerging from the candidate hypothesis and then select the one that is most similar to the observed outcome. Their approach outperforms COMeTEmb+GPT2 on the αNLI task and achieves 72.2 on the test set. (Ji et al., 2020) proposed GRF, which is based on GPT-2 and dynamic multi-hop reasoning for multi-relational paths extracted from ConceptNet for αNLG.

REFLECTIVE DECODING (West et al., 2020) is an unsupervised text generation algorithm for text infilling that uses two pre-trained forward and backward language models. This algorithm outperforms all unsupervised methods, but is still significantly behind the fine-tuned model of COMeTEmb+GPT2 in abductive generation.

3 Method

Abductive reasoning can be formulated using a single observation as a premise and generating a hypothesis. However, following (Qin et al., 2020) we formulate abductive commonsense language generation as the task of generating a hypothesis H given two observations, O_1 and O_2 that happen at times t_1 and t_2, respectively, in which $t_2 > t_1$. The hypothesis H happens between t_1 and t_2.

This shows abductive and temporal reasoning is closely related to each other (Verdoolaege et al., 2000). More specifically, abductive reasoning requires temporal reasoning about the consequences of events (what typically occurs after them) and the reasons behind them (what may happen prior to or trigger them).

Commonsense knowledge graphs (CSKB) are knowledge graphs containing many commonsense facts about the world that help to understanding and reasoning about events, social interactions and physical entities. ATOMIC2020 (Hwang et al., 2020) is the largest CSKB having 1.33M tuples about entities and events of inferential knowledge and introduces 23 relation types. In this paper we focus on two classes of these relations: *before* relations and *after* relations. *before* relations are those that take place before the observation or trigger them, such as: `isBefore`, `Causes`, `xEffect`, `xReacts`, `xIntents`, and `xWants`. *after* relations are those that occur after the observation, such as: `isAfter`, `oReact`, `oWant`, `oEffect`, `xReason`. Neural Knowledge Graphs are models trained on CSKB tuples and are able to generate tails given the new heads. For instance, to predict the tail of the tuple (X votes for Y, `xIntents` ?) is to generate "to give support". We use the state-of-the-art pretrained Bidirectional and Autoregressive Transformer (BART) (Lewis et al., 2020) named Comet that is trained on ATOMIC2020.

For temporal commonsense augmentation, we generate n *after* relation facts for O_1 and n *before* relation facts for O_2. The $Comet(O, R)$ is the function that generates the commonsense for observation O for the relation R. If R_A and R_B are the *after* and *before* relations, then the following commonsense responses are generated:

$$C^A = Comet(O_1, R_A) \tag{1}$$

$$C^B = Comet(O_2, R_B) \tag{2}$$

However, not all *after* and *before* relations are relevant for every situation. The generated commonsense facts should be filtered out based on the context. For each commonsense relation, we chose the most likely fact based on the semantic similarity to the other observation. More specifically, the most likely *after* (*before*) fact for O_1 (O_2) based on the similarity to O_2 (O_1) is chosen:

$$c^A = \underset{c_i}{\operatorname{argmax}}\, Sim(O_2, C_i^A) \tag{3}$$

$$c^B = \underset{c_i}{\operatorname{argmax}}\, Sim(O_1, C_i^B) \tag{4}$$

where Sim is the cross-encoder (Reimers and Gurevych, 2019) based on BERT that calculates the similarity of two input texts. Figure 1 shows the pipeline for temporal commonsense generation and contextual filtering. This is similar to how we consider possible conclusions from the observations. We try to limit these based on how well they correspond to other observations in hand (Paul and Frank, 2021).

Given the observations $O_1 = \{t_1^{O_1} \dots t_m^{O_1}\}$, $O_2 = \{t_1^{O_2} \dots t_n^{O_2}\}$ and hypothesis $H = \{t_1^H \dots t_l^H\}$ as a sequence of tokens, we can augment the input with the commonsense knowledge from the previous step $K = \{c^A, c^B\}$ as a sequence of tokens $K = \{t_1^K \dots t_q^K\}$. The Abductive Commonsense Language Generation can be formulated by minimizing the following negative log-likelihood:

$$\mathcal{L} = -\sum_{i=1}^{N} \log P(t_i^H \mid t_{<i}^H, O_1, O_2, K) \tag{5}$$

Training: We trained three different models for αNLG - CoGen$_{LG}$, CoGen$_{MD}$ and CoGen$_{SM}$ by fine-tuning three GPT-2 models of sizes large, medium, and small, respectively. We used an embedding size of 512 for all models with a maximum token size of 128. The learning rate was set to $5e{-}4$ with a weight decay of 0.01. We stopped training after 5 epochs before overfitting to the training set occurrs.

We also propose the fine-tuned CoGen$_{RB}$ model for αNLI, which is based on the large RoBERTa (Liu et al., 2019) model. We set the

Model	BERT-Score	BLEURT	BLEU	TER	METEOR	ROUGE	Human
COMeT-Emb+GPT2	88.25	-1.07	3.22	106.31	9.74	17.42	44.56
CoGen$_{LG}$	88.74	-1.12	28.80	**123.47**	21.62	26.75	52.00
CoGen$_{MD}$	**89.75**	**-0.83**	**37.15**	104.19	**22.56**	**30.58**	**69.2**
CoGen$_{SM}$	88.14	-0.99	10.25	103.40	11.50	20.62	43.2

Table 1: The automatic evaluations of generative models on the *test* set of ART Dataset (Bhagavatula et al., 2019)

first 20% for the warm-up with the learning rate of $1e-5$ and after that decrease it linearly by a ratio of 0.01.

Inference: For inference, we use beam search decoding with a beam size of 5. We chose this search as it works best with controllable language generation (Zandie and Mahoor, 2021). For each pair of observations, multiple hypotheses are generated and then filtered out based on entailment. We use the pre-trained semantic entailment BERT cross-encoder (Reimers and Gurevych, 2019), trained on SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018), to filter out each generated hypothesis H, if the $O_1 \rightarrow H$ or $H \rightarrow O_2$ is a contradiction. Using this technique we can remove undesired hypotheses that are incompatible with the given observations.

4 Result

We report BERT-Score (Zhang* et al., 2020), BLEURT (Sellam et al., 2020), BLEU (Papineni et al., 2002), TER (Snover et al., 2006), METEOR (Banerjee and Lavie, 2005) and ROUGE (Lin, 2004) for automatic evaluation of our model. The results in Table 2 show that both CoGen$_{MD}$ and CoGen$_{LG}$ outperform the best model in (Bhagavatula et al., 2019), which is COMeT-Emb+GPT2 model on all metrics on the *test* set of the ART dataset (Bhagavatula et al., 2019). Additionally, CoGen$_{MD}$ performs the best among all the models.

We assessed human evaluations on 100 randomly selected results from the *test* set. The evaluation was completed by five graduate students unrelated to our research, providing us with unbiased data. These evaluations, shown in Table 2, are consistent with previous automatic results. These results show that CoGen$_{MD}$ generates better results compared to the base model (COMeT-Emb+GPT2) and the Real Hypothesis in most cases. Also, CoGen$_{LG}$ outperforms the base model.

Finally, we show the results of αNLI task from different models in Table 3. This table displays that CoGen$_{RB}$ surpasses the previous model used (LMI + MTL) (Paul and Frank, 2021) by a substantial margin. The results of αNLI show the importance of temporal reasoning and contextual filtering along with RoBERTa.

Model	<	Neutral	<	Comparator
CoGen$_{LG}$	48.00	22.20	29.80	RH
CoGen$_{LG}$	37.00	17.00	**46.00**	CM
CoGen$_{MD}$	30.60	32.80	**36.40**	RH
CoGen$_{MD}$	23.80	23.40	**52.40**	CM
CoGen$_{SM}$	56.60	24.20	19.00	RH
CoGen$_{SM}$	42.00	32.20	25.80	CM

Table 2: Human Judgements of CoGen as compared to the comparators - Real Hypothesis (RH) and COMeT-Emb+GPT2 (CM)."Neutral" means our model is equally good to the comparator. The left and right columns to Neutral means the model is worse and better than comparator respectively.

Model	Dev Acc (%)	Test Acc (%)
ESIM+ELMo	58.20	58.80
BERT$_{Large}$	69.10	68.90
COMeT-Emb+GPT2	69.40	69.10
LMI + MTL	72.90	72.20
CoGen$_{RB}$	**82.90**	**83.26**

Table 3: Results on αNLI task. Last row in bold shows the performance of CoGen$_{RB}$ based on RoBERTa

5 Conclusion

We present CoGen, a novel approach to generate abductive reasoning given incomplete observations in three different sizes. This integrates temporal reasoning, context filtering, and semantic entailment to complete the base GPT-2 model for better reasoning. Both human and automatic evaluations assessed in this study show that CoGen outperforms previous methods used for abductive reasoning. Our approach sets a new state-of-the-art for αNLI and αNLG tasks on ART dataset.

Limitations

The CoGEN model introduced in this paper uses temporal relations as a process of abductive reasoning. Although, temporal relations have been shown to be very useful in abductive reasoning (Verdoolaege et al., 2000), the measure of the effectiveness of other types of relations about an observation have not been evaluated in this paper. In addition, because of the unavailability of a large number of human evaluators, we randomly selected 100 selected results as opposed to the entire result which would have been ideal.

References

Henning Andersen. 1973. Abductive and deductive change. *Language*, pages 765–793.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. *arXiv preprint arXiv:2010.05953*.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. *arXiv preprint arXiv:2009.11692*.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Debjit Paul and Anette Frank. 2021. Generating hypothetical events for abductive inference. *arXiv preprint arXiv:2106.03973*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. *arXiv preprint arXiv:1909.04076*.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. 2020. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. *arXiv preprint arXiv:2010.05906*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *ACL*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *arXiv preprint arXiv:1705.09655*.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Sven Verdoolaege, Marc Denecker, and Frank Van Eynde. 2000. Abductive reasoning with temporal information. *arXiv preprint cs/0011035*.

Douglas Walton. 2014. *Abductive reasoning*. University of Alabama Press.

Chenguang Wang, Xiao Liu, and Dawn Song. 2020. Language models are open knowledge graphs. *arXiv preprint arXiv:2010.11967*.

Peter West, Ximing Lu, Ari Holtzman, Chandra Bhagavatula, Jena Hwang, and Yejin Choi. 2020. Reflective decoding: Beyond unidirectional generation with off-the-shelf language models. *arXiv preprint arXiv:2010.08566*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Rohola Zandie and Mohammad H Mahoor. 2021. Topical language generation using transformers. *arXiv preprint arXiv:2103.06434*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6

☒ A2. Did you discuss any potential risks of your work?
Not relevant to this research

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

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

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ 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)?
No response.

□ 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?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ 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.
No response.

C ☑ Did you run computational experiments?

Section 3

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Computational infrastructure used was pretty general, and nothing out of ordinary there to report

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?
Section 3

☒ 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?
Not relevant for this research

☑ 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.)?
Section 4

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 4

☒ 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.?
The instructions given to annotators were straight-forward

☒ 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)?
We recruited student volunteers

☒ 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?
The annotators were verbally explained that the data they were using were open-source data

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not relevant to this research

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not relevant to this research