Responsible NLP Checklist

Paper title: *Rule Discovery for Natural Language Inference Data Generation Using Out-of-Distribution Detection*
Authors: *Juyoung Han, Hyunsun Hwang, Changki Lee*

How to read the checklist symbols:

☑ the authors responded 'yes'

☒ the authors responded 'no'

N/A the authors indicated that the question does not apply to their work

☐ the authors did not respond to the checkbox question

For background on the checklist and guidance provided to the authors, see the Responsible NLP Checklist page at ACL Rolling Review.

☑ A. Questions mandatory for all submissions.

☑ A1. Did you describe the limitations of your work?
This paper has a Limitations section.

☒ A2. Did you discuss any potential risks of your work?
While Section 6 (Limitation) addresses concerns about data quality and validation, the paper does not comprehensively discuss potential risks. Key risks that should be considered include: potential misuse of the method to create deceptive training examples, accessibility issues due to computational resource requirements, and the possibility of introducing or amplifying biases in the generated data. This paper needs a broader discussion of these risks and specific mitigation strategies beyond the current limitations section.

☑ B. Did you use or create scientific artifacts? (e.g. code, datasets, models)

☑ B1. Did you cite the creators of artifacts you used?
In the Related Work (section2) and throughout the Methods (section3). This paper appropriately cites all major artifacts used in the research, including the SNLI dataset (Bowman et al., 2015), BERT model (Devlin et al., 2019), e-SNLI dataset (Camburu et al., 2018), and various tools such as WordNet, Gensim, and ConceptNet (Miller, 1992; Rehurek and Sojka, 2011; Speer et al., 2017). We also used GPT-4o-mini model for data generation.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
In Appendix E.1 (Data Generation Process), we specify that the SNLI dataset (version 1.0) used in our research is publicly available for research purposes under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) License. We also note that this license allows sharing and adaptation of the dataset with proper attribution and distribution under the same terms. Additionally, we use GPT-4o-mini model for data generation, with proper citation to related work (Madaan et al., 2023).

☑ 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)?
In section 3 and section 4. This paper discusses using SNLI and e-SNLI datasets for their intended

purpose of NLI research and data generation. The generated data is also intended for NLI research purposes, consistent with the original datasets.

☒ 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?
This paper does not explicitly discuss checking for or handling personally identifiable information or offensive content in the generated data. The SNLI dataset is based on sentence pairs generated from an anonymous dataset and does not contain personally identifiable information such as names, addresses, or phone numbers. Therefore, since the data I generated involves extracting premise sentences from SNLI and generating hypothesis sentences accordingly, it is also considered anonymous data and does not include personally identifiable information. Based on this reasoning, I did not conduct a separate review for personally identifiable information.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In section 3 and section 4. This paper provides comprehensive documentation of the artifacts used and created, including the use of English language data from the SNLI dataset, detailed descriptions of premise-hypothesis pairs and transformation rules, characteristics of generated training examples, and the complete evaluation framework with metrics. We also describe the process of rule discovery and data generation in detail.

☑ 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?
In section 4.2-4.4 and section Appendix E-G. This paper reports comprehensive statistics about the data used and created, including the SNLI training set size of 550,152 examples, 50,000 OOD premise-hypothesis pairs identified, 10,000 clusters generated, and varying numbers of samples generated per rule (ranging from 300 for uniform distribution to 4-2,307 for distribution-aware approach). We also report test set size of 10,000 examples and provide detailed performance metrics across different dataset sizes (2k, 10k, 50k, 550k).

☑ **C. Did you run computational experiments?**

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In section 4.3 NLI Performance Analysis. This paper reports that we used BERT-base model for fine-tuning, and experiments were performed with specific hyperparameters (batch size=32, learning rate=3e-5). While detailed computing infrastructure specifications and total computational costs (e.g., GPU hours) were not explicitly described in the main text, they were provided in the supplementary materials. Our computing infrastructure consists of an NVIDIA RTX A6000 (48GB VRAM) GPU, Intel Core i9-9820X (10 cores) CPU, and 16GB RAM. Using this setup, we trained on the SNLI training set with 5 different random seeds for 25 epochs each. With approximately 550,000 examples, each epoch took around 30 minutes to complete. In contrast, when using a subset of 5,000 examples, each epoch completed in less than one minute. Although we did not measure the total computational costs and time for all experiments, we plan to supplement this information with detailed calculations of computational resources and associated costs through additional experiments.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In Section 4.3. This paper provides a detailed description of the experimental setup. We used the BERT-base model trained with NLL loss and Adam optimizer, with specific hyperparameters (batch size=32, learning rate=3e-5) that were optimized on the SNLI validation set. The hyperparameters were tuned using grid search on the SNLI validation set to ensure optimal performance. Training

was conducted for 25 epochs, and all experiments were repeated with 5 different random seeds to ensure reliability of the results.

☑ 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?
In Section 4.3. This paper reported comprehensive descriptive statistics throughout our experimental results. All performance metrics are presented with standard deviations (e.g., 89.85 0.36), and results are averaged across 5 random seeds to ensure reliability. We provided both best and average performances on validation and test sets, along with detailed performance comparisons across different dataset sizes (2k, 10k, 50k, 550k).

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
We conducted experiments using a customized implementation based on the BERT-base model from the transformers library, specifically modified for OOD detection. While we did not detail the specific code implementation in the main text, focusing instead on the theoretical methodology, the complete code is provided in the supplementary materials. As our primary research objective was to identify patterns and rules after OOD detection, we provided detailed explanations of our rule discovery process and data generation methods, rather than focusing on the technical implementation details.

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

N/A 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.?
Not applicable. This study did not involve human annotators or research participants.

N/A 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)?
Not applicable. This study did not involve human annotators or research participants.

N/A D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?
Not applicable. This study did not involve human annotators or research participants.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. This study did not involve human annotators or research participants.

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. This study did not involve human annotators or research participants.

☒ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

☒ E1. If you used AI assistants, did you include information about their use?
We did not use AI assistants like ChatGPT or Copilot for research, coding, or writing. While our study utilized GPT-4o-mini as a research tool for generating NLI training data, this was part of our experimental methodology rather than an AI assistant used in the research process itself. The LLM was used specifically for implementing the Chain-of-Thought prompting technique to generate Premise-Hypothesis-Label triples, as detailed in Section 4.2 and Appendix C.2.