CS4248 Project Report: Enhancing Natural Language Inference with Machine-Generated Explanations: A Comparative Study on Explanation Effectiveness in NLI Tasks

A0216040Y, A0226576W, A0238913Y, A0194490X, A0221977W
Group 38
Mentored by Esther Gan
{e0538141,e0638862,e0773511,e0376920,e0559509}@u.nus.edu

Abstract

Research in Natural Language Inference (NLI) plays a pivotal role in advancing our understanding of human language comprehension, which is crucial for applications like machine translation, sentiment analysis, and question answering. A key challenge in NLI is elucidating the decision-making processes of models. The e-SNLI dataset (Camburu et al., 2018) features human-annotated explanations that justify model decisions. However, the creation of these explanations is both time-consuming and labor-intensive. Our study explores whether machine-generated explanations, created using a GPT-2 model, can match or surpass the effectiveness of human-annotated explanations in enhancing the performance of NLI tasks. We assess this impact using metrics such as accuracy, precision, F1-score, and BLEU score, revealing insights into the potential and limitations of machine-generated explanations.

1 Introduction

The field of Natural Language Inference (NLI) is fundamental to improving machine comprehension of human language. NLI involves classifying relationships between premises and hypotheses into categories such as Entailment, Neutral, and Contradiction. Understanding these semantic relationships and logical coherences is crucial for enhancing Natural Language Understanding (NLU), which supports a variety of applications from semantic search to interactive dialogue systems. This research investigates the potential of machine-generated explanations in NLI, aiming to streamline and possibly enhance the explanatory process that supports decision-making in AI systems.

1.1 Motivation

The importance of NLI extends beyond theoretical research; it is crucial for enhancing AI's capacity to process and understand human language accurately. The quality of explanations in NLI not only affects the transparency but also the trustworthiness and reliability of AI decisions. Given the labor-intensive nature of crafting human explanations, our motivation is to explore whether AI can autonomously generate high-quality explanations that are both accurate and helpful for improving NLI models. This could significantly reduce human effort and enable more scalable NLI solutions, crucial for real-world applications in sectors like education, healthcare, and customer service

1.2 Key contributions

Despite existing research, there is still unexplored potential in understanding the utility of machine-generated explanations. This study addresses this gap through comprehensive experiments aimed at improving the generation and application of these explanations in Natural Language Inference (NLI) tasks. Our key contributions are:

1. Introducing a framework that leverages a GPT-2 model to generate explanations for NLI tasks, assessing whether AI can match human performance in explanation quality.

2. Providing a comparative analysis of machine-generated versus human-annotated explanations, using metrics such as accuracy, precision, F1-score, and BLEU score to measure effectiveness.

3. Highlighting the current limitations and potential future applications of AI in automating explanation generation, setting the stage for further improvements and wider adoption in practical AI applications.

2 Related Work / Background

The original paper that published the e-SNLI dataset (Camburu et al., 2018) leveraged Amazon Mechanical Turk to gather annotations and focused on encouraging annotators to provide explanations that highlight subtle elements influencing relationships between sentences. They ensured annotation quality through in-browser validation tools and a two-step process requiring annotators to first highlight pivotal words and then craft detailed explanations. Constraints varied depending on the type of relationship, which helped maintain high standards in responses. This methodological rigor sets a benchmark in the field of explanation generation in NLI.

Subsequent studies have explored the utility of explanations in enhancing various NLP tasks. For instance, (Rajani et al., 2019) integrated human explanations into the Common sense Question Answering (CQA) dataset (Talmor et al., 2019) and released their dataset (CoS-E). Their approach demonstrated a notable improvement, boosting CommonsenseQA task performance by 10%. This work underscores the potential of explanations in increasing the robustness of NLP models.

Further research by (Narang et al., 2020) focused on creating quality explanations while (Thorne et al., 2019), (Rajagopal et al., 2021) focused on the explainability of generated explanations. Their work highlights ongoing challenges in ensuring that explanations genuinely reflect and justify model reasoning, an area that continues to offer significant opportunities for innovative research.

Another piece of related work on the e-SNLI dataset is by (Zhou et al., 2023), employing a two-step methodology for generating explanations followed by fine-tuning a classifier using an explanation-aware prompt-based method. Their findings revealed that while the method holds promise, many generated explanations still fell short in justifying the classification decisions adequately, signaling a significant gap in the quality of generated explanations.

3 Corpus Analysis

3.1 Data exploration

The training dataset comprises 549,367 entries, each consisting of a hypothesis, a premise, and an accompanying explanation. These entries are categorized into three distinct labels: entailment, contradiction, and neutral. These labels delineate the nature of the relationship between the premise and the hypothesis—specifically, whether the hypothesis entails, contradicts, or is neutral regarding the premise. The distribution of these categories is relatively balanced with 183,416 instances of entailment, 183,187 of contradiction, and 182,764 of neutral.

During the initial data processing, we identified 25 entries with missing explanations. Given the minimal impact of these missing entries on the overall dataset—representing less than 0.005% of the total data—they were excluded from further analysis. This decision ensures the integrity and consistency of our training data, which is crucial for maintaining the reliability of our model's performance evaluations.

4 Methodology

4.1 Classifier

In this study, we employed the RoBERTa[1] model as our primary classifier for this natural language inference (NLI) task.

4.1.1 Model Rationale

The choice of RoBERTa was predicated on its robust pre-trained architecture and enhanced capacity for processing context and semantics over its predecessors, such as BERT (Bidirectional Encoder Representations from Transformers). While both RoBERTa and BERT are built on the transformer architecture, RoBERTa is trained with a larger corpus and for a longer duration, enabling it to excel in tasks requiring deep contextual understanding.

4.1.2 Model Architecture & Implementation

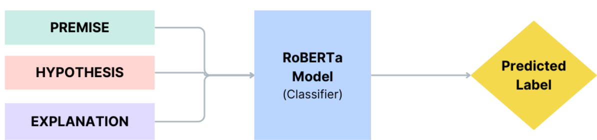

Figure 1: RoBERTa Model Architecture

The initial stage of our pipeline involved tokenization using the RobertaTokenizerFast, which efficiently handles the conversion of text into tokens that the model can process. Our input structure was carefully designed to maximize the contextual relations between the premise, hypothesis, and explanation. We used the template: "Given that [premise], it is hypothesized that [hypothesis]. [explanation]." This format ensures that the model recognizes and processes the logical flow intended in NLI tasks, where understanding the causal and

[1]Hugging Face RoBERTa model

contradictory elements between statements is crucial.

We chose to freeze all layers of RoBERTa except for the classification head during fine-tuning. This strategy is particularly effective because it leverages the deep, context-aware embeddings learned during RoBERTa's extensive pre-training across a vast corpus and variety of tasks, allowing the model to retain its pre-learned high-quality representations while focusing training on the high-level task of discerning entailment, contradiction, or neutrality.

Additionally, we modified the classifier by downsizing the originally larger classification layer to a smaller, more efficient one. This adjustment was primarily aimed at reducing the computational load and enhancing the speed of the model. By minimizing the size of the final layer, we maintain the model's ability to make fine-grained distinctions without the excessive computational cost typically associated with larger models. All code can be found in our GitHub[2] repository.

4.1.3 Evaluating predictions

In evaluating our classifier model, we utilized a comprehensive set of metrics to ensure a balanced assessment of its performance in multi-class classification settings. The Macro F1 Score treats all classes equally by averaging the individual F1 scores, providing fairness across class representation. The Micro F1 Score aggregates outcomes across all classes to reflect overall precision and recall, useful for assessing performance in dominant classes. The Weighted F1 Score adjusts each class's F1 score according to its frequency, offering a realistic view of performance based on class prevalence. Additionally, we used Accuracy for its straightforward depiction of the model's overall correctness. This multi-metric approach not only enriches our understanding of the model's effectiveness across varied scenarios but also helps in fine-tuning the model's robustness and reliability across a spectrum of scenarios.

4.2 Explanation Generator

We chose GPT-2, which is a pre-trained large language model [3], as a medium to generate the machine-explanations using the premise, hypothesis and respective label as input.

[2] GitHub repository
[3] GPT-2 Model Documentation

4.2.1 Model Rationale

The rationale for choosing GPT-2 model was mainly because it employs a multi-layered transformer architecture that enables bidirectional context understanding as well as efficient processing of sequential data, which is important for the task at hand to generate meaningful explanations. Moreover, being trained on diverse text data, it can capture a range of linguistic patterns and semantic relationships to provide coherent text generation capabilities. Also, GPT-2 allows a flexible architecture to fine-tune parameters with a self-attention mechanism, which can be useful to experiment around while monitoring model performance.

4.2.2 Model Architecture & Implementation

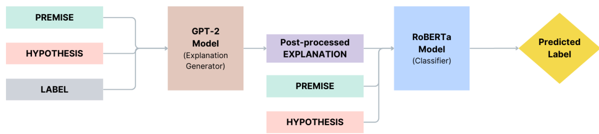

Figure 2: GPT Model Architecture

Before training, the hypothesis and premise were converted to lowercase and the punctuation marks were removed and a prompt is crafted using the premise and hypothesis with regard to its respective label. With various trial-and-error to craft an appropriate prompt, the below conditional prompts were used to generate a reasoning:

- Entailment: "Explain why [premise] has to be true when [hypothesis] is true?"
- Contradiction: "Explain why [premise] cannot be true when [hypothesis] is true and vice versa?"
- Neutral: "Explain why there is no evidence that if [premise] is related to [hypothesis]?"

Each prompt was then encoded using the toeknizer: *GPT2Tokenizer* to get an input sequence, which was processed in an attention mask, which matches the shape of the input sequence so that all tokens get processed equally in the assigned order. Afterwhich, the model *GPT2LMHeadModel* was initialised which generates the explanations accordingly with various parameters like maximum text length, output diversity, randomness, token limits, etc. The generated explanations were then post-processed to remove redundant preceding words for clarity, check for spelling and grammatical accuracy to exceute a coherence check for logical flow of the generated explanation. Finally, these machine-generated explanations were combined with premise and hypothesis to feed in the baseline *RoBERTa* model for classification.

Overall, we chose 2 variations of the GPT-2 models (GPT-2 Base & GPT-2 Medium) with the motivation to improvise with the "medium" variant as it has larger capacity for learning and capturing complex patterns as compared to the base GPT-2 model. One of the quantitative differences between the two variants is that GPT-2 Medium loads more parameters with deeper understanding of each, which increases the likelihood of better performance but augments to the additional computational requirements.

4.2.3 Evaluating explanations

In evaluating the machine generated explanations, we used a variety of metrics designed to assess both the semantic and syntactic alignment with human generated explanations. BLEU was chosen for its effectiveness in measuring the precision of n-grams, providing a basic gauge of lexical similarity. METEOR was included for its ability to account for synonymy and sentence structure, offering a more nuanced assessment of linguistic and semantic alignment. ROUGE-1 and ROUGE-L were utilized to evaluate the recall of content and the fluency of the explanations, respectively, reflecting both detail retention and coherence. The BERT Score (covering Precision, Recall, and F1) provided insights into the deep semantic similarity by using contextual embeddings, ensuring that the explanations are semantically coherent with the references. Lastly, Word Mover's Distance (WMD) was employed to measure the semantic distance between word embeddings in the generated and reference texts, capturing the overall semantic alignment more effectively. These metrics together enabled a comprehensive evaluation of how well machine generated explanations mimic human reasoning in NLI tasks.

5 Experiments

5.1 Classifier Hyperparameter Tuning

The Ray[4] library was used to conduct hyperparameter tuning. The parameters adjusted included the learning rate, optimizer, training batch size, and weight decay, with the following values:

- Learning Rate: [1e-6, 1e-5, 1e-4, 1e-3]
- Weight Decay: [0.01, 0.03, 0.05, 0.08, 0.1]
- Optimizer: ["AdamW", "SGD", "Adam"]
- Train Batch Size: [16, 32, 64]

The best parameter set determined from our tuning efforts was 1e-3, 0.01, AdamW and 32 for the learning rate, weight decay, optimizer and train batch size respectively.

5.2 Explanation Generator Hyperparameter Tuning

For the explanation generator GPT-2 model, we prepared 3 versions of the model:

- Model v1: base *'gpt-2'* model
- Model v2: fine-tuned base *'gpt-2'* model
- Model v3: fine-tuned *'gpt-2-medium'* model

Among these model variations, we standardised the attention mask tensor as well as the number of outputs generated per input to 1 (num_return_sequences=1) so that one good quality explanation can be generated at each instance. On the other hand, we tested the effect of various parameters like maximum length, generation of n-grams, token limits based on probability, randomness. The tuning was performed on small batches of the training set for the following values to monitor their effect on the sensibility and relevance of each explanation:

- Maximum length (max_length): [90,120,150]
- Randomness (temperature): [0.7,0.8,0.9]
- Token Limit using highest probabilities (top_k): [10,50,90]
- Token Limit using cumulative probabilities (top_p): [0.5, 0.95]
- N-grams (no_repeat_ngram_size): [1,2,3]

From the above settings, it was observed that the combination of temperature, top_k and top_p values affected the diversity of the answers as they affect the sampling strategy of the model. Whereas, the length and n-gram parameters affected the level of detail in each output. After testing these values, we found the following best parameters in each model variant for max_length, temperature, top_k, top_p, no_repeat_ngram_size respectively:

[4] Ray

- GPT-2 Base (v2): 90,0.7,50,0.95,2
- GPT-2 Medium (v3): 120,0.7,50,0.95,2

One of the reasons that this setting worked the best was since the maximum length and randomness of the text generated was limited to a smaller threshold value, creating a smaller window for new tokens, allowing enough diverse outputs yet not making it too big to allow extremely random outputs which can reduce the coherence. While the token limit parameters (top_k and top_p) were expanded to a higher values since it allowed more top-performing words with higher probability to be added. Lastly, for the n-grams, setting to 2, that is, bigram worked best since it is avoiding repeating two consecutive tokens rather than the individual tokens. This increases the diversity yet avoiding any unnecessary repetitions or redundant phrases.

5.3 Explanation Generator Evaluation Results

In an attempt to enhance the clarity, grammatical accuracy, and overall quality of the machine-generated explanations, we conducted post-processing on the generated explanations on all 3 versions of machine generated explanations. We employed several methods including spelling correction and the removal of redundant words. Here, we present the evaluation results using multiple metrics: BLEU (BL), METEOR (MT), ROUGE-1 (R1), ROUGE-L (RL), BERT Precision (BP), Recall (BR) and F1 Score (BF), and Word Mover's Distance (WMD).

Table 1: Evaluating explanations before Post-Processing

Version	BL	MT	R1	RL	BP	BR	BF	WMD
v1 (raw)	0.005	0.093	0.065	0.046	0.803	0.849	0.825	1.150
v2 (raw)	0.030	0.159	0.176	0.139	0.831	0.860	0.846	0.985
v3 (raw)	0.024	0.162	0.151	0.114	0.826	0.859	0.842	1.006

Table 2: Evaluating explanations after Post-Processing

Version	BL	MT	R1	RL	BP	BR	BF	WMD
v1 (spelling)	0.005	0.100	0.065	0.046	0.797	0.845	0.820	1.150
v1 (redundant)	0.005	0.096	0.067	0.046	0.802	0.846	0.823	1.149
v2 (spelling)	0.031	0.161	0.177	0.140	0.827	0.859	0.843	0.987
v2 (redundant)	0.030	0.159	0.176	0.139	0.831	0.860	0.846	0.985
v3 (spelling)	0.024	0.162	0.151	0.114	0.826	0.859	0.842	1.006
v3 (redundant)	0.024	0.162	0.151	0.114	0.826	0.859	0.842	1.006

As shown in Table 8, the application of spelling corrections (as observed in v1 and v2) showed a slight improvement in METEOR and BLEU scores, indicating better lexical accuracy and alignment with reference texts. However, the impact on BERT scores and WMD was minimal, suggesting that while spelling improvements increase surface-level quality, they do not significantly alter the semantic content or the perceived distance between generated and reference explanations.

The removal of redundant words did not significantly alter the performance metrics across all versions. This outcome suggests that while redundancy reduction may improve readability, it does not substantially impact the metrics used for evaluating the quality of explanations in terms of their alignment with human-generated references.

Among the versions, v2 consistently showed better performance across most metrics compared to v1, which struggled particularly in terms of coherence and linguistic accuracy as indicated by lower ROUGE and METEOR scores. v3 showed a moderate performance, balancing between lexical richness and semantic coherence.

The post-processing steps, particularly spelling correction, have demonstrated their utility in slightly improving the textual quality of generated explanations. However, the minimal impact on deeper semantic metrics like BERT Scores and WMD suggests that future work should explore more sophisticated techniques for enhancing the relevance and depth of content in generated explanations. These could include more advanced linguistic models, better context integration, and learning-based approaches to post-processing.

5.4 Classifier Experiments

In our classifier experiments, we conducted fine-tuning on two different datasets: one containing the original data and another supplemented with machine-generated explanations. These explanations were generated using two versions: v2 and v3. The latter showed a slight performance increase of 1-2% over v2. However, due to the substantial computational resources and memory required by the v3's underlying GPT-2 medium model, we opted for v2 for its computational efficiency despite the minor performance drop.

We evaluated these models against a baseline, which was fine-tuned solely on the original dataset without any explanations. This comparison was crucial to assess the impact of explanations on the model's ability to make accurate predictions. According to the results captured in Table 3, it became clear that while human-generated explanations significantly enhance prediction accuracy, the machine-generated explanations, specifically from v2, actually deteriorated the performance of the model. This outcome underscores the variable influence that the quality and source of explanations can have on NLI tasks, highlighting the importance of selecting appropriate explanation sources to optimize model performance.

Table 3: Classifier Results

Expt.	Weighted F1	Micro F1	Macro F1	Acc.
1	0.937	0.937	0.936	0.94
2	0.456	0.452	0.448	0.45
3	0.549	0.498	0.446	0.50
4	0.543	0.534	0.524	0.53
Baseline	0.713	0.712	0.711	0.71

1. **Fine-tune on original dataset, test on original test set**: This setup achieved the highest scores across all metrics, indicating robust model performance when both trained and tested on human-curated data. The high scores reflect the model's ability to adapt to the nuances and specific linguistic patterns present in the original dataset.

2. **Fine-tune on original dataset, test on machine-generated explanations**: There was a substantial decrease in performance metrics, likely due to the linguistic discrepancies between the training data (human-generated) and the test data (machine-generated). This indicates challenges in generalization when the test data introduce new linguistic features not present during training. Additionally, machine-generated text may have idiosyncrasies such as repetitive phrases or less natural syntax, which are not typically captured during training on human-curated content.

3. **Fine-tune on machine-generated dataset, test on original test set**: In the experiment where the classifier was fine-tuned on a machine-generated dataset and tested on the original test set, we observed a significant drop in performance. This decline can be attributed to several factors:

    - The machine-generated dataset used for training was considerably smaller than the original dataset, creating a severe imbalance. This size discrepancy likely led to inadequate training, as the smaller dataset did not provide enough diversity and did not cover the full spectrum of features and complexities that the original dataset has. Consequently, this limitation could have resulted in the model not being adequately equipped to handle the richer linguistic variety in the original test set.
    - Secondly, training exclusively on machine-generated data may have predisposed the model to learn patterns and dependencies that are specific to the generation algorithms rather than those intrinsic to natural human language. This can cause the model to develop biases or overfit to artificial characteristics that do not translate well when confronted with human-generated text.

4. **Fine-tune on machine-generated dataset, test on machine-generated test set**: This experiment resulted in a performance similar to experiment 3, but an improvement over experiment 2. This suggests that while the model could handle machine-generated content somewhat better when both trained and tested on such data, it still struggles due to the inherent limitations in the training data.

5.5 Classifier Results on Post-Processed Explanations

We also explored how before and after post-processing the explanations, it affected the performance of a classifier that was fine-tuned on original human-generated data and tested on machine-generated explanations. This evaluation is crucial as it explores the classifier's adaptability to variations in explanation quality, which is key in applications like automated content generation and evaluation. Given computational constraints, our tests were limited to a sample size of 500 for each explanation type.

In terms of classifier performance, the results were mixed. While v3 of the generated explanations, which featured the highest intrinsic quality, benefited from spelling corrections with an improvement in accuracy and F1 scores, v1 and v2 showed minimal or no benefit from post-processing. This indicates that the underlying quality of the generated explanations is a more critical factor for

Table 4: Classifier Performance for Generated Explanations - before and after Post-Processing

Version	Weighted F1	Micro F1	Macro F1	Accuracy
v1 (raw)	0.380	0.341	0.308	0.34
v1 (spelling)	0.411	0.339	0.272	0.34
v1 (redundant)	0.383	0.343	0.307	0.34
v2 (raw)	0.494	0.492	0.491	0.49
v2 (spelling)	0.483	0.478	0.474	0.48
v2 (redundant)	0.495	0.494	0.493	0.49
v3 (raw)	0.519	0.494	0.466	0.49
v3 (spelling)	0.511	0.512	0.512	0.51
v3 (redundant)	0.521	0.494	0.463	0.49

classifier performance than the application of superficial text corrections.

These findings suggest that for developing robust NLI systems, greater emphasis should be placed on generating high-quality, coherent explanations right from the start, rather than relying on post-processing to correct minor flaws. Moreover, classifiers should be designed to be adaptive to variations in explanation quality to ensure consistent performance across different real-world scenarios where the quality of text can vary significantly. This approach would not only improve the reliability of AI systems in NLI tasks but also enhance their applicability in diverse applications.

6 Discussion

Our experiments have highlighted the critical role that explanation quality plays in the performance of models tasked with understanding and interpreting relationships between texts. It became clear that human-generated explanations, which are meticulously vetted for relevance and coherence, consistently outperform machine-generated explanations from models like GPT-2. This discrepancy can largely be attributed to the self-attention mechanism of RoBERTa, which, when presented with inaccurate explanations, was "distracted," leading to incorrect inferences. To illustrate the potential for improvement, we generated additional examples using GPT-4 with the same prompt template. These examples showed a marked improvement in the quality of explanations over those generated by GPT-2, indicating advancements in model capabilities for generating more contextually relevant explanations [5]. For detailed comparisons of explanations for identical premise-hypothesis pairs, see Appendix A.

In an attempt to quantify the relevance and usefulness of explanations in relation to the premise and hypothesis, we employed sentence embeddings. By concatenating the premise and hypothesis and encoding this combined text using a pretrained sentence-BERT model[5], we obtained a unified embedding vector. A similar process was applied to the explanations to generate a second vector, after which we calculated the cosine similarity between the two. This method was applied to the test set with v2 explanations, and the summary statistics for both correct and incorrect predictions were compiled [6].

A subsequent T-test revealed no significant difference between the means, challenging our initial hypothesis that sentence embeddings and cosine similarity could effectively measure the utility of an explanation. This unexpected result might be explained by two factors:

- High cosine similarity scores do not necessarily correlate with logical or factual correctness. An explanation might echo the vocabulary and context of the premise and hypothesis accurately yet still derive incorrect conclusions.

- Conversely, a correct explanation might hinge on a few pivotal terms from the premise and hypothesis, guiding the model to the correct answer but resulting in a lower than anticipated cosine similarity score for its sentence embedding.

Despite these findings, the pursuit of methods to evaluate the utility of generated explanations remains worthwhile. Establishing a metric for immediate evaluation of explanation quality can enable more efficient improvements in model training and performance, bypassing the need for extensive downstream testing. This approach not only enhances the understanding of how explanations impact model decision-making but also contributes to the development of more reliable and transparent AI systems.

7 Conclusion

In this project, we tackled Natural Language Inference (NLI) on the e-SNLI dataset using advanced models like RoBERTa and GPT-2. Complementing the knowledge we gained throughout the course, this project helped us explore fundamental NLP concepts such as tokenization, minimum edit distance while encouraging deeper exploration into transformer architectures, attention mechanisms

[5] https://sbert.net/docs/pretrained_models.html

and sequence generation further. The project also allowed us to delve into the mechanics of sequence generation and to understand various evaluation metrics deeply.

Through systematic experimentation, we assessed the capabilities and limitations of the models used. This process highlighted several practical challenges and areas for potential improvement in NLI systems.

7.1 Challenges & Limitations

The project faced significant computational and methodological challenges:

- The extensive dataset required substantial computational power, which limited the frequency and scope of our experiments. This was a critical bottleneck in testing and optimizing the models comprehensively.
- Due to memory constraints, we were limited to using only the base and medium variations of the GPT-2 model. Larger models, which might improve performance due to their greater capacity, were not feasible within our resource limits.
- The GPT model functioned as a "black box," making it challenging to predict or understand how changes in prompts might affect the output. This unpredictability necessitated a trial-and-error approach to optimize prompt design and model tuning.

7.2 Future Improvements

In response to the challenges and limitations encountered in our current study, we propose the following strategic improvements to enhance our research and application:

- Optimizing Computational Resources: To manage the high computational demand observed, we propose the implementation of more efficient data processing and model training techniques. Utilizing distributed computing and scalable cloud-based GPU resources can help in mitigating computational constraints. Additionally, adopting mixed precision training could be a strategic move to decrease memory usage while speeding up training times, without significant performance trade-offs.
- Exploring Larger Model Variants: Given the constraints in exploring larger GPT-2 models due to resource limitations, future initiatives should focus on securing funding or forming partnerships that provide access to enhanced computational facilities. This would enable us to explore the potential benefits of larger models such as GPT-2 Large and XL. A phased scaling strategy—starting from smaller models and incrementally moving to larger ones—will allow for efficient resource use and optimal model tuning.
- Enhancing Model Interpretability and Prompt Engineering: To better understand the underlying mechanisms of the GPT model's text generation, we will integrate interpretability tools such as feature visualization and attention mapping. This will assist in refining our prompt engineering strategies. Furthermore, automating the prompt generation and testing process will streamline the trial-and-error method, thus improving the overall efficiency and effectiveness of model outputs.
- Expanding and Diversifying the Dataset: Our dataset will be expanded to include more diverse sources such as CoS-E and ECQA, which contain a variety of explanation lengths and formats. This expansion will aid in generalizing model performance across broader datasets. Additionally, implementing data augmentation strategies will simulate a larger dataset, providing deeper insights into model behaviors across diverse textual contexts.
- Implementing Incremental Learning: We aim to incorporate incremental learning techniques that allow the model to adapt to new data continuously without losing previously acquired knowledge. This approach is essential as we expand our dataset and integrate evolving data types, thus maintaining a robust learning trajectory.
- Benchmarking and Comparative Analysis: Regular benchmarking against state-of-the-art models will be conducted to ensure our models remain competitive and effective. Comparative analysis will further allow us to understand the performance variations across different GPT model configurations and align our strategies accordingly.

References

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. E-SNLI: Natural language inference with natural language explanations.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *CoRR*, abs/2004.14546.

Dheeraj Rajagopal, Vidhisha Balachandran, Eduard H Hovy, and Yulia Tsvetkov. 2021. Selfexplain: A self-explaining architecture for neural text classifiers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. In *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics. [link].

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Generating token-level explanations for natural language inference. *CoRR*, abs/1904.10717.

Yangqiaoyu Zhou, Yiming Zhang, and Chenhao Tan. 2023. Flame: Few-shot learning from natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Acknowledgements

We would like to extend our gratitude towards Prof. Min and Prof Christian for giving us this opportunity to work on such a wholesome project. Also, Esther Gan, our mentor, who gave us insights on how to approach the project and clarified our doubts from time-to-time, enriching our project experience.

Statement of Independent Work

1A. Declaration of Original Work. By entering our Student IDs below, we certify that we completed our assignment independently of all others (except where sanctioned during in-class sessions), obeying the class policy outlined in the introductory lecture. In particular, we are allowed to discuss the problems and solutions in this assignment, but have waited at least 30 minutes by doing other activities unrelated to class before attempting to complete or modify our answers as per the class policy.

Signed,[A0216040Y, A0226576W, A0238913Y, A0194490X, A0221977W], e0538141, e0638862, e0773511, e0376920, e0559509@u.nus.edu

Appendix

Labels	Precision	Recall	F1
Entailment	0.92	0.92	0.92
Neutral	0.78	0.96	0.86
Contradiction	0.96	0.73	0.83

Table 5: Test results using GPT-4 generated explanations

type	count	mean	std
Wrong label	253	0.469	0.153
Correct label	247	0.472	0.145

Table 6: Semantic similarity summary statistics

Table 7: Evaluating explanations before Post-Processing

Version	**BL**	**MT**	**R1**	**RL**	**BP**	**BR**	**BF**	**WMD**
v1 (raw)	0.005	0.093	0.065	0.046	0.803	0.849	0.825	1.150
v2 (raw)	0.030	0.159	0.176	0.139	0.831	0.860	0.846	0.985
v3 (raw)	0.024	0.162	0.151	0.114	0.826	0.859	0.842	1.006

Table 8: Evaluating explanations after Post-Processing

Version	**BL**	**MT**	**R1**	**RL**	**BP**	**BR**	**BF**	**WMD**
v1 (spelling)	0.005	0.100	0.065	0.046	0.797	0.845	0.820	1.150
v1 (redundant)	0.005	0.096	0.067	0.046	0.802	0.846	0.823	1.149
v2 (spelling)	0.031	0.161	0.177	0.140	0.827	0.859	0.843	0.987
v2 (redundant)	0.030	0.159	0.176	0.139	0.831	0.860	0.846	0.985
v3 (spelling)	0.024	0.162	0.151	0.114	0.826	0.859	0.842	1.006
v3 (redundant)	0.024	0.162	0.151	0.114	0.826	0.859	0.842	1.006

Table 9: Classifier Performance for Generated Explanations - before and after Post-Processing

Version	**Weighted F1**	**Micro F1**	**Macro F1**	**Accuracy**
v1 (raw)	0.380	0.341	0.308	0.34
v1 (spelling)	0.411	0.339	0.272	0.34
v1 (redundant)	0.383	0.343	0.307	0.34
v2 (raw)	0.494	0.492	0.491	0.49
v2 (spelling)	0.483	0.478	0.474	0.48
v2 (redundant)	0.495	0.494	0.493	0.49
v3 (raw)	0.519	0.494	0.466	0.49
v3 (spelling)	0.511	0.512	0.512	0.51
v3 (redundant)	0.521	0.494	0.463	0.49

Table 10: Sample Explanations Generated by v2

Premise	**Hypothesis**	**Machine Explanation**	**GPT-4 Explanation**	**Human Explanation**
A couple walk hand in hand down a street	A couple is walking together	The answer is that the two are walking hand-in-hand (1) The two walk hands in the same direction. (2) They are both walking hands- in opposite directions. The fact that they are standing in a straight line is not a coincidence. It is a fact	Walking together as a couple inherently includes walking hand in hand if it is specified they are doing so; thus, if it's stated they are walking together, the scenario of them walking hand in hand fits naturally within that broader description.	If a couple is walking hand in hand it is implied that they are walking together.
A person dressed in a dress with flowers and a stuffed bee attached to it is pushing a baby stroller down the street	A person outside pushing a stroller	The answer is that the person pushing the strollers is not a child (emphasis added)	Pushing a stroller outside aligns directly with the more detailed scenario of a person in a specific outfit pushing a baby stroller. The additional details of the dress and bee do not affect the truth of the basic action of pushing the stroller.	Pushing a stroller down the street implies that the person is outside pushing the stroller.
Several women are playing volleyball	They women are hitting a ball with their arms	The answer is that the women who are doing the hitting are not playing the ball. They are just playing with the arms (The women in the video are actually playing a volleyball.),:"The video shows that women have a tendency to hit the balls with a lot	Playing volleyball involves hitting a ball with their arms, which is a central action in volleyball. Therefore, women hitting a ball with their arms directly supports that they are playing volleyball.	For women to play volleyball, they must hit the ball with their arms.

Premise	**Hypothesis**	**Machine Explanation**	**GPT-4 Explanation**	**Human Explanation**
A person in a red shirt is mowing the grass with a green riding mower	A person in red is swimming in the olympics	The answer is that the person who is in an orange shirt and is not in any other color is a mow the lawn with the green mowers. (See also: "Why is it that a man in blue is wearing	Mowing the grass with a mower and swimming in the Olympics are activities that are entirely distinct and mutually exclusive based on both the nature of the activities and their settings. If a person is swimming in the Olympics, they cannot simultaneously be mowing the lawn, and vice versa.	One can't be mowing the grass and swimming at the same time.
The girls walk down the street	Girls were going to the park	The answer is that the answer to that question is not clear (1) The girls are walking down a street. The park is a public place. It is the place where the children are going. If the kids walk up the sidewalk, they are not going down. They are just	Walking down the street does not necessarily mean the girls are going to the park; they could be headed anywhere, and without specific information linking their walk to a destination like the park, the relationship between these two scenarios remains speculative.	Girls walking down the street are not always moving to park.