Improving ExpBERT: Representation Engineering with GPT-Generated, Natural Language Explanations

Yash Parikh
Princeton University
yparikh@princeton.edu

Evan Wang
Princeton University
evanwang@princeton.edu

Nikhil Ajjarapu
Princeton University
nikhila@princeton.edu

Abstract

Recent work has suggested manual feature creation can be replaced with natural language understanding (Murty et al., 2020). Murty's approach involves using a set of explanations interpreted by a BERT model to improve performance on relationship extraction tasks. We study the importance of these explanations through ablations to the original work. After replicating the baseline results of the paper, we further study the Disease dataset, ablated to use a BERT instead of SciBERT model. We then study the effects of changes in the quantity and quality (source) of the explanations to see how these affect model performance. More explanations are found to improve model performance, as are higher quality explanations. We find that increasing the amount of training data (in this dataset) does not improve F1 as much as adding even a small amount of explanations. Through our exploration, our work establishes the validity of using other LLMs to provide natural language explanations as a way to improve performance.

1 Introduction and Motivation

In the contemporary setting of neural networks, it is unclear how to incorporate manual feature engineering into neural architecture. This tradeoff has led to powerful neural network-based models, but has also decreased the tangible control and understanding we have over what these models learn, particularly in the field of Natural Language Processing where textual comprehension is crucial. Thus, we examine one method of regaining control and understanding in neural architecture via explanations. Specifically, consider the following task: given a sentence containing a chemical and a disease, identify whether or not the chemical *causes* the disease. By introducing explanations-such as "Symptoms of {disease} appeared after exposure to {chemical}" as features to be used in the classifier model, we are able to increase our control and understanding of the model. In this project, we reproduce Murty et al. 2020's paper "ExpBERT: Representation Engineering with Natural Language Explanations." Murty et al. test various kinds of explanation features on model performance, namely the F1 score of identifying causal chemical-disease relations. These features are produced by feeding explanations through a BERT model with the training objective of sequence-entailment classification. The BERT output is then fed to a classifier network that predicts the entities' relationship. The various features include a baseline of no explanations (No-Exp), semantic-parser based explanations (Sem-Parser), n-gram patterns as explanations (Patterns), explanation-entailment probabilities (Prob), and finally natural language explanations (ExpBERT). The paper finds that the higher-level, nuanced explanations of the ExpBERT approach yielded the best performances. Through our reproduction of Murty et al.'s paper and subsequent ablations, we strive to achieve the following five goals:

1. Replicate Murty et al.'s results, verifying and correcting their codebase as necessary.
2. Understand how the **quality** and **quantity** of explanations affect model performance.
3. Investigate to what extent explanations can be a **replacement** for large-scale pre-training and fine-tuning of a model on a specific task.
4. Evaluate the potency of **LLM-generated explanations** in producing relevant features.
5. Determine if the usage of explanations can reduce the amount of training required to achieve adequate model performance.

As a general overview, we were successfully able to reproduce Murty et al.'s results. We found that the quantity of explanations does matter, albeit the effect plateaus, and quality may matter

to some extent. In addition, we find that explanations were successful in substituting large-scale pre-training/fine-tuning, and that GPT-generated explanations were able to match the performance of human-generated explanations. Furthermore, we find that explanations provide promise in reducing the amount of training data required for adequate model performance.

2 Related Work

Traditional feature engineering approaches use rule-based approaches to define a feature (Zhou et al., 2005). In the modern era of deep learning, previous approaches have relied on using semantic parsers to transform natural language explanations into executable logical forms, such as regular expressions (Srivastava et al., 2017), and into labeling functions (Hancock et al., 2018). However, these are still rule-based approaches that do not rely on the same properties of deep learning that make today's neural networks so successful.

To solve this, (Murty et al., 2020) uses distributed language representations – namely BERT (Devlin et al., 2018), fine-tuned on the natural language inference dataset MultiNLI (Williams et al., 2018) – not semantic parsers, to generate features from natural language representations that augment the neural representation with inductive biases. This method generates much more expressive learned features, resulting in higher F1 scores across each dataset examined.

Similar work, such as (Camburu et al., 2018) uses instance-level explanations instead of creating a global set of features from language as (Murty et al., 2020) does. Although the researchers do not explicitly compare the effect of implementing either global or instance features, it is clear that generating instance-level explanations manually would be much more labor intensive. Already, (Murty et al., 2020) cites that writing a set of global explanations is time-intensive, even though a small number of explanations "can significantly and disproportionately reduce the number of labeled examples required."

With the number of, and quality of human-generated explanations being the bottleneck, automation of these explanations is an area of interest, especially considering the advent of Large Language Models.

(Liu et al., 2022) uses GPT-3 to develop examples for its NLI dataset in a similar step that (Murty et al., 2020) manually took to write their explanations. However, to our knowledge, there is not any analysis of how varying the quantity or quality of these auto-generated explanations might affect a task that is downstream of NLI, such as classification of relationships. Our approach will explore these unanswered questions by using multiple GPT models of varying performance to generate examples for a relationship extraction task to measure their effect on classification performance.

3 Data

The work by Murty et al. focused on three different datasets, each with their own corresponding, author generated explanations as well.

Dataset	Train	Val	Test	Exps.
Spouse	22055	2784	2680	40
Disease	6667	773	4101	29
TACRED	68124	22631	15509	128

Table 1: Comparison of Models for Spouse and Disease. Exps. means explanations

3.1 Descriptions

The spouse dataset involves classifying two different entities to see if they are married. The disease dataset involves classifying is a chemical is a cause of a disease, and the TACRED Dataset involves classifying two entities into a set of 41 different categories (ranging from siblings to age to city of death). The explanations were generated by the authors by randomly sampled 50 training examples, and then creating explanations based on the labels for each of them. The authors specify that each explanation took around a minute to generate.

3.2 Choice for Ablations

We choose to focus on the Disease dataset, for a number of reasons. The spouse dataset took 4 hours to just generate features for, let alone train and test, using Princeton's Ionic Cluster. Additionally, the Disease dataset, we reason, is something that is more likely to come up during BERT's pre-training process on WikiCorpus data (Devlin et al., 2018)

This means that BERT already is likely to have an understanding of spousal relationships, while it may not have as strong of an understanding of cause-and-effect with respect to disease. This is validated by the fact that SciBERT on the Disease

dataset with no explanations performs worse in the initial paper than BERT on the spouse dataset (52.9 vs 49.7). This makes the disease dataset an ideal space to conduct research on how well explanations can guide relation understanding. In addition, the paper uses SciBERT instead of BERT on the disease dataset, but for our ablations we use BERT only. The justification is two-fold. First, in terms of feasibility, the codebase did not make their SciBERT interpreter (fine-tuned on a natural langauge inference dataset) available. Second, by using BERT instead of SciBERT, we are able to eliminate SciBERT's advantage of being pretrained on a large multi-domain corpus of scientific publications, and thus better isolate the influence of explanations in creating relevant classification features. In other words, in using BERT with the disease task, we can examine to what extent explanations can replace large-scale model re-configuring via fine-tuning/pretraining which SciBERT utilizes, contributing to our **third goal** outlined in the introduction.

It is also important to note here that `Disease` is a primarily negative dataset (it consists of 79.2% negative examples), and thus, throughout both the author's and our work, we use F1 score as a metric to track performance.

3.3 Examples

The following examples, from (Hancock et al., 2018)

1. Young women on replacement **estrogens** for ovarian failure after cancer therapy may also have increased risk of **endometrial carcinoma** and should be examined periodically

2. Both cohorts showed signs of **optic nerve toxicity** due to **ethambutol**.

Previous examples would have used explanations such as literally parsing out that "risk of" occurs before disease in order to classify these as examples of diseases caused by chemicals.

3.4 State-of-the-art

When Li et al., 2016 released their chemical-disease relation task in 2016, the best F-score was 62.80 (Wei et al., 2016). Since then, Yang et al., 2018 show that the best models have improved to F-scores of around 70-such methods take advantage of convectional neural networks/incorporating semantic dependency parsing and graph-based prior knowledge as features. These graph-based prior knowledge feature require extensive databases that already have entity-relations mapped, as well as additional effort to determine graph properties (path types, number of paths, etc.) that complicate the model (Zhou et al., 2018). Thus, we see explanations as a simpler, more feasible option for many tasks that has the potential to aid in relation identification without requiring the complications of convolution/graph data.

With state-of-the-art techniques achieving around 70 F1, we note that disease entity relation is not an easy task. We will use these numbers to place our f-scores in context, which will be discussed further in the results section.

4 Model Architecture

This work reproduces and expands on Murty et al. 2020's work, so our model is based upon theirs.

For the task of relationship extraction, we must classify the relationship $y \in \mathcal{Y}$ between entities o_1 and o_2 given $x = (s, o_1, o_2)$, where s is a sentence. $\mathcal{Y}$ includes a relationship pertaining to "no relationship" as well. The explanations are a global set of natural language explanations $\mathcal{E} = e_1, e_2, ..., e_n$ that might indicate a possible relationship y, but are not tied to any single one.

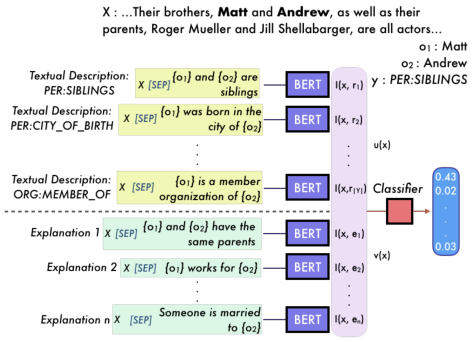

Figure 1: Model Architecture from Murty

ExpBERT defines an *interpreter* $\mathcal{I}$ as a function which takes input x and an explanation e_j and generates an $\mathbb{R}^d$ feature vector. On the `Disease` dataset, Murty et al. uses SciBERT as $\mathcal{I}$ on an input in the form of the explanation e_j and premise x separated by a [SEP] token, with a [CLS] token that will represent the data as a 768-dimensional vector.

Concretely,

$$\mathcal{I}(x, e_j) = \text{SciBERT}([\text{CLS}], s, [\text{SEP}], e_j). \quad (1)$$

For n explanations, these are concatenated to form what the authors name the *explanation representation*:

$$v(x) = [\mathcal{I}(x, e_1), \mathcal{I}(x, e_2), ..., \mathcal{I}(x, e_n)] \in \mathbb{R}^{768n}. \quad (2)$$

Additionally, the authors define a *input representation* by mapping a textual description r_i to each class $y \in \mathcal{Y}$, applying $\mathcal{I}$ to each pair of x, r_i, and concatenating them to get:

$$u(x) = [\mathcal{I}(x, r_1), \mathcal{I}(x, r_2), ..., \mathcal{I}(x, r_{|\mathcal{Y}|})] \in \mathbb{R}^{768|\mathcal{Y}|}. \quad (3)$$

The authors then train a MultiLayer Perceptron:

$$f_\theta x = \text{MLP}[u(x), v(x)]. \quad (4)$$

5 Implementation Details

5.1 Technical Information

We conducted our experiments in either 1) a local conda environment utilizing a GeForce RTX 2060 GPU, or 2) a NVIDIA Tesla K80 GPU accessed through nodes501 and 502 on Princeton's Ionic computing cluster. We found that is takes approximately one and a half hours on the cluster to build features, train and evaluate one model on the disease dataset. On the spouse dataset, we found that it takes up to 6 hours, which is one of the reasons we opted to perform our ablations on the disease dataset.

5.2 Codebase Changes

In order to reproduce the paper, we utilized the features in the codebase, and coded an arg_parser.ipynb notebook in order to parse their hyperparameters for each respective experiment from the codebase's checkpoint directory-we then matched our hyperparameters accordingly. This is because the commands on how to run the methods resulted in all negative predictions– we hypothesis that this is because the authors went to put a general example of how to run the code but neglected some relevant hyperparameter (in our experience with the codebase, this frequently occured when dropout was too high, for example). We additionally needed to fix the BERT interpreter imports– one of the BERT models was loaded from an absolute path from the author's computer. Furthermore, Murty et al. 2020's provided code did not have the full method of creating BERT-produced features from 0 explanations, so we created our own interpreter method, batch_interpret _noexp() function in feature_factory/interpreter.py, as well as a new create_features_noexp.py file to call this function noexp. Ultimately, these additions were used to run the create features based on 0 explanations to feed to the classifier network, in both reproducing the paper and running our own ablations. Finally, the Murty et al. 2020's original codebase either had buggy/outdated code in feature_factory/bert _utils.py, which is called by feature_factory/interpreter.py and in turn create_features.py to generate features. Specifically, the run_bert() function had several run-time errors related to its usage of interpreters-for the output of BERT, the function failed to actually obtain the 768-dimensional vector at the [CLS] token as the sequence representation. Our modified code in run_bert() achieves this goal and overall is able to create features from explanations.

6 Baselines

First, as baselines, we reproduced Murty et al. 2020's results on both the spouse and disease datasets with their baseline models and explanation variants. We did not reproduce results with the TACRED dataset as it is not easily publicly available; moreover, the paper did not focus much on TACRED with respect to the reasoning behind their experimentation. For spouse and disease, the paper examined NoExp, BERT + ProbExp, BERT + LangExp, BERT + Patterns, and ExpBert. We replicated all of these using the feature weights provided by the authors, and we validated their results creating our own feature vectors. Results are shown below:

	Disease		Disease	
Model	Ours	Murty	Ours	Murty
NoExp	54.03	52.90	48.05	49.70
ProbExp	60.10	58.30	52.13	49.70
LangExp	53.73	53.60	50.14	49.10
Patterns	55.62	53.30	48.77	49.00
ExpBERT	62.43	63.50	54.34	52.40

Table 2: Comparison of Models for Spouse and Disease. We reproduced the paper's values to within 5% error.

7 Methods

Our work is centered around our research questions, as outlined in the introduction.

We hypothesize that implementing explanations using GPT models will create additional high-quality, quickly-generated explanations that will improve baseline BERT performance on the `Disease` dataset towards SciBERT performance.

We do this through a number of different ablations and changes to the original paper, namely:

1. We use generate features for and train and evaluate a model that uses BERT instead of SciBERT to featurize the explanations in an effort to better evaluate the effectiveness of explanations.

2. We test the model on 0, 10, 29 (29 being the full set) author-generated explanations to test the impact of **quantity** of explanations on model performance.

3. We create 87 novel explanations using GPT-3, GPT-3.5, and GPT-4 to test the impact of **quality** of explanations on model performance.

4. We vary the number of each of these generated explanations to see how the incremental change in explanations affects results for each type of model.

5. We augment the author's explanations with a GPT4's explanations in an effort to see how the model performance with a large number of high quality explanations.

6. We train NoExp and ExpBERT with ChatGPT-generated explanations with varying amounts of data to test to what degree explanations reduce the need for more training data.

Note that in for the MultiLayer Perceptron's hyperparameters, we matched the corresponding values from Murty et al., i.e. we used the hyperparameters they used for ExpBERT on the disease dataset.

8 Results

Our replication of the author's metrics are detailed in Table 2. Table 3 contains the evaluation metrics for running the same model as the authors, but with different amounts of explanations used to generate the features. The explanations were all balanced by the amount of positive and negative explanations. The data for how the amount of training data affects F1 performance is detailed in table 4, and table 5 gives the confusion matrix for our best trial: combining Human and GPT4 Explanations.

9 Discussion

9.1 General Performance

Our best performance of 53.94 F1 was reached by using the full set of human-generated explanations along with the full set of GPT 4 explanations (Table 5). This is only marginally higher than the results from all of the human explanations, and we hypothesize that this might be the case becuase of the way we prompted: we had GPT4 created examples resembling the authors. We note that this is lower than the state-of-the-art values of around 70 F1. However, we think that given these state-of-the-art techniques' reliance on heavier models that use complicated convolutions or extensive graphs/dependency parsers, explanations give a simpler, more feasible way to increase model performance when other options may not be available.

9.2 BERT vs SciBERT

In achieving our third goal of comparing the power of explanations to large-scale pre-training and fine-tuning, we can examine our default BERT based results juxtaposed with Murty et al.'s SciBERT results. SciBERT, without explanations, reached an F1 score of 49.70 (48.05 in our reproduction), while BERT without explanations reached 35.83, which shows a clear dropoff. This agrees with our hypotheses, as default BERT is more perplexed by the highly technical nature of the disease dataset. We see that without explanations, the SciBERT model with vast amounts of scientific corpus training significant outperforms the default trained BERT model. On the other hand, with 29 explanations, default BERT was able to match SciBERT's F1 of 52.40 in the paper (54.34 in our reproduction): human-generated explanations achieved 53.01 F1. Thus, introducing explanations to both SciBERT and BERT were able to "even the playing field" between the two transformer models. We see this as promising evidence that explanations can serve as a substitute for large-scale pre-training that may not be feasible due to computational constraints or lack

[1] results do not converge

Explanations	Author	GPT-3	GPT-3.5	GPT-4
0	35.82	35.82	35.82	35.82
10	51.94	44.79	47.97	40.11
20	52.75	46.91	50.00	51.65
29	53.01	49.08	51.92	52.68

Table 3: Number of Explanations vs. F1 score

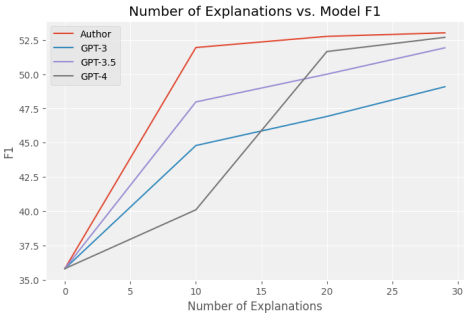

Figure 2: Number of Explanations vs. F1 score

% of Data	Baseline (0 exps)	Author (29 exps)	GPT 3.5 (29 exps)
10	31.39	N/A [1]	0.4325
20	34.96	43.74	46.08
30	38.50	44.17	47.72
40	34.58	44.41	49.30
50	36.41	47.53	49.79
60	33.30	46.70	50.23
70	35.84	48.96	51.72
80	32.27	51.18	52.64
90	33.27	52.13	52.66
100	35.82	53.16	51.87

Table 4: Results for the amount of data

Human and GPT4 Explanations Combined
F1 Score: **53.94**

		True Class	
		Positive	Negative
Predicted Class	Positive	680	905
	Negative	256	2260

Table 5: Best Results: 58 explanations (29 Human, 29 GPT4)

of labelled corpus, especially for niche scientific fields.

9.3 Quality of Explanations

Generating the Explanations

In order to generate the explanations, we use the author's explanations as a part of prompt asking to generate positive (and negative) explanations. We chose to do it this way in an effort to most closely replicate the structure of the other explanations, given the models' propensity to be verbose. We also prompt separately for positive and negative examples so that there is no confusion about whether or not the explanations should be for a sentence belonging to class or not belonging to the class.

Impact of Quality Changes

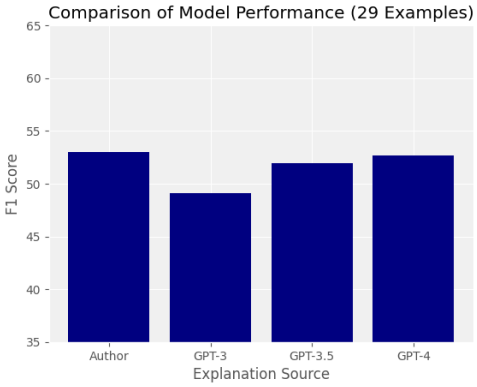

Figure 3: Explanation source vs. F1 score. The number of explanations are constant at 29, and all model features are generated using BERT. Author indicates the human-crafted explanations used in Murty et al.

We find that the quality of explanations encoded by BERT increases the model's performance on `Disease`, with GPT-4 having an F1 Score just .33 lower than the author-generated explanantions. We see a trend across all of the models– when using a sufficiently large amount of explanations (such as 20 or 29), each successive GPT model provides a higher F1 score. Our results corroborate what we

would suspect– GPT3 is quite verbose and thus it does not help the model create patterns that would be easily recognized. As an example, one of the explanations from GPT3 is the following:

A correlation between ingestion of {e1}, and occurrence of symptoms related to, has been established by research on patients suffering from, respectively, {e1}, and {e2}.

This example is a poor quality explanation: it uses e1 as both a chemical and a complication in the same sentence.

9.4 Quantity

Quantity of Human Explanations

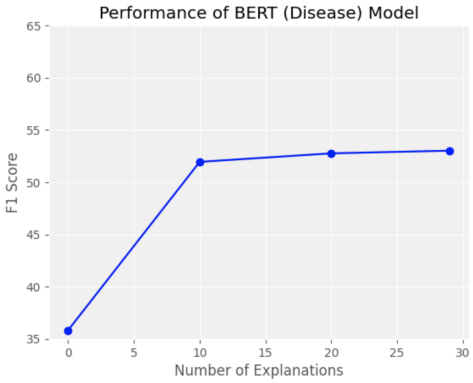

Figure 4: Number of Explanations vs. F1 Score. The explanations are randomly chosen from Murty et al.'s set of 29, and the features are generated via BERT.

We note that in the base case, our BERT model (F1 score of 35) performs worse than SciBERT (F1 score of 49.7). We attribute this to the fact that the base BERT model is not trained on science-specific information, and as a result, cannot understand the highly technical dataset. As seen in Figure 4, however, it does not take many explanations, however, for the model to improve its F1 score. Within 10 explanations the improvement is near 15 points. With each additional explanation, the F1 score improves, but it improves less.

Quantity of Generated Explanations

The results become more interesting when considering the effects of quantity changes across all of the models.

We find that at lower amounts of explanations, performance is very depedent on the model itself. As an example, GPT3.5 with 10 explanations performed better than GPT3 with 20 explanations, but worse than GPT4 with 20 explanations. Similarly, 10 GPT4 explanations led to the lowest F1 performance across the board, despite GPT4 providing the best explanations when they exist in high quanities. We hypothesize that this is becuase the explanations are materially different from one another and thus with just a few explanations, how representative of the broader class each one is matters more.

One of the most important trends from the data is that as the number of explanations from each source increases, performance strictly increases. This confirms our hypothesis that more explanations allow for better model performance. However, the path that this follows is different (once again alluding to the issue of the quality when there are low amounts). GPT 4 explanations offer only a 5 point in F1 with 10 explanations, while 10 of the authors lead to a 16 point increase.

The data around quality and quantity show us that: as quality increases, performance increases, and likewise with quantity. However, at lower quantity levels, quality matters more.

9.5 Amount of Training Data

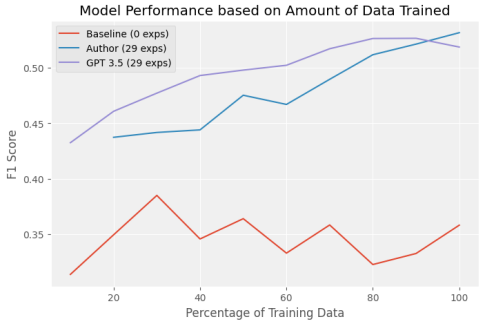

Figure 5: Percentage of Training Data used vs. F1 score, shown for all explanation sources.

Looking towards our fifth goal of examining how explanations affect the amount of training required, Figure 5 yields key insights. First, it is evident that even when explanation-based models train on just 10% of the data, the no-explanation model was outperformed even when trained on the entire dataset. Moreover, generally, if we have a specific f1 we want to reach, we see that explanations have the potential to drastically decrease the amount of training data required to achieve that goal. This property means time and resources saved not only on computation, but also on labelling data.

10 Conclusions

10.1 Limitations & Future Work

First and foremost, our results are limited to this dataset. While we hypothesize that our results can be generalized across datasets, we have yet to show that empirically. Additionally, our results use the same hyperparameters that the authors use– given more time, we would have found those hyperparameters ourselves in an effort to maximize each F1 performance. Prompting is an area we feel we could have improved, as we fed examples into the model instead of asking GPT to generate sentences explaining chemical-disease relationships. Future work would need to study other datasets and expand our work with other sets of explanations: nonetheless, we find the following conclusions compelling

10.2 Conclusions

Through our novel study into the importance of explanations in feature generation for relationship extraction tasks, we find that the explanations matter in both quality and quantity. At low quantity levels, quality matters more, but both are correlated with increasing F1 score. We note that automatically generating explanations from LLMs is a viable method for creating explanations, and could save a large amount of time for researchers. Additionally, we've shown that explanations can be more important that additional training data, offering another avenue for researchers to save time and resources when working with relationship extraction tasks.

11 Acknowledgements

We'd like to thank Danqi Chen, Karthik Narasimhan, Howard Chen, and the remainder of the course staff for their guidance and support with this project. This document was created using the ACL 2023 template.

12 Code

Our code can be viewed at https://www.dropbox.com/sh/wdmilcam1oh9nvy/AACtJGYqa3fPDv0o_UnD7IYma?dl=0.

References

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. Training classifiers with natural language explanations. *CoRR*, abs/1805.03818.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan P Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database : the journal of biological databases and curation*, 2016:baw068.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: worker and AI collaboration for natural language inference dataset creation. *CoRR*, abs/2201.05955.

Shikhar Murty, Pang Wei Koh, and Percy Liang. 2020. Expbert: Representation engineering with natural language explanations. *CoRR*, abs/2005.01932.

Shashank Srivastava, Igor Labutov, and Tom Mitchell. 2017. Joint concept learning and semantic parsing from natural language explanations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1527–1536, Copenhagen, Denmark. Association for Computational Linguistics.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan P Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2016. Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. *Database : the journal of biological databases and curation*, 2016:baw032.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Zhifang Yang, Wei Wang, Zhongqiang Wu, and Yongsheng Zhang. 2018. Neural word segmentation with rich pretraining. *Computational Linguistics*, 44(3):427–438.

GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 427–434, Ann Arbor, Michigan. Association for Computational Linguistics.

Huiwei Zhou, Shixian Ning, Yunlong Yang, Zhuang Liu, Chengkun Lang, and Yingyu Lin. 2018. Chemical-induced disease relation extraction with dependency information and prior knowledge. *Journal of Biomedical Informatics*, 84:171–178.

A Appendix

A.1 Explanations

{e1} is a cause of {e2}
The symptoms of {e2} appeared after the administration of {e1}
{e2} developed after {e1}
Patients developed {e2} after being treated with {e1}
{e1} contributes indirectly to {e2}
{e1} has been associated with the development of {e2}
Symptoms of {e2} abated after withdrawal of {e1}
A greater risk of {e2} was found in the {e1} group compared to a placebo
{e2} is a side effect of {e1}
{e2} has been reported to occur with {e1}
{e2} has been demonstrated after the administration of {e1}
{e1} caused the appearance of {e2}
Use of {e1} can lead to {e2}
{e1} can augment {e2}
{e1} can increase the risk of {e2}
Symptoms of {e2} appeared after dosage of {e1}
{e1} is a chemical
{e2} is a disease
{e1} induces {e2}
administering {e1} causes {e2} to worsen
{e1} has an effect on {e2}
{e1} is used for the treatment of {e2}
{e1} is known to reduce the symptoms of {e2}
{e1} is used for the prevention of {e2}
{e1} ameliorates {e2}
{e1} causes a disease other than {e2}
{e1} is an organ
{e1} is effective for the treatment of {e2}
{e1} has an attenuating effect on {e2}

Table 6: Human generated positive and negative explanations, delimited by a horizontal line

{e1} triggers the onset of {e2}
{e1} contributes directly to the development of {e2}
{e2} is a known complication of {e1} use
{e1} has been implicated in the causation of {e2}
{e2} manifests as a result of exposure to {e1}
{e1} enhances the likelihood of developing {e2}
{e1} is a risk factor for {e2}
{e2} occurs as a consequence of {e1} exposure
{e1} is associated with the emergence of {e2} symptoms
{e2} arises from the use of {e1}
{e1} leads to the development of {e2} symptoms
{e1} exacerbates {e2}
{e2} is an adverse event associated with {e1} administration
{e1} is a contributing factor to the onset of {e2}
{e2} occurs as a result of exposure to {e1} over time
{e1} influences the manifestation of {e2}
{e2} is a potential complication of using {e1}
{e1} is linked to the emergence of {e2} in some patients
{e2} arises as a result of {e1} exposure and/or use
{e1} triggers a cascade of events that culminate in {e2}
{e2} is a known sequelae of {e1} exposure
{e1} is a chemical compound that is not linked to the cause of {e2}
{e1} is not associated with the development of {e2}
Although {e1} has medicinal properties, it is not a contributing factor in the onset of {e2}
{e1} is a safe chemical that is not known to induce {e2}
The use of {e1} does not result in the occurrence of {e2}
{e1} has no known harmful effects on human health, including the development of {e2}
{e1} is a non-carcinogenic chemical that does not lead to the formation of {e2}
{e1} is a chemically stable substance that does not promote the growth of {e2}

Table 7: GPT3.5 generated positive and negative explanations, delimited by a horizontal line

Exposure to {e1} has been linked to the onset of {e2}
{e1} increases the likelihood of developing {e2}
Prolonged use of {e1} has been correlated with {e2}
The presence of {e1} has been implicated in the progression of {e2}
Introducing {e1} can result in the manifestation of {e2}
{e1} has been shown to exacerbate {e2}
The occurrence of {e2} is connected to the usage of {e1}
{e1} has been identified as a contributing factor to {e2}
Ingestion of {e1} may lead to the development of {e2}
The likelihood of {e2} is heightened by the presence of {e1}
Exposure to {e1} can trigger the onset of {e2}
The chemical {e1} is known to provoke {e2}
{e1} has been observed to cause {e2} in some cases
{e2} is a known consequence of {e1} exposure
Prolonged contact with {e1} has been related to {e2}
{e1} is a potential catalyst for {e2}
The development of {e2} is associated with the use of {e1}
{e1} is known to be a risk factor for {e2}
Symptoms of {e2} are known to be triggered by {e1}
{e1} can be a causal agent for {e2}
{e1} may initiate the progression of {e2}
{e1} serves as a therapeutic agent for {e2}
{e1} helps in alleviating the signs of {e2}
{e1} is employed to hinder the onset of {e2}
{e1} mitigates the severity of {e2}
{e1} is associated with a different disease, not {e2}
{e1} is a chemical compound, not an organ related to {e2}
{e1} demonstrates efficacy in addressing {e2}
{e1} has a diminishing impact on the progression of {e2}

Table 8: GPT4 generated positive and negative explanations, delimited by a horizontal line

{e1} is a factor in the development of {e2}
The use of {e1} has been linked to an increased risk of {e2}
Patients exposed to {e1} have experienced symptoms related to {e2}
A correlation between exposure to {e1} and the onset of {e2} has been observed
An association between intake of {e1} and occurrence of {e2} has been established
Ingestion of {e1} can lead to the manifestation of {e2} symptoms
Prolonged contact with {e1} may result in the emergence of {e2} signs
Long-term exposure to {e1} is associated with an elevated risk for developing {e2}
Studies suggest that there is a connection between taking {e1} and having signs or symptoms indicative of {e2}
{e1} has been found to be a contributing factor in the development of {e2}
Research indicates that there is an association between intake of {e1} and the onset of {e2}
It has been observed that patients who are exposed to {e1} may experience symptoms related to {e2}
The use of {e1} has been linked with an increased risk for developing {e2}
Prolonged contact with {e1} may result in the emergence of signs or symptoms indicative of {e2}
Long-term exposure to {e1} is associated with an elevated risk for having signs or symptoms suggestive of {e2}
Studies suggest that there is a connection between taking {e1} and having signs or symptoms indicative of {e2}.
There is evidence suggesting that prolonged contact with {e1} can lead to the manifestation of {e2}.
A correlation between ingestion of {e1}, and occurrence of symptoms related to, has been established by research on patients suffering from, respectively, {e1}, and {e2}.
Patients treated with {e1} have developed {e2} as a side effect .
After administration , {e1} was found to cause {e2} in some cases .
Ingestion of {e1} can increase the likelihood for developing {e2} .
{e1} is not the cause of {e2}
{e1} does not trigger {e2}
The presence of {e1} does not lead to the development of {e2}
There is no evidence that suggests that {e1} causes or contributes to the onset of {e2}
It has been established that there is no causal relationship between {e1} and {e2}
The use of {e1} does not result in an increased risk for developing {e2}
No correlation exists between exposure to {e1} and incidence rates for {e2}
Research indicates that there is no link between ingestion/exposure to {e1} , and occurrence of symptoms associated with {e2}

Table 9: GPT3 generated positive and negative explanations, delimited by a horizontal line

A.2 Prompts

Positive Prompt:
I am creating natural language explanations that indicate that the first entity, denoted as {e1}, is a chemical and is a cause of the second entity, denoted as {e2}, which is a disease. Here is my current list:

Come up with 21 additional explanations. These explanations may be paraphrasings of ones in my current list, but you may not repeat any verbatim in your generated list. You must also include {e1} and {e2} in each explanation. In your response, give a well-formed output and number each explanation.

Negative Prompt:
I am creating natural language explanations that indicate that the first entity, denoted as {e1}, is a chemical and is not a cause of the second entity, denoted as {e2}, which is a disease. Here is my current list:

Come up with 8 additional explanations. These explanations may be paraphrasings of ones in my current list, but you may not repeat any verbatim in your generated list. You must also include {e1} and {e2} in each explanation. In your response, give a well-formed output and number each explanation.