Enhancing Natural Language Understanding: A Practical Evaluation using e-SNLI and ChatGPT

Marijana Petojevic 12017529
Scientfic writing Witersemester 2023/24
marijana.petojevic@tuwien.ac.at

Abstract

In the dynamic field of natural language processing (NLP), the assessment of models against diverse and challenging datasets plays a pivotal role in unraveling their capacities and limitations. This paper builds upon the fundamental work presented in paper *e-SNLI: Natural Language Inference with Natural Language Explanations*(1) by Oana-Maria Camburu. This paper embarks on a practical exploration of NLP models. This paper will focus on conducting tests on the e-SNLI(2) dataset—a natural language inference benchmark that extends the influential Stanford Natural Language Inference (SNLI) dataset. The e-SNLI dataset introduces a layer of human-annotated explanations, providing an avenue to investigate the interpretability and robustness of NLP models.

0.1 Methodology

Our study employs ChatGPT, a state-of-the-art language model, to scrutinize the e-SNLI dataset. The methodology involves the careful construction of prompts that mirror the dataset's structure, encapsulating premises and hypotheses characteristic of natural language inference tasks. The goal is to assess the model's proficiency in inferring relationships—be it entailment, contradiction, or neutrality—aligning with the annotations provided by human evaluators in the e-SNLI dataset.

0.2 Results and Analysis

The results of our experimentation offer a perspective on ChatGPT's performance. Instances of accurate inferences are compared side by side with challenges faced by the model, highlighting their skill in handling the complexities of human language. We're not just looking at how accurate the model is. We're digging into how it thinks and finding patterns in what it says. This helps us connect what the model says with how people understand things, making NLP systems easier to understand and trust.

0.3 Real-Life Applications

Beyond the technical evaluation, this paper seeks to draw connections between the insights gained from testing e-SNLI with ChatGPT and the practical applications of NLP in real-life scenarios. We explore the potential impact of such models in domains such as customer support, content understanding, and information retrieval.

0.4 Conclusion

In conclusion, our contribution extends beyond the theoretical realm to the practical applications of NLP models. Leveraging the e-SNLI dataset and ChatGPT, our investigation delves into the model's functionality and its potential real-world applications. We underscore the importance of models that offer not just accurate

predictions but also comprehensible justifications. In a world increasingly reliant on advanced technologies, the ability to trust and comprehend the decisions made by these models is paramount.

1 Introduction - Natural Language Processing Motivation

Natural Language Processing (NLP) is a field at the intersection of computer science, artificial intelligence, and linguistics. It's concerned with the interactions between computers and human (natural) languages. The primary goal of NLP is to enable computers to understand, interpret, and respond to human language in a valuable and meaningful manner. NLP systems can process both written text and spoken words, but for our research, we are going to be focusing on processing written text.

1.1 Steps in Natural Language Processing

The process of NLP involves several crucial steps to understand and generate human language:

- **Tokenization:** Breaking down the text into smaller units (like words, phrases, or sentences), known as tokens.
- **Normalization:** Includes converting all characters to lowercase, removing punctuation, or standardizing date formats.
- **Part-of-Speech Tagging and Parsing:** Tokens are analyzed to understand their parts of speech (like nouns, verbs, adjectives) and the grammatical structure of sentences.
- **Semantic Analysis:** Involves understanding the meanings of individual words in the context of the sentence and grasping the actual meaning or semantics behind the sentence.
- **Context Understanding:** Advanced NLP systems use context for a better understanding of the text, recognizing that words or phrases can have different meanings depending on their use. For example, the word *bank* can have different meanings depending on whether it's used in a financial or environmental context.
- **Machine Learning and Deep Learning:** These techniques are utilized to automatically learn patterns in languages from a large amount of data, using models like neural networks.

Lastly, output from NLP can be used in various applications like translation services (*Google Translate*), chatbots, sentiment analysis (understanding emotions in text), speach recognition (*Bixby* on Samsung, *Siri* on Apple devices or *Alexa* - Amazon's representative), information extraction and more.

1.2 Natural Language Inference (NLI)

In the field of NLP, Natural language Inference is a specific task, that should be able to determine the relationship between two text peaces, typically be classifying them as **entailment**, **contradiction** or **neutral**. NLI focuses on how the truth of one statement (*the hypothesis*) follows from another statement (*the premise*). Classification can be determined as follows:

1. **Entailment:** This label is used when the first sentence logically implies the second one. For instance, if sentence A is "The cat is sleeping," and sentence B is "An animal is resting," then A entails B.
2. **Contradiction:** This occurs when the first sentence contradicts or negates the second. For example, if sentence A is "It is raining heavily," and sentence B is "The sun is shining brightly," then A contradicts B.
3. **Neutral:** This label is applied when the relationship between the two sentences is neither of entailment nor contradiction. Essentially, the second sentence might or might not be true based on the information in the first sentence.

The purpose of these labels is to train machine learning models in understanding and interpreting human language. By analyzing sentence pairs and their relationships, these models learn to process and reason about natural language, which is crucial for tasks like text analysis, sentiment analysis, and automated question answering.

1.3 SNLI and e-SNLI

The Stanford Natural Language Inference (SNLI)(3) dataset is a collection of 570,000 sentence pairs written by humans, each labeled with one of three relations: entailment, contradiction, or neutral. The pairs consist of a premise and a hypothesis, and the goal is to determine the relationship between them. This dataset is widely used for training and testing natural language processing models on the task of textual inference. SNLI has played a crucial role in advancing research in NLI, providing a large, diverse, and well-annotated resource that supports the development of models capable of understanding complex language relationships.

In research published in 2018, Gururangan et al. (4) raised questions about the actual learning process of models trained on the SNLI dataset. They suggested that these models might not be fully understanding language but rather focusing on certain patterns or "artifacts" within the data. For instance, they observed that specific words in the hypotheses often strongly indicate the label: words like "friends" or "old" frequently appear in neutral hypotheses, while "animal" or "outdoors" are common in entailment contexts, and "nobody" or "sleeping" are often found in contradictions. Their findings were underscored by the performance of a hypothesis-only model, which achieved a 67% accuracy rate on the test set. This indicates a reliance on these artifacts rather than a genuine comprehension of the text. The study further demonstrated that generating explanations based on these patterns is significantly more challenging than merely producing labels.

The e-SNLI dataset extends the original SNLI dataset by adding human-annotated natural language explanations to each of the 570,000 sentence pairs. These explanations provide insight into why a particular pair is classified as entailment, contradiction, or neutral. This addition aims to enhance model interpretability in natural language inference tasks, allowing not just for classification, but also for understanding the reasoning behind these classifications. While existing NLI models are proficient in classification tasks (identifying entailment, contradiction, or neutrality between sentence pairs), they typically do not provide insights into their reasoning processes. This opacity limits their practical utility and trustworthiness, particularly in applications requiring transparency, such as legal or medical domains. This makes e-SNLI a valuable resource for developing more transparent and explainable AI models in the field of NLP. See Figure 1 for some e-SNLI examples.

Premise: An adult dressed in black holds a stick. Hypothesis: An adult is walking away, empty-handed. Label: contradiction Explanation: Holds a stick implies using hands so it is not empty-handed.
Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her. Hypothesis: A young mother is playing with her daughter in a swing. Label: neutral Explanation: Child does not imply daughter and woman does not imply mother.
Premise: A man in an orange vest leans over a pickup truck. Hypothesis: A man is touching a truck. Label: entailment Explanation: Man leans over a pickup truck implies that he is touching it.

Figure 1: Figure shows some examples from e-SNLI. Annotators were given the hypothesis, the premise and label. They highlighted the words they considered essential for the label and provided explanations.

2 Collecting explanations for e-SNLI

The e-SNLI dataset was created using Amazon Mechanical Turk[1], aiming to explain the reasoning behind the relationships (entailment, neutrality, or contradiction) between sentence pairs. The methodology focused on encouraging annotators to identify critical, non-obvious elements determining these relationships, avoiding mere repetition of the premise in the hypothesis.

For quality control, a two-step process was implemented: highlighting key words in the sentences and then formulating explanations using these words. This approach, combined with in-browser checks and specific guidelines for each relation type, ensured the collection of meaningful, self-contained explanations. To ensure the quality of the e-SNLI dataset, an extensive analysis and refinement process was conducted. A random sample of 1,000 explanations was manually evaluated for correctness, with a scoring system ranging from 0 (incorrect) to 1 (correct). Explanations that were overly generic, repetitive, or template-like were marked as uninformative and incorrect. The team identified common unhelpful templates and filtered out explanations that closely matched these. This process revealed certain challenges, particularly in entailment pairs, where explanations were often incomplete or simply reiterations of the premise.

The final analysis showed varying error rates across different types of pairs, with entailment pairs posing the most significant challenge (error rate was 19.55%, where as for neutral 7.26% and 9.38% for contradiction, showing total error rate of 9.62%). The dataset includes explanations for each sentence pair in the training set and three explanations per pair in the validation and test sets. An extensive review process was conducted to ensure the quality and relevance of the explanations provided.

3 Experiments on e-SNLI

In her research, Camburu et al. (1) demonstrates models ability to utilize explanations for improving NLI tasks. The experiments explore various aspects:

1. **`PremiseAgnostic`**: Examines if models relying on data artifacts in SNLI can generate correct explanations. The model is given hypothesis only and is prompted to predict the entailment label. Idea of this experiment is to test undersoce the finding of previously metioned research by Gurugaran et al. that a neutral network which is given hypothesis only predicts the correct label 67% of time. Results show that only 66 out of the first 100 test examples were predicted correctly.
2. **`PredictAndExplain`**: Investigates models predicting labels and generating corresponding explanations. The experiment indicates that models can maintain label prediction accuracy while adding explanation generation capabilities. For this experiment InferSent (5) architecture was used. The e-inferSent model (trained on e-SNLI) obtained test accuracy of 83.96%.
3. **`ExplainThenPredict`**: Focuses on generating explanations first, then predicting labels. This approach results in higher quality explanations but with a slight drop in label prediction accuracy. Accuracy of the model in this experiment was shown to be 81.71%.
4. **`Represent - Universal Sentence Representations`**: Researchers aimed to refine universal sentence representations by developing an encoder that generates semantically meaningful, fixed-length outputs for phrases or sentences. This approach is critical for enhancing performance when dealing with limited labeled data. Unlike the established practice in computer vision, where pretrained ImageNet-based encoders (10) provide standard image feature extractors, the NLP community has not yet agreed upon a standard encoder. Building on the insight that supervised learning on NLI datasets outperforms traditional unsupervised methods, Cabaru et al. introduce the e-inferSent model. This model advances the InferSent approach by incorporating natural language explanations, aiming to improve the learning process.

[1]Amazon Mechanical Turk (MTurk) is a crowdsourcing platform operated by Amazon. It connects businesses or individuals who need tasks completed (known as Requesters) with workers who can perform these tasks (known as Workers). Tasks on MTurk, often referred to as Human Intelligence Tasks (HITs), can include data entry, surveys, content moderation, and more. The platform is especially useful for tasks that require human judgment or intuition, which can't be easily automated.

5. **`Transfer`**: Tests the model's ability to transfer learning to out-of-domain NLI datasets. While accuracy improvements are minimal, the model provides insights into its decision-making process through explanations.

4 Testing the e-SNLI dataset on OpenAI API

In this study, we aim to evaluate the capabilities of OpenAI's API in conducting experiments similar to those presented by Camburu et al. We will focus on how effectively the API can generate labels and explanations, as well as perform both tasks simultaneously, using selected sentence pairs from the e-SNLI dataset. Our analysis will be constrained to a subset of examples due to the rate limits imposed by the OpenAI API, offering a focused investigation into its NLI performance. For scripts used to generate resulting datasets and test OpenAI on e-SNLI dataset, please refer to our GitHub page[2].

4.1 Experiment 1: `PredictLabelGPT`

For the first experiment, we will utilize the e-SNLI dataset's sentence pairs, which include a premise and a hypothesis, along with highlighted words that are key to the relationship between the sentences. The objective is to assess how well OpenAI's API, specifically ChatGPT-3.5-Turbo, can predict the correct entailment label (entailment, contradiction, or neutral) based on given information. This will be a test of ChatGPT's understanding of natural language inference and its ability to use contextual clues to make accurate predictions.

OpenAI API Prompt To prompt OpenAI API following instruction was used, where `Sentence1` is the premise from e-SNLI dataset, `Sentence2` the hypothesis and `Sentence1_marked_1` and `Sentence2_marked_1` the marked words in both sentences which are believed to be crucial for determining the entailment relationship, see code listing 1:

```
 input_text = (
        "Based on premise and hypothesis, \n"
        "find the entailment label between them: entailment, contradiction or neutral."
        "Your answer must be one word only.\n"
        f"Premise: {row['Sentence1']} Hypothesis: {row['Sentence2']} \n"
        f"Marked Words Sentence 1: {row['Sentence1_marked_1']} Marked Words Sentence 1: {row['Sentence2_marked_1']}"
    )
```

Listing 1: Prompt for experiment "PredictLabelGPT"

Results Upon examining the initial 100 test cases, it was observed that ChatGPT successfully predicted the correct label in 66 instances. Breaking down the results by category - entailment, contradiction, and neutral - ChatGPT demonstrated varying levels of accuracy. For entailment, the accuracy was remarkably high at approximately 90.91%, and it achieved perfect accuracy (100%) in predicting contradiction. However, the performance markedly dropped for neutral relations, with an accuracy of only around 11.43%. These findings echo the hypothesis posited by Gururangan et al. (4) that a neural network, relying on data artifacts, can predict the correct label with about 67% accuracy. This variance in performance across different categories underscores the complexity and challenges in neural network-based NLI tasks. The Table 1 summarizes the results.

Table 1: Results of Experiment 1: PredictLabelGPT

Metric	**Performance**
Overall Alignment	66.0%
Accuracy for Entailment	90.91%
Accuracy for Contradiction	100.00%
Accuracy for Neutral	11.43%

[2]`https://github.com/mpetojevic/Scientific-Writing`

4.2 Experiment 2: `ExplainGPT`

In the second experiment, we will utilize OpenAI's API to explore its capacity for explanation generation. Here, ChatGPT will be given a premise, a hypothesis, followed by marked words in both, and the correct label from the e-SNLI dataset. The task for the AI will be to generate a coherent and contextually appropriate explanation that justifies the given label. AI is instructed to give one sentence explanations, short and precise, not mentioning any of the marked words or gold labels (*entailment, contradiction, neutral*). The aim is to assess the semantic alignment of ChatGPT's explanations with the human-annotated e-SNLI explanations.

OpenAI API Prompt Here OpenAI API was prompted to generate an explanation based on following input text. Human annotated label from e-SNLI dataset is named `gold_label`, see code listing 2:

```
input_text = (
        "Based on premise, hypothesis, gold explanation and marked words in sentences provide an explanation of the "
        "relationship between two sentences.\n"
        "Explanation can be at most 1 sentence, should be short and precise, not repeating gold label, "
        "hypothesis or the premise. Your explanation mustn't mention marked words! You should just explain"
        "what situation could be going on based on the two sentences.\n"
        f"Premise: {row['Sentence1']} Hypothesis: {row['Sentence2']} Gold Explanation: {row['gold_label']}\n"
        f"Marked Words Sentence 1: {row['Sentence1_marked_1']} Marked Words Sentence 1: {row['Sentence2_marked_1']}"
    )
```

Listing 2: Prompt for experiment "ExplainGPT"

Results The analysis involved a qualitative review of explanations. Given the complexities of natural language, this examination focused on the semantic similarity - whether the core meanings and implications of the explanations aligned, rather than a mere textual match. Due to the rate limits imposed by the OpenAI API, our analysis was restricted to a dataset of 60 entries. The first 20 entries were examined in detail to establish an initial understanding of the patterns in ChatGPT's responses compared to e-SNLI.

In the initial 20 cases, we observed a mixture of alignment and deviation in the explanations. There were instances where ChatGPT's explanations closely aligned with those in the e-SNLI dataset, as well as cases where the interpretations diverged significantly. This variation emphasizes the inherent complexity in natural language understanding, particularly in scenarios requiring nuanced interpretation. To illustrate these findings, a selection of instances showcasing both matches and differing interpretations is presented in Table1 2.

Table 2: Comparison of Explanations between e-SNLI and ChatGPT

e-SNLI Explanation	ChatGPT Explanation	Match Assessment
Jumping on skateboard is the same as doing tricks on skateboard	The boy on the skateboard is performing a trick.	Good Match
One cannot be smiling and frowning at the same time.	The hypothesis contradicts the gold explanation.	Good Match
The person is not necessarily training his horse	The person is preparing their horse for an event.	Not a Match
Just because they are smiling and waving at a camera doesn't mean they are tourists.	The children are smiling at their parents.	Not a Match

Notably, despite specific instructions to avoid using single-word responses such as 'Contradiction', 'Entailment', or 'Neutral' as explanations, ChatGPT occasionally defaulted to these terms. This deviation from the expected detailed sentence format highlights an aspect of unpredictability in AI response generation. In cases where ChatGPT adhered to the guidelines and provided more elaborate explanations, we observed a significant semantic alignment with the e-SNLI annotations, indicating a robust understanding of the contextual relationships between sentence pairs. However, the instances of deviation, where the model resorted to using just the labels as explanations, underscore the challenges in aligning AI models' outputs with specific instructional nuances, particularly in complex language interpretation tasks.

This analysis underscores the evolving capabilities and limitations of current NLP models in tasks requiring deep semantic comprehension. The variation in explanation alignment illustrates the ongoing challenge of ensuring AI models not only understand language at a superficial level but also grasp deeper contextual and inferential nuances. The limitation posed by the API's rate limit also highlights the practical challenges in conducting extensive NLP research using current AI technologies.

The full output of the Experiment can be found on GitHub. [3]

4.3 Experiment 3: **`PredictAndExplainGPT`**

In the third experiment, mirroring the PredictAndExplain approach, we will engage OpenAI's API in a two-step task. Initially, the API will be tasked with predicting the entailment label (mirroring Experiment 1 **`PredictLabelGPT`**), utilizing the provided premise and hypothesis. In this task, due to token limitations of OpenAI API, the prompting text will not include the highlighted words in sentences. The exact input text can be seen in code listing 3. Subsequently, it will attempt to generate an explanation for its predicted label. See how API was prompted in code listing 4. This experiment is designed to evaluate the integrated capabilities of the AI in both inference and explanation generation. It aims to assess not only the accuracy of the AI's predictions but also its proficiency in articulating these predictions in a clear and understandable manner. Due to the rate limits imposed by the OpenAI API, this experiment will be conducted on a limited set of only 20 examples.

```
label_prompt = (f"Predict the entailment label for the given premise and hypothesis"
      f" (entailment, contradiction or neutral)."
      f"\nPremise: {row['Sentence1']}\nHypothesis: {row['Sentence2']}")
```

Listing 3: Prompt for the first subtask of experiment "PredictAndExplainGPT"

```
explanation_prompt = (f"Explain why the relationship between the following premise and hypothesis is {predicted_label}.
Explanation in max. 1 sentence in natural language"
f"\nPremise: {row['Sentence1']}\nHypothesis: {row['Sentence2']}")
```

Listing 4: Prompt for the second subtask of experiment "PredictAndExplainGPT"

Results A significant finding of this experiment was that in 70% of the cases, the semantic meaning of ChatGPT's explanations aligned with those provided in the e-SNLI dataset. This result indicates a substantial degree of understanding by the AI model in interpreting and justifying the relationships between sentence pairs. This result aligns with the one from the first experiment (**`PredictLabelGPT`**) although the sample size was significantly smaller (only 20 sentence pairs were tested, where as in **`PredictLabelGPT`** 100 labels were predicted).

The model achieved a remarkable accuracy of 100% in predicting ***entailment*** labels, demonstrating its proficiency in identifying clear logical relationships where the premise directly leads to the hypothesis. In contradiction cases, the accuracy was approximately 57.14%, suggesting challenges in recognizing scenarios where the premise and hypothesis are mutually exclusive or logically inconsistent. In neutral relationships, the model also achieved an accuracy of about 57.14%. This indicates a level of difficulty in cases where the hypothesis neither directly follows from nor contradicts the premise, reflecting the complexity of interpreting subtleties in human language. These findings are illustrated in Table 3.

It also needs to be noted that there were some disagreements in the label prediction of **`PredictAndExplainGPT`** and **`PredictLabelGPT`** experiments. These variations can be attributed not only to differences in sample size, prompt structure, and the inherent challenges of model predictions across

[3] `https://github.com/mpetojevic/Scientific-Writing`

contexts but also to the operational constraints encountered during the `PredictAndExplainGPT` task. Specifically, the necessity to remain within the API's rate limits required a cautious approach to token usage across the two-step process, where a label prediction was followed by a request for an explanation. This constraint potentially influenced the depth and detail of the prompts used, thereby affecting the model's performance and the comparability of results between the two experiments. The first experiment, with a larger sample size, showed a high accuracy in contradiction predictions but struggled with neutral labels, whereas the second experiment, despite its smaller size, indicated improved performance for neutral predictions. The statistical variability inherent in smaller samples, the nuanced approach required for understanding neutral relations, and the specific structure and content of prompts could have significantly influenced the model's focus and interpretation, impacting its prediction accuracy. Additionally, the complexity of neutral relations and the potential biases in the model, stemming from its training data, further complicate accurate prediction across different contexts. These factors collectively highlight the variability and challenges in using NLP models for nuanced tasks like label prediction in natural language inference, especially when operational constraints such as token limits are in play.

Table 3: Results of Experiment 3: PredictAndExplainGPT

Metric	**Performance**
Overall Semantic Alignment	70.0%
Accuracy for Entailment	100.0%
Accuracy for Contradiction	57.14%
Accuracy for Neutral	57.14%

Through a manual review, it was observed that whenever ChatGPT accurately predicted the label, the corresponding explanations it generated were also semantically aligned with the human-annotated explanations in the e-SNLI dataset. This alignment is a testament to ChatGPT's ability not only to predict labels accurately but also to generate explanations that are contextually and semantically coherent.

It is important to note that ChatGPT was specifically instructed to generate explanations based on the labels it predicted. Consequently, in instances where the label prediction was incorrect, the subsequent explanations were inherently misaligned with the actual relationship between the premise and hypothesis. This outcome underlines a critical aspect of AI models in NLI tasks: the accuracy of explanations is intrinsically tied to the correctness of the initial label prediction.

To further illustrate the efficacy of ChatGPT in generating meaningful explanations, a selection of examples where there was a strong semantic match between ChatGPT's explanations and the e-SNLI dataset will be presented in an upcoming table. These examples will showcase instances of both the model's accuracy in label prediction and its proficiency in explanation generation. See Table 4.

To provide a balanced perspective, a selection of examples where ChatGPT failed to predict the correct label and thus provided incongruent explanations will be showcased in the Table 5. The analysis revealed that ChatGPT exhibited a particular challenge when tasked with identifying neutral relationships within the sentence pairs. This was evidenced by the model's tendency to confuse neutral labels more frequently than others, predominantly misclassifying them as entailment. The data indicates that in instances where the correct label should have been neutral, ChatGPT was more inclined to predict the label as entailment. This trend suggests a bias in the model's inference process towards finding a direct logical connection between the premise and the hypothesis. This tendency to misinterpret neutral scenarios as entailment points to an area of potential improvement in the model's understanding of subtleties in human language. Neutral statements, which neither directly follow from nor contradict the premise, require a nuanced recognition of implicit or unstated information, an area that appears to be challenging for the current iteration of ChatGPT.

Table 4: Comparison of Correct Labels and Explanations between e-SNLI and ChatGPT

e-SNLI label	**e-SNLI Explanation**	**ChatGPT Predicted Label**	**ChatGPT Explanation**
Entailment	A broken down airplane is outdoors.	Entailment	Because the hypothesis is a logical consequence of the premise.
Contradiction	One cannot be smiling and frowning at the same time.	Contradiction	The hypothesis contradicts the premise by stating that the children are not doing what is described in the premise, which is smiling and waving.
Neutral	Hugging does not imply being on vacation.	Neutral	The premise states that the women are hugging, while the hypothesis claims they are on vacation, which is not a direct inference.
Contradiction	People cannot be at a restaurant and in school at the same time.	Contradiction	The premise states that the people are in a restaurant, while the hypothesis states that they are in a school setting, which contradicts each other.

Table 5: Cases where ChatGPT failed to predict correct label and thus the appropriate explanation.

e-SNLI label	**e-SNLI Explanation**	**ChatGPT Predicted Label**	**ChatGPT Explanation**
Neutral	the person is not necessarily training his horse	Entailment	The hypothesis is entailed by the premise because the action of jumping over a broken down airplane suggests that the person is training their horse, which would be necessary for a competition.
Contradiction	One cannot be on a jumping horse cannot be a diner ordering food.	Neutral	The premise and hypothesis are neutral because they are describing two separate scenarios that are not related to each other.
Neutral	Just because they are smiling and waving at a camera does not imply their parents or anyone is anyone behind it.	Entailment	TThe hypothesis is entailed by the premise because children smiling and waving at the camera can be seen as evidence that they are smiling at their parents.
Neutral	Just because the boy is jumping on a skateboard does not imply he is wearing safety equipment	Entailment	The fact that the boy is jumping on a skateboard on a red bridge necessarily implies that he is wearing safety equipment.

5 Alternative Approach - KACE

Recent efforts in NLP have focused on explaining model behaviors using generative models. However, a common limitation is the emphasis on explaining "WHY A" predictions rather than the contrastive "WHY A NOT B" scenarios. Chen et al. (9) introduced a novel Knowledge-Aware Contrastive Explanations framework (**KACE**) for NLI. The framework identifies key phrases in input sentences, utilizes them as perturbations for generating counterfactual examples, and employs a knowledge-aware generative pre-trained language model to produce contrastive explanations. Experimental results highlight the benefits of contrastive explanations in clarifying differences between predicted answers and alternatives. The proposed approach, when applied to a BERT-large based NLI model, achieves an accuracy of 91.9% on SNLI, showcasing improvements over other methodologies. Workflow of conservative explanation is shown in Figure 5.

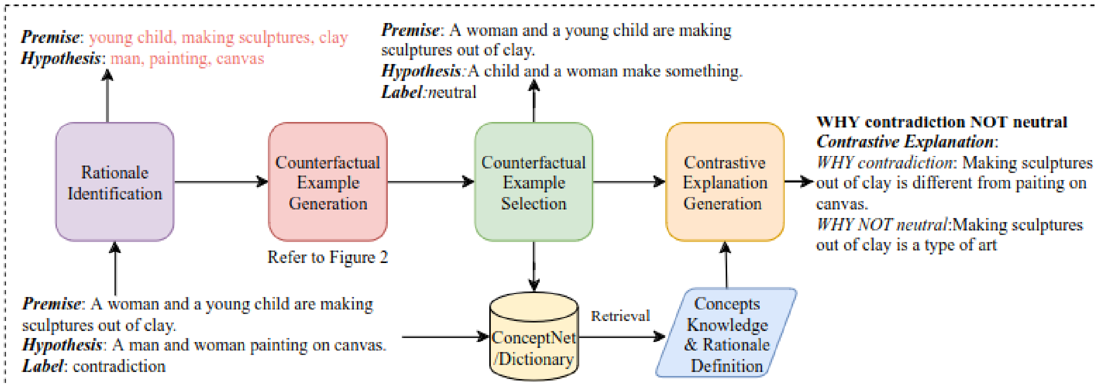

Figure 2: The overall workflow of contrastive explanation generation, which contains rationale identification, coun- terfactual example generation (as described in Figure 2) and selection, and knowledge-aware contrastive explana- tion generation. In our "WHY A NOT B" paradigm, we will generate explanations for A and each other class B (i.e., we will generate "WHY NOT neutral" and "WHY NOT entailment" in this example). The counterfactual example selection aims to select one most qualified for any other class B.

The methodology employed by KACE represents a departure from the approach taken by e-SNLI, particularly in its focus on contrastive explanations. Whereas e-SNLI enriches the Stanford Natural Language Inference (SNLI) dataset with natural language justifications for each label decision, KACE advances this concept by prioritizing explanations that delineate the reasoning behind selecting one label in preference to others. This contrastive explanation framework is designed to deepen the interpretability of model decisions. Nevertheless, it is important to note that the e-SNLI dataset plays a crucial role within the KACE framework. Specifically, KACE leverages the annotated highlights within both the hypothesis and premise of e-SNLI entries to identify rationales, and employs the dataset's human-annotated neutral language explanations to fine-tune its "WHY A" generator. This integration showcases how e-SNLI's resources contribute to enhancing the explanatory depth and accuracy of the KACE framework.

The process of generating conservative explanations involves serval steps:

1. **Rationale Identification**: Key phrases (rationales) are identified from the input sentences. These rationales are crucial for generating counterfactual examples.First input sequence for premise p and hypothesis h are constructed as $S^h = \langle s\rangle Label\langle s\rangle Premise\langle s\rangle$ and $S^h = \langle s\rangle Hypothesis\langle s\rangle$ where $\langle s\rangle$ is a special token that separates the components. We represent y as relation between S^p and S^h where $y \in enatailment, contradiction, neutral$.
2. **Counterfactual Example Generation**: The framework generates counterfactual examples based on the identified rationales. These examples are hypothetical variations of the original text, designed to lead to different inference results, thereby helping to understand the impact of specific phrases.
3. **Contrastive Explanation Generation**: The final step is generating explanations. The framework utilizes a knowledge-aware model, which not only provides reasons for the actual inference outcome but also contrasts it with what would have happened if the counterfactual examples were true.

KACE demonstrated the efficacy of integrating contrastive explanations with external knowledge in Natural Language Inference systems. It outperformed baseline models, particularly in generating richer and more contextually informed explanations. This approach led to advancements in the interpretability of NLI predictions, setting a new benchmark for the field. Performance of the KACE model can be found in Figure 5.

Figure 3: Human evaluation of contrastive and baseline explanations on 100 SNLI test samples. Average score of two annotators (%)

Model	Explanations Quality
NILE:post-hoc	81.5
LIREx-base	88.5
Contrastive Exp	90.5

6 Utilizing NLP and NLI Tasks in Everyday Use

NLP and NLI a are transformative technologies with far-reaching impacts across numerous sectors. Their greatness lies in their ability to bridge the gap between human language and computer understanding, enabling machines to interpret, analyze, and respond to human language in meaningful ways.

NLP and NLI find successful applications in a variety of domains, including:

Legal Systems: Processing legal documents, predicting legal outcomes, and providing legal assistance.

As per Zhong et al. (6) NLP serves as a pivotal tool, reshaping the legal landscape. NLP enhances the efficiency of legal document analysis, streamlines the parsing of complex texts, and aids in summarizing judgements. Its significant contribution lies in predicting legal outcomes through analysis of past cases, a function that revolutionizes legal decision-making. Furthermore, NLP's ability to extract specific entities enhances legal reasoning. The development of NLP-driven question-answering systems democratizes legal knowledge, broadening accessibility. This comprehensive exploration affirms NLP's potential to significantly advance the legal domain, driving it towards greater efficiency and accessibility.

Customer Service: Automating responses and understanding customer queries through chatbots.

In the context of customer service and chatbots, Upreti et al. (7) provides a framework for understanding how NLP enhances conversational AI. It emphasizes the roles of entity extraction and intent classification in understanding and responding to user queries. By analyzing various NLP algorithms, the thesis contributes to the development of AI assistants that are more efficient and user-friendly, capable of offering personalized and accurate responses in customer service scenarios. This research underscores the importance of advanced NLP techniques in creating effective digital customer service solutions.

Medical Domain and Translation: Interpreting clinical notes and patient interactions for better diagnostics and care.

Noll et al. (8) provides crucial insights into how NLP is utilized in the fields of translation and medicine. The paper discusses the challenges and advancements in translating medical terminologies, primarily using NLP and Machine Translation (MT) techniques. It emphasizes the need for accurate translation of medical terms to ensure consistent communication in healthcare across different languages. This highlights NLP's critical role in overcoming language barriers in medicine, enhancing global research collaboration, and improving patient care by facilitating clearer and more accurate medical communication.

7 Conclusion

In conclusion, the fields of natural language processing (NLP) and natural language inference (NLI) play a pivotal role in advancing our understanding of human-computer interactions. NLP empowers machines to comprehend and respond to human language, while NLI focuses on the intricate relationships between textual elements.

Our experiments employing OpenAI's ChatGPT API on the e-SNLI dataset have demonstrated the AI's capability to predict labels and generate explanations. However, it is imperative to acknowledge the variability in accuracy observed during these tasks, underscoring the ongoing challenges within the current landscape of NLP models. Our findings reveal that ChatGPT exhibits a high degree of accuracy in label prediction, especially in categories like entailment and contradiction, while facing more challenges in neutral instances and often confusing them with entailment. This underscores the complexities inherent in NLI tasks and highlights the nuanced understanding required for accurate inference. The experiment focusing on explanation generation shed light on the model's ability to articulate reasoning, with varying degrees of alignment with human-annotated explanation

In comparing our findings with the novel KACE approach, we observe the evolving landscape of NLI. The contrastive explanations provided by KACE, emphasizing the "Why A not B" paradigm, offer a different perspective on model interpretability, further enriching the field of NLP.

As the field of NLP continues to advance, the integration of models like ChatGPT and frameworks like KACE into practical applications will undoubtedly continue to grow. From customer service and legal systems to the medical field and beyond, the potential uses of these technologies are vast and impactful.

As society increasingly relies on advanced technologies, the establishment of trust and comprehensibility in NLP and NLI models remains of paramount importance. It is essential for these fields to confront challenges and continually strive for improvements in transparency, interpretability, and precision within NLP and NLI systems.

References

[1] *e-SNLI: Natural Language Inference with Natural Language Explanations, Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, Phil Blunsom* `https://proceedings.neurips.cc/paper_files/paper/2018/file/4c7a167bb329bd92580a99ce422d6fa6-Paper.pdf`

[2] *e-SNLI dataset available on Github:* `https://github.com/OanaMariaCamburu/e-SNLI`

[3] *Stanford Natural Language Inference SNLI* `https://nlp.stanford.edu/projects/snli/`

[4] *Annotation Artifacts in Natural Language Inference Data, Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, Noah A. Smith* `https://aclanthology.org/N18-2017.pdf`

[5] *Supervised learning of universal sentence representations from natural language inference data, onneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). CoRR, abs/1705.02364.* `https://arxiv.org/pdf/1705.02364.pdf`

[6] *How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence, Zhong, Xiao, Tu, Zhang, Liu, Sun, 2020* `https://arxiv.org/pdf/2004.12158.pdf`

[7] *A Comparative Analysis of NLP Algorithms for Implementing AI Conversational Assistants, Aanchal Upreti, October 2023* `https://mdh.diva-portal.org/smash/get/diva2:1803935/FULLTEXT01.pdf`

[8] *Machine translation of standardised medical terminology using natural language processing: A scoping review, Noll, Frischen, boeker, Storf, Schaaf*`https://www.sciencedirect.com/science/article/pii/S1871678423000432`

[9] *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 2516–2527 August 1–6, 2021. ©2021 Association for Computational Linguistics 2516KACE: Generating Knowledge-Aware Contrastive Explanations for Natural Language Inference, Chen, Ji, Zeng, Li, Zhang, Zhang* `https://aclanthology.org/2021.acl-long.196.pdf`

[10] *ImageNet: A large-scale hierarchical image database, J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li, Li Fei-Fei, 009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.* `https://image-net.org/static_files/papers/imagenet_cvpr09.pdf`