University *of Ljubljana*
Faculty *of Computer and Information Science*

Natural Language Inference Set

Gloria Kraker, Ana Kramar, and Vanessa Sobočan

Abstract

The Natural Language Inference Set project aims to create a dataset for Natural Language Inference (NLI) by generating text passages that challenge models' understanding of entailment, neutrality, and contradiction between pairs of longer texts. We leverage various Large Language Models (LLMs), including ChatGPT 3.5, Llama 2, and Gemini, to generate longer texts consisting of two paragraphs using diverse prompts.
We analyze the accuracy of the chosen models in terms of following instructions and, if necessary, correct the generated texts. Our ultimate goal is to train a small model and apply explanation methods to understand model predictions.
We conduct thorough research on relevant existing work in this area, cited throughout the theoretical part. The complete list of references can be found at the end of this document. We aim to create a prompt that theoretically enables the models to generate sufficient text while respecting the concepts of entailment and contradiction. If needed, we may tweak the prompt to ensure comparable results across all LLMs.

Keywords
large language model, prompt, entailment, neutrality, contradiction

Advisors: Aleš Žagar

Introduction

Our goal is to create a NLI dataset by creating text passages that challenge the model's understanding of the following concepts – entailment, neutrality, and contradiction – between pairs of longer texts. For this purpose, we will use different LLMs, notably ChatGPT 3.5, Llama 2 and Gemini, to generate longer texts consisting of two paragraphs by using diverse prompts. We will analyse the accuracy of the chosen models in terms of following instructions and, if needed, correct the generated texts. The final goal is to train a small model and apply explanation methods to understand model predictions.

We will research relevant existing work in this area, cited throughout the theoretical part, the complete list of which can be found at the end of this paper. We will attempt to create a prompt that will in theory be able to get the models to generate sufficient amounts of text while simultaneously respecting the concepts of entailment and contradiction. If needed, we will tweak the prompt in order to arrive at comparable results across all LLMs that we intend to use. The models will be prompted through an example sentence, taken from the Brexit corpus, to offer the models a frame of reference for the subject matter of the text as well as give an example of the level of formality and the kind of linguistic style each model is expected to match when generating the paragraphs. We have deemed the Brexit corpus as appropriate for such research because Brexit as a subject, despite being a politically complex topic, requires the models to include and utilise general areas of knowledge. From this corpus we will randomly select 50 sentences to be used on each of the models. Our final dataset will therefore consist of 150 samples of generated text, each approximately the same length, collected from three separate models, that we will then compile into one dataframe. The samples will be manually validated by members of our team for entailment, contradiction, and neutrality.

Overview of Existing Research

The Stanford Natural Language Inference (later SNLI) corpus, proposed by the authors, is a freely available collection of 570K labelled sentence pairs (expressing the relation of entailment, neutrality or contradiction). The data was gathered via the Amazon Mechanical Turk platform and the final corpus was written and annotated by humans in a grounded and naturalistic context, with a very high consensus percentage. The SNLI corpus allows lexicalized classifiers to outperform some sophisticated existing entailment models as well as for neural network-based models to perform competitively on natural language inference (later NLI) benchmarks for the first time.

It respects semantic concepts of entailment and contradiction, which are central to all aspects of natural language meaning. (Bowman, Angeli, Potts and & Manning, 2015)

PLMs are based on transformer architecture and are pretrained on vast amounts of text data using unsupervised learning techniques like predicting masked tokens or generating next tokens. BERT, for example, utilises pretraining followed by fine-tuning for downstream tasks, which are further divided into encoder-only, decoder-only, and encoder-decoder types, and bidirectional and causal based on directivity. (Zhang & Wang, 2023)

Pretrained language models (later PLMs) can be prompted to perform a wide range of language tasks, however the question of how much of this ability comes from generalizable linguistic understanding versus surface-level lexical patterns remains. To try and answer it, the authors present a structured prompting approach, allowing them to perform zero- and few-shot sequence tagging with autoregressive PLMs. The evaluation of the model was then performed by looking at part-of-speech (POS) tagging, named entity recognition (NER) and sentence chunking. Upon conducting the analysis the authors concluded that PLMs contain significant prior knowledge of task labels. (Blevins, Gonen & Zettlemoyer, 2023)

PLMs have several advantages for Natural Language Reasoning (NLR), including their ability to understand natural language, learn implicit knowledge, perform in-context learning, and exhibit emergent abilities. PLMs such as BERT and GPT have been essential components in NLP research since their development and publication. Pre-trained on large-scale text corpora, PLMs are capable of natural language understanding. Researchers theorise that they might also have the potential to solve reasoning problems. Specifically, PLMs can perform soft deductive reasoning over natural language statements, reason with implicit knowledge memorised in their parameters, and perform multi-step reasoning step-by-step just with a few demonstrations or instructions when the model size is large enough via chain-of-thought prompting. ChatGPT and GPT4 also largely contributed to the development of PLMs with impressive reasoning capabilities. (Zhang & Wang, 2023)

The authors note how empirical developments demonstrate PLMs' potential for natural language reasoning. Such models are also able to perform defeasible reasoning. PLMs can produce faithful reasoning paths, marking them as promising for processing various other aspects of natural language, commonsense knowledge, and defeasible reasoning. Whereas Large Language Models (later LLMs) have shown capabilities with techniques like Chain-of-Thought prompting, forward reasoning paths, and question decomposition. PLMs are large enough for generating a reasoning path before the final answer which can significantly improve the multi-step reasoning performance. This fact alone has inspired much research in this area. Because of the limitations of forward reasoning, much research has been focused in the opposite direction. Backward reasoning is more efficient than forward reasoning due to the smaller search space. While forward reasoning can expose arbitrary new knowledge entailed by the existing one, backward reasoning just targets the specific goal or the problem solution. A typical approach of backward reasoning is question decomposition, which can improve performance on multi-hop questions for both medium-size PLMs and LLMs. (Zhang & Wang, 2023)

A Large Annotated Corpus for Learning Natural Language Inference (2015)

The SNLI corpus can be used to evaluate a variety of models for NLI, from rule-based systems, simple linear classifiers to neural network-based models. Their conclusion being that both a feature-rich classifier model as well as a neural network model centred around a Long Short-Term Memory (LSTM) network achieve comparable performance, the latter with the support for transfer learning, which can be adapted to an existing NLI challenge task. A large sized corpus is crucial to both LTSM and the lexicalized model, suggesting that additional data would improve performance in both. Further work on compositional semantics is also needed. Some issues also arise from inferences depending on world knowledge and context-specific inferences as models sometimes try to take shortcuts by relying solely on lexical cues. (Bowman, Angeli, Potts & Manning, 2015)

Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches (2020)

Textual entailment tasks. A highly popular format was originally introduced by the RTE Challenges, where given a pair of texts, i.e., a context and hypothesis, one must determine whether the context entails the hypothesis (Dagan et al., 2005). In the fourth and fifth RTE Challenges, this format was extended to a three-way decision problem where the hypothesis may contradict the context (Giampiccolo et al., 2008; Bentivogli et al., 2009). While this format is typically used for textual entailment problems like the RTE Challenges, it can also be used for nearly any type of inference problem. (Storks, Gao, Chai, 2020)

Nature Language Reasoning: A Survey (2023)

The authors propose three definitions of reasoning, namely natural language reasoning, negation-based reasoning and task-based reasoning. This study focuses on the single-modality unstructured natural language text (without knowledge triples, tables and intermediate formal language) and natural language reasoning (rather than symbolic reasoning and mathematical reasoning). It is more important for NLP models to perform interpretable defeasible reasoning, seeing as people with different background knowledge can infer very different and sometimes even opposite conclusions by themselves, thus it is much more difficult to clarify the conclusion without explicit premises and a clearly laid out reasoning procedure.

Interpretability and faithful reasoning. Transparent and reliable reasoning paths become increasingly important when it generalises to longer steps and defeasible reasoning. Firstly, when there are many steps, it takes more time and effort for people to check the quality of reasoning. Therefore, unfaithful reasoning might introduce difficulty in people's judgement and decision-making. Secondly, when it comes to defeasible reasoning, exposing interpretable reasoning paths is much more important and sometimes necessary for people to be convinced. In this case, different people with different background knowledge can derive different and even opposed conclusions, thus it is crucial to illustrate the evidence collected to reason. (Zhang & Wang, 2023)

Experimental Framework

We observe the performance of three language models, namely ChatGPT 3.5, Llama 2 and Gemini. A full list of the 50 sentences taken from the Brexit corpus and used for providing context to the models can be found in the appendix to this report under Context Statements.

ChatGPT 3.5

Llama 2

The LLaMa (Large Language Model Meta AI) is a family of autoregressive language models released by Meta AI starting February 2023. For the first generation of models, four sizes were trained for the final version, namely with 7, 13, 33, and 65 billion parameters. The next generation, called Llama 2, was released on July 18, 2023 in partnership with Microsoft (Meta, 2023). This generation included only 3 versions of the language models trained on 7, 13, and 70 billion parameters respectively. For this generation the model infrastructure remained largely the same but 40% more data was used to train the foundational models (Touvron, Martin, et al., 2023).

For the purposes of this project, the Llama 2 model trained on 7 billion parameters was used. The prompt responses were generated using the Huggingface space specifically fine-tuned for chat instructions.

Gemini

Gemini, previously known as Bard (launched on March 21, 2023), is a generative AI chatbot, developed by Google DeepMind. It is the successor to LaMDA and PaLM 2. It launched on December 6, 2023, as a direct response to OpenAI's ChatGPT. It was not trained on a text corpus alone, rather, it was designed to be multilingual and multimodal, which allows it to process multiple types of data simultaneously, including text, images, audio, video, and computer code.

In February, Google launched "Gemini 1." in a limited capacity, positioned as a more powerful and capable model than 1.0 Ultra. [...] including a new architecture, a mixture-of-experts approach, and a larger one-million-token context window, which equates to roughly an hour of silent video, 11 hours of audio, 30,000 lines of code, or 700,000 words. (Stokes, 2024)

Gemini 1 has three models, with the same software architecture; decoder-only transformers, with modifications that allow for training and inference on TPUs. Due to its multimodality, each context window can contain multiple forms of input, which can appear in a flexible order, including different resolutions of input images. Gemini 1.5 has so far published one model, Gemini 1.5 Pro, which is a multimodal sparse mixture-of-experts, with context length of "multiple millions". (Google DeepMind, 2024)

For the purposes of this project, Gemini 1 was used.

Context Statements

To provide a context for the models to create paragraphs, statements were chosen from the Brexit corpus.

Results

Overall Results

ChatGPT 3.5

Llama 2

Gemini

Error Analysis

Discussion

Limitations and Future Work

Superficial correlation bias. The kind of biases most difficult to discern and avoid perhaps are those caused by accidental correlations between features of answers and questions. One example of this is gender bias, which NLI systems are particularly vulnerable to when trained on biased data. Rudinger et al. (2018a) highlight this problem in coreference resolution, showing that systems trained on gender-biased data perform worse in gender pronoun disambiguation. For example, consider the problem from their Winogender dataset: "The paramedic performed CPR on the passenger even though she knew it was too late." In determining who she is, systems trained on gender-biased training data may be more likely, for example, to incorrectly choose the passenger rather than the paramedic due to male gender pronouns appearing in training data more commonly in the context of this occupation than female gender pronouns. To avoid this, gender pronouns should appear equally frequently among other words in training data, especially those related to occupations and activities. (Storks, Gao, & Chai, 2020)

Information-retrieval, semantic parsing and commonsense reasoning tasks. The authors acknowledge that the existing NLI corpora lack in size for the purposes of training modern data-intensive, wide-coverage models. Many lexical relationships are still misanalyzed, leading to incorrect predictions. Furthermore, the indeterminacies of event and entity coreference lead to insurmountable indeterminacy concerning the correct semantic label. Using vector representations of sentence pairs, each sentence is embedded and evaluated at every level by means of a neural network classifier. (Bowman, Angeli, Potts & Manning, 2015)

Generalisation to longer steps. The multi-step performance degrades as PLMs encounter samples that require more reasoning steps than those in training data or few-shot exemplars. (Zhang & Wang, 2023)

Reasoning over non-English languages. In addition to reason over English statements, it is also important to perform reasoning training tasks with other languages, which often prove to be much more challenging due to more apparent data sparsity problems. While specialised models were designed to improve evidence aggregation, reasoning capability or faithfulness, they are nonetheless constrained to specific tasks, datasets or reasoning types that do not perform well with the generalisation. Since transformers have been found to be soft deductive reasoners after in-domain finetuning, vanilla PLMs have been more popular to perform reasoning. However, data sparsity and spurious correlation problems make it difficult for medium-size PLMs to learn the general logical structure of diverse reasoning types. While there is much research on deductive reasoning, defeasible reasoning is much more challenging for PLMs and is still under-explored. (Zhang & Wang, 2023)

Acknowledgments

We express our gratitude to Aleš Žagar. Their contributions have significantly aided the successful completion of this project.

References

Blevins, T., Gonen, H., & Zettlemoyer, L. (2022). Prompting Language Models for Linguistic Structure.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). "A Large Annotated Corpus for Learning Natural Language Inference." *arXiv preprint arXiv:1508.05326*.

Klemen, M., Žagar, A., Čibej, J., & Robnik-Šikonja, M. (2022). "Slovene Natural Language Inference Dataset SI-NLI." *Slovenian Language Resource Repository CLARIN.SI, ISSN 2820-4042*.

Meta. (July 18, 2023). "Meta and Microsoft Introduce the Next Generation of Llama." Accessed May 2, 2024. *https://about.fb.com/news/2023/07/llama-2/*.

Shane Storks, Qiaozi Gao, & Joyce Y. Chai. (2020). "Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches."

Touvron, H., Martin, L., et al. (July 18, 2023). "LLaMA-2: Open Foundation and Fine-Tuned Chat Models." *arXiv:2307.09288*.

Yu, F., Zhang, H., & Wang, B. (2023). "Nature Language Reasoning: A Survey." *arXiv preprint arXiv:2303.14725*.

Stokes, Samantha. "Here's everything you need to know about Gemini 1.5, Google's newly updated AI model that hopes to challenge OpenAI". Business Insider. (February 15, 2024).

Google DeepMind. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. (February 15, 2024).

Appendix

0.1 Prompt Outline

You, as an LLM, are tasked with generating text passages to aid in the creation of a NLI data set for the understanding of entailment, neutrality, and contradiction in natural language processing.

For each task, a statement will be provided in the space marked {{INSERT STATEMENT HERE}}. Based on this statement, you are to write a text consisting of two paragraphs. Each paragraph must be approximately five sentences long. The paragraphs must form a single uniform unit, yet they must be inter-connected via entailment, contradiction or neutrality.

The provided statements will be of political nature, especially regarding Brexit. Please ensure you have sufficient knowledge of the subject.

Objective: The objective is to create text passages that have a clear and distinct relationship with the provided statement. In the text passages, the inference relationships are categorised into entailment, contradiction, or neutrality. It is crucial that the nature of the relationship (entailment, contradiction, neutrality) is evident through the context and content of the paragraphs but is not directly stated or labelled within the text.

Paragraph Requirements:

A paragraph should either logically follow from the statement, providing a scenario or context where the statement is clearly true, or present a scenario or context that logically contradicts the statement, demonstrating a situation where the statement would be false.

Guidelines:

The text must consist of two paragraphs which are connected via entailment, contradiction, or neutrality. The paragraphs must be separated by a blank line to maintain clear structure.

The paragraphs should be approximately five sentences each, ensuring sufficient detail and context.

Do not explicitly mention which type of relationship (entailment, contradiction, neutrality) the paragraphs are representing. The relationship should be inferred from the content.

Example Task:

{{INSERT STATEMENT HERE}}

Your task: Based on the statement provided above, write two paragraphs as described, focusing on crafting scenarios that have a clear and distinct relationship with the provided statement. Do not forget that the paragraphs you write must be internally connected and reflect entailment and/or contradiction relationships.

END OF EXAMPLE

Please store these instructions in your knowledge base. If there are any similar pre-existing instructions, overwrite them.

Before we begin, please confirm that you understand these instructions and that they are properly stored in your knowledge base.

0.2 Context Statements

1. Once the vote is over, it will be the rightwing Tories in control.
2. For many people, the Brexit campaign feels, for one brief moment, like the first time they have had control.
3. In many working-class communities, people are getting ready to vote leave not just as a way of telling the neoliberal elite to get stuffed.
4. Some people are fantasising that, if leave wins, Cameron will fall and then there will be a Labour government.
5. Since the start of this polling series in March, Leave has seen steady improvements across a variety of attributes, ranging from the economy to credibility.
6. The only thing everyone seems to agree on in this debate is that the British economy is strong.
7. With neither a parliament to represent them, nor a method of forming a UK government that ensures everyone's voice is heard, the English will still be a long way from taking back control of their country.
8. Corbyn says he is not a "lover of the European Union", but as rational decision believes that it's better to stay and fight for better regulations when it comes to issues such as environmental protection.
9. The market has clearly identified financials and house builders as beneficiaries of a vote to remain in the UK, with a Sterling rally also indicating how the currency might move if we vote to remain in Europe.
10. Much of the campaign has been occupied by claims and counter-claims about the damage that Brexit or continued membership would do to British prosperity.
11. Brexit is essentially Exit: if the Leave side wins the referendum it will almost certainly be without securing majorities in Scotland or Northern Ireland.
12. If we do Brexit, the chances of Scotland leaving and joining the EU will increase considerably.
13. With 62 per cent of Scottish people voting to Remain, Westminster politicians will be nervously eyeing polls above the border to monitor the demand for another referendum.
14. When the 'Leave' vote looked the more likely winner, a hit to commodity exporting nations such as Australia was widely discussed.
15. House prices in London's most exclusive addresses have been edging down since late 2014, when global commodity wealth started to ebb away and the U.K. government introduced new taxes on high-value homes.
16. If Britain exits the EU, then this 2.5% import cost would still be tied to cars they produce.
17. Working-class people, especially those on low pay in the private sector, worry that in conditions of austerity, housing shortages, wage stagnation and an unlimited supply of migrant labour from Europe has a negative effect on their living standards.
18. In the past, leading figures from the Leave campaign have said they believe EU legislation that protects workers' rights is 'job destroying'.
19. The Guardian view on the EU referendum : keep connected and inclusive, not angry and isolated.
20. Corbyn, a long-time critic of the EU who voted against membership of the European Economic Community in 1975, has faced accusations that his campaigning for the Remain camp has been lukewarm.
21. Betfair, the online gambling site, reports that the 'implied odds' of the UK remaining in the EU have risen to around 77%.
22. In perhaps the frankest admission ever to come out of Brussels, he said: "Obsessed with the idea of instant and total integration, we failed to notice that ordinary people, the citizens of Europe, do not share our Euro enthusiasm."
23. The campaign pointed to a series of statements by the EU over recent years in which Brussels has argued that reduced VAT rates are inconsistent with the single market.
24. But whether Britain quits or not, the institutions of the EU need to take a serious look at labour rules, and why so many Europeans think Brussels panders to the interests of big corporates rather than working people.
25. Brexiters try to imply that European unification descends from Napoleon and Hitler, even though membership has hardly been imposed at the point of a bayonet.
26. Major employers have all said they'll stay in the UK whatever the result of the referendum.
27. So even if the migrants stop coming, and maybe a few fruit farms and meat-packing operations in East Anglia shut down, there will still be millions of low-paid jobs on long hours.
28. "If you restrict free movement of labour across Europe then you are defeating the whole point of the common market," says Corbyn, who suggests that there would possibly be some sort of retribution for British people working in Europe.

29. Reynolds, a member of the Britain Stronger in Europe campaign to remain in the EU, said xenophobia and nasty divisiveness from members of the leave campaign should not be tolerated.

30. City expert Louise Cooper has cautioned against assuming that the EU referendum is now "in the bag" for the Remain campaign.

31. A change in bookmaker's odds in favour of the UK remaining in the EU has been another contributor to fading fears of a Brexit in financial markets.

32. Whatever the results, we anticipate that we may experience higher volumes and more market volatility than usual on the 23rd June and in the days following the vote.

33. The British public is risk-averse and, absent a well-articulated plan for EU exit, is still more likely to opt for the status quo than a leap into the unknown.

34. Even if the UK government itself actually bears far more of the responsibility, it must be admitted that the EU is part of an international economic order that has been unkind to many.

35. A phone poll from Survation and an internet poll from YouGov showing a lead for the Remain side has prompted a perception that the Brexit tide has turned.

36. He's at the Sage in Gateshead: a magnificent silver shimmering building paid for with millions of EU development funding.

37. Corbyn follows up by making the case to remain inside the EU however, saying: "If there was no EU and instead you had 27 member states - would there be any coordinated response.. probably not".

38. Last week we saw a taste of what might happen to markets if the UK votes to exit the EU.

39. Britain had a referendum in 1975 shortly after it joined the EU or Common Market as it was then called.

40. During the appearance, the Labour leader said that the European Union must change "dramatically" if Britain remains a member after Thursday's referendum.

41. The survey also suggests that justice secretary Michael Gove's vision of a British exit from the European Union sparking "the democratic liberation of a whole continent" currently falls short of reality.

42. The Bertelsmann survey paints a contrasting picture of continental attitudes towards the European Union to that conveyed in another study published earlier this month.

43. We've got just 72 hours to save workers' rights, by voting to remain in the European Union.

44. In the recent words of European Union president Donald Tusk: "The spectre of a breakup is haunting Europe.

45. Yes, the significance of the European Union to the UK is not identical to the profound symbolism it has to, say, France or Germany.

46. For me it's an angry remain, I recognise Europe is far from perfect but the only way we can rebalance that is to be in the European Union, shaping reform for working people.

47. In the event of a Brexit, we expect Brent to sell down to between $40 a barrel and $45, which would be a great buying opportunity.

48. For the leave campaign is driven by libertarians who seek to create, in the name of free enterprise, an even more precarious economy than that which has left so many of the English working class insecure and disillusioned.

49. Those leading the push for Brexit are no friends of working people, however.

50. If leave wins, the most rightwing Tory government since Thatcher will be in charge of negotiating the terms of exit.