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

Towards Generalizable NLP Models: A Prompt-Guided Inference Dataset with Integrated Reasoning Rules

Blaž Lipar and Meta Kokalj

Abstract

This paper proposes a novel approach to training Natural Language Inference models capable of handling diverse reasoning patterns. The core of our method lies in strategically chosen scenario categories for training data. These categories - historical events, scientific explanations, everyday situations, news reports, and fictional stories - expose the model to a comprehensive range of reasoning styles. Further, we employ a combination of iterative prompting and curated Few-Shot Learning with Chain of Thought Prompting to enhance the model's reasoning abilities. Our analysis across these categories demonstrates the effectiveness of transitive reasoning within prompting techniques. This method empowers language models to generate informative and nuanced responses.

Keywords

natural language inference (NLI), reasoning, inference scenario creation, paragraph-level inference, prompt engineering, chain-of-thought prompting

Advisors: Aleš Žagar

Introduction

Natural Language Inference (NLI) tasks serve as a valuable benchmark for assessing the capabilities of new models within Natural Language Understanding (NLU) research. This is particularly significant as NLU plays a crucial role in the success of many actively studied NLP problems, including question answering, translation, and dialogue [1].

Interestingly, the field of Natural Language Processing (NLP) often overlooks the role of paragraphs despite their importance in discourse and text generation. The majority of NLI datasets and research focused only on sentence-level inference, where the premises and hypotheses are expressed in single, typically brief sentences. They are therefore inappropriate for use in other open-ended NLP situations. This is particularly crucial in tasks aimed at verifying the factual accuracy of summaries of documents [2]. Paragraphs, often described as "units of thought," go beyond the grammatical structure of individual sentences (syntax) [3]. Analyzing how sentences within a paragraph connect through cohesion (referential links) and coherence (logical flow of ideas) unlocks a deeper understanding of meaning [4, 5, 6]. This analysis of relationships between sentences is crucial for tasks like discourse generation, where sentences need to be arranged in a way that creates a clear and cohesive flow of information. Incorporating paragraph-level semantics into NLP models could significantly improve their ability to handle real-world language tasks.

In addition, many crowdsourced NLI datasets, while convenient, may not be well-suited for tackling real-world NLP problems due to their creation in isolation from specific downstream tasks and due to inherent annotator biases [7, 8]. Specifically, as shown in the case of Multi-Genre Natural Language Inference (MultiNLI), crowd workers sometimes relied on certain strategies to form hypotheses of a specific label like adding negators for contradiction labels, and introducing bias into the dataset. This approach creates an unrealistically easy task, failing to capture the true complexity of real-world human reasoning [8]. Many analyses of the SNLI dataset conducted in recent years have revealed that models trained on it often lack significant compositionality in their representations. Instead, these models primarily rely on straightforward heuristics, such as word overlap, when making inferences (this phenomenon is commonly recognized as the "lexical overlap heuristic," as articulated by McCoy et al.(2019) [9]. For instance, a high overlap in words between premises and hypotheses typically predicts entailment, while contradictory sentence pairs usually exhibit minimal or no word overlap

and/or the presence of negation words. This underscores the importance of developing models that can effectively incorporate external knowledge for more nuanced natural language inference tasks [10, 11].

What is emerging NLI's most pressing problem is their propensity for overfitting. While these models demonstrate impressive performance on the datasets they are trained upon, potentially exceeding human benchmarks, their ability to generalize to unseen data remains severely limited. This phenomenon, well-established in machine learning, is referred to as the generalization problem [12]. An underlying cause might be the focus on deductive reasoning in training data creation. Careful examination of crowd worker instructions for popular NLI datasets reveals an emphasis on logically guaranteed inferences, neglecting inductive reasoning, which involves concluding unseen data.

To address these challenges and generate more realistic, nuanced data, we propose a novel paragraph scope scenario-inference pair elicitation method where each scenario is accompanied by multiple potential inferences. By incorporating various reasoning modes within the prompt creation stage, we aim to train an NLI model that can handle a broader spectrum of reasoning patterns, ultimately leading to a more robust and generalizable understanding of natural language inference.

Recent work

Pioneering datasets like The Stanford Natural Language Inference (SNLI) [13] and the Multi-Genre Natural Language Inference (MNLI) [1] established a benchmark with their large scale, balanced data, and clear labels. However, as the field matures, researchers are constantly innovating, recognizing limitations in existing datasets such as reliance on human-generated sentences and potential biases.

Cornerstone Datasets

The Stanford Natural Language Inference (SNLI) dataset, a cornerstone for NLI research, offers a large-scale, balanced dataset with clear labels, facilitating robust models for diverse logical relationships [13]. However, limitations exist. The simplistic sentence generation process from simple, static scenes limits reasoning style diversity, and the reliance on affirmative image captions restricts the ability to assess negation, a crucial aspect of natural language [14]. Building upon SNLI, the Multi-Genre Natural Language Inference (MNLI) dataset extends the scope by incorporating sentence pairs from various genres and domains (e.g., fiction, telephone conversations) [1]. This inclusion of diverse genres promotes model robustness by exposing them to a wider range of language styles and reasoning patterns. Additionally, MNLI facilitates transfer learning tasks, where models trained on this dataset can be adapted to perform well on different NLP applications. While MNLI offers significant advancements, it inherits limitations from SNLI, such as reliance on human-generated sentences, and suffers from genre imbalance within the dataset. Gururangan et al. [15] questioned whether SNLI-trained models are achieving genuine language comprehension or simply capitalizing on coincidental patterns (artifacts) within the dataset.

Domain-Specific and Challenging Datasets

There are many existing NLI training datasets, each with its strengths and weaknesses. SciTail [7] focuses on scientific language, training models for reasoning specific to scientific contexts, and encouraging understanding of contextual relevance but is limited in size. FEVER (Fact Extraction and VERification) tackles real-world fact-checking challenges with claims and evidence from Wikipedia. While valuable for these tasks, FEVER can be susceptible to biases and noise from imperfect evidence retrieval [16]. BioNLI is a dataset in biomedical natural language inference [17]. Interestingly, this research addresses the automatic creation of meaningful negative examples by using rule-based and neural-based counterfactuals, potentially removing the need for expert involvement.

The development of NLI datasets is a continuous process. The paper of Liu et al. [18] introduced a new dataset called ConTRoL for contextual reasoning in natural language inference. It is a passage-level NLI dataset that focuses on complex reasoning types and is more challenging than previous benchmarks. Rather than evaluating a hypothesis versus a short passage, Koreeda and Manning [19] introduced an evaluation of the hypothesis against a whole document. The system is provided with a contract and a set of hypotheses, which could be statements about the contract's obligations, rights, or other key points. This approach applied to contracts, requires models to understand the broader context of the document. Worth noting is the study by Nie et al. (2020) which presented a large-scale NLI benchmark dataset collected through an iterative human-in-the-loop approach, resulting in a dataset that surpasses existing benchmarks in difficulty. Beyond the dataset itself, the authors proposed a novel data collection method that functions as a "moving target" for NLU. Unlike static benchmarks that models can quickly overfit on, this approach acts as a never-ending learning scenario that allows for continuous improvement and adaptation [20].

HANS (Heuristic Analysis for NLI Systems) [21] and ANLI (Adversarial NLI) [20] both aim to improve the robustness of NLI models. HANS achieves this by exposing model biases through pre-defined templates, while ANLI uses specially crafted adversarial examples.

Parlay creation

This project seeks to construct a comprehensive inference dataset for training a NLP model. The proposed dataset (Parlay) will be designed to promote model generalizability by encompassing five distinct categories, ranging from historical events to fictional narratives as inference pairs. To ensure the dataset's further diversity, each category will include inference pairs drawn from a wide range of contexts within its topic. Clear and concise prompts will be employed to introduce scenarios within each category. These scenarios will provide a brief description of a situation or event, offering

essential context for the model to perform various inference tasks effectively.

The selection of scenario categories – historical events, scientific explanations, everyday life situations, news reports, and fictional stories – is strategically chosen to provide a comprehensive training ground for various reasoning modes. Each category offers unique advantages. Examining historical events allows the model to analyze complex cause-and-effect relationships, temporal reasoning (e.g., order of events), reasoning about counterfactuals ("what if" scenarios), and reasoning by analogy (using the law of recurring patterns and motivations that influence similar situations) [22, 23, 24, 25, 26, 27, 28]. Scientific explanations necessitate the understanding of deductive logical reasoning rules like modus ponens and modus tollens [29]. Everyday scenarios provide a fertile environment for practical reasoning. The model encounters incomplete information, necessitating inductive reasoning based on patterns, experiences, and common sense [30]. Additionally, everyday scenarios require analogical reasoning, where the model identifies similarities to use knowledge from one situation to understand another [31]. This approach equips the model with transferable reasoning skills, allowing the model to adapt to novel situations. News reports naturally require the model to exercise critical reasoning skills [32]. This involves evaluating source credibility, identifying potential biases, and distinguishing factual information from opinion. Fictional stories will provide a platform for exploring narrative reasoning, including understanding characters' motivations, and emotions (theory of mind) [33, 34]. Furthermore, because stories often involve social interactions and dynamics the model can learn to reason about how characters behave in social contexts and predict their actions based on social norms (social reasoning) [35, 36].

We opted for an alternative approach utilizing prompt-driven scenario generation. While inspired by natural texts and ideas as prompt inputs, the scenarios themselves will be newly created based on those prompts. This approach addresses key challenges inherent in natural text-based datasets: citation management and authorship attribution difficulties. This can ultimately impact the credibility of datasets constructed from natural texts. Our prompts allow for greater control over the information presented, mitigating bias.

The dataset will be built in Slovenian, contributing to the overall linguistic diversity of NLI resources. Each data point within the dataset will consist of two key components: a scenario paragraph to establish the context for a specific situation and the corresponding paragraph generated based on entailment, contradiction, or neutral labels. Crucially, each data point will be further augmented with explanations detailing the specific reasoning that informs the categorization of each pair into entailment, neutral, or contradiction. This approach goes beyond simply labeling the data and provides valuable insights into the LLM's underlying logic for each categorization.

These detailed explanations play a critical role in enabling efficient learning with few examples, also known as "few-shot learning"[37]. Analyzing the explanations allows us to gain a deeper understanding of the underlying logic the LLM used to establish the relationship between the paragraphs. This knowledge allows us to effectively guide the LLM when creating new scenario-inference pairs, even with a limited number of starting examples. The explanations essentially act as a training aid, helping the LLM apply similar reasoning patterns to novel scenarios. Additionally, by analyzing the explanations, we can identify broader reasoning principles that can be applied beyond specific scenarios. This empowers the LLM to generalize its reasoning skills and handle a wider range of inference tasks.

By utilizing prompt-driven scenarios, we will ensure the creation of a high-quality and well-controlled foundation for generating diverse inference pairs. We will achieve this through the following methods:

Reasoning Rule-Guided Prompt Engineering. We will develop prompts that incorporate established reasoning rules, such as modus ponens, modus tollens, causal reasoning, textual entailment, and predictive inference (Liu, Teng, Ning, Liu, Zhou, Zhang, 2023). Our focus will be on crafting effective prompts that integrate factual hints and guide the model towards specific inferences (entailment, contradiction, neutral). Factual hints refer to pieces of information embedded within the prompts that provide the model with essential knowledge about the scenario. These hints serve as anchors for the model's reasoning process, allowing it to make inferences based on established facts. While factual hints provide a strong foundation, real-world inferences often rely on broader reasoning patterns beyond established rules. To address this, we will incorporate heuristics within our prompts. Heuristics can capture these general reasoning patterns, allowing our prompts to be more adaptable to diverse scenarios and inferences encountered in real-world situations. We will achieve this by exploring templates or conditional statements within the prompts based on the reasoning rule we want to target.

Chain of Thought Reasoning. We will investigate the application of chain-of-thought reasoning, where the model explicitly reveals its reasoning steps as it arrives at an inference. This will allow us to gain deeper insights into the model's decision-making process, especially in relation to the applied reasoning rules, and further refine our prompt design for optimal performance [38].

We will detail the specifics of this approach and its advantages in the following section.

Preliminary Results

Iterative Prompting and Human Expertise

This section of the report details the process undertaken by team member 1 to create a Natural Language Inference (NLI) dataset utilizing the large language model GPT-3.5.

The prompting process began by familiarizing the LLM with the concept of an NLI dataset. The team member inquired

about the concept and requested an example. Following GPT-3.5's initial attempt (which exhibited issues regarding the "neutral" label), the team member clarified the concept of neutrality within the examples. To establish a consistent structure for the dataset, the team member presented an example from the Stanford NLI (SI-NLI) dataset. This served as a foundation for prompting GPT-3.5 to generate its examples. The structure of the generated examples was refined to achieve a consistent format encompassing premise, hypothesis, particular and generalized explanation, outlining the reasoning process. The primary focus was on prompting GPT-3.5 to create diverse NLI examples across various categories. On a few occasions, the team member provided prompts based on specific quotes or ideas but primarily relied on GPT-3.5's generation capabilities.

The most problems regarding the content arose when creating the 'neutral' part of the example, for it often understood neutral as "the subject of the premise not being in favour of one or the other option". That is why often the team member had to manually correct the examples to make an example functional or clearer. There were, however, examples, in which the link that was created by the LLM at first seemed a bit unclear, e.g. the link between the premise and the hypothesis seemed somewhat hard to detect, but upon further investigation the underlying reasoning provided by GPT-3.5 was recognized as sound and logical in a more direct sense, taking into consideration the meaning of each word and the connection between them. That is why these examples were retained as well because it is an addition to the diversity of the dataset.

Curated Few-Shot Learning with Chain of Thought Prompting

The second approach followed a two-stage methodology: first, compiling high-fidelity data points using the large language model Gemini, and second, training and inferencing the LLM using chain-of-thought prompting (again using LLM Gemini).

The project's foundation was built on crafting high-quality data points. Team member 2 carefully manually chose well-known historical occurrences, everyday life situations, and natural phenomena or processes as the core themes for these data points. To make sure they effectively prompted the LLM, each data point went through a rigorous process of refinement using a trial-and-error approach. This involved creating, testing, and then revising the data points until they consistently produced good results. Data points were crafted to incorporate rich contextual information surrounding the concept. This might have involved narrative framing and the inclusion of multiple perspectives within the scenario. This selection encouraged diverse reasoning applications by encompassing concepts with varying domains, complexities, and reasoning technique requirements. The second stage explored the application of one-shot or few-shot learning techniques for LLM training. This approach involved training the LLM on a limited set of thoroughly prepared data points. The goal of this technique was to leverage the rich information content embedded within each data point to enable the LLM to effectively generalize its learnings. Following training, a prompting technique termed chain of thought prompting was employed. This approach involved crafting prompts that explicitly guide the LLM through a series of logical steps. The prompts referenced the underlying factors and relationships that define the concepts presented in the data points. Conditioned by these prompts, the LLM was then tasked with inferring related concepts or situations. Crucially, the LLM was also expected to clarify the reasoning behind its inferences, thereby revealing the connections between the original concept and the inferred one. This prompting technique employed a crucial feedback loop. The initial prompts were crafted, and the LLM's responses were analyzed. Based on these outcomes, the prompts were iteratively refined to better guide the LLM's reasoning process and elicit more insightful explanations. This back-and-forth process ensured the prompts became progressively more effective in unlocking the LLM's ability to perform chain-of-thought reasoning.

To optimize the LLM's training process, we translated the data points and prompts into English (and finally translated the outcomes back into Slovenian). This decision stemmed from the fact that languages with fewer resources require more tokens to represent the same information. This increased tokenization could have compromised the LLM's ability to effectively simulate understanding of the concepts. Another interesting observation was that LLMs often favor short, two-sentence outputs. To address this, we implemented techniques that encourage the generation of longer and more coherent paragraphs.

Discussion

Metrics

In order to give extra value to the dataset we have, besides equipping it with explanations, we decided to collect a wide array of output examples (elements of our NLI dataset), created by LLMs. Instead of tending to it manually to achieve grammatical correctness and clear interference, we have decided for a different approach. We have decided to keep all the output examples (the ones that would maintain the fixed structure and the length of the paragraphs that form the example). In that way, we aim for a dataset of higher diversity that still conveys a classification value. For the purpose of maintaining the latter, we have created an evauation metric, Outcome Quality Assessment Composed Metrics (OQACM). The OQACM serves as an evaluation tool, with which we can decide which of the output examples is more valuable, that is which of the examples bears a bigger learning potential. The evaluation is based on the human factor, which can provide a good understanding of the relation between the premise and the hypothesis sentences, and can evaluate how well the text in the paragraphs reads. The human evaluator is asked to give a score from the scale of 1 to 5 in each of the categories to evaluate the fluency, the coherence and

the NLI succes of the example. "NLI success" refers to the original objective of an NLI dataset: expressing whether the hypothesis is true (entailment), untrue (contradiction) or none of these conditions (neutral) in relation to the premise. The human evaluator is simply asked to evaluate how well the paragraphs express these relations, and how well the LLM expresses the reasons for creating such a paragraph ("Explanation"). However, the inference quality has a higher weight than the explanation, meaning that we give more importance to it. The inference quality ("NLI success") has, in fact, the highest weight also when compared to the fluency and the coherence. While these two, the latter consisting of the evaluation categories "Completeness", "Conciseness", "Clarity", "Organisation" and "Focus", and the former of the evaluation categories "Readability", "Grammatical accuracy", "Style", have been assigned the weights 0,83 (fluency) and 0,5 (coherence) for each of the mentioned categories, the categories of "NLI success" have weights of 1,5. The value of the weights comes in question when summing up to the total score of the evaluation metrics. Namely, for every single category, in which the evaluator will give his score of 1 to 5, the score will be multiplied by the weight of the category. These products will then sum up into the total score.

Simple descriptive statistics in the category of scientific explanations shows that the average total score for GPT-4o outcomes is 43.37, with a standard deviation of 2.88, while for Gemini outcomes, the average total score is 41.33 with a standard deviation of 4.83. The t-test results indicate a t-statistic of 1.14 and a p-value of 0.27, suggesting no significant difference in mean total scores between the two systems. Therefore, there is insufficient evidence to conclude that one system consistently outperforms the other based on these scores.

As it is not the main objective of this project to evaluate the performance of the LLM, we do not intend to discuss the statistics in detail. Though, it is worth mentioning that there were differences between the generated paragraphs, on the level of the inference elements (entailment, contradiction, neutral) as well as on the level of categories. In some categories, the examples are visibly more diverse, and the inference relation is more intricate, thus contributing more to the general quality of the dataset. In others, they are simpler and more loyal to one pattern, meaning that they are more similar. To a certain extent, these differences are also noted in the evaluation metrics, giving a slightly better score to the more intricate and creative examples as opposed to the examples where a relation between the premise and the hypothesis is established mainly through summarization, negation or generalization, depending on the inference.

What is worth noting (and is also represented in the evaluation metrics) is GPT-4o'and Gemini's capability of creating fluent, coherent, clear, and predominantly grammatically correct paragraphs. Though there were, as mentioned above, cases where inference is not expressed with great success, in all these cases there is little to no need to correct the grammar or the style of paragraphs. That is especially important given that the dataset is entirely in Slovenian language.

Final Dataset

Seeing as the prompting process, used in the familiarisation process, proved to be successful in creating a consistent structure of examples, created by GPT-4o, team member 1 decided to pursue the same approach in the process of creation. To achieve the objective of creating an NLI dataset, an inquiry was made to find out whether the LLM is familiar with the concept of an NLI dataset. After ensuring that the LLM was indeed familiar with the concept, it was given one of the examples, created during the familiarisation process, and asked to create an example with the same structure, but a different topic. Receiving an output of correct structure, team member 1 took notice of the length of the paragraphs not sufficing the general goal of 'around five sentences'. Because of that, another prompt was given to the LLM, asking for longer paragraphs. That proved to be the last structuring prompt, creating a structure that would never be changed by the LLM. However, all of the following prompts would reference the structure and advise the LLM to maintain the structure as it is and just create a premise, the pertaining hypotheses, and the explanations on a certain topic ('Create another example with the same structure. Topic: ***') The topic would be chosen by team member 1 and given as a reference to the LLM, which would then have the freedom of generating the paragraphs without other limitations. The only exception is the category "News reports", in which a title of an article and a link to the webpage where it was posted would be given as a reference to the LLM.

Team member 2 employed the approach of transitive reasoning to prompt language models across different scenarios. Transitive reasoning, a fundamental cognitive process, proved instrumental in guiding the Gemini language model to provide insightful responses across various scenarios. By structuring prompts to initiate transitive reasoning, we facilitated the generation of intermediate steps, leading to a deeper comprehension of the subject matter. Through an examination of our approach in each category, this report demonstrates the versatility and efficacy of transitive reasoning in prompting language models to provide informative and nuanced responses. The methodology involved structuring prompts to apply transitive reasoning, tailored to each category. While creating news reports we designed prompts that focus on extracting the core message or central theme of the scenario, facilitating Natural Language Inference (NLI) inference. In the realm of fictional stories, prompts delve into character psychology. Asking about emotions, motives, and expectations nudges LLM to connect the dots between a character's actions, their past experiences with other characters (observed relationships), and how those experiences might influence their current behavior. When dealing with historical narratives, prompts aimed to pinpoint the pivotal event within a sequence. By asking LLM to target the main event, we encourage LLM to identify the cause-and-effect relationships between historical occurrences. Through explicit prompts for creating scientific explanations

we required language models to elucidate the main idea or phenomenon under investigation, enhancing understanding of scientific concepts or natural phenomena. Finally, for understanding everday scenarios, prompts focus on common-sense reasoning. Here, prompts encouraged LLM to consider the observed social cues, past experiences with similar circumstances (known relationships), and how those experiences might influence the likely responses of the people involved.

References

[1] Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2018.

[2] Wenpeng Yin, Dragomir Radev, and Caiming Xiong. Docnli: A large-scale dataset for document-level natural language inference. *arXiv preprint arXiv:2106.09449*, 2021.

[3] Wlodek Zadrozny and Karen Jensen. Semantics of paragraphs. *Computational Linguistics*, 17(2):171–210, 1991.

[4] Hesham Suleiman Alyousef. Text cohesion in english scientific texts written by saudi undergraduate dentistry students: A multimodal discourse analysis of textual and logical relations in oral biology texts. *Sage Open*, 11(3), 2021.

[5] Wolfram Bublitz. Discursive pragmatics. In Jan Zienkowski, Jan-Ola Östman, and Jef Verschueren, editors, *Handbook of Pragmatics Highlights 8*, pages 37–49. 2011.

[6] Advaith Siddharthan. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109, 2006.

[7] Tushar Khot, Ashish Sabharwal, and Peter Clark. Scitail: A textual entailment dataset from science question answering. In *Proceedings of AAAI*, pages 5189–5197, 2018.

[8] Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kuebler, and Lawrence S. Moss. Ocnli: Original chinese natural language inference. *arXiv preprint arXiv:2010.05444*, 2020.

[9] R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*, 2019.

[10] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *arXiv preprint arXiv:1812.01193*, 2018.

[11] Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J. Gershman, and Noah D. Goodman. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*, 2018.

[12] Reto Gubelmann, Ioannis Katis, Christina Niklaus, and Siegfried Handschuh. Capturing the varieties of natural language inference: A systematic survey of existing datasets and two novel benchmarks. *Journal of Logic, Language and Information*, 33:21–48, 2024.

[13] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[14] Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. An analysis of natural language inference benchmarks through the lens of negation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online, November 2020. Association for Computational Linguistics.

[15] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[16] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[17] Mohaddeseh Bastan, Mihai Surdeanu, and Niranjan Balasubramanian. Bionli: Generating a biomedical nli dataset using lexico-semantic constraints for adversarial examples. *arXiv preprint arXiv:2210.14814*, 2022.

[18] Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. Natural language inference in context – investigating contextual reasoning over long texts. *arXiv preprint arXiv:2011.04864*, 2020.

[19] Yuta Koreeda and Christopher Manning. ContractNLI: A dataset for document-level natural language inference for contracts. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[20] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online, July 2020. Association for Computational Linguistics.

[21] Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. Generating data to mitigate spurious correlations in natural language inference datasets. *arXiv preprint arXiv:2203.12942*, 2022.

[22] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

[23] Marta Tatu and Munirathnam Srikanth. Experiments with reasoning for temporal relations between events. In *COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference*, Manchester, UK, August 2008.

[24] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems. *arXiv preprint arXiv:1209.2355*, 2013.

[25] Lucius E. J. Bynum, Joshua R. Loftus, and Julia Stoyanovich. A new paradigm for counterfactual reasoning in fairness and recourse. *arXiv preprint arXiv:2401.13935*, 2024.

[26] Zhihong Deng, Jing Jiang, Guodong Long, and Chengqi Zhang. Causal reinforcement learning: A survey. *arXiv preprint arXiv:2307.01452*, 2023.

[27] Roman Abel and Martin Hänze. Generating causal relations in scientific texts: The long-term advantages of successful generation. *Frontiers in Psychology*, 10:199, 2019.

[28] Jie Yang, Soyeon Caren Han, and Josiah Poon. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 64:1161–1186, 2022.

[29] Ronald J. Brachman and Hector J. Levesque. *Knowledge Representation and Reasoning*. Elsevier, 2004.

[30] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[31] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983.

[32] Danjie Chen, Yan Zheng, Liqun Ma, and Fen Qin. An ontology-based knowledge representation method for typhoon events from chinese news reports. *ISPRS International Journal of Geo-Information*, 12(9):383, 2023.

[33] Renate LEP Reniers, Rhiannon Corcoran, Richard Drake, Nick M Shryane, and Birgit A Völlm. The qcae: A questionnaire of cognitive and affective empathy. *Journal of Personality Assessment*, 93(1):84–95, 2011.

[34] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, NY, 2011.

[35] Debjit Paul and Anette Frank. Social commonsense reasoning with multi-head knowledge attention. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2969–2980, Online, November 2020. Association for Computational Linguistics.

[36] Peng Wu, Xiaotong Li, Chen Ling, Shengchun Ding, and Si Shen. Sentiment classification using attention mechanism and bidirectional long short-term memory network. In *Proceedings of the xxx International Conference on Artificial Intelligence (AAAI)*, 2018.

[37] Archit Parnami and Minwoo Lee. Learning from few examples: A summary of approaches to few-shot learning.

[38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2023.

APPENDIX

Demonstration of a prompt for Scientific Explanations

"Create an NLI premise, one contradicting one entailing and one neutral sentence and their labels. Then make premise and hypotheses 5-6 sentences (not less). Apply general principles or theories that are widely accepted in the scientific community. So this way we will have an NLI dataset with not sentences but paragraphs (both hypothesis and premise). That way there will be a common premise and 3 separate hypothesis paragraphs. Add the explanation to each label (why is this an entailment, contradiction or neutral). Along with providing the premise, also mention the main idea or phenomenon being addressed in the present case. To get entailment, you would apply the general principles or theories from the premise to a specific case, such as the behavior of a particular species. The resulting explanation would logically follow from the general principles, providing a scientifically entailed understanding of the phenomenon. For contradiction, ensure that the newly formed hypothesis is logically inconsistent with the explanation in the premise by negating it. One way is to explain the same idea or phenomenon in the opposing way. Neutral should be something that does is not made more or less likely because of the premise. Let the premise be a scientific explanation, that is clear and focuses on the underlying scientific principles or theories, and to get entailment, contradiction causal reasoning should be applied (in case of entailment) or denied (in case of contradiction). Be careful about the neutral hypothesis. It should be something in the context but not

really related to the main idea. In a way seems like relevant to premise but not in the core idea.

Let the phenomenon be sea water density."