ScoNe: Benchmarking Negation Reasoning in Language Models With Fine-Tuning and In-Context Learning[*]

Jingyuan Selena She
Haverford College
jshe@haverford.edu

Christopher Potts
Stanford University
cgpotts@stanford.edu

Samuel R. Bowman
New York University & Anthropic, PBC
bowman@nyu.edu

Atticus Geiger
Stanford University
atticusg@stanford.edu

Abstract

A number of recent benchmarks seek to assess how well models handle natural language negation. However, these benchmarks lack the controlled example paradigms that would allow us to infer whether a model had learned how negation morphemes semantically scope. To fill these analytical gaps, we present the **Sc**oped **Ne**gation NLI (ScoNe-NLI) benchmark, which contains contrast sets of six examples with up to two negations where either zero, one, or both negative morphemes affect the NLI label. We use ScoNe-NLI to assess fine-tuning and in-context learning strategies. We find that RoBERTa and DeBERTa models solve ScoNe-NLI after many shot fine-tuning. For in-context learning, we test InstructGPT models and find that most prompt strategies are not successful, including those using step-by-step reasoning. To better understand this result, we extend ScoNe with ScoNe-NLG, a sentence completion test set that embeds negation reasoning in short narratives. Here, InstructGPT is successful, which reveals the model can correctly reason about negation, but struggles to do so on prompt-adapted NLI examples outside of its core pretraining regime.

1 Introduction

Negation is a ubiquitous but complex linguistic phenomenon that poses a significant challenge for NLP systems. A diverse array of benchmarks focused on negation have appeared in recent years, many of which contain families of contrasting examples that provide a local view of the model decision boundary (Gardner et al., 2020). For instance, Cooper et al. (1996), McCoy and Linzen (2018), Wang et al. (2019), Ettinger (2020), Hartmann et al. (2021), and Kassner and Schütze (2020) all conduct evaluations with minimal pairs of examples that are identical except for a negative morpheme. These examples reveal whether the presence of negation has a causal impact on model predictions.

However, negation is not simply present or absent in a sentence. Rather, negation morphemes are semantic operators that take scope in complex ways, as we see in clear contrasts like *the person who was at the talk wasn't happy* and *the person who wasn't at the talk was happy*. The recent CondaQA benchmark of Ravichander et al. (2022) includes minimal pairs aimed at determining whether models are sensitive to these differences in scope.

With the current paper, we seek to provide an even fuller picture of the complexities of negation and semantic scope. We introduce the English-language **Sco**ped **Ne**gation Natural Language Inference Benchmark (ScoNe-NLI). ScoNe-NLI extends the negated portion of the Monotonicity NLI dataset (Geiger et al., 2020) such that each of the 1,202 examples is now a contrast set with six examples in which zero, one, or two negations are present and each negation may or may not have a semantic scope such that the NLI label is impacted by its presence. These six conditions offer a rich picture of how negation affects NLI reasoning, and they allow us to determine whether models are truly able to handle nested negation and scope or whether they have found simplistic solutions.

We evaluate models on ScoNe-NLI using many-shot fine-tuning as well as a wide range of in-context learning strategies. For fine-tuning approaches, we find that RoBERTa and DeBERTa models both solve ScoNe-NLI. For in-context learning, we evaluate the latest InstructGPT model with a variety of prompt strategies. We find that these models perform well on sections of ScoNe-NLI where the negation morphemes can simply be ignored, but they systematically fail in conditions where exactly one negative morpheme has semantic scope such that its presence changes the NLI label. In other words, these models fail to learn in context how negation actually takes scope.

To better understand this result, we introduce a sentence completion test set (ScoNe-NLG) contain-

[*] https://github.com/selenashe/ScoNe

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1803–1821
July 9-14, 2023 ©2023 Association for Computational Linguistics

Split	Premise	Rel.	Hypothesis	Examples
No negation	The cowboy fell off a horse at the competition	$\sqsupset$	The cowboy fell off a racehorse at the competition	1,202
One Not Scoped	The cowboy did not fear anything, until he fell off a horse at the competition	$\sqsupset$	The cowboy did not fear anything, until he fell off a racehorse at the competition	1,202
Two Not Scoped	The cowboy, who was not very old, was not proud that he fell off a horse at the competition	$\sqsupset$	The cowboy, who was not very old, was not proud that he fell off a racehorse at the competition	1,202
Two Scoped	There is no way that the cowboy did not fall off a horse at the competition	$\sqsupset$	There is no way that the cowboy did not fall off a racehorse at the competition	1,202
One Scoped	The cowboy did not fall off a horse at the competition	$\sqsubset$	The cowboy did not fall off a racehorse at the competition	1,202
One Scoped, One not Scoped	The cowboy did not fall off a horse, but the competition was not too important	$\sqsubset$	The cowboy did not fall off a racehorse, but the competition was not too important	1,202

(a) A six-example contrast set from ScoNe-NLI.

No Negation
Glen is a fan of learning math. When he sees that his new high school requires that he take a calculus course, he
Negation
Glen is not a fan of learning math. When he sees that his new high school requires that he take a calculus course, he
Non-Scoping Negation
Glen isn't just a fan of learning math, he's obsessive. When he sees that his new high school requires that he take a calculus course, he

(b) A three-example contrast set from ScoNe-NLG.

Table 1: Two contrast sets from the ScoNe Benchmark

ing examples that seem better aligned with what we can infer about the training data used for InstructGPT models. In each ScoNe-NLG example, negation reasoning is needed to provide a coherent ending to an incomplete narrative (see Figure 1b). ScoNe-NLG contains minimal triplets of examples where negation is absent, present with relevant scope, or present without relevant scope. InstructGPT is successful on ScoNe-NLG. When considered alongside our negative result for ScoNe-NLI, this finding seems to show that these models *can* learn in-context about how negation takes scope, but only when the examples are hand-tailored to be aligned with the training data and aligned with known strengths of these models. Thus, when used together, ScoNe-NLI and ScoNe-NLG serve as a clear diagnostic for exploring useful prompting strategies and assessing the capacity of language models to reason about negation and scope.

2 A Brief Review of Negation in NLI Benchmarks

A diverse array of benchmarks and diagnostic experiments have included negation reasoning in recent years (Nairn et al., 2006; McCoy and Linzen, 2018; Wang et al., 2019; Ettinger, 2020; Hartmann et al., 2021; Kassner and Schütze, 2020; Ravichander et al., 2022).

Hossain et al. (2022) analyze a variety of natural language understanding benchmarks and find that negation is underrepresented, and that when negation is present it often has no impact on the example label. Hossain et al. (2020) address this issue by manually adding negation to the premise-hypothesis pairs in MNLI (Williams et al., 2018), SNLI (Bowman et al., 2015), and RTE (Dagan et al., 2007; Cooper et al., 1996).

Yanaka et al. (2019a) introduce the crowdsourced MED dataset, which has many NLI examples where negation generates inferences. Monotonicity NLI (MoNLI; Geiger et al. 2020) consists of modified SNLI sentences that have gold labels impacted by lexical entailments in affirmative contexts (PMoNLI) and lexical entailments reversed by a negation (NMoNLI). BERT fine-tuned on SNLI and MNLI fails to generalize to both of these datasets, but succeeds with further fine-tuning on MED/MoNLI. Some automatically generated NLI datasets also include negation reasoning (Geiger et al., 2019; Richardson et al., 2020; Yanaka et al., 2019b, 2021).

3 ScoNe-NLI

ScoNe-NLI is an extension of MoNLI (Geiger et al., 2020). MoNLI was generated by randomly selecting a sentence from SNLI and replacing a noun with a hypernym (more general term) or

Fine-tuning Datasets	No Negation	One Not Scoped	Two Not Scoped	Two Scoped	One Scoped	One Scoped, One not Scoped
MAF-NLI	82.0	86.0	81.5	91.0	5.0	5.0
MAF-NLI+ MoNLI (Geiger et al., 2020)	96.2	87.5	99.5	8.9	100.0	100.0
MAF-NLI+ MED (Yanaka et al., 2020)	84.8	83.5	82.0	58.9	99.5	97.0
MAF-NLI+ Neg-NLI (Hossain et al., 2020)	91.3	88.5	83.0	70.4	37.0	29.0
MAF-NLI+ MoNLI + ScoNe-NLI	100.0	100.0	100.0	100.0	100.0	100.0

Table 2: DeBERTa fine-tuning results on ScoNe-NLI. MAF-NLI stands for on MNLI, ANLI, and Fever-NLI.

Conditional Q	Is it true that if **Premise**, then **Hypothesis**?
Hypothesis Q	Assume that **Premise**. Is it then definitely true that **Hypothesis**? Answer yes or no.
Conditional Truth	If **Premise**, then **Hypothesis**. Is this true?
Brown et al.	P: **Premise**\n Q: **Hypothesis**\n Yes, No, or Maybe?
Structured	P: **Premise**\n H: **Hypothesis**\nL:
Reasoning	Logical and commonsense reasoning exam.\n\n Explain your reasoning in detail, then answer with Yes or No. Your answers should follow this 4-line format:\n\n Premise: <a tricky logical statement about the world>.\n Question: <question requiring logical deduction>.\n Reasoning: <an explanation of what you understand about the possible scenarios>\n Answer: <Yes or No>.\n\n Premise: **Premise**\n Question: **Hypothesis**\n Reasoning: Let's think logically step by step. The premise basically tells us that

Table 3: Prompts used to adapt a 2-way NLI example (**Premise**, **Hypothesis**). Newlines are indicated with \n. Full prompts with few-shot variants are in Appendix E.

hyponym (less general term). The original and edited sentences are then used to form two premise–hypothesis pairs, one with the label *entailment* and the other with the label *neutral*. In about half of the examples, this replacement is in an affirmative context with no negation (PMoNLI). In the other half, it is under the scope of a single negation (NMoNLI).

The authors generated ScoNe-NLI by using each example of NMoNLI to create a contrast set of six examples where gold labels are impacted by the scope of zero, one, or two negations, as in Table 1.

To succeed across all sections of ScoNe, models need to attend to the presence of negation as well as the way it scopes semantically. Table 1a shows an actual example of how ScoNe extends MoNLI. We use the train–test split of MoNLI where substituted lexical items are disjoint across training and testing data. Appendix C provides further details.

Fine-Tuning on ScoNe-NLI We used publicly available weights on HuggingFace for the DeBERTa-v3-base models already fine-tuned on MNLI, Fever-NLI, and Adversarial-NLI (Laurer et al., 2022; He et al., 2021). Appendix B contains comparable results for the RoBERTa model (Liu et al., 2019). Fine-tuning results are in Table 2.

Fine-tuning on existing NLI datasets is insufficient for good performance on ScoNe-NLI: DeBERTa-v3-base fine-tuned on existing NLI datasets, even those that focus on negation, systematically fails. Thus, it seems that ScoNe-NLI captures novel aspects of negation reasoning.

In contrast, fine-tuning on MoNLI and ScoNe-NLI training data results in near perfect performance on ScoNe-NLI test data. This shows that DeBERTa can learn negation reasoning and generalize to new lexical items.

In-context Learning on ScoNe-NLI We evaluated InstructGPT using OpenAI's API with *text-davinci-002* and *text-davinci-003* engines and a temperature of 0.0 (Brown et al., 2020). We ask InstructGPT to infer NLI labels given the premise and hypothesis using prompts. All prompts are constructed such that if the response contain "yes" (case-insensitive), then the label *entailment* is predicted, else the label *neutral* is predicted. We use six prompts (Table 3). For each prompt, we implemented both zero-shot and few-shot inference experiments. Appendix E provides the full prompts.

InstructGPT makes systematic errors similar to a baseline that ignores negation entirely. The best results are for the few-shot reasoning prompt with *davinci-003*. While its overall accuracy of 82% may initially appear to be a success, further analysis reveals otherwise. InstructGPT succeeds only on the sections of ScoNe-NLI where zero or two negations take scope, namely, no negation (99%), one not scoped (97%), two not scoped

		No Negation	One Not Scoped	Two Not scoped	Two Scoped	One Scoped	One Scoped, One not Scoped	Overall
Zero-shot	Structured	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	Brown et al.	0.74	0.70	0.74	0.55	0.44	0.45	0.60
	Conditional Q	0.79	0.84	0.80	0.50	0.52	0.44	0.65
	Conditional Truth	0.98	0.86	0.80	0.43	0.66	0.47	0.70
	Hypothesis Q	0.69	0.90	0.70	0.51	0.62	0.42	0.64
	Reasoning	0.90	0.88	0.94	0.72	0.52	0.46	0.73
Few-shot	Structured	0.50	0.50	0.50	0.50	0.50	**0.50**	0.50
	Brown et al.	0.86	0.66	0.80	0.83	0.36	0.28	0.63
	Conditional Q	0.92	0.85	0.90	0.62	0.34	0.34	0.66
	Conditional Truth	0.94	0.90	0.94	0.64	0.36	0.37	0.69
	Hypothesis Q	0.98	0.96	0.94	0.83	0.51	0.40	0.77
	Reasoning	**0.99**	**0.97**	**0.98**	**0.89**	**0.69**	0.43	**0.82**
	Ignore-Negation	1.00	1.00	1.00	1.00	0.00	0.00	0.66

Table 4: In-context learning results on ScoNe-NLI for InstructGPT (*davinci-003* engine; see Appendix F for corresponding results for *davinci-002*, which are uniformly lower). Zero-shot results are given in the first group of rows, with the best results in that condition underlined. Few-shot results are given in the second group, with the best results for this condition (and overall) in bold. The bottom row specifies a simple, idealized Ignore-Negation baseline that makes predictions as if negations were absent. The baseline shows that the seemingly solid Overall results of these models are driven largely by conditions for which negation can be ignored. Conversely, models are often at or below chance where negation is critical in some way.

	No Negation	One Scoped	One Not Scoped	Overall
Zero-shot	0.99	0.90	0.88	0.92
Few-shot	0.93	1.00	0.93	0.95

Table 5: Results for ScoNe-NLG using `davinci-003`. The three conditions correspond to those of ScoNe and test the essential scope-taking properties of negation.

(98%), and two scoped (89%). InstructGPT performs much worse on sections where exactly one negation takes scope, namely one scoped (69%), one scoped/one not (48%). An idealized baseline entirely ignoring the presence of negation (last row of Table 4) succeeds and fails on the same sections, indicating a systematic flaw in InstructGPT.

4 ScoNe-NLG

InstructGPT fails to reason about negation when given NLI examples that must be adapted to natural language generation (NLG) with prompts. We hypothesized that InstructGPT may correctly reason about negation when evaluated on examples hand tailored to its pretraining objective, because there is no need for prompt engineering (Liu et al., 2021; Wei et al., 2022; Kojima et al., 2022).

Dataset ScoNe-NLG is a natural language generation dataset that contains 74 contrasting triplets of examples of half-completed naturalistic narratives that have different coherent completions depending on the presence and scope of a negation. InstructGPT fails on the sections of ScoNe-NLI examples containing only one negation, so we opt for contrast sets with three examples that require knowledge of a lexical entailment in an affirmative context without negation, an affirmative context with non-scoping negation, and an negative context with scoping negation, respectively. See Table 1b.

In-context Learning on ScoNe-NLG We used InstructGPT to complete the partial sentence inputs with the *text-davinci-003* engine (temperature of 0.0). In the zero-shot setting, the prompt consists of the ScoNe-NLG example. In the few-shot setting, four demonstrations from ScoNe-NLG are given one with no negation, two with scoping negation, and one with non-scoping negation. See Appendix E.13 for the complete prompts.

To evaluate, the authors went through the responses by hand and determined whether the generated text is coherent and compatible with the initial narrative. The authors agreed on these annotations for 216/222 of the zero-shot responses with a Fleiss kappa of 0.84 and 220/222 of the few-shot responses with a Fleiss kappa of 0.91. These agreement rates are so high that we evaluate InstructGPT only for the cases where the annotators agree. Here, InstructGPT is successful but not perfect, achieving 95% and 92% accuracy in the few and zero-shot settings, respectively. We do not observe the systematic failures seen on ScoNe-NLI.

```
SCONE-BOOL($\mathbf{p}, \mathbf{h}$)
lexrel $\leftarrow$ GET-LEXREL($\mathbf{p}, \mathbf{h}$)
neg1 $\leftarrow$ FIRST-SCOPE($\mathbf{p}, \mathbf{h}$)
neg2 $\leftarrow$ SECOND-SCOPE($\mathbf{p}, \mathbf{h}$)
if (neg1 $\oplus$ neg2)):
    return REVERSE(lexrel)
return lexrel
```

(a) An interpretable program that solves ScoNe-NLI by computing two Boolean variables that encode whether the first and second negation scope and reversing entailment if exactly one is true.

```
SCONE-COUNT($\mathbf{p}, \mathbf{h}$)
lexrel $\leftarrow$ GET-LEXREL($\mathbf{p}, \mathbf{h}$)
count $\leftarrow$ COUNT-SCOPED($\mathbf{p}, \mathbf{h}$)
if count == 1:
    return REVERSE(lexrel)
return lexrel
```

(b) An interpretable program that solves ScoNe-NLI by counting the scoped negations and reversing entailment if there is exactly one.

```
IGNORE-SCOPE($\mathbf{p}, \mathbf{h}$)
lexrel $\leftarrow$ GET-LEXREL($\mathbf{p}, \mathbf{h}$)
count $\leftarrow$ COUNT-NEG($\mathbf{p}, \mathbf{h}$)
if count == 1:
    return REVERSE(lexrel)
return lexrel
```

(c) A flawed heuristic program: we count the negations and reverse entailment if there is a single negation, which is equivalent to ignoring the scope of negation.

```
IGNORE-NEGATION($\mathbf{p}, \mathbf{h}$)
lexrel $\leftarrow$ GET-LEXREL($\mathbf{p}, \mathbf{h}$)
return lexrel
```

(d) A flawed heuristic program for ScoNe-NLI that outputs the lexical relation and ignores negation entirely.

Figure 1: Four human-interpretable algorithms for ScoNe-NLI. The first two solve the task perfectly, and the other two implement flawed heuristics that a model might learn to implement. The function GET-LEXREL retrieves the relation between the aligned words in the premise and hypothesis, COUNT-SCOPED counts scoped negations, COUNT-NEG counts negations regardless of scope, and GET-FIRST returns true if the first negation scopes, while GET-SECOND returns true if there is a second negation and it scopes.

5 Future Work on Interpretability

ScoNe is based in naturalistic examples, but it also has a controlled structure that offers valuable opportunities to move beyond simple behavioral testing and more deeply understand *how* models solve tasks related to lexical entailment and negation.

The theory of causal abstraction provides a framework for interpretability (Geiger et al., 2023a), where a neural model can be understood to implement the intermediate variables and internal structure of a program or algorithm (Geiger et al., 2021, 2022; Wu et al., 2022b,a; Huang et al., 2022; Geiger et al., 2023b). In fact, the MoNLI dataset and the technique of interchange interventions (which is the primary technique in causal abstraction analysis) were jointly introduced in Geiger et al. 2020, where interchange interventions were used to investigate whether a BERT model implements a simple, human-interpretable algorithm that can perfectly label MoNLI using a variable representing lexical entailment and a variable representing the presence of negation.

With ScoNe, we can ask even deeper interpretability questions of this form. To encourage future work in this direction, we present a range of algorithmic solutions in Figure 1. Two of these solutions solve ScoNe and could perhaps explain neural models that learn the task perfectly, and two others implement flawed heuristics that could explain neural models with poor task performance.

Figure 1a and Figure 1b present two intuitive and correct algorithms that solve ScoNe, but have distinct intermediate variables and internal structure. The first computes two Booleans representing whether each negation scopes, and the second computes a count of how many negations scope.

Figure 1d is the flawed heuristic that ignores negation that we discussed in Section 3 as a hypothesis about how models fail at our task. Figure 1d is a second flawed heuristic that counts the number of negations present but ignores scope.

Using the toolkit of causal abstraction, we can assess models not only behaviorally, but also evaluate whether they implement an interpretable algorithm. The results of Geiger et al. (2023b) begin to show how such analyses could be extended to in-context learning with LLMs, as in Section 4.

6 Conclusion

We introduced ScoNe, a benchmark for fine-tuning and in-context learning experiments on negation. ScoNe is challenging for NLI models fine-tuned on other datasets, even those designed for negation reasoning, but modest amount of fine-tuning on ScoNe leads to success. For in-context learning, we find that that InstructGPT models fail dramatically on ScoNe. However, we also introduce ScoNe-NLG, which uses more narrative-like examples to probe models' capacity to handle negation, and show that InstructGPT is successful with zero-shot and few-shot prompts for this task. These results show that ScoNe supports fine-grained assessments of whether models can reason accurately about natural language negation, and our discussion in Section 5 suggests that ScoNe can be a powerful tool for discovering *how* models reason semantically.

Limitations

We are releasing ScoNe as a diagnostic tool for conducting controlled scientific experiments. This is our primary intended use, and we advise against uncritical use of ScoNe for real-world applications, as we have not audited the dataset for such purposes.

As a diagnostic tool, ScoNe's primary limitation is its focus on English. Cross-linguistically, we find many strategies for expressing negation. The English-language strategy of using mostly adverbial modifiers for sentential negation is not the only one by any means, and we would expect to see quite different results for languages in which negation is expressed, for example, with verbal suffixes. This highlights the value of potential future efforts extending ScoNe to other languages.

By the same token, we acknowledge that many linguistic phenomena interact with negation even internal to English. ScoNe restricts to negation in the context of lexical entailment, and mostly uses "not" as the negative morpheme. This excludes a wide range of negation morphemes and negation strategies that ultimately need to be brought into the picture.

Finally, we note that there may be undesirable biases in ScoNe that could interact with biases in the models. ScoNe is in part derived from SNLI, which is known to contain gaps, social biases, and artifacts (Poliak et al., 2018; McCoy et al., 2019; Belinkov et al., 2019; Gururangan et al., 2018; Tsuchiya, 2018), and ScoNe may inherit some of these.

References

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891, Florence, Italy. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework. Technical report, LRE 62-051 D-16, The FraCaS Consortium.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2007. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*.

Allyson Ettinger. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1307–1323. Association for Computational Linguistics.

Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. 2019. Posing fair generalization tasks for natural language inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4475–4485, Stroudsburg, PA. Association for Computational Linguistics.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586.

Atticus Geiger, Chris Potts, and Thomas Icard. 2023a. Causal abstraction for faithful interpretation of AI models. ArXiv:2106.02997.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online. Association for Computational Linguistics.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. 2022. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7324–7338. PMLR.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. 2023b. Finding alignments between interpretable causal variables and distributed neural representations. Ms., Stanford University.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *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. Association for Computational Linguistics.

Mareike Hartmann, Miryam de Lhoneux, Daniel Hershcovich, Yova Kementchedjhieva, Lukas Nielsen, Chen Qiu, and Anders Søgaard. 2021. A multilingual benchmark for probing negation-awareness with minimal pairs. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 244–257, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. 2022. An analysis of negation in natural language understanding corpora. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 716–723, Dublin, Ireland. Association for Computational Linguistics.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Jing Huang, Zhengxuan Wu, Kyle Mahowald, and Christopher Potts. 2022. Inducing character-level structure in subword-based language models with Type-level Interchange Intervention Training. Ms., Stanford University and UT Austin.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2022. Less annotating, more classifying – addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys (CSUR)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

R. Thomas McCoy and Tal Linzen. 2018. Non-entailed subsequences as a challenge for natural language inference. *CoRR*, abs/1811.12112.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Rowan Nairn, Cleo Condoravdi, and Lauri Karttunen. 2006. Computing relative polarity for textual inference. In *Proceedings of the Fifth International Workshop on Inference in Computational Semantics (ICoS-5)*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Abhilasha Ravichander, Matt Gardner, and Ana Marasović. 2022. Condaqa: A contrastive reading comprehension dataset for reasoning about negation.

Kyle Richardson, Hai Hu, Lawrence S. Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8713–8721. AAAI Press.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Zhengxuan Wu, Karel D'Oosterlinck, Atticus Geiger, Amir Zur, and Christopher Potts. 2022a. Causal Proxy Models for concept-based model explanations. ArXiv:2209.14279.

Zhengxuan Wu, Atticus Geiger, Joshua Rozner, Elisa Kreiss, Hanson Lu, Thomas Icard, Christopher Potts, and Noah Goodman. 2022b. Causal distillation for language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4295, Seattle, United States. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, and Kentaro Inui. 2020. Do neural models learn systematicity of monotonicity inference in natural language? In *Annual Meeting of the Association for Computational Linguistics*.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019a. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019b. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 250–255, Minneapolis, Minnesota. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, and Kentaro Inui. 2021. SyGNS: A systematic generalization testbed based on natural language semantics. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 103–119, Online. Association for Computational Linguistics.

Appendices

A Experimental Details

A.1 Fine-tuning Protocol

For our fine-tuning experiments, we used a learning rate of 1e-5, batch size of 4, gradient accumulation steps of 6 for a total of 10 epochs. We used these default hyperparameters as they were successful in fine-tuning on ScoNe. We implemented these experiments with Pytorch (Paszke et al., 2019) and used the scikit learn package (Pedregosa et al., 2011).

A.2 Hugging Face Models

We test RoBERTa[1] and DeBERTa[2] in these experiments. We used the roberta-large model fine-tuned on MNLI[3] with 354 million parameters, 500K steps, and trained on 1,024 V100 GPUs (Liu et al., 2019). DeBERTa-v3-base-mnli-fever-anli model[4] was fine-tuned on MNLI, Fever-NLI,[5] and ANLI.[6]

RoBERTa weights link:
`https://huggingface.co/roberta-large-mnli`

Deberta weights link:
`https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli`

A.3 Fine-Tuning Datasets

We further fine-tuned our model on the datasets MoNLI,[7] Negation-NLI, [8] MED. [9]

B RoBERTa Results

Fine-tuning Datasets	No Negation	One Not Scoped	Two Not Scoped	Two Scoped	One Scoped	One Scoped, One not Scoped
MAF-NLI	96.5	97.0	97.0	96.5	3.0	5.0
MAF-NLI+ MoNLI (Geiger et al., 2020)	85.4	100.0	100.0	4.5	100.0	100.0
MAF-NLI+ MED (Yanaka et al., 2020)	85.1	92.0	89.5	44.6	85.5	81.5
MAF-NLI+ Neg-NLI (Hossain et al., 2020)	93.1	97.5	93.0	73.2	20.5	17.5
MAF-NLI+ MoNLI + ScoNe-NLI	100.0	100.0	100.0	100.0	100.0	100.0

Table 6: RoBERTa fine-tuning results on ScoNe-NLI. MAF-NLI stands for on MNLI, ANLI, and Fever-NLI.

C ScoNe Dataset Details

For some examples, we modified the lexical items replaced. Consider the NMoNLI sentence pair ‘a man is not tossing anything’-‘a man is not tossing socks’ (entailment), and non-scoping counterpart ‘a man not here is tossing something’-‘a man not here is tossing socks’ (neutral). Here, ’anything’ must be replaced by ’something’. The positive and negative examples in MoNLI *do not* come in minimal pairs, so the examples in ScoNe-NLI with no negation are *not* from PMoNLI.

[1] released under the MIT license

[2] released under the MIT license

[3] released under the MIT license

[4] released under the MIT license

[5] released under the Creative Commons Attribution-ShareAlike License (version 3.0)

[6] released under the Attribution-NonCommercial 4.0 International license

[7] released under the Creative Commons Attribution Share Alike 4.0 International license

[8] released under the MIT license

[9] released under the Creative Commons Attribution Share Alike 4.0 International license

D Prompting Methods

The experimental runs reported in the paper were conducted on January 11, 2023. We used InstructGPT[10] models with 1.3 billion parameters and 6 billion parameter. The exact cost of constructing the InstructGPT models is not public, but the pre-training protocol involves (1) fine-tuning a GPT3 model on an instruction following dataset, (2) fine-tuning a GPT3 model to rank different answers to the instruction following dataset, and (3) using reenforcement learning to combine these two models. We use a temperature parameter of 0.0 for all experiments. If the response contains "yes" (case-insensitive), then we infer the label `entailment`, else we infer `neutral`. Across experiments, the only thing that varies is the nature of the prompt function.

E In-Context Learning Prompts

We have indicated all actual newlines with \n. The newlines in the formatting are just to make them intuitive to read.

E.1 Conditional Question Prompt

Prompt example

Is it true that if we didn't eat pizza, then we didn't eat food?

E.2 Few-Shot Conditional Question Prompt

Prompt example

Q1: Is it true that if a not so tall person reading a paper is not currently sitting inside a building, then a not so tall person reading a paper is not currently sitting inside a club?\n

A1: Yes\n

\n

Q2: Is it true that if the man does not own a dog and does not own a cat, then the man does not own a retriever and does not own a cat?\n

A2: Yes\n

\n

Q3: Is it true that if a not so tall person reading a paper is not currently sitting inside a cabin, then a not so tall person reading a paper is not currently sitting inside a building?\n

A3: Maybe\n

\n

Q4: Is it true that if a not so tall person reading a paper is not currently sitting inside a casino, then a not so tall person reading a paper is not currently sitting inside a building? A4: Maybe\n

\n

Q: Is it true that if we didn't eat pizza, then we didn't eat food?\n

A:

E.3 Hypothesis Question Prompt

Prompt example

Assume that we didn't eat pizza. Is it then definitely true that we didn't eat food? Answer Yes or No.

[10] information on terms of use is available at: https://openai.com/terms/

E.4 Few-Shot Hypothesis Question Prompt

Prompt example

Q1: Assume that a not so tall person reading a paper is not currently sitting inside a building. Is it then definitely true that a not so tall person reading a paper is not currently sitting inside a casino? Answer Yes or No.\n
A1: Yes\n
\n
Q2: Assume that the girl will not get a stuffed dog as a gift, but not because she failed the exam. Is it then definitely true that the girl will not get a stuffed pinscher as a gift, but not because she failed the exam? Answer Yes or No.\n
A2: Yes\n
\n
Q3: Assume that the girl will not get a stuffed shetland as a gift, but not because she failed the exam. Is it then definitely true that the girl will not get a stuffed dog as a gift, but not because she failed the exam? Answer Yes or No.\n
A3: No\n
\n
Q4: Assume that a not so tall person reading a paper is not currently sitting inside a monastery. Is it then definitely true that a not so tall person reading a paper is not currently sitting inside a building? Answer Yes or No.\n
A4: No\n
\n
Q: Assume that we didn't eat pizza. Is it then definitely true that we didn't eat food? Answer Yes or No.\n
A:

E.5 Conditional Truth Evaluation Prompt

Prompt example

If we didn't eat pizza, then we didn't eat food. Is this true?

E.6 Few-Shot Conditional Truth Evaluation Prompt

Prompt example

```
C1: If the man does not own a dog and does not own a cat, then the man does not own a shetland and does not own a cat. Is this true?\n
A1: Yes\n
\n
C2: If a not so tall person reading a paper is not currently sitting inside a building, then a not so tall person reading a paper is not currently sitting inside a house. Is this true?\n
A2: Yes\n
\n
C3: If the man does not own a collie and does not own a cat, then the man does not own a dog and does not own a cat. Is this true?\n
A3: Maybe\n
\n
C4: If the man does not own a corgi and does not own a cat, then the man does not own a dog and does not own a cat. Is this true?\n
A4: Maybe\n
\n
C:If we didn't eat pizza, then we didn't eat food. Is this true?\n
A:
```

E.7 Brown Et Al Style Prompt

Prompt example

```
C: We didn't eat pizza\n
Q: We didn't eat food. Yes, No, or Maybe?
```

E.8 Few-Shot Brown Et Al Style Prompt

Prompt example

```
C1: The man, who's eyes are not open, is not steering a car.\n
Q1: The man, who's eyes are not open, is not steering a sedan. Yes, No, or Maybe?\n
A2: Yes\n
\n
C2: A dog not on the playground did not catch any ball.\n
Q2: A dog not on the playground did not catch any volleyball. Yes, No, or Maybe?\n
A3: Yes\n
\n
C3: the man does not own a collie and does not own a cat.\n
Q3: the man does not own a dog and does not own a cat. Yes, No, or Maybe?\n
A4: Maybe\n
\n
C4: A not so tall person reading a paper is not currently sitting inside a inn.\n
Q4: A not so tall person reading a paper is not currently sitting inside a building. Yes, No, or Maybe?\n
A5: Maybe\n
\n
C: We didn't eat pizza\n
Q: We didn't eat food. Yes, No, or Maybe?\n
A:
```

E.9 Structured Prompt

Prompt example

```
P: We didn’t eat pizza\n
H: We didn’t eat food\n
L:
```

E.10 Few-Shot Structured Prompt

Prompt example

```
P1: The players who did not score did not have a ball.\n
H1: The players who did not score did not have a baseball.\n
L1: entailment\n
\n
P2: the man does not own a dog and does not own a cat.\n
H2: the man does not own a poodle and does not own a cat.\n
L2: entailment\n
\n
P3: the man does not own a terrier and does not own a cat.\n
H3: the man does not own a dog and does not own a cat.\n
L3: neutral\n
\n
P4: the man does not own a husky and does not own a cat.\n
H4: the man does not own a dog and does not own a cat.\n
L4: neutral\n
\n
P: We didn’t eat pizza\n
H: We didn’t eat food\n
L:
```

E.11 Reasoning Prompt

Prompt example

```
Logical and commonsense reasoning exam.\n
\n
Explain your reasoning in detail, then answer with Yes or No. Your answers should follow this 4-line format:\n
\n
Premise: <a tricky logical statement about the world>.\n
Question: <question requiring logical deduction>.\n
Reasoning: <an explanation of what you understand about the possible scenarios>.\n
Answer: <Yes or No>.\n
\n
Premise: we didn’t eat pizza\n
Question: Can we logically conclude for sure that we didn’t eat food?\n
Reasoning: Let’s think logically step by step. The premise basically tells us that
```

E.12 Few-shot Reasoning Prompt

For this prompt, we insert two demonstrations right before the test example. These are of the correct type for the test example, and they exemplify each of the two labels. The demonstrations are from a fixed set of examples, which we include here:

E.12.1 No Negation

Prompt example

```
Here are some examples of the kind of reasoning you should do:\n
\n
Premise: The students ate pizza\n
Question: Can we logically conclude for sure that the students ate food?\n
Reasoning: Let's think logically step by step. The premise basically tells us that pizza is a type of food. Therefore, the premise that the students ate pizza entails that the students ate food.\n
Answer: Yes\n
\n
Premise: The students ate food\n
Question: Can we logically conclude for sure that the students ate pizza?\n
Reasoning: Let's think logically step by step. The premise basically tells us that pizza is a type of food. Therefore, the premise that the students ate food does not allow us to conclude that the students ate pizza. They might have eaten something else.\n
Answer: No\n
\n
```

E.12.2 One Scoped

Prompt example

```
Here are some examples of the kind of reasoning you should do:\n
\n
Premise: The students didn't eat any pizza\n
Question: Can we logically conclude for sure that the students didn't eat any food?\n
Reasoning: Let's think logically step by step. The premise basically tells us that pizza is a type of food. Therefore, the premise that the students didn't eat any pizza does not allow us to conclude that the students didn't eat any food. They might have eaten something else.\n
Answer: No\n
\n
Premise: The students didn't eat any food\n
Question: Can we logically conclude for sure that the students didn't eat any pizza?\n
Reasoning: Let's think logically step by step. The premise basically tells us that pizza is a type of food. Therefore, the premise that the students didn't eat any food entails that the students didn't eat any pizza.\n
Answer: Yes\n
\n
```

E.12.3 One Not Scoped

Prompt example

```
Here are some examples of the kind of reasoning you should do:\n
\n
Premise: The students who weren’t in class ate pizza\n
Question: Can we logically conclude for sure that the students who weren’t in class ate food?\n
Reasoning: Let’s think logically step by step. The premise basically tells us that pizza is a type of food. Therefore, the premise that the students who weren’t in class ate pizza entails that the students who weren’t in class ate food.\n
Answer: Yes\n
\n
Premise: The students who weren’t in class ate food\n
Question: Can we logically conclude for sure that the students who weren’t in class ate pizza?\n
Reasoning: Let’s think logically step by step. The premise basically tells us that pizza is a type of food. Therefore, the premise that the students who weren’t in class ate food does not allow us to conclude that the students who weren’t in class ate pizza. They might have eaten something else.\n
Answer: No\n
\n
```

E.12.4 One Scoped, One Not Scoped

Prompt example

```
Here are some examples of the kind of reasoning you should do:\n
\n
Premise: The students who weren’t in class didn’t eat any pizza\n
Question: Can we logically conclude for sure that the students who weren’t in class didn’t eat any food?\n
Reasoning: Let’s think logically step by step. The premise basically tells us that pizza is a type of food. Therefore, the premise that the students who weren’t in class didn’t eat any pizza does not allow us to conclude that the students who weren’t in class didn’t eat any food. They might have eaten something else.\n
Answer: No\n
\n
Premise: The students who weren’t in class didn’t eat any food\n
Question: Can we logically conclude for sure that the students who weren’t in class didn’t eat any pizza?\n
Reasoning: Let’s think logically step by step. The premise basically tells us that pizza is a type of food. Therefore, the premise that the students who weren’t in class didn’t eat any food entails that the students who weren’t in class didn’t eat any pizza.\n
Answer: Yes\n
\n
```

E.12.5 Two Not Scoped

Prompt example

Here are some examples of the kind of reasoning you should do:\n

\n

Premise: The students who weren’t in class ate pizza that wasn’t hot\n

Question: Can we logically conclude for sure that the students who weren’t in class ate food that wasn’t hot?\n

Reasoning: Let’s think logically step by step. The premise basically tells us that pizza is a type of food. Therefore, the premise that the students who weren’t in class ate pizza that wasn’t hot entails that the students who weren’t in class ate food that wasn’t hot.\n

Answer: Yes\n

\n

Premise: The students who weren’t in class ate food that wasn’t hot\n

Question: Can we logically conclude for sure that the students who weren’t in class ate pizza that wasn’t hot?\n

Reasoning: Let’s think logically step by step. The premise basically tells us that pizza is a type of food. Therefore, the premise that the students who weren’t in class ate food that wasn’t hot does not allow us to conclude that the students who weren’t in class ate pizza that wasn’t hot. They might have eaten something else.\n

Answer: No\n

\n

E.12.6 Two Scoped

Prompt example

Here are some examples of the kind of reasoning you should do:\n

\n

Premise: It is not the case that the students didn’t eat any pizza\n

Question: Can we logically conclude for sure that it is not the case that the students didn’t eat any food?\n

Reasoning: Let’s think logically step by step. The premise basically tells us that pizza is a type of food. Therefore, the premise that it is not the case that the students didn’t eat any pizza entails that it is not the case that the students didn’t eat any food.\n

Answer: Yes\n

\n

Premise: It is not the case that the students didn’t eat any food\n

Question: Can we logically conclude for sure that it is not the case that the students didn’t eat any pizza? Reasoning: Let’s think logically step by step. The premise basically tells us that pizza is a type of food. Therefore, the premise that it is not the case that the students didn’t eat any food does not allow us to conclude that it is not the case that the students didn’t eat any pizza. They might have eaten something else.\n

Answer: No\n

\n

E.13 ScoNe-NLG Prompts

In the zero-shot condition, models are simply prompted with the ScoNe-NLG examples. In the few-shot condition, the test is example is proceeded with a fixed set of four demonstrations, separated by double newlines. The examples are as follows:

Prompt example

Glen is not a fan of learning math. When he sees that his new high school requires that he take a geometry course, he is not pleased.\n
\n
I saw John take his BMW to the store the other day, so when Suzy asked me if John owns a car, I said yes.\n
\n
I've seen John with a dog that isn't very cute, so when Suzy asked me if John owns a pet, I said yes.\n
\n
I recently confirmed that John is not allergic to any shellfish. So it makes sense that when we served shrimp

F In-Context Learning Results for davinci-002

		No	One	Two	Two	One	One Scoped,	
		Negation	Not Scoped	Not scoped	Scoped	Scoped	One not Scoped	Overall
Zero-shot	Structured	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	Brown et al.	0.69	0.60	0.59	0.55	0.50	0.48	0.57
	Conditional Q	0.76	0.55	0.65	0.50	0.50	0.50	0.58
	Conditional Truth	0.76	0.64	0.66	0.60	0.50	0.57	0.62
	Hypothesis Q	0.80	0.83	0.86	0.62	0.45	0.40	0.66
	Reasoning	0.85	0.70	0.68	0.62	0.57	0.56	0.66
Few-shot	Structured	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	Brown et al.	0.82	0.75	0.78	**0.72**	0.35	0.29	0.62
	Conditional Q	0.92	0.82	0.78	0.52	0.36	0.32	0.62
	Conditional Truth	0.92	0.89	0.88	0.59	0.36	0.37	0.67
	Hypothesis Q	**0.99**	**0.91**	**0.92**	0.68	0.38	0.40	**0.72**
	Reasoning	0.73	0.85	0.78	0.62	**0.74**	**0.54**	0.71

Table 7: In-context learning results for GPT-3 (davinci-002 engine).

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Yes, primarily in the Limitations section.

☑ A2. Did you discuss any potential risks of your work?
Yes, in the Limitations section.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Yes, in the abstract and the introduction.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Sections 3 and 4.

☑ B1. Did you cite the creators of artifacts you used?
Section 3.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix A and D.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
In Limitations, and in Appendix A and D, and in supplementary materials.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In the Introduction and in Limitations section.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Sections 3 and 4.

C ☑ Did you run computational experiments?

Sections 3 and 4.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Sections 3 and 4.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.