Language models are not naysayers: An analysis of language models on negation benchmarks

Thinh Hung Truong[1] **Timothy Baldwin**[1,3] **Karin Verspoor**[2,1] **Trevor Cohn**[1,*]
[1]University of Melbourne [2]RMIT University [3]MBZUAI
hungthinht@student.unimelb.edu.au tb@ldwin.net
karin.verspoor@rmit.edu.au trevor.cohn@unimelb.edu.au

Abstract

Negation has been shown to be a major bottleneck for masked language models, such as BERT. However, whether this finding still holds for larger-sized auto-regressive language models ("LLMs") has not been studied comprehensively. With the ever-increasing volume of research and applications of LLMs, we take a step back to evaluate the ability of current-generation LLMs to handle negation, a fundamental linguistic phenomenon that is central to language understanding. We evaluate different LLMs — including the open-source GPT-neo, GPT-3, and InstructGPT — against a wide range of negation benchmarks. Through systematic experimentation with varying model sizes and prompts, we show that LLMs have several limitations including insensitivity to the presence of negation, an inability to capture the lexical semantics of negation, and a failure to reason under negation.

1 Introduction

Despite being a core linguistic phenomenon, negation remains a major stumbling block for modern NLP architectures (Kassner and Schütze, 2020; Hossain et al., 2022). A reason for this could be that texts containing negation are underrepresented in training data of language models, as humans tend to express themselves using affirmative rather than negative expressions (Ettinger, 2020). Regardless, negation has been shown to be challenging even for humans to correctly interpret due to the diversity of forms across domains (Truong et al., 2022a). For instance, in clinical documents, many acronyms are used to denote negation such as *NAD (no abnormality detected)*, and implicit negation abounds, such as *normal chest x-ray scan*, which implies the absence of an abnormality. Even more complex is the use of negation in combination with other linguistic phenomena such as quantifiers, gradable adjectives (*not unattractive* does not imply *attractive*) (Truong et al., 2022b); licensing context (negative polarity items, e.g. *any, either, yet*, normally appear in certain negative grammatical contexts Warstadt et al. (2019)); downward entailment (*A man owns a dog* entails *A man owns an animal* but *A man does not own a dog* does not entail *A man does not own an animal*) (Geiger et al., 2020).

Traditionally, negation has been treated as a standalone problem, e.g. as negation detection (Chapman et al., 2001). The investigation of the impact of negation in various downstream tasks (Hossain et al., 2022; Hossain and Blanco, 2022a), or through probing (Ettinger, 2020) has revealed several limitations of modern large language models ("LLMs") in handling negation. Given that LLMs are being adopted in an ever-growing range of tasks and have been shown to display emergent abilities for high-level tasks that require complex reasoning (Wei et al., 2022a), we are interested in exploring how the handling of negation has progressed.

In this work, we investigate the performance of auto-regressive language models on different negation-focused benchmarks. Instead of just looking at samples containing negation in common NLP datasets, we consider datasets in which negation plays an important role in making the correct judgement. In particular, we classify the benchmarks into three categories corresponding to the requisite negation reasoning abilities: (1) sensitivity to negation through cloze completion (fill-in-the-blank) queries of factual statements; (2) lexical semantics of negation through classification of antonym/synonym relationships; and (3) ability to reason with negation through language inference tasks.

We conduct extensive experiments using prompt-based learning to facilitate zero- and few-shot evaluation of LLMs, and find the following:

- larger LMs are more insensitive to negation compared to smaller ones (Section 3);

[*]Now at Google DeepMind.

*Proceedings of the The 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 101–114
July 13-14, 2023 ©2023 Association for Computational Linguistics

Benchmark	Task	# Samples	Example
MKR-NQ	Completion	3360	Query: *Iburofen isn't a kind of [MASK].* Wrong completions: *NSAID, painkiller, drug, medicine.*
MWR	Completion	27546	Query: *Demand is an antonym of [MASK].* Wrong completions: *necessitate, demands, request, requirement, imposition, need, demand.*
SAR	NLI	2000	Word 1: *Superiority* / Word 2: *Inferiority* / Label: Antonym
NegNLI	NLI	4500	P: *They watched me constantly for weeks.* / H: *They did not leave me on my own for weeks.* / Label: Entailment
NaN-NLI	NLI	258	P: *Not all people have had the opportunities you have had.* / H: *Some people have not had the opportunities you have had.* / Label: Entailment
MoNLI	NLI	200	P: *The man does not own a dog.* / H: *The man does not own a mammal.* / Label: Not Entailment

Table 1: Summary of the negation-related benchmark datasets used in this paper.

- LLMs lack lexical semantic knowledge about negation, yielding almost random performance for synonym/antonym classification (Section 3);
- LLMs have limited ability to reason under negation, performing worse than random across most NLI datasets (Section 3). Only with the latest instruction fine-tuned model (Ouyang et al., 2022; Chung et al., 2022) do we observe above-chance performance (Section 3);
- For each dataset, we also experiment with prompt variations and find that in most cases, providing more information (context, instruction, simple wording) leads to a degradation in performance.

2 Experimental settings

In this section, we outline the settings that , including benchmark datasets, models to evaluate, and the prompts that were used. Our code is available at `https://github.com/joey234/llm-neg-bench`.

2.1 Benchmarks

We use a range of benchmark datasets that exhibit the effects of negation across a wide range of tasks, in the form of either cloze completion or classification tasks. An overview of the datasets is presented in Table 1, categorized according to purpose and the type of negation they contain. Specifically, we focus on: (1) investigating whether LLMs are sensitive to the presence of negation in factual statements; (2) testing whether LLMs capture negation in lexical semantics relations (synonym/antonym relations); and (3) investigating whether LLMs are able to reason under negation through multiple natural language inference benchmarks. We discuss the datasets in greater detail in Section 3.

2.2 Models

For the LLMs, we primarily focus on open-source auto-regressive LLMs with up to 6.7B parameters, including GPT-Neo (Black et al., 2021), and OPT (Zhang et al., 2022), which are claimed to be comparable in performance to similar-sized GPT-3 class models. Architecture-wise, they are both decoder-only PLMs pre-trained with a causal LM objective, with the main difference being in their pre-training corpora: GPT-neo was trained solely on the Pile dataset (Gao et al., 2020) consisting of 22 sub-datasets spanning different sources, whereas OPT was trained on the combination of datasets used in RoBERTa (Liu et al., 2019), Pile, and PushShift Reddit (Baumgartner et al., 2020). We use the official model checkpoints from HuggingFace hub,[1] as detailed in Appendix A. We experiment with smaller-sized variants of these two classes of models to observe the effect of scaling on their performance over different benchmarks.

We also consider base GPT-3 (175B) (Brown et al., 2020), and its instruction fine-tuned variant InstructGPT (Ouyang et al., 2022), as well as a strong open-source instruction-tuned model FLAN-T5-XXL (11B) (Chung et al., 2022), to examine how recent commercial LLMs perform on negation.

[1] `https://huggingface.co/models`

Task	Prompt name	Example
MKR-NQ	Default	An expectorant isn't a type of ___
	Contrasting	**An expectorant is a type of medicine.** An expectorant isn't a type of ___
	Discourse	**An expectorant is a type of medicine. Therefore,** an expectorant isn't a type of ___
	Mask	An **[MASK]** is a type of medicine. An **[MASK]** isn't a type of ___
MWR	Default	Greed is an antonym of ___
	Quote	**The word "**greed**"** is an antonym of **the word "** ___
SAR	Default	Choose the correct answer: bad and good are antonyms or synonyms? Answer: ___
	Simple	**Choose the correct answer:** bad and good are **opposite** or **similar**? **Answer:** ___
	Negation	**Antonyms are words with opposite meaning. Synonyms are words with similar meaning. Choose the correct answer:** bad and good are antonyms or synonyms? **Answer:** ___
NLI	Default	Not all people have had the opportunities you have had. **Question:** Some people have not had the opportunities you have had. True, False, or Neither? **Answer:** ___
	Negation	**The question requires reasoning about negation.** Not all people have had the opportunities you have had. **Question:** Some people have not had the opportunities you have had. True, False, or Neither? **Answer:** ___

Table 2: Prompts used for each task

2.3 Prompts

We adopt prompt-based learning to enable zero- and few-shot evaluation of LLMs (Radford et al., 2019). Given that LLMs have been found to be sensitive to prompt variation (Wei et al., 2021), and that more natural-sounding prompts correlate with model performance (Gonen et al., 2022), we also experiment with different types of prompts (see Table 2).

We use GPT-3 style prompts (Brown et al., 2020) as the *Default* setting. As handling negation plays an important role in all tasks, we also design prompts to prime the LLMs to focus more on the negation context, by introducing modifications specific to each task. In detail, for the cloze completion task MKR-NQ, we investigate whether a given model can detect the difference between two contrasting sentences (with/without negation). To achieve this, we prepend the prompt with the corresponding sentence without negation (*Contrasting* prompt). In addition, we also evaluate alternative prompts where we connect the two sentences with a discourse marker (*Discourse* prompt), or mask the main subject to encourage the model to attend more to negation cues (*Mask* prompt).

For antonym/synonym-related tasks (MWR, SAR), we also experiment with simplifying the prompt and use descriptive terms rather than the formal names of the relations (e.g. *antonyms, synonyms* $\rightarrow$ *opposite of, similar to*), based on the intuition that these terms will appear more frequently in the pre-training data.

Finally, for classification tasks, we propose negation-aware prompting (*Negation* prompt) by modifying the prompts to explicitly state that the task involves reasoning about negation. Note that we explicitly include class options in the prompts to help reduce the effect of the surface form competition causing LLMs to assign lower probabilities to the correct answers (Holtzman et al., 2021).

For datasets with an accompanying training set (SAR, MoNLI), we also experiment with few-shot evaluation formulated as *in-context learning* by prepending the input prompts with 10 random samples from the training set.

2.4 Metrics

To evaluate cloze completion tasks, we employ *Weighted Hit Rate (WHR)* (Jang et al., 2022b), which measures the number of matches between the top-k predicted words and a given set of target wrong predictions, taking into account the prediction probabilities:

$$WHR_k(x, W) = \frac{\sum_{i=1}^{k} c_i \times \mathbb{1}(w_i \in W_x)}{\sum_{i=1}^{k} c_i} \quad (1)$$

where W_x is the wrong prediction set of the input query x, and w_i is the top i-th prediction with confidence score c_i, obtained by taking the softmax of log probabilities $p(w_i|x)$ from the LM. Note that the model performance is better if there are fewer matches between models' predictions and wrong completions, *WHR* is an error metric (lower is better). One problem with the *WHR* metric is that we can only evaluate using a fixed set of wrong predictions. Regardless, we believe the relative performance numbers are indicative of model performance.

For classification tasks, we evaluate using *Accuracy*, noting that all datasets are reasonably balanced.

3 Main findings

We summarize the main findings in this section. In general, the performance of GPT-neo and OPT follows a similar trend across all benchmarks (we present GPT-neo results; results of OPT models are in Appendix B).

Finding 1: Larger LMs are more insensitive to negation

MKR-NQ (Jang et al., 2022b) Masked Knowledge Retrieval – Negated Query (MKR-NQ) is a negated version of the LAMA dataset (Petroni et al., 2019), which contains lexicalized statements of triples in ConceptNet (Speer et al., 2017). This dataset contains factual statements with verbal negations (i.e. negators *not, don't* are associated with the main verb of the sentence), e.g. *Iburofen is a type of medicine.* $\rightarrow$ *Iburofen isn't a type of medicine*.

Each sample contains the query along with a set of wrong word completions, supporting the evaluation of the sensitivity of the model to negation by measuring how likely a model will generate incorrect completions. Note that MKR-NQ only considers sample sentences that contain a single verb, making it trivial to negate the original sentences.

Findings From Figure 1, which is based on LLMs with a negated factual statement (*Default* prompt), we observe relatively low hit rates ($<$ 0.15) across all model sizes, and a clear inverse scaling trend between model sizes and their performance. The smallest variant (`GPT-neo-125M`) has the best performance, which is comparable to that of masked language model of a similar size (`BERT-base`, 110M parameters) (Jang et al., 2022b). This phenomenon reflects the finding that larger models tend to memorize the training data more (McKenzie et al., 2022; Jang et al., 2022a). Moreover, higher hit rates for top-1 predictions suggest that models predict wrongly with high confidence.

For *Contrasting* prompts, in which we prepend the negated statement with its non-negated version, we notice a drastic increase in *WHR*, showing that models are prone to repeating what is presented in the prior context, confirming the finding of Kassner and Schütze (2020). When a discourse term is added to connect the two sentences (*Discourse* prompt), we do not observe any improvement, and the performance of the largest model is even worse. To investigate whether this phenomenon is attributable to models not being able to detect the presence/absence of negation, we experiment with masking out the main noun/verb of the queries (*Mask* prompt). We observed even higher *WHR*, especially for the top-1 prediction in this setting. The results suggest that repetitions are caused more by LLMs being easily primed by repeating what is present in the previous context, than by generating words that are closely associated with the main subject of interest. This again shows that the models cannot differentiate between identical contexts, differing only on whether negation is present or absent (i.e., outputs tend to be similar with or without negation).

To further analyze the outputs, we calculate the perplexity (PPL) of the generated predictions to determine their plausibility (Wilcox et al., 2020). Here, we choose the model with the best *WHR*$_5$ score on the MKR-NQ benchmark, and calculate the mean perplexity over all queries for each prompt type (5 completions for each query). PPL is calculated as the exponentiated average negative log-likelihood of a sequence, with exponent base e. As a point of reference, we calculated the average perplexity of the provided completion of the original non-negated dataset (denoted *Corpus*). From the reported perplexities (Table 3), we can see that *Default* output are the most plausible (with PPL markedly lower than *Corpus*), while *Contrast-*

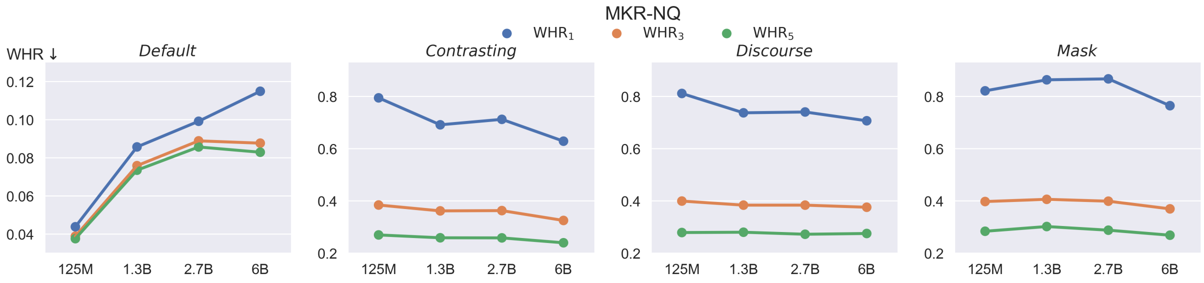

Figure 1: Zero-shot performance of GPT-neo on MKR-NQ using different prompts under the Weighted Hit Rate (WHR) metrics (lower scores are better). Note the different scale for the left-most plot.

Setting	Example	Mean PPL↓
Corpus	[Baseball is a type of sport.]	434.42
Default	[Baseball isn't a type of sport.]	288.94
Contrasting	Baseball is a type of sport. [Baseball isn't a type of sport.]	533.56
Discourse	Baseball is a type of sport. Therefore, [baseball isn't a type of sport.]	477.44
Mask	MASK is a type of sport. [MASK isn't a type of sport.]	448.23

Table 3: Mean perplexity (PPL) calculated using the `GPT-J-6B` model. Only the strings enclosed in square brackets are considered during calculation in order to provide a fair comparison with similar token length. For Corpus, PPL is calculated using the provided gold completion.

ing is the least natural. The remaining prompts types (*Discourse, Mask*) are comparable to *Corpus*. These results show that LLMs can indeed generate plausible and human-like output for this task.

Finding 2: LMs fail to capture synonym/antonym lexical relations

MWR (Jang et al., 2022b) To test the ability of LMs to capture negative lexical semantics, we use MWR dataset, where models are asked to predict the antonym/synonym of a target word. The dataset was constructed by using the most frequent nouns, adjectives, and adverbs that appear in SNLI (Bowman et al., 2015), then choosing their corresponding synonyms and antonyms from ConceptNet (Speer et al., 2017). The dataset also contains different wordings for antonym-asking and synonym-asking queries (e.g. *is the opposite of, is different from* and *is similar to, is a rephrasing of*) to test model sensitivity to prompt variations.

Findings From Figure 2, we can observe the same inverse scaling trend as for MKR-NQ using

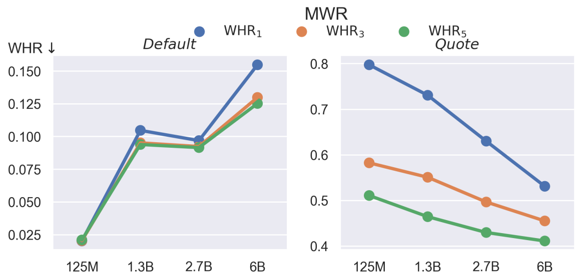

Figure 2: Zero-shot performance of GPT-neo on MWR using different prompts (WHR metrics; lower is better)

Query	Wrong completions	Top-5 predictions
Greed is an antonym of	*greed, avarice, desire, greeds, gluttony*	*altruism, self-sacrifice, self-denial, self-abnegation,* ***gods***
Finale is an antonym of	*conclusion, finish, finales, finale*	*last,* ***epiphany****, finality,* ***anti-climax****,* ***anti-climactic***

Table 4: Example output of `GPT-J-6B` on MWR. **bolded** words are related to target words, but are neither synonyms nor antonyms. underlined are wrong antonyms but are not in the given set of wrong completions.

the *Default* prompt, where the hit rate of the smallest model is around 0.02, better than previously-reported SOTA results (Jang et al., 2022b). With a more natural query with more focus on the target words via quotation marks (*Quote* prompt), surprisingly, we noticed a drastic jump in hit rates. However, MWR may not be a good indicator of model performance due to how the task is framed. One problem is that models can generate words that are not in the given wrong prediction set, but are also irrelevant, and are also neither antonyms nor synonyms of the given target words, as demonstrated in Table 4.

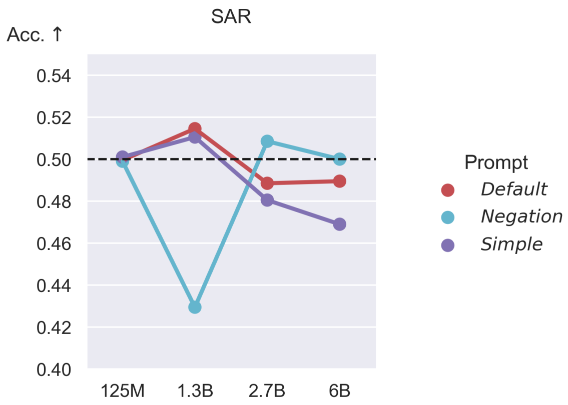

Figure 3: Zero-shot performance of GPT-neo on SAR dataset using different prompts (accuracy metric; higher is better)

SAR (Jang et al., 2022b) To further investigate the ability of LLMs to capture negative lexical semantics, we consider the antonym/synonym relation classification task (SAR). Different from the MWR cloze-style synonym/antonym prediction task, this benchmark is framed as a binary classification task of predicting the correct antonym or synonym relationship between two given words. Data is once again taken from ConceptNet, where triplets with synonym and antonym relations are extracted in equal numbers (1000 samples for each relation).

Findings In contrast to the high results for MWR, we find that for this task, model performance is equivalent to random, with accuracy fluctuating around 0.5 (Figure 3). For prompt variants, we do not observe any meaningful improvement, in that *Simple* follows a similar trend to *Default* and *Negation* performs better for larger models (2.7B and 6B). This is a huge degradation from previous fully fine-tuned results over encoder models. For instance, Jang et al. (2022b) reported that $\text{BERT}_{\text{large}}$ achieves 92.5% accuracy on SAR. We argue that this is a specific task that is not captured in the next token prediction training objective of LLMs and thus, requires explicit supervision.

Finding 3: LLMs are unable to reason under negation

NegNLI (Hossain et al., 2020) NegNLI contains 4500 premise–hypothesis pairs with *important* negation, where negation is essential in making the correct judgement about the relationship between the premise–hypothesis pairs. Samples are extracted from the commonly-used NLI datasets (RTE Dagan et al. (2005), SNLI Bowman et al. (2015), MNLI Williams et al. (2018)), then the negator *not* is added to the main verb either in the premise, hypothesis, or both. Here, we consider each subset separately, as the number of classes are not the same, and denote them SNLI-neg, MNLI-neg, RTE-neg.

MoNLI (Geiger et al., 2020) MoNLI is an NLI dataset focused on lexical entailment and negation. Specifically, the dataset investigates the downward monotonicity property where negation reverses entailment relations (e.g. *dance* entails *move*, but *not move* entails *not dance*). MoNLI was created by extending samples from SNLI by substituting the nouns by their hypernyms/hyponyms from WordNet (Miller, 1998).

NaN-NLI (Truong et al., 2022b) NaN-NLI is a test suite which focuses on sub-clausal negation, in which only part of the sentence's meaning is negated, thus making it harder to correctly determine the correct negation scope (e.g. in *Not the first time that they pulled that off* the negation scope is only *Not the first time* and the main clause of the sentence *they pulled that off* is not negated). Each premise–hypothesis pair is constructed so that the corresponding hypotheses are constructed to reflect different interpretations that the negated instance in the premise are likely to be misunderstood for.

Findings Similar to the antonym/synonym classification task, the performance for most negation-focused NLI benchmarks is low. In particular, for all NLI datasets, the performance is generally lower than baseline. As shown in Figure 4, scaling up model size has almost no effect, and the largest model performs worse in many cases, even when the prompt explicitly states that the task requires reasoning about negation (*Negation* prompt). For datasets which include a training set (SAR, MoNLI), we also experimented with few-shot learning but did not observe any noticable improvement (Figure 5). One exception is that the 2.7B model seems to pick up some signal from the provided MoNLI training samples, but falls back again when we increase the model size to 6B.

Even with general NLI datasets, zero-shot applications of LLMs were previously shown to be roughly equivalent to a random baseline (Wei et al., 2021). When negation is involved, the task becomes even more complex. As pointed out in Brown et al. (2020), one possible reason that LLMs

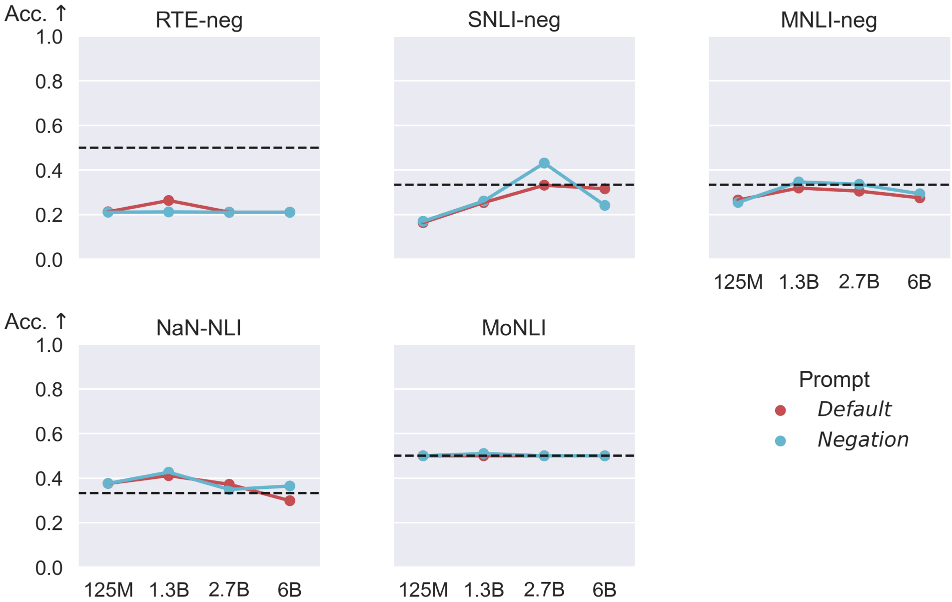

Figure 4: Zero-shot performance of GPT-neo on NLI datsets using different prompts (higher is better). The dashed line denotes a random baseline. Note that RTE-neg and MoNLI are 2-way classification tasks while the rest are 3-way.

Benchmark		GPT-J-6B	GPT-3	InstructGPT	InstructGPT w/ *Neg.* prompt	FLAN-T5-XXL w/ *Neg.* prompt
MKR-NQ	*WHRs* ↓	**0.083**	0.172	0.195	NA	NA
MWR	*WHRs* ↓	**0.125**	0.488	0.504	NA	NA
SAR	*Accuracy* ↑	0.490	0.501	0.687	**0.780**	0.507
SNLI-neg	*Accuracy* ↑	0.316	0.267	0.640	**0.673**	0.477
MNLI-neg	*Accuracy* ↑	0.275	0.359	0.548	**0.625**	0.354
RTE-neg	*Accuracy* ↑	0.211	0.525	0.767	**0.807**	0.770
NaN-NLI	*Accuracy* ↑	0.298	0.469	0.647	**0.682**	0.376
MoNLI	*Accuracy* ↑	0.500	**0.540**	0.470	0.400	0.500

Table 5: Zero-shot results on the different benchmarks. “NA” denotes that *Negation* prompts are not applicable to MKR-NQ and MWR. The best results are bolded for each task (row).

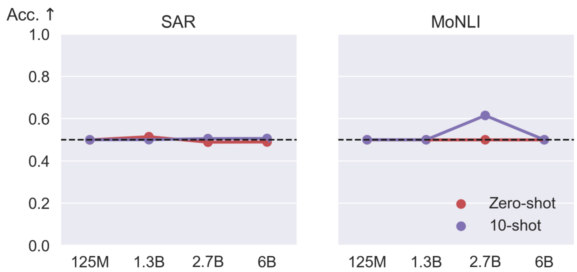

Figure 5: 10-shot performance of GPT-neo on SAR and MoNLI using *Default* prompt (higher is better)

struggle with NLI is that the samples consist of two disjoint sentences, which are unlikely to appear naturally in standard training corpora. We hypothesise that NLI is a generally hard task that requires substantially more supervision in order for models to detect meaningful patterns.

Finding 4: Instruction fine-tuning improves reasoning under negation

We further evaluate with GPT-3 class models of significantly larger scale (175B), which have been shown to achieve strong results in zero- and few-shot settings across a wide range of tasks (Brown et al., 2020). In detail, we benchmark the largest GPT-3 model (`text-davinci-001`: Brown et al. (2020)) and its variant InstructGPT, which is trained to follow human instructions using reinforcement learning (`text-davinci-003`: Ouyang et al. (2022)). The results can be found in Table 5.

For the base GPT-3 model, the results over most benchmarks are no better than much smaller language models (`GPT-neo-125M`). For cloze-completion tasks, consistent with the earlier-

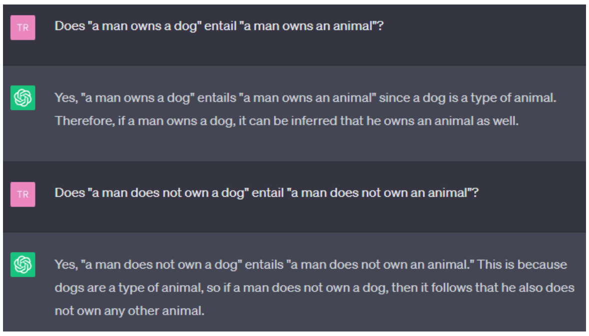

Figure 6: A ChatGPT-generated output of a failed negative monotonicity reasoning sample. The output was generated using ChatGPT Feb 13 Version

observed trend of larger models performing worse, we observe higher (worse) *WHR* scores compared to that of smaller language models, confirming our finding that larger models are more *in*sensitive to the presence of negation. Results get even worse with using the instruction fine-tuned model.

On the other hand, for most classification tasks, InstructGPT achieves better zero-shot results than other models. In addition, using this model in combination with explicit instruction about negation (*Negation* prompt) further improves performance, which we did not observe for other LLMs. It is, however, unclear what data the instruction-tuning process was performed on. Thus, the huge gain in performance could be attributed to the existence of similar patterns in the training set (i.e. explicit supervision over similar tasks). Interestingly, InstructGPT performance on MoNLI did not increase (it underperfomed other models). We hypothesize that this is due to an inductive bias from model's ability to reason with hypernymy. For instance, the model can understand that "*dog* is an *animal*" (and therefore *own an animal* entails *own a dog*), but incorrectly generalizes this logic to a similar sample containing negation (*not own a dog* entails *not own an animal*). This is indeed true when we look at the explanation generated by ChatGPT, the subsequent model to InstructGPT (Figure 6).

We also experiment with the instruction-tuned FLAN-T5-XXL model (Chung et al., 2022) and find that the results are better than GPT-3 for most NLI tasks, despite being ∼16x smaller. These results suggest that instruction fine-tuning has much greater impact than model scaling in terms of models developing the ability to perform reasoning tasks under negation.

4 Related work

Our work builds upon previous research on negation. In particular, we were inspired by the pioneering works of Kassner and Schütze (2020) and Ettinger (2020), which reveal that pre-trained language models have a major issue in being insensitive to the presence of negation, based on evaluation over a set of cloze-style queries. Following this line of research, Jang et al. (2022b) also explored negation in a cloze completion context by negating factual statements extracted from ConceptNet and come to a similar finding.

In a broader context, Hossain et al. (2020, 2022) investigated the performance of BERT-based methods on samples containing negation in the GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) datasets. Their main finding is that the results for the subsets containing only negation are lower than those without, as well as the whole test set, showing that models struggle with negation, even when fine-tuned on relevant training data. Ravichander et al. (2022) proposed the challenging CONDAQA dataset to test the ability of models to reason about the implications of negation. The authors conducted comprehensive analysis of different types of LLMs under different settings, and found that the best-performing models were still well below human performance. Negation has also been investigated as part of psycholinguistic probing datasets (Lialin et al., 2022; Jumelet et al., 2021; Staliūnaitė and Iacobacci, 2020). Contrasting previous finding, Gubelmann and Handschuh (2022) found that the ability to understand negation of LMs is underestimated in previous studied. Through designing a controlled dataset with minimal pairs varying in syntactic structure, gender, profession, and first name, they concluded that the models are indeed sensitive to negation and thus, their struggle comes more from the contextualization of the tasks.

As part of the analysis on emergent abilities of LMs, negation has been shown to be one of the tasks that displays a flat scaling curve (Wei et al., 2022a) or even inverse-scaling (McKenzie et al., 2022). This behaviour was later shown to be alleviated by instruction fine-tuning (Wei et al., 2022b). The effectiveness of instruction fine-tuning is further supported in Jang and Lukasiewicz (2023). The authors investigated the logical consistency of ChatGPT and found that ChatGPT understands negation and antonyms much better than previous models.

Beside probing and evaluation, there have also been works on making language models more robust to negation, including unlikelihood training (Hosseini et al., 2021), adaptive pre-training on relevant data (Truong et al., 2022a), leveraging affirmative interpretations from negation (Hossain and Blanco, 2022b), and learning better representation of negation through contrastive learning (Jiang et al., 2022; Wang et al., 2022).

5 Conclusion

We have shown that LLMs still struggle with different negation benchmarks through zero- and few-shot evaluations, implying that negation is not properly captured through the current pre-training objectives. With the promising results from instruction-tuning, we can see that rather than just scaling up model size, new training paradigms are essential to achieve better linguistic competency. Through this investigation, we also encourage the research community to focus more on investigating other fundamental language phenomena, such as quantification, hedging, lexical relations, and downward entailment.

6 Limitations

First, regarding the experimental settings, the *WHR* metrics used to evaluate cloze completion tasks are imperfect, as we discussed. Framing cloze completion tasks in the style of multiple-choice question answering to limit the options that models are evaluated on would be a good direction to follow (Robinson et al., 2022). In addition, the prompt engineering in this work is in no way exhaustive, and could be extended using different prompt engineering strategies such as soft prompt tuning (Lester et al., 2021), or mining- and paraphrasing-based methods to generate high quality prompts (Jiang et al., 2020).

Second, due to computational constraints, we could not perform an extensive set of experiments for larger models like PaLM (with up to 540B parameters) (Chowdhery et al., 2022). Recent work by Wei et al. (2022b) has shown that the inverse scaling trend on several benchmarks can be alleviated using the large instruction fine-tuned models such as FLAN-PaLM-540B, which is largely in line with our findings regarding InstructGPT and FLAN-T5. With a small-scale experiment, we found that ChatGPT displayed strong performance on challenging samples in the investigated benchmark, so the main findings of the paper may not hold true for newer LLMs.

Finally, this work only considers negation in the English language. There is every reason to believe that negation is an equally challenging problem in other languages. As this is a linguistically-intensive task, and requires native speakers to conduct thorough analysis of the results, we leave this for future work.

Acknowledgement

The authors would like to thank the anonymous reviewers for their detailed, kind, and constructive reviews. This research was conducted by the Australian Research Council Training Centre in Cognitive Computing for Medical Technologies (project number ICI70200030) and funded by the Australian Government. This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200.

References

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The PushShift Reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

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, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The Pascal Recognising Textual Entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

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.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

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.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. 2022. Demystifying prompts in language models via perplexity estimation. *arXiv preprint arXiv:2212.04037*.

Reto Gubelmann and Siegfried Handschuh. 2022. Context matters: A pragmatic study of PLMs' negation understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4602–4621, Dublin, Ireland. Association for Computational Linguistics.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Md Mosharaf Hossain and Eduardo Blanco. 2022a. Leveraging affirmative interpretations from negation improves natural language understanding. *arXiv preprint arXiv:2210.14486*.

Md Mosharaf Hossain and Eduardo Blanco. 2022b. Leveraging affirmative interpretations from negation improves natural language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5833–5847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. 2021. Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1301–1312, Online. Association for Computational Linguistics.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2022a. Can large language models truly understand prompts? a case study with negated prompts. *arXiv preprint arXiv:2209.12711*.

Myeongjun Jang and Thomas Lukasiewicz. 2023. Consistency analysis of chatgpt. *arXiv preprint arXiv:2303.06273*.

Myeongjun Jang, Frank Mtumbuka, and Thomas Lukasiewicz. 2022b. Beyond distributional hypothesis: Let language models learn meaning-text correspondence. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2030–2042, Seattle, United States. Association for Computational Linguistics.

Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. 2022. PromptBERT: Improving BERT sentence embeddings with prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8826–8837, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language models use monotonicity to assess NPI licensing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4958–4969, Online. Association for Computational Linguistics.

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.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vladislav Lialin, Kevin Zhao, Namrata Shivagunde, and Anna Rumshisky. 2022. Life after BERT: What do other muppets understand about language? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3180–3193, Dublin, Ireland. Association for Computational Linguistics.

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*.

Ian McKenzie, Alexander Lyzhov, Alicia Parrish, Ameya Prabhu, Aaron Mueller, Najoung Kim, Sam Bowman, and Ethan Perez. 2022. Inverse scaling prize: Round 1 winners.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Abhilasha Ravichander, Matt Gardner, and Ana Marasović. 2022. CONDAQA: A contrastive reading comprehension dataset for reasoning about negation. *arXiv preprint arXiv:2211.00295*.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2022. Leveraging large language models for multiple choice question answering. *arXiv preprint arXiv:2210.12353*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Ieva Staliūnaitė and Ignacio Iacobacci. 2020. Compositional and lexical semantics in RoBERTa, BERT and DistilBERT: A case study on CoQA. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7046–7056, Online. Association for Computational Linguistics.

Thinh Truong, Timothy Baldwin, Trevor Cohn, and Karin Verspoor. 2022a. Improving negation detection with negation-focused pre-training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4188–4193, Seattle, United States. Association for Computational Linguistics.

Thinh Hung Truong, Yulia Otmakhova, Timothy Baldwin, Trevor Cohn, Jey Han Lau, and Karin Verspoor. 2022b. Not another negation benchmark: The NaN-NLI test suite for sub-clausal negation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 883–894, Online only. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Hao Wang, Yangguang Li, Zhen Huang, Yong Dou, Lingpeng Kong, and Jing Shao. 2022. Sncse: Contrastive learning for unsupervised sentence embedding with soft negative samples.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. 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 2877–2887, Hong Kong, China. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Yi Tay, and Quoc V Le. 2022b. Inverse scaling can become U-shaped. *arXiv preprint arXiv:2211.02011*.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, page 1707–1713.

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.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Model checkpoints

For open-sourced LMs, we consider the official released checkpoints on the HuggingFace hub at:

- `https://huggingface.co/EleutherAI/x`
- `https://huggingface.co/facebook/y`

where x in {*gpt-neo-125M,gpt-neo-1.3B,gpt-neo-2.7B,gpt-j-6B*}, and y in {*opt-125m,opt-350m,opt-1.3b,opt-2.7b,opt-6.7b*}.

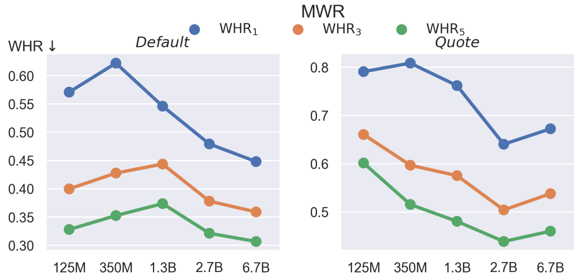

Figure 7: Zero-shot performance of OPT on MWR using different prompts

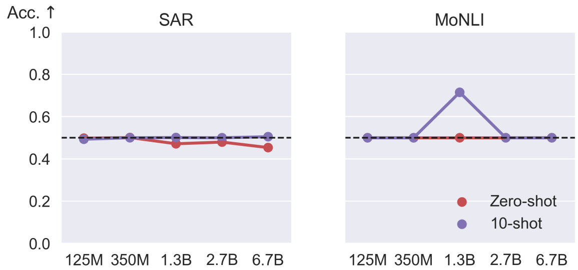

Figure 8: Zero-shot performance of OPT on SAR using different prompts

For GPT-3 models, we access them through the official API at `https://openai.com/api/`, using the *Text completion* endpoint. The considered model identifiers along with their sizes are:

- `text-ada-001`: 350M
- `text-babbage-001`: 1.3B
- `text-curie-001`: 6.7B
- `text-davinci-001`: 175B
- `text-davinci-003`: 175B

B OPT results

For MWR, although we observe improvements with increasing model sizes, the WHR scores are much higher than those of GPT-neo, showing that OPT is worse at predicting antonyms and synonyms of words. The gap in performance may lie in differences in training data between the two types of models.

C Model outputs

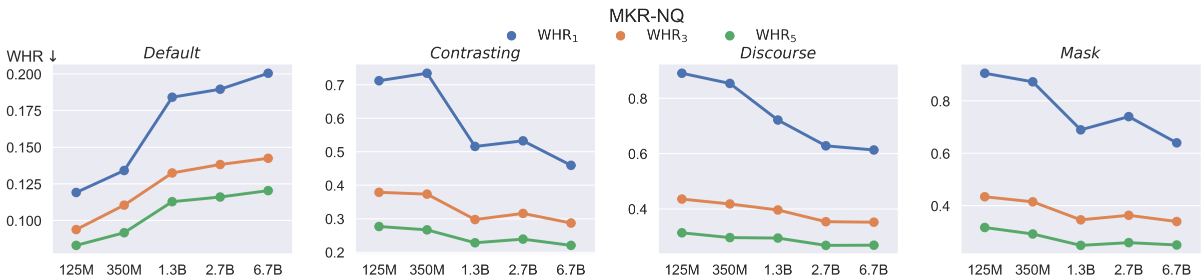

Figure 9: Zero-shot performance of OPT on MKR using different prompts

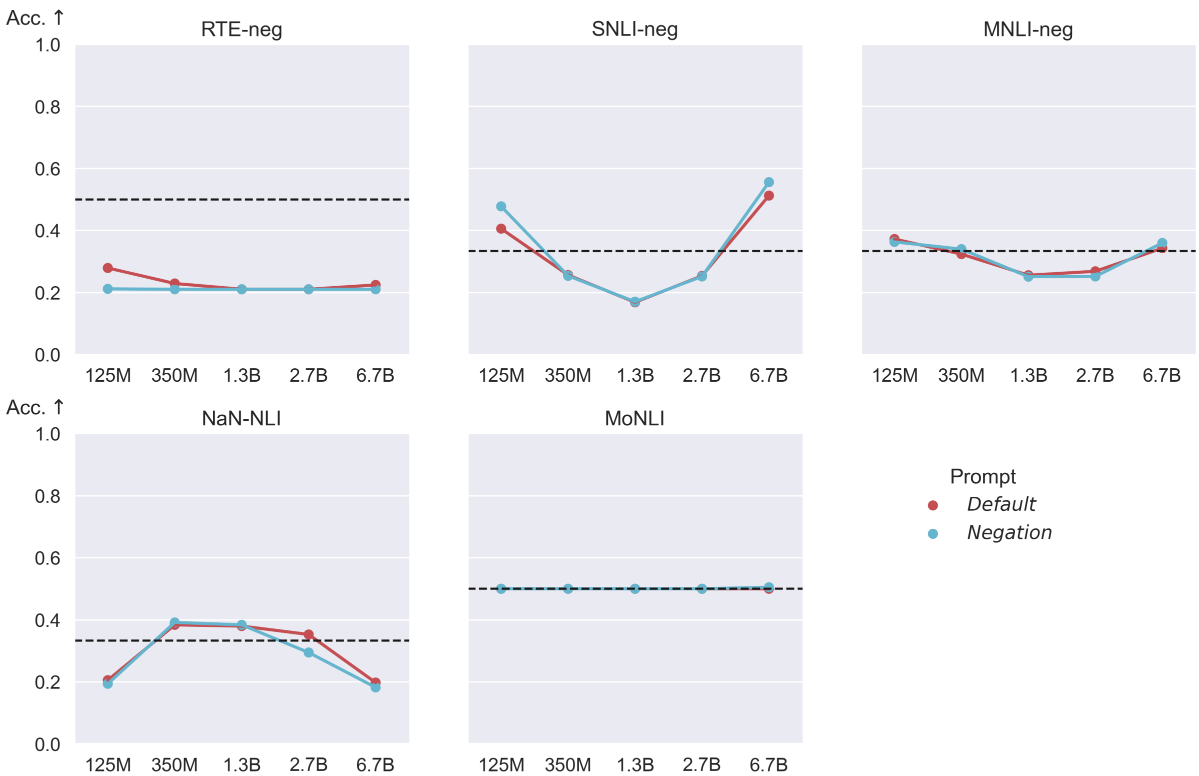

Figure 10: Zero-shot performance of OPT on NLI tasks using different prompts

Prompt	Model	Output
Paracetamol isn't a kind of	GPT-neo-125M	*muscle*
	GPT-J-6B	painkiller
	OPT-125M	pain
	OPT-6.7B	medicine
	GPT-3	medication
	InstructGPT	NSAID
Entrance is an antonym of	GPT-neo-125M	*interest*
	GPT-J-6B	entrance
	OPT-125M	entrance
	OPT-6.7B	exit
	GPT-3	departure
	InstructGPT	entrance
Choose the correct answer: flimsy and sturdy are synonyms or antonyms?	GPT-neo-125M	Synonyms
	GPT-J-6B	Synonyms
	OPT-125M	Antonyms
	OPT-6.7B	Synonyms
	GPT-3	Antonyms
	InstructGPT	Antonyms
I can not think of a few reasons for the allergy to substance. Question: There are not reasons why there's an allergy. True, False, or Neither? Answer:	GPT-neo-125M	True
	GPT-J-6B	True
	OPT-125M	True
	OPT-6.7B	Neither
	GPT-3	False
	InstructGPT	Neither
The man does not own a dog. Question: the man does not own a mammal. True or Not true? Answer:	GPT-neo-125M	True
	GPT-J-6B	True
	OPT-125M	True
	OPT-6.7B	True
	GPT-3	True
	InstructGPT	Not True

Table 6: Example outputs of models. Wrong answers are highlighted