Beyond Positive Scaling:
How Negation Impacts Scaling Trends of Language Models

Yuhui Zhang[*][†] **Michihiro Yasunaga**[*] **Zhengping Zhou**[*] **Jeff Z. HaoChen**[*]
James Zou **Percy Liang** **Serena Yeung**
Department of Computer Science
Stanford University

Abstract

Language models have been shown to exhibit positive scaling, where performance improves as models are scaled up in terms of size, compute, or data. In this work, we introduce NeQA, a dataset consisting of questions with negation in which language models do not exhibit straightforward positive scaling. We show that this task can exhibit inverse scaling, U-shaped scaling, or positive scaling, and the three scaling trends shift in this order as we use more powerful prompting methods or model families. We hypothesize that solving NeQA depends on two subtasks: question answering (task 1) and negation understanding (task 2). We find that task 1 has linear scaling, while task 2 has sigmoid-shaped scaling with an emergent transition point, and composing these two scaling trends yields the final scaling trend of NeQA. Our work reveals and provides a way to analyze the complex scaling trends of language models.

1 Introduction

Language models have been shown to exhibit *positive scaling*, where task performance improves as models are scaled up in terms of size, compute, or data, like the blue curve in Figure 1 (Kaplan et al., 2020; Brown et al., 2020; Rae et al., 2021; Chowdhery et al., 2022; Srivastava et al., 2022; Liang et al., 2022). However, there are exceptions. Recent works show that some tasks can exhibit *inverse scaling* (McKenzie et al., 2022), where the performance degrades as models are scaled up (green curve), or *U-shaped scaling* (Wei et al., 2022b), where the performance degrades first but then improves as models are scaled up (red curve). Analyzing tasks that exhibit different scaling trends, such as inverse and U-shaped scaling, is therefore useful for better understanding the behaviors of language models, identifying their limitations, and guiding future development.

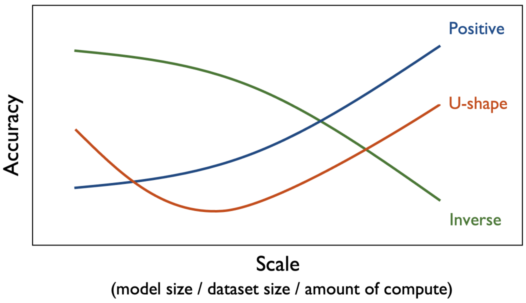


Figure 1: Illustration of three types of scaling trends.

In this work, we introduce NeQA, a new task of answering multiple-choice questions containing negation words, constructed by transforming questions from OBQA (Mihaylov et al., 2018) and NegatedLAMA (Kassner and Schütze, 2020). We conduct experiments on this task using 4 language model families and 3 prompting methods, and show that large language models do not follow straightforward positive scaling on this task. Specifically, as we use more powerful prompting methods or model families, NeQA exhibits a gradation from inverse scaling to U-shape to positive scaling. This result provides a unified view of when the three types of scaling trends (inverse, U-shaped, and positive scaling) occur for language models. Our result indicates that the development of large language models' capability to process negation may be a complex and nuanced problem.

To further understand this nuanced scaling trend of the NeQA task, we decompose the task into two subtasks: question answering (task 1) and negation understanding (task 2). Our empirical results show that task 1 has linearly positive scaling, while task 2 has sigmoid-shaped scaling with an emergent transition point, where the transition point is influenced by the prompt method and model family. Combining these two scaling trends yields the final scaling trend observed in NeQA. The task decomposition provides a new way to think of the scaling on a task in terms of a combination of its component skills.

In summary, our contributions are (1) the NeQA

[*] Equal contributions. [†] Correspondence to: Yuhui Zhang <yuhuiz@stanford.edu>.

Findings of the Association for Computational Linguistics: ACL 2023, pages 7479–7498
July 9-14, 2023 ©2023 Association for Computational Linguistics

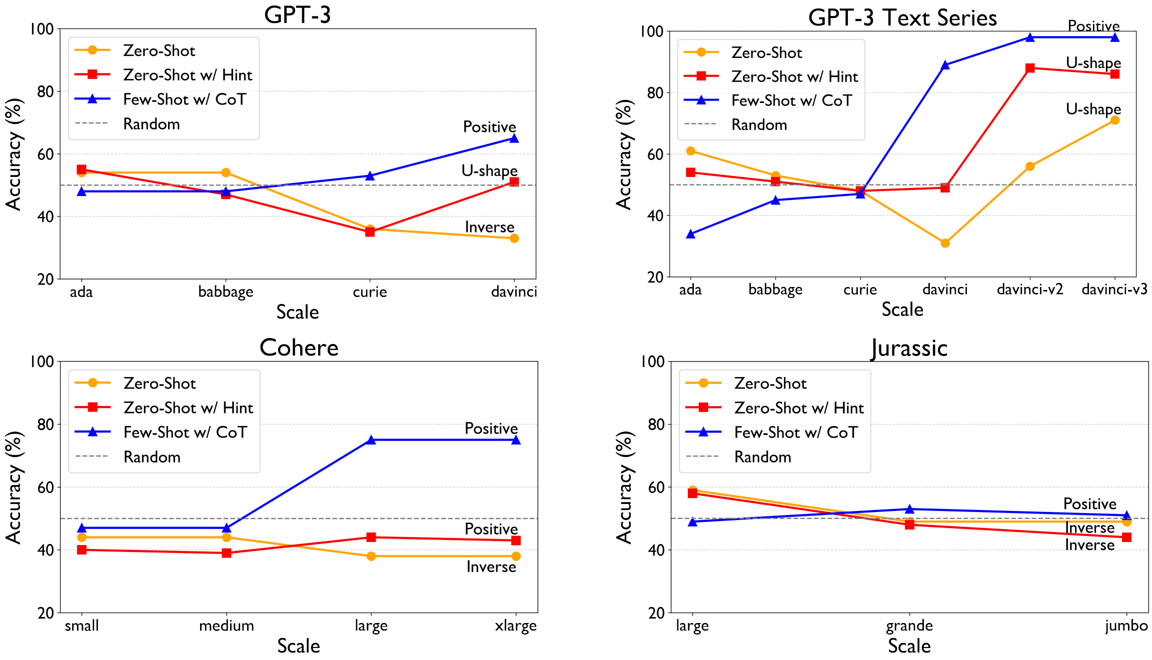


Figure 2: Scaling trends of various language models on the NeQA dataset. As we use more powerful prompting methods or model families, we observe a gradation from inverse scaling to U-shape to positive scaling. More details in §C.

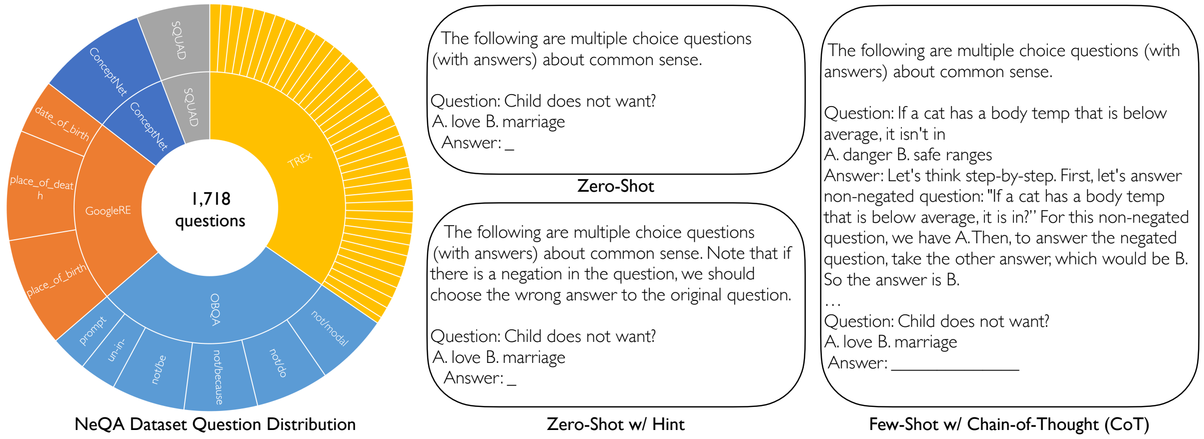
dataset that contains diverse distributions of texts about negation; (2) an evaluation of different large language models on the NeQA dataset, which exhibits different scaling trends; (3) a task decomposition analysis explaining the above scaling trends.

2 Dataset: NeQA

We develop NeQA, a question answering dataset designed to evaluate the ability of models to process negation in natural language. Each example of the dataset consists of a negated question and two answer choices, one correct and one incorrect. An example of NeQA looks like: (question “Child does not want?”, correct choice “marriage”, incorrect choice “love”). To construct this, we leveraged NegatedLAMA (Kassner and Schütze, 2020) and OBQA (Mihaylov et al., 2018).

The NegatedLAMA dataset includes negated questions from four subsets: ConceptNet, GoogleRE, SQuAD, and TREx. Each subset comprises multiple files that represent different question distributions, such as questions about different entity relations. Each question is associated with a negated question, an answer, and a misprimed question (i.e., a wrong answer followed by the question). For instance, when “Child wants?” is the original question, “Child does not want?” can be its associated negated question, “love” can be the answer, and “Marriage? Child wants?” can be its misprimed question. We turn it into a multiple choice question by setting the negated question as the question, and setting the wrong answer in the misprimed question in conjunction with the correct answer as the two choices. For instance, in the above example, we get “Q: Child does not want? A. love B. marriage” (Appendix Table 3). To ensure diversity and representativeness, we randomly selected at most 50 questions from each file.

To be able to analyze the impact of different negation types, we also created additional data by applying diverse rules to transform questions in OBQA (Mihaylov et al., 2018) into negated ones. We defined six types of negation transformations: action verb negation (e.g., “cause” → “does not/doesn’t cause”), linking verb negation (e.g., “is” → “is not/isn’t”), modal verb negation (e.g., “can” → “can not/can’t”), conjunction negation (e.g., “because” → “not because”), negation prefix (e.g., “able” → “unable”), and negation prompt (e.g., add “choose the wrong answer”). For each type, we collected 50 questions by applying a rule-based transformation, sampling an incorrect answer as the correct answer, and treating the correct answer as the incorrect answer. For example, “Pushing on a pedal is an example of” is an original question in OBQA with the correct answer “force” and one of the incorrect answers “speed”. We apply the rule-based transformation to change the verb “is” to “isn’t” and get “Q: Pushing on a pedal isn’t an



NeQA Dataset Question Distribution

The following are multiple choice questions (with answers) about common sense.

Question: Child does not want?
A. love B. marriage
Answer: _

Zero-Shot

The following are multiple choice questions (with answers) about common sense. Note that if there is a negation in the question, we should choose the wrong answer to the original question.

Question: Child does not want?
A. love B. marriage
Answer: _

Zero-Shot w/ Hint

The following are multiple choice questions (with answers) about common sense.

Question: If a cat has a body temp that is below average, it isn't in
A. danger B. safe ranges
Answer: Let's think step-by-step. First, let's answer non-negated question: "If a cat has a body temp that is below average, it is in?" For this non-negated question, we have A. Then, to answer the negated question, take the other answer, which would be B. So the answer is B.
. . .
Question: Child does not want?
A. love B. marriage
Answer: ______________

Few-Shot w/ Chain-of-Thought (CoT)

Figure 3: (Left) Statistics of NeQA data sources. (Right) Three prompting methods that yield shifts in scaling trends.

example of? A. speed. B. force", where "A" is the answer (Appendix Table 3).

We employ post-processing techniques such as redistributing labels evenly between "A" and "B" and balancing the use of negation words such as "not" and "n't". The validity of each question is ensured through manual examination and editing. Our dataset comprises a total of 1718 questions sourced from ConceptNet (150 questions), GoogleRE (374 questions), SQuAD (100 questions), TREx (594 questions), and OBQA (500 questions), providing a diverse range of negation types, text distributions, and prompts. We believe that this dataset serves as a valuable benchmark for assessing the ability of language models to process negation. Data distributions are shown in Figure 3.

Out of the 1718 questions, we define a set of 944 questions from ConceptNet, TREx, and a subset of OBQA that exhibit clear positive scaling on the corresponding original (non-negated) questions. For our experiments (§3), we randomly select 100 questions from this positive set in order to make the scaling more obvious during our analysis.

3 Results

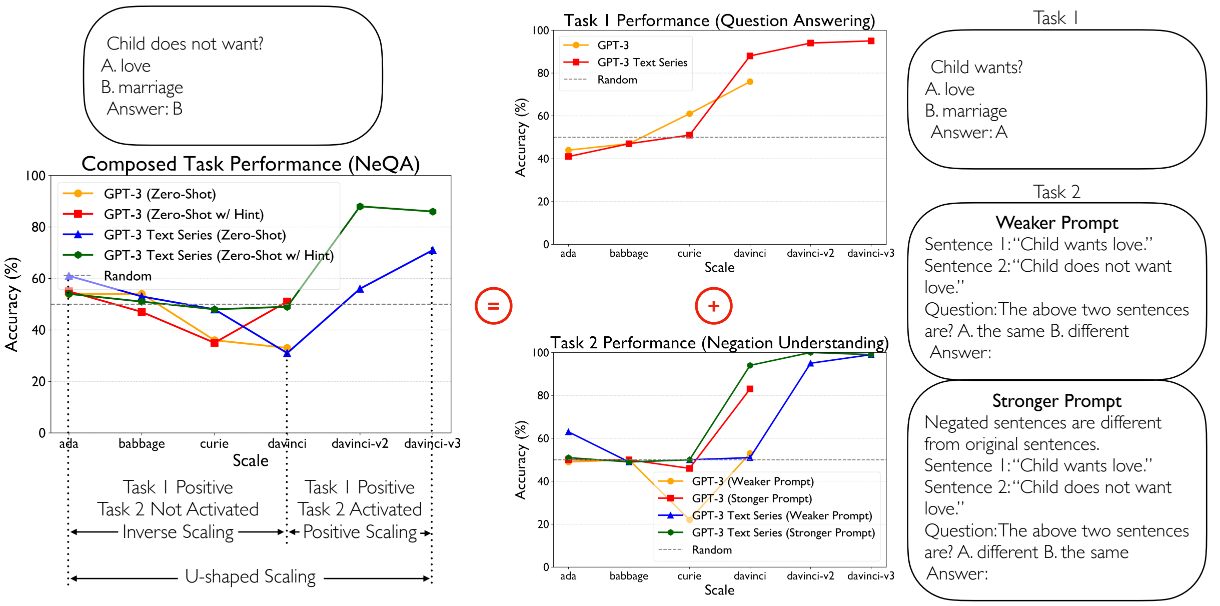
3.1 Evaluation Setup: Models and Prompts

We evaluated four different language model families on NeQA: GPT-3 (Brown et al., 2020), GPT-3 Text Series (Ouyang et al., 2022), Cohere (co:here), and Jurassic (AI21Labs) (model details in §C.4). We employed three different prompting methods: zero-shot, zero-shot with hint (Kojima et al., 2022), and few-shot with chain-of-thought (CoT) (Wei et al., 2022c), as illustrated in Figure 3.

For zero-shot and zero-shot with hint evaluation, we follow the evaluation protocol of the MMLU paper (Hendrycks et al., 2021). We generate a prompt composed of a question and multiple choice options, where the options are labeled "A" and "B". For example, a prompt may be "Question: Child does not want? A. love B. marriage Answer:". We then generate one token from the language model and rank the probability of the model selecting option "A" or "B". For few-shot with CoT, we follow the evaluation protocol of CoT paper (Wei et al., 2022c) by generating sentences until reaching the end and parsing the answer using regular expressions. As our metric, we report the accuracy of the model predictions, where the chance accuracy is 50% as NeQA is a balanced two-choice dataset.

3.2 Scaling Trends

Our evaluation reveals that the scaling trends of language models on the NeQA task vary depending on the prompting method and model family used (Figure 2). We found that the scaling trends of all language model families can be altered by different prompts. For example, zero-shot prompting resulted in inverse scaling in 3 out of 4 model families, whereas few-shot CoT prompting consistently resulted in positive scaling. As the prompt becomes stronger (i.e., more information, like rationales and demonstrations, is provided for the language model), we observed a transition from inverse scaling, to U-shaped scaling, to positive scaling. For instance, GPT-3 exhibited inverse scaling, U-shaped scaling, and positive scaling, respectively, with these prompting methods. Additionally, we discovered that switching to a stronger model family can alter the scaling shape. For example, transitioning from GPT-3 to GPT-3 Text Series, which was further trained to align with human values on multiple tasks compared to GPT-3, resulted



Child does not want?
A. love
B. marriage
Answer: B

Child wants?
A. love
B. marriage
Answer: A

Sentence 1: "Child wants love."
Sentence 2: "Child does not want love."
Question: The above two sentences are? A. the same B. different
Answer:

Negated sentences are different from original sentences.
Sentence 1: "Child wants love."
Sentence 2: "Child does not want love."
Question: The above two sentences are? A. different B. the same
Answer:

Figure 4: Task decomposition analysis of the NeQA task. NeQA task can be decomposed into two subtasks: question answering (task 1) and negation understanding (task 2). Our empirical results show that task 1 has linear scaling, while task 2 has sigmoid-shaped scaling with an emergent transition point, where the transition point is influenced by the prompt method and model family. Combining these two scaling trends yields the final scaling trend observed in NeQA.

in a shift from inverse scaling to U-shaped scaling when the same prompting (e.g., zero-shot) is used.

In conclusion, stronger prompts or model families lead to a transition from inverse scaling, to U-shaped scaling, to positive scaling on the NeQA task. We may also make the following interpretation: the overarching scaling trend of language models for NeQA is U-shaped, and if the model is weak (i.e., weaker prompt or model family), the left part of "U"/inverse slope is observed; if the model is strong, the right part of "U"/positive slope is observed.

3.3 Task Decomposition Analysis

We conducted further empirical analysis on the reasons why the scaling trends can be inverse, U-shaped, or positive and can transition with different prompts or model families. We decomposed the NeQA task into two subtasks: task 1 is to answer the original non-negated questions, and task 2 is to "understand negation". In Figure 4, we show the scaling of task 1 and task 2 performance with GPT-3 and GPT-3 Text Series families. The task 1 performance is measured by the accuracy of answering original non-negated questions, and the task 2 performance is measured by the accuracy of differentiating original questions from negated questions. The task examples are shown in Figure 4 right. Both tasks are evaluated in a zero-shot way.

Our experiments showed that task 1 scales mostly linearly in a positive direction, whereas task 2 scales like a sigmoid shape with an emergent transition point, analogous to the Grokking curve (Power et al., 2022). Before this transition point, models do not "understand negation", and achieve low accuracy in differentiating original questions from negated questions, which results in outputting the same answer to both the original and negated questions. It is worth noting that the labels for the composed task NeQA are essentially the inverse of the non-negated QA labels for task 1. Therefore, the positive scaling in task 1 results in inverse scaling for the composed task NeQA, because the predictions remain unchanged while the ground-truth labels are inverted. After the transition point, models start to "understand negation" and predict opposite answers to the original questions, resulting in positive scaling. When the transition point never happens within the sizes available in the model family, the overall scaling looks inverse; when the transition point happens before the smallest model, the overall scaling looks positive. When the transition point is in the middle, the overall scaling looks U-shaped. We provide further explanations of the composed performance curve in §A and §B.

Interestingly, we found that the transition point can be moved earlier with stronger prompting methods or model families. For example, both GPT-

3 and GPT-3 Text Series show that the transition point happens much earlier when using the stronger prompt compared to the weaker prompt (see Figure 4). Furthermore, GPT-3 Text Series has an earlier transition point than the GPT-3 models. This can explain why using stronger prompts or stronger model families results in a transition from inverse scaling, to U-shaped, to positive scaling.

By decomposing a task and studying the scaling trends of the individual subtasks, our analysis offers a new way to understand the complexity of language model scaling trends. This analysis could be applied to various tasks beyond NeQA, especially tasks that consist of multiple subtasks, each of which may be of different levels of difficulty. This analysis can provide a deeper understanding of the strengths and weaknesses of different language models and offer useful insights into the development of better models and training/prompting methods.

4 Related Works

Scaling trends. Recent years have seen significant scaling of language models, such as scaling from GPT-1 to GPT-3, which has led to tremendous improvements in their performance and capabilities in natural language processing (Radford et al., 2018, 2019; Brown et al., 2020). Researchers have begun to investigate the scaling trends of language models to capture the relationship between model performance and model scale, including the parameter count and amount of training data/compute used (Kaplan et al., 2020). While most scaling papers show positive scaling trends where larger models perform better on various tasks (Brown et al., 2020; Rae et al., 2021; Chowdhery et al., 2022; Srivastava et al., 2022; Liang et al., 2022), it is important to also investigate tasks that exhibit other trends such as inverse scaling, which can shed light to limitations in current language model development and guide future improvements. For instance, TruthfulQA (Lin et al., 2022) was one of the earliest tasks that exhibit inverse scaling, where they find larger language models are prone to hallucination and generate more untrue answers. A recent competition, the Inverse Scaling prize (McKenzie et al., 2022), called for tasks that cause inverse scaling. In the first round, four tasks, including NeQA, redefine math, quote repetition, and hindsight neglect, showed inverse scaling. Wei et al. (2022b) then found that some of these tasks show U-shaped scaling after further scaling up language models. In this work, we unify the above findings and provide a holistic picture of scaling trends, including the transition from inverse to U-shaped to positive scaling across model families and prompting methods, and empirical explanations behind these scaling trends.

Negation understanding. Negation is a fundamental aspect of natural language understanding (Ackrill et al., 1975; Blanco and Moldovan, 2011). Existing works have found that NLP models can struggle in processing negation in text (Jiménez-Zafra et al., 2020). For example, these works investigate models' abilities to process negation through natural language inference tasks (Cooper et al., 1996; Dagan et al., 2006; Hossain et al., 2020; Geiger et al., 2020), machine translation (Fancellu and Webber, 2015; Hossain et al., 2022), language model prompting (Kassner and Schütze, 2020; Ettinger, 2020; Jang et al., 2022), contrastive reading comprehension (Ravichander et al., 2022), and probing model activations (Burns et al., 2022). In response, existing works have also studied methods to improve the abilities of NLP models to process negation, such as leveraging datasets about negation (Kim et al., 2019; Jiang et al., 2021), auxiliary training objectives/tasks (Khandelwal and Sawant, 2020; Moore and Barnes, 2021; Hosseini et al., 2021; Truong et al., 2022), and neuro-symbolic reasoning modules (Yasunaga et al., 2021, 2022). While these existing works typically study a fixed size or type of models, our work provides the first studies into the effect of negation on the *scaling trends* of language models. We find that negation can exhibit nuanced scaling trends, e.g., U-shaped scaling with increased model size and improved model families and prompting methods. This finding offers a more comprehensive insight into how to improve the abilities of language models to understand negation, e.g., the model size, training algorithm, and prompting method all matter.

5 Conclusion

We introduced NeQA, a new question answering dataset that yields different scaling trends of language models than traditional positive scaling. We then proposed task decomposition analysis, a general idea to decompose the task to better understand the complex scaling trends and their transitions. We hope that these insights can facilitate the understanding and development of language models.

Limitations

This work introduced NeQA, a question answering dataset for evaluating the ability of large language models to process negation. While our NeQA attempted to cover diverse types of negation (e.g., different negation phrases and positions) and multiple data sources (e.g., OBQA, LAMA), it is possible that the dataset construction misses some types of negation or domains of text. Our future work will extend the dataset to cover more comprehensive types of negation and domains of text, beyond OBQA and LAMA. Additionally, NeQA is an English dataset, and it would be interesting to extend it to non-English languages and conduct a more comprehensive evaluation of language models, including multilingual ones.

Another potential limitation is sensitivity in language model prompting. Language model performance is known to be influenced by the specific prompt used to query the model (e.g., a rephrased prompt may lead to different model outputs), and prompt engineering—finding the "right" prompt—may be needed to obtain reasonable outputs from the language models (Jiang et al., 2020; Ruis et al., 2022; Wang et al., 2022). As our language model evaluation protocol uses prompting (§3), the evaluation results may inherit such prompt sensitivity. It would be an interesting future work to incorporate techniques to mitigate prompt sensitivity in language model evaluation (e.g., Burns et al. 2022).

Ethics Statement

Our work offers benchmarks and insights to help develop language models that understand negation. Developing language models that understand negation is crucial to the society in many ways.

First, as language models are being used in various real-world applications, including fields like finance, healthcare, and law, it is important to ensure that they understand negation and make correct predictions. If they do not understand negation, they may output the opposite of what we actually want and may make harmful decisions for humans.

Negation is also a fundamental aspect of natural language understanding, and a language model that does not understand negation correctly may not be able to truly process natural language. This can undermine trust and confidence in the outputs of the model, ultimately undermining its utility.

Understanding negation correctly is therefore crucial for the development of reliable language models. We hope that our benchmark and evaluation results provide insights into the behavior of current language models and inspire the future development of language models that understand negation.

Reproducibility Statement

We provide our datasets and implementations at https://github.com/yuhui-zh15/NeQA. The implementations will enable researchers to reproduce datasets and results described here, as well as apply our negation transformations to other datasets and run their own analyses.

Acknowledgments

We greatly thank members of Stanford NLP, P-Lambda, and MARVL groups for providing valuable feedback. M.Y. is supported by Microsoft Research PhD Fellowship. J.Z. is supported by a Sloan Fellowship and a Chan-Zuckerberg Investigator Award. P.L. is supported by a PECASE Award. S.Y. is supported by a Chan-Zuckerberg Investigator Award.

References

John L Ackrill et al. 1975. *Categories and De interpretatione*. Clarendon Press.

AI21Labs. Jurassic language model.

Eduardo Blanco and Dan Moldovan. 2011. Semantic representation of negation using focus detection. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–589, Portland, Oregon, USA. Association for Computational Linguistics.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *ArXiv preprint*, abs/2108.07258.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *ArXiv preprint*, abs/2212.03827.

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*, abs/2204.02311.

co:here. Cohere language model.

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, Technical Report LRE 62-051 D-16, The FraCaS Consortium.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. 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.

Federico Fancellu and Bonnie Webber. 2015. Translating negation: A manual error analysis. In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 2–11, Denver, Colorado. Association for Computational Linguistics.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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. 2022. Can large language models truly understand prompts? a case study with negated prompts. *ArXiv preprint*, abs/2209.12711.

Liwei Jiang, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. 2021. "I'm not mad": Commonsense implications of negation and contradiction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4380–4397, Online. 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.

Salud María Jiménez-Zafra, Roser Morante, María Teresa Martín-Valdivia, and L. Alfonso Ureña-López. 2020. Corpora annotated with negation: An overview. *Computational Linguistics*, 46(1):1–52.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv preprint*, abs/2001.08361.

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.

Aditya Khandelwal and Suraj Sawant. 2020. NegBERT: A transfer learning approach for negation detection and scope resolution. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5739–5748, Marseille, France. European Language Resources Association.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. 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 preprint*, abs/2205.11916.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *ArXiv preprint*, abs/2211.09110.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 6781–6792. PMLR.

Ian McKenzie, Alexander Lyzhov, Alicia Parrish, Ameya Prabhu, Aaron Mueller, Najoung Kim, Sam Bowman, and Ethan Perez. 2022. Announcing the inverse scaling prize. *Lesswrong*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *ArXiv preprint*, abs/2202.12837.

Andrew Moore and Jeremy Barnes. 2021. Multi-task learning of negation and speculation for targeted sentiment classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2838–2869, Online. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *ArXiv preprint*, abs/2203.02155.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *ArXiv preprint*, abs/2201.02177.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. *OpenAI Blog*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv preprint*, abs/2112.11446.

Abhilasha Ravichander, Matt Gardner, and Ana Marasović. 2022. Condaqa: A contrastive reading comprehension dataset for reasoning about negation. *ArXiv preprint*, abs/2211.00295.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2022. Large language models are not zero-shot communicators. *ArXiv preprint*, abs/2210.14986.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Nimit Sharad Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. 2020. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv preprint*, abs/2206.04615.

Thinh Truong, Timothy Baldwin, Trevor Cohn, and Karin Verspoor. 2022. 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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *ArXiv preprint*, abs/2203.11171.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Yi Tay, and Quoc V Le. 2022b. Inverse scaling can become u-shaped. *ArXiv preprint*, abs/2211.02011.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022c. Chain of thought prompting elicits reasoning in large language models. *ArXiv preprint*, abs/2201.11903.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy Liang, and Jure Leskovec. 2022. Deep bidirectional language-knowledge graph pretraining. In *Neural Information Processing Systems (NeurIPS)*.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

A Task Decomposition Simulation: Composing Subtask Scaling Trends Yields U-shape Scaling

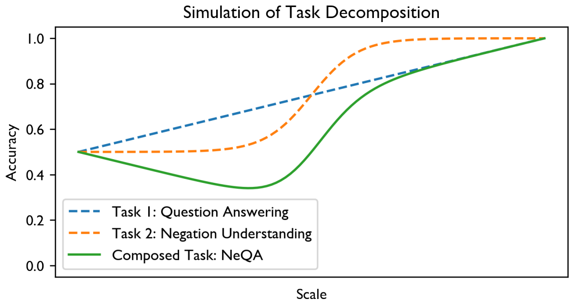


Figure 5: Task decomposition simulation illustrates the U-shape of composed task when given individual tasks.

In this section, we present a simple simulation to demonstrate how the U-shape scaling trends of a composed task can be obtained through the scaling trends of each decomposed task. Let's assume that the accuracy of Task 1 (Question Answering) is represented by $t_1(x)$ and has a linear shape with an initial performance of 0.5 (random performance) and a final performance of 1.0 (perfect performance). Similarly, the accuracy of Task 2 (Negation Understanding) is represented by $t_2(x)$ and has a sigmoid-like shape with an initial performance of 0.5 (random performance) and a final performance of 1.0 (perfect performance), where x represents the scale (a combination of model size, data size, training computation, and prompting method). We define the score of negation understanding as $s_2(x) = (t_2(x) - 0.5)/0.5$, which represents the probability that the model will treat a negated sentence differently from the original sentence. For the composed task, NeQA, it will have an accuracy of $t(x) = t_1(x)s_2(x) + (1 - t_1(x))(1 - s_2(x))$ given scale x.

Figure 5 shows the plots of these three curves, $t_1(x)$, $t_2(x)$, and $t(x)$. The simulated performance curve of NeQA, $t(x)$, indeed exhibits a U-shape.

Discussion of Task Decomposition Validity and Generalizability to Other Tasks. We first clarify that task decomposition analysis is not intended to derive scaling laws (i.e., predict the exact performance of language model scaling). Instead, our analysis aims to explain scaling trends (inverse, U-shape, positive). For example, translation performance may not be simple addition of generation performance and word translation performance but should be positively correlated. Furthermore, while this exact decomposition structure might not hold in more complex tasks, our proposed decomposition analysis is a pioneering attempt to explain scaling trends on a task other than vanilla language modeling. Investigating the applicability of decomposition to other tasks is an essential future direction, and we hope our work will inspire others to push these boundaries. Lastly, we believe that our work's focus on negation is already a well-scoped and significant research contribution, as negation is one of the most common linguistic phenomena. To study negation, we collected the NeQA dataset, which exhibits inverse/U/positive scaling. To explain this, we propose this decomposition intuition, which works well because answering negated questions requires first answering the original questions and then flipping the answers.

B Fine-tuning Simulation: Training Data Attributes and Training Computes Also Impact Scaling Trends

In addition to the prompting methods and model families that we studied in §3, we are also interested in studying other factors that may contribute to scaling trends, specifically those related to the *training* process. However, most large language models are not publicly available and training/reproducing them from scratch would require excessive computational resources. In light of this, here we conduct experiments using synthetic data and small-size language models to simulate and analyze the language model learning process.

We adapt the SST-2 dataset (Socher et al., 2013) for our simulation. For each sentence s in SST-2, with probability $1-x$, we modify it to "s. This does suggest it is good/bad (depending on the label)", and with probability x, we change it to "s. This does not suggest it is good/bad". Then, we fine-tune different sizes of GPT-2 (Radford et al., 2019) on this synthetic corpus with the standard causal language modeling objective. We vary the numbers of epochs t and negation ratio x to understand their effect to scaling trends.

To evaluate the fine-tuned language models, we use the language model to complete "s. This does suggest it is _" for the original sentiment classification task (similar to task 1 in the main paper), and use the language model to complete "s. This does not suggest it is _" for the negated sentiment classification task (similar to the composed task NeQA in the main paper). We report accuracy on

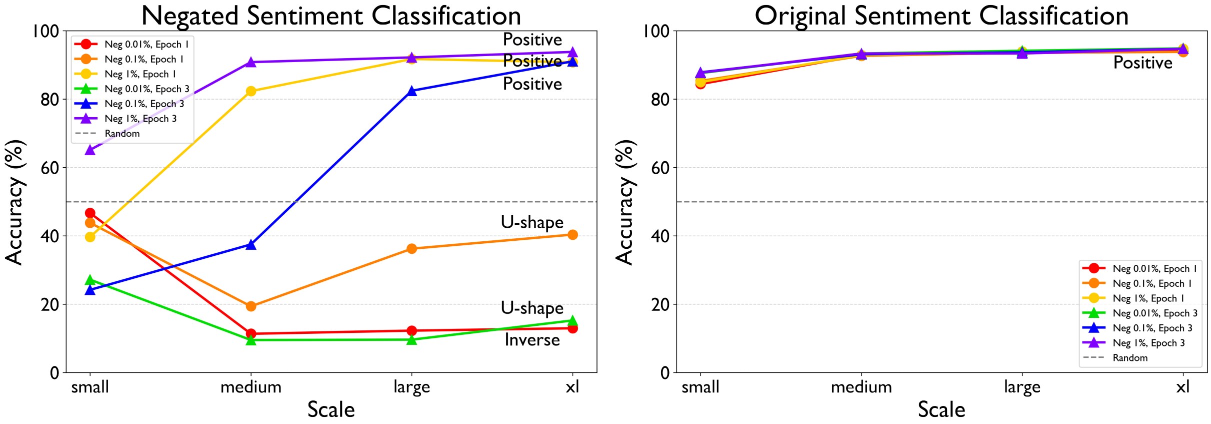


Figure 6: Fine-tuning simulation reveals dataset attributes and training computes also impact scaling trends. We fine-tune different-sized GPT-2 models on a transformed SST-2 sentiment classification dataset with the casual language modeling objective. When the negation ratio in the dataset or fine-tuning epoch increases, we observe a shift from inverse scaling to U-shape to positive scaling on negated sentiment classification task. The original sentiment classification task always shows positive scaling.

the original sentiment classification and negated sentiment classification.

Our simulation demonstrates that the scaling trends on negated sentiment classification are influenced by the negation ratio x and training epoch t (Figure 6). With the same number of training epochs $t = 1$, increasing the negation ratio x from 0.01%, to 0.1% and then to 1% causes the scaling to shift from inverse scaling, to U-shape, then to positive scaling. Additionally, increasing the number of training epochs from 1 to 3 causes the scaling trend to shift from inverse scaling to U-shape when the negation ratio is $x = 0.01\%$, and from U-shape to positive when the same negation ratio is $x = 0.1\%$.

This simulation highlights that factors in the *training* process, such as dataset attributes (e.g., negation ratio) and training compute, also have significant impacts on the scaling trends. Together with the *inference* factors, such as prompting methods and model families discussed in the main paper, we provide a comprehensive understanding of the complexity of scaling trends and how different factors can influence them.

The transition of the scaling trends can also be explained by task decomposition, where Task 1 (original sentiment classification) is always positively scaled, while Task 2 (negation understanding) is also positive but is shaped like a sigmoid, with the transition point controlled by the number of negation examples seen by the language model. The number of negations seen can be modified by using a larger negation ratio or more training epochs. The composition of these subtask scaling trends yields the final scaling curves.

The reason why Task 1 has a more linear shape, while Task 2 has a more sigmoid-like shape, can be understood with the intuition of deep learning processes. Empirical risk optimization (ERM) optimizes for average performance, and since negated sentences are significantly underrepresented in comparison to non-negated sentences in the training data, they are ignored at the beginning of training (Sagawa et al., 2020; Sohoni et al., 2020; Liu et al., 2021). As a result, the performance for negated sentences lags behind the average. However, as the majority of the training examples are learned, ERM finally starts to optimize for the underrepresented groups, leading to improved performance for negated sentences. This intuition adds new insights into the emergence of language models (Bommasani et al., 2021; Wei et al., 2022a), and we leave more rigorous analyses to future works.

C Experimental Details

C.1 Results

The performance of various models on different tasks that generate Figure 2 and Figure 4 can be found in Table 1 and Table 2.

C.2 Data

In Table 3, we provide examples showing the data generation process of the NeQA dataset that was introduced in §2.

In Table 4 and 5, we present a list of 100 data samples from NeQA that were utilized throughout

Model	Prompting	Shape	ada	babbage	curie	davinci
GPT-3	Zero-shot	Inverse	0.54	0.54	0.36	0.33
	Few-shot	Inverse	0.51	0.55	0.51	0.22
	Zero-shot w/ Hint	U-Shape	0.55	0.47	0.35	0.51
	Few-shot w/ CoT	Positive	0.48	0.48	0.53	0.65

Model	Prompting	Shape	ada	babbage	curie	davinci	davinci-v2	davinci-v3
GPT-3 Text Series	Zero-shot	U-Shape	0.61	0.53	0.48	0.31	0.56	0.71
	Few-shot	U-Shape	0.52	0.44	0.45	0.09	0.79	0.80
	Zero-shot w/ Hint	U-Shape	0.54	0.51	0.48	0.49	0.88	0.86
	Few-shot w/ CoT	Positive	0.34	0.45	0.47	0.89	0.98	0.98

Model	Prompting	Shape	small	medium	large	xlarge
Cohere	Zero-shot	Inverse	0.44	0.44	0.38	0.38
	Few-shot	Inverse	0.51	0.52	0.08	0.08
	Zero-shot w/ Hint	U-Shape	0.40	0.39	0.44	0.43
	Few-shot w/ CoT	Positive	0.47	0.47	0.75	0.75

Model	Prompting	Shape	large	grande	jumbo
Jurassic	Zero-shot	Inverse	0.59	0.49	0.49
	Few-shot	U-Shape	0.52	0.39	0.45
	Zero-shot w/ Hint	Inverse	0.58	0.48	0.44
	Few-shot w/ CoT	Positive	0.49	0.53	0.51

Table 1: Table of performance corresponding to Figure 2 and performance of few-shot prompting.

Task	Model	Prompting	Shape	ada	babbage	curie	davinci	d-v2	d-v3
Task 1	GPT-3	Zero-shot	Positive	0.44	0.47	0.61	0.76	-	-
	GPT-3 Text Series	Zero-shot	Positive	0.41	0.47	0.51	0.88	0.94	0.95
Task 2	GPT-3	Zero-shot	Sigmoid	0.49	0.50	0.22	0.53	-	-
	GPT-3	Zero-shot w/ Hint	Sigmoid	0.50	0.50	0.46	0.83	-	-
	GPT-3 Text Series	Zero-shot	Sigmoid	0.63	0.49	0.50	0.51	0.95	0.99
	GPT-3 Text Series	Zero-shot w/ Hint	Sigmoid	0.51	0.49	0.50	0.94	1.00	0.99

Table 2: Table of performance corresponding to Figure 4. Task 1 is question answering and task 2 is negation understanding. d-v2 and d-v3 are abbreviations of davinci-v2 and davinci-v3, respectively.

the paper to examine scaling behaviours and task decomposition.

C.3 Prompts

The specific prompts utilized for various prompting methods and tasks are outlined in Table 6.

C.4 Models

In Table 7, we present a list of all the models used in this work, including 4 model families and 17 models. Model details are from Liang et al. (2022).

D Additional Analyses

D.1 Few-Shot Prompting

Few-shot in-context learning has been demonstrated to be an effective method for adapting pre-trained language models to specific tasks. We experimented with few-shot prompting (not few-shot chain-of-thought prompting) but didn't include the results in the main paper because the scaling shapes were often the same as zero-shot prompting across 3 out of 4 studied model families (inverse for GPT-3 and Cohere, U-shape for GPT-3 Text-Series; only Jurassic changes from inverse to U-shape). We provide the few-shot prompting results in Table 1.

Several recent works can explain why few-shot prompting doesn't alter the scaling curve shape. For example, Min et al. (2022) and Xie et al. (2021) show that in-context learning can be viewed as a Bayesian inference process, with the model learning more about input-output format than input-output mapping. When providing demonstrations of negated question-answer pairs, the model fails to learn the mapping between them and predicts the same answer as without demonstrations.

D.2 Prompt Variations

Due to the sensitivity of language model performance to prompts (Jiang et al., 2020; Ruis et al., 2022; Wang et al., 2022) (also discussed in limitations), we experimented with various prompts and found:

1. Minor changes like word substitution or paraphrasing result in similar scaling shapes;

2. Major prompt changes can alter curve shape, e.g., adding 'For example, "isn't", "is not", "not because ", "do not" are strong signs of negation' to zero-shot w/ hint prompting changes GPT-3 from inverse to a weak U-shape. This can be seen as increasing CoT strength by providing more hints/rationales;

3. Varying CoT information levels affects the shape. Intermediate-level information in CoT prompts shows a scaling shape between U-shape (zero-shot w/ hint; weakest CoT version) and strong positive (few-shot CoT; strongest CoT version).

D.3 Dataset Validity

NeQA is curated by applying rule-based transformations on existing QA datasets. To ensure the dataset quality, we carefully design and verify the transformation rules through manual inspection of the transformed examples. We found that the transformation rules generally work well and only removed a few questions due to grammatical errors after adding negation.

Furthermore, as part of the submission for the inverse scaling prize (McKenzie et al., 2022), the organizers have done crowdsourcing experiments to demonstrate the validity of our dataset. Specifically, they validated labels by crowdsourcing 50 random examples from NeQA, and found the average agreement between workers and gold labels is 100% with no confusing questions.

D.4 Subset Selection

The NeQA dataset is composed of five subsets: ConceptNet, GoogleRE, SQUAD, TREx, and OBQA. For the purpose of this analysis, we only include ConceptNet, TREx, and OBQA. Our goal is to examine the scaling trends, so we aim for steeper scaling. However, GPT-3 does not exhibit strong positive scaling and inverse scaling on the original and negated GoogleRE and SQUAD datasets (Figure 7), so these subsets were not included in the analysis.

Furthermore, these scaling trends of NeQA subsets provide additional verification of our task decomposition analysis. When language models fail to understand negation (Task 2), a stronger positive scaling on the original dataset (Task 1) causes a stronger inverse scaling on the negated data (Composed Task).

D.5 Negation Category

In Figure 8 (left), we find that negating by adding "un-/in-" prefix to a word or negating modal verbs

Dataset	Original Data	Transformed Data
NegatedLAMA	***Original Question:*** Child wants?	***Question:*** Child does not want? *(take **Negated Question**)*
(including ConceptNet, GoogleRE, SQuAD, TREx)	***Original Answer:*** love	***Choices:*** [love, marriage] *(take **Original Answer** and misprime in **Misprimed Question**)*
	Negated Question: Child does not want?	***Answer:*** marriage *(take misprime in **Misprimed Question**)*
	Misprimed Question: Marriage? Child wants?	
OBQA	***Original Question:*** Pushing on a pedal is an example of?	***Question:*** Pushing on a pedal isn't an example of? *(apply rule-based transformation on **Original Question**)*
	Original Choices: [patching, force, practice, speed]	***Choices:*** [force, speed] *(take **Original Answer** and sample one incorrect **Original Choices**)*
	Original Answer: force	***Answer:*** speed *(take sampled incorrect **Original Choices**)*

Table 3: Data generation process of NeQA dataset. NeQA is constructed by transforming two existing QA datasets: NegatedLAMA and OBQA. All the fields of the original and transformed questions are shown.

(e.g., "can" to "cannot") does not show clear inverse scaling in zero-shot prompting. We suspect that the difference is because these negation categories replace a word instead of adding an additional word "not". We leave the further analysis to future work.

D.6 Wrong Choice

This experiment aimed to understand whether more confusing choices will change the scaling (Figure 8 (middle)). For example, given the question "Apple is not made by", the wrong choice can be "Microsoft" (high-ranked, more confusing), or "air" (low-ranked, less confusing), or a random word "China". We find that the wrong choice has little impact on the scaling trends.

D.7 Mispriming

Following Kassner and Schütze (2020), we put the wrong choice (i.e., the correct choice before negation) before the question (e.g., change "iPhone is not made by" to "Apple? iPhone is not made by"). Mispriming makes inverse scaling stronger on negated questions in zero-shot prompting setting (Figure 8 (right)). Interestingly, we also note a phase change happens in small-size models. While this is a very interesting finding, mispriming might not be frequent in real-world applications of language models, so we are not including this in the NeQA dataset.

Question	Choices	Answer
Eden Roc Renaissance Hotel Miami Beach is not a?	[hotel, download]	1
The Betty Hutton Show was not originally aired on?	[youtube, cbs]	0
Hot 8 Brass Band does not play ___ music?	[jazz, friedrich]	1
Crying is not part of?	[life, love]	1
Melbourne International Film Festival is not located in?	[april, melbourne]	0
Mitsubishi G4M is not developed by?	[mitsubishi, vinci]	1
A man grows flowers in the winter using his greenhouse. The flowers blossom there, despite the cold, freezing weather and lack of sunshine, not because	[they are regularly chilled, they are carefully protected]	0
The main reason people want to use wind as an energy source is not because there is	[make cookies, an everlasting supply]	0
A light isn't refracted when	[it shines at a mirror, it shines at a gem]	0
John Selden did not die in?	[manila, london]	0
adenine does not consist of?	[chocolate, nitrogen]	0
Kolkata Book Fair is not located in?	[chinatown, kolkata]	0
Fall is not?	[coming, fearing]	1
Maurice Joly is not a ___ by profession?	[barrister, lawyer]	0
Sojitra is not a?	[village, town]	1
Avoiding the use of fossil fuel is not an example of	[Ecosystem Vulnerability, Conservationism]	0
The native language of Jean-Paul Sartre is not?	[french, greek]	1
Birds cannot?	[fly, ulster]	1
Worms are excellent decomposers not because	[worms enjoy eating dirt and decay, they enjoy eating carcasses and expel waste back to earth]	0
The capital of Azerbaijan is not?	[baku, kidnapped]	1
An electric car uses less gasoline than a regular car not because	[gas is unable to be used, power is prioritized over gas]	0
chicken nugget does not consist of?	[beef, chicken]	0
The native language of Eugene Kaspersky is not?	[russian, korean]	1
Digital Audio Tape is not owned by?	[sony, parsons]	1
Lizards do not have a?	[tail, song]	1
When the moon, sun, and earth line up, tides aren't at their	[Happy, Max]	0
Jean Nicolas Pierre Hachette did not die in?	[navarro, paris]	0
Rottendorf is not located in?	[germany, croatia]	1
David Sainsbury, Baron Sainsbury of Turville never worked in?	[london, director]	1
ground beef is not a subclass of?	[beef, pigs]	1
President of Ireland is not a legal term in?	[irish, ireland]	0
The official language of Ruovesi is not?	[wrestled, finnish]	0
A creature can hide in plain sight from predators not because	[it misdirects, it burns]	1
Information is not a?	[beauty, commodity]	0
The native language of Jules Roy is not?	[french, polish]	1
disruptive coloration is not a subclass of?	[insects, camouflage]	0
Organisms cannot?	[teatro, die]	0
Abdel Sattar Sabry does not play in ___ position?	[wing, midfielder]	0
Some animals get caught easily by other animals not because they are slow while the predators are	[manipulative, speedy]	0
Ano Liosia Olympic Hall is not located in?	[strikeouts, athens]	0
Esko Aho does not work for?	[nintendo, nokia]	0
Riga and ___ are not twin cities?	[tallinn, augsburg]	1
OKK Beograd is not located in?	[homosexual, belgrade]	0
Canjuers is not located in?	[bordeaux, france]	0
Child does not want?	[lab, love]	0
Tungsten is a metal which is found in large amounts in filaments, which are used in incandescent light bulbs. These light bulbs work not because	[the filament glows warmly, light bulbs get very warm]	1
Apple Thunderbolt Display is not developed by?	[intel, apple]	0
Funkadelic does not play ___ music?	[strips, funk]	0
Morice River is not a?	[muscle, river]	0
one inherited characteristic of fur isn't it's	[hue, weight]	1

Table 4: (1/2) Data samples from NeQA we used for scaling trends and task decomposition analyses. Note: labels are balanced.

Question	Choices	Answer
Salvadoran Football Federation is not a member of?	[fifa, uefa]	1
Biblioteka Dlya Chteniya was not written in?	[russian, helpful]	1
Pete Lesperance does not play?	[guitar, trombone]	1
Ohio House of Representatives is not a legal term in?	[delaware, ohio]	0
Jonathan Haidt does not work in the field of?	[music, psychology]	0
Internet Archive was not written in?	[english, polish]	1
Teres Ridge is not located in?	[antarctica, europe]	1
Joseph Clay was not born in?	[ponce, philadelphia]	0
One of the reasons some species go extinct is not because predators	[go dancing with them, murder too many of them]	0
Cans cannot be?	[impressed, recycled]	0
Lev Naumov does not play?	[piano, harp]	1
Tom Cruise is not affiliated with the ___ religion?	[scientology, islam]	1
Anti-Oedipus does not work in the field of?	[philosophy, algebra]	1
Two wheeled conveyances are not held together by	[glue, helically ridged pins]	0
Louisiana Voodoo was not founded in?	[nagoya, louisiana]	0
Animator.ru was not created in?	[russia, argentina]	1
Dan Sealey does not play?	[guitar, pipe organ]	1
Football Association of Brunei Darussalam is not a member of?	[fifa, application]	1
The leading cause of soil and rock erosion is not	[NaCl, H2O]	0
Cyprus is not an?	[accessory, island]	0
Cape Town and ___ are not twin cities?	[johannesburg, jerusalem]	1
The Earth's closest heat source isn't	[our celestial fireball, gamma rays]	1
Cars are big polluters not because they release toxins from the gas into the air making it	[space, unhealthy to breathe]	0
Italy does not maintain diplomatic relations with?	[lebanon, insights]	1
Heribert of Cologne has not the position of?	[mayor, archbishop]	0
Parippally is not located in?	[propagation, india]	0
Which of these isn't less likely to cause pollution	[chopper, mountain bike]	0
National Film Board of Canada does not work in the field of?	[animation, art]	1
fermented milk product does not consist of?	[syntax, milk]	0
Quartz crystals aren't made up of	[hexagons, majic]	1
nickel silver does not consist of?	[shrine, copper]	0
Rats cannot?	[sadness, bite]	0
A landslide is not likely to occur on a slope as	[downward movement is facilitated, there is more debris]	1
Tire is not ___ wheels?	[rubber, culminated]	1
A person wanting to find a live bear in a forest will have difficulty not because bears	[are friendly, avoid humanity]	0
Corn and wheat aren't among the favorite foods of	[humans, fish]	1
Percussion instruments is not made of?	[wood, temple]	1
Chevrolet Trax is not produced by?	[specialist, chevrolet]	0
The original language of The Bitch is not?	[english, breton]	1
Bill Gates does not work for?	[microsoft, ibm]	1
Rajesh Khanna not used to communicate in?	[hindi, yiddish]	1
Edinburgh Airport is not named after?	[edinburgh, bourne]	1
William of Hirsau has not the position of?	[abbot, republic]	1
Austrian Football Association is not a member of?	[lecturer, fifa]	0
Daniele Franceschini was not born in?	[telegraph, rome]	0
The native language of Louis Barthou is not?	[french, mortar]	1
A thing's position is not altered when	[thing feels moved emotionally, the thing adjusts its location]	0
Name is not proper?	[noun, address]	1
Which isn't recyclable	[Styrofoam food container, soda can]	0
The Jazztet does not play ___ music?	[jazz, classical]	1

Table 5: (2/2) Data samples from NeQA we used for scaling trends and task decomposition analyses. Note: labels are balanced.

Zero-Shot	The following are multiple choice questions (with answers) about common sense. Question: Child does not want? A. love B. marriage Answer:
Zero-Shot w/ Hint	The following are multiple choice questions (with answers) about common sense. Note that if there is a negation in the question, we should choose the wrong answer to the original question. Question: Child does not want? A. love B. marriage Answer:
Few-Shot w/ CoT	The following are multiple choice questions (with answers) about common sense. Question: If a cat has a body temp that is below average, it isn't in A. danger B. safe ranges Answer: Let's think step-by-step. First, let's answer non-negated question: "If a cat has a body temp that is below average, it is in?" For this non-negated question, we have A. Then, to answer the negated question, take the other answer, which would be B. So the answer is B. Question: As the barometer reading goes lower there is not a greater chance of A. sunshine B. getting wet Answer: Let's think step-by-step. First, let's answer non-negated question: "As the barometer reading goes lower there is a greater chance of?" For this non-negated question, we have B. Then, to answer the negated question, take the other answer, which would be A. So the answer is A. Question: Coral is a type of living organism which cannot be identified in A. saltwater locations that are open B. any where with liquid Answer: Let's think step-by-step. First, let's answer non-negated question: "Coral is a type of living organism which can be identified in?" For this non-negated question, we have A. Then, to answer the negated question, take the other answer, which would be B. So the answer is B. Question: Child does not want? A. love B. marriage Answer:
Task 1	The following are multiple choice questions (with answers) about common sense. Question: Child wants? A. love B. marriage Answer:
Task 2 (Weaker Prompt)	Sentence 1: "Child wants love." Sentence 2: "Child does not want love." Question: The above two sentences are? A. the same B. different Answer:
Task 2 (Stronger Prompt)	Negated sentences are different from original sentences. Sentence 1: "Child wants love." Sentence 2: "Child does not want love." Question: The above two sentences are? A. the same B. different Answer:

Table 6: Specific prompts for various prompting methods and tasks. Note: for Task 2 prompts, we randomly swap labels “the same” and “different” to balance the distribution.

Family	Model	Details
GPT-3	ada babbage curie davinci	Original GPT-3 (350M parameters) autoregressive language model. Original GPT-3 (1.3B parameters) autoregressive language model. Original GPT-3 (6.7B parameters) autoregressive language model. Original GPT-3 (175B parameters) autoregressive language model.
GPT-3 Text Series	ada babbage curie davinci-v1 davinci-v2 davinci-v3	text-ada-001 model that involves supervised fine-tuning on human-written demonstrations. text-babbage-001 model that involves supervised fine-tuning on human-written demonstrations. text-curie-001 model that involves supervised fine-tuning on human-written demonstrations. text-davinci-001 model that involves supervised fine-tuning on human-written demonstrations. text-davinci-002 model that involves supervised fine-tuning on human-written demonstrations. Derived from code-davinci-002. text-davinci-003 model that involves reinforcement learning (PPO) with reward models. Derived from text-davinci-002.
Cohere	small medium large xlarge	Cohere small v20220720 (410M parameters). Cohere medium v20220720 (6.1B parameters). Cohere large v20220720 (13.1B parameters). Cohere xlarge v20220609 (52.4B parameters).
Jurassic	large grande jumbo	Jurassic-1 Large (7.5B parameters). Jurassic-1 Grande (17B parameters) with a few tweaks to the training process. Jurassic-1 Jumbo (178B parameters).

Table 7: List of models used in this work, including 4 model families and 17 models. Note that the publicly available GPT-3 Text Series model APIs used in this paper differ from those described in the original InstructGPT paper (Ouyang et al., 2022), and OpenAI does not provide information on the training procedure and appropriate model sizes (Liang et al., 2022).

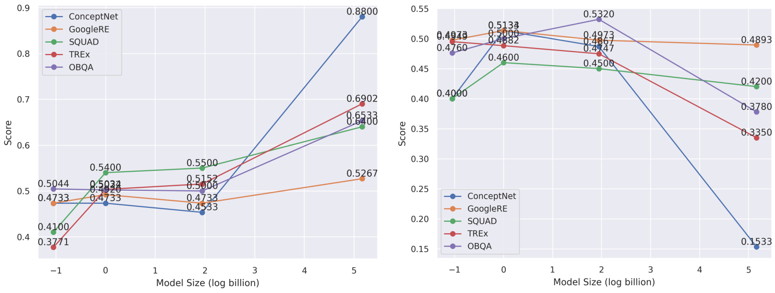


Figure 7: Scaling trend of GPT-3 on five subsets in NeQA using zero-shot prompting. Left: positive scaling (i.e., Task 1) on original questions. Right: inverse scaling on negated questions. Since GPT-3 does not exhibit strong positive scaling and inverse scaling on the original and negated GoogleRE and SQUAD datasets, these subsets were not included in the analysis.

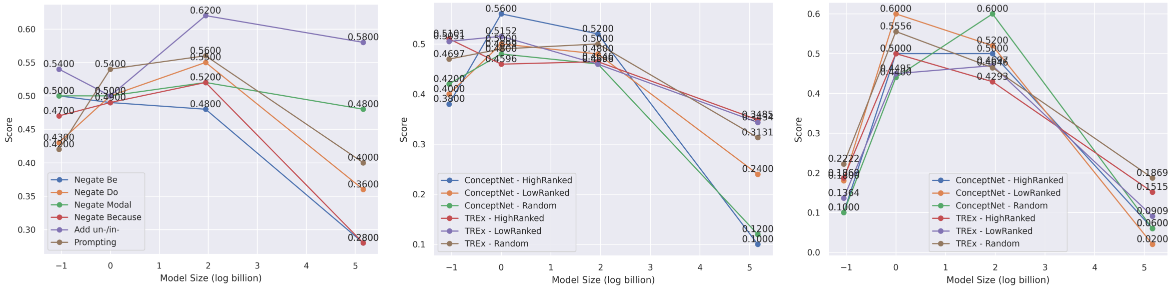


Figure 8: The impact of negation category, wrong choice, and mispriming on scaling trend. These experiments are done using GPT-3 and zero-shot prompting.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Page 6.

☑ A2. Did you discuss any potential risks of your work?
Page 6.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Page 1.

☑ A4. Have you used AI writing assistants when working on this paper?
Used ChatGPT and Grammarly to check and improve grammar.

B ☑ Did you use or create scientific artifacts?

Section 2.

☑ B1. Did you cite the creators of artifacts you used?
Section 2.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Our data will be made publicly available and can be used for research purposes.

☑ 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)?
Section 2.

☒ 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?
Our dataset does not contain names or sensitive information.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2.

☑ 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.
Section 2.

C ☑ Did you run computational experiments?

Section 3.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3.

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?
Section 3.

□ 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?
Not applicable. Left blank.

□ 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.)?
Not applicable. Left blank.

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

Section 2.

☒ 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.?
We annotated the data by ourselves.

☒ 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)?
We annotated the data by ourselves.

☒ 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?
We annotated the data by ourselves.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our data does not have ethics concern.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We annotated the data by ourselves.