How Important is the Truth?

Stanford CS224N Custom Project

Joseph Tan
Department of Computer Science
Stanford University
Prepared data, Developed Evaluation, Prompt Engineered
dghosef@stanford.edu

Rehaan Ahmad
Department of Computer Science
Stanford University
Setup Model Pipeline, Designed Experiment, Analyzed Data
rehaan@stanford.edu

Abstract

In context learning is a powerful technique in natural language processing that allows models to specialize to tasks without requiring any finetuning or adjustment of weights; instead, large language models are able to learn how to solve tasks by observing examples given during inference. While this is a known phenomena, exactly why this works and the exact features and characteristics of the provided examples that the model leverages is still unknown. In this paper, we explore the importance of having "correct" examples in modern open source models for both generative and classification tasks. Our findings suggest that the importance of having correct examples varies very much per task. For some tasks, correctness matters little. For others, it is better to have no demonstrations than incorrect ones. This provides novel insight into how language models learn from context for different kinds of tasks.

1 Key Information

- Mentor: Archit Sharma
- External Collaborators (if you have any): None
- Sharing project: None

2 Introduction

In-context learning has been a defining feature of LLMs which allows models to specialize on a specific task without actually having to update its weights. Fine-tuning can be a powerful technique to adapt to new tasks, but with exceedingly large models and limited data points this can be extremely impractical. Our goal is to learn more about how in-context learning affects models. More specifically, we do a deep dive into understanding the role of demonstrations and if the actual (x,y) pairings matter in a demonstration or the individual distributions of x and y are more important. The paper our final project is based on, "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?"Min et al. (2022b), evaluates precisely this. However, this paper was published two years ago and we are curious about how the results of this paper would extend to a wider variety of more recent LLMs and datasets. We specifically aim to evaluate the applicability of these results to a number of state of the art open source models with varying sizes. We believe these will be valuable datapoints in ongoing experiments.

3 Related Work

There has been plenty of recent research regarding in-context learning from which we base our work on top of. A cornerstone paper for LLM in-context learning is "Language Models are Few-Shot Learners" from OpenAI Brown et al. (2020). In-context learning is just one of the many ways to solve the problem of few-shot learning and this paper reveals the emergent properties that arise in LLMs when it comes to this problem. Few-shot learning is the problem where a model has to adapt to a new tasks with only a few examples. To address few-shot learning, there have been many techniques tried that use more traditional meta-learning techniques – this paper claims that as models grow exceedingly large, something special happens. The models show emergent properties in their ability to deal with new in-context provided information. Rather than trying to generate new set of weights to address the new data provided, the paper asserts that models of sufficient size can actually improve by just directly feeding this information into model context.

From here many papers emerged building on the efficacy of in-context learning, a few that we looked at were: Zhao et al. (2021), Holtzman et al. (2022), and Min et al. (2022a). Zhao et al show that with GPT-2 and GPT-3 there can be an associated instability with in-context learning and claim that calibration is needed to avoid this instability. These papers fit a similar theme to the paper we are covering for our final project – namely, they attempt to probe how & why in-context learning works. Min's paper that we cover for our final project performs an interesting analysis that gives some insight into how such LLMs could work – specifically, Min performs ablations where the demonstration ground truth values are altered, but still in-distribution. She notices that the actual accuracy drop-off of these in-context demonstrations is not large even though the x,y pairings are completely wrong. The assertion here is that the distribution of the individual questions and answers is more important than the actual pairings – i.e. the LLM isn't actually "learning" anything about the specific pairings but more so learning about the structure and distribution of output answers.

Another great reference point about in-context learning was this 2022 survey put out by researchers at Peking University: Dong et al. (2023). This survey discussed some of the open research questions regarding why in-context learning works: like Min discusses in her paper, a lot of research works consider the importance of the distribution of inputs/outputs rather than the actual mapping. However, some works also assert that transformers end up encoding learning algorithms that mimick least squares regression. There's no question that both of these lines of reasoning are correct and play a role in making in-context learning: the question is just how much of in-context learning is due to the demonstration vs the actual mapping – and Min asserts that much more of the success of in-context learning can be attributed to the distribution.

4 Approach

The paper our final project is based on is "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?", which evaluates the effect of demonstrations over 12 models in the GPT family. More specifically, the paper asserts that when doing in-context learning, the distribution of inputs and outputs are more important than the actual mapping of a corresponding input to a corresponding output. In our final project, we aim to evaluate if this claim holds on more recent open & closed-source models and across other datasets. For each task, we consider the following three prompt setups

- {input question}
- {exampleq1, examplea1}, {exampleq2, examplea2}, ..., {exampleq1n examplean} {input question}
- {exampleq1, randomea1}, {exampleq2, randomea2}, ..., {exampleq1n randoman} {input question}

To test this claim we set up an inference and evaluation pipeline using Modal. The inference pipeline with Modal would spin up an A100 "Stub" that we would load a Llama or Mixtral model onto and conduct sequential inference (no batching was done here since the model is small enough and didn't take too much time for inference). From here we ran the model with just the initial input question, N pairs of correct demonstration data for the model in the context, and N pairs of randomly matched demonstration data in the context. As aforementioned, evaluation of the models is not

straight forward so we piped all output to a separate text file. Based on the task we then ran a suite of different evaluation functions to extract the intended answer from the LLM. For the closed-source models we replaced the Modal "stub" function with the corresponding completions API provided. An important note is that we did spent a fair amount of the 230 dollar budget conducting large scale prompt tuning experiments to figure out which prompts lead to the best results for the model. This meant that we exhausted the Modal credits and had to turn to alternative ways of getting evaluation metrics from the open-source Llama model. We learned that there were commercial providers of open source models like Together AI which we ended up using to more affordably get experiment results. Once we converged on optimal prompts for each of the separate datasets, we ran a final set of experiments using together.ai that we report here.

5 Experiments

5.1 Data

We ran our experiments on a total of 12 datasets. Their names and a brief description are listed below. For each of them, we tested a few prompts and picked the one that led to the highest accuracy.

- **xnli**: Given 2 phrases in Bulgarian, determine whether or not the phrases entail each other, contradict each other, or do neither. Conneau et al. (2018)
- **winograd**: Given a sentence and a pronoun, determine the noun that the pronoun refers to. Levesque et al. (2012)
- **swag**: Given a text and 4 possible endings, determine which ending would make the most sense to complete the text. Zellers et al. (2018)
- **news**: Classify a news headline as 'world', 'sports', 'business' or 'tech.' Zhang et al. (2015)
- **paws**: Determine if two sentences are paraphrases of each other(convey the same information and have the same meaning) or not Zhang et al. (2019)
- **emotions**: Determine the emotion that most closely describes a piece of text out of 'sadness', 'joy', 'love', 'anger', 'fear', 'surprise.' Saravia et al. (2018)
- **commonsense qa**: Answer 4-choice multiple choice questions on common sense Talmor et al. (2019)
- **gsm8k**: Answer difficult middle school math word problems Cobbe et al. (2021)
- **ethos**: Classify text as 'hate speech' or 'not hate speech' Mollas et al. (2020)
- **quora**: Given 2 quora questions, determine whether or not they are duplicates Sharma et al. (2019)
- **race**: Answer 4-choice reading comprehension multiple choice questions about news articles Lai et al. (2017)
- **boolq**: Answer yes or no questions about a piece of text Clark et al. (2019)

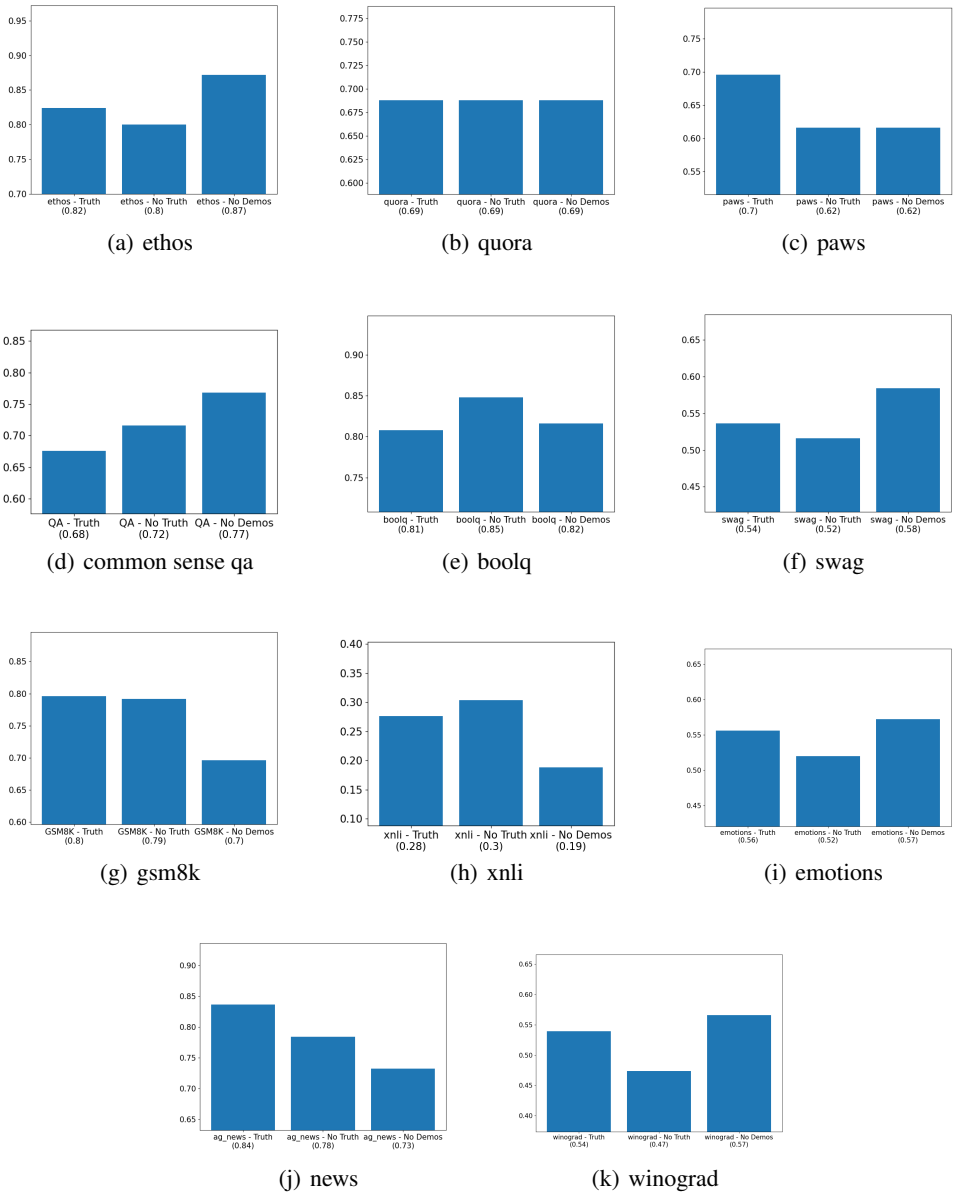
5.2 Evaluation method

Our evaluation method was to carefully match the output of the LLM for each task with the correct answer. Because it is hard to enforce even the best closed-source LLMs to match an exact output pattern without a guardrails-type package we had to be creative with evaluation. First off was to prompt tune and play around with the prompt to ensure we are getting as close to a consistent output pattern as possible. However, this often times was not enough. A good example of this is with GSM8K where we would specifically prompt the model like so: "Make sure to answer this question by first explaining your reasoning followed by four hashes and then a single numerical answer." However often times the model would not follow this rule. Some failure modes would be not putting 4 #s and giving an answer like "18 apples" instead of "18". In this case, the simple heuristic we applied was a regex to fetch the last numerical entity in the answer and treat that as the LLM's predicted answer. This had to be done meticulously on a case-by-case basis for each of the dozen tasks we tried for these experiments. Had we had more compute units/resources we would evaluate some of the harder-to-evaluate tasks by passing the output through another LLM to extract the true answer.

5.3 Experimental details

For each of the tasks, we evaluated Llama-3-13b three times on the testing data. On one trial, we provide only the question prompt. In the others, in the question before the prompt we provide 16 random question-answer pairs from the training data. First, we made sure the question-answer pairs matched(baseline ground-truth demonstrations) and then we matched questions with different answers. All experiments were run on Nvidia A100 GPUs through Modal Labs and all random numbers were generated with python's random.randint and the default seed.

5.4 Results



(a) ethos (b) quora (c) paws

(d) common sense qa (e) boolq (f) swag

(g) gsm8k (h) xnli (i) emotions

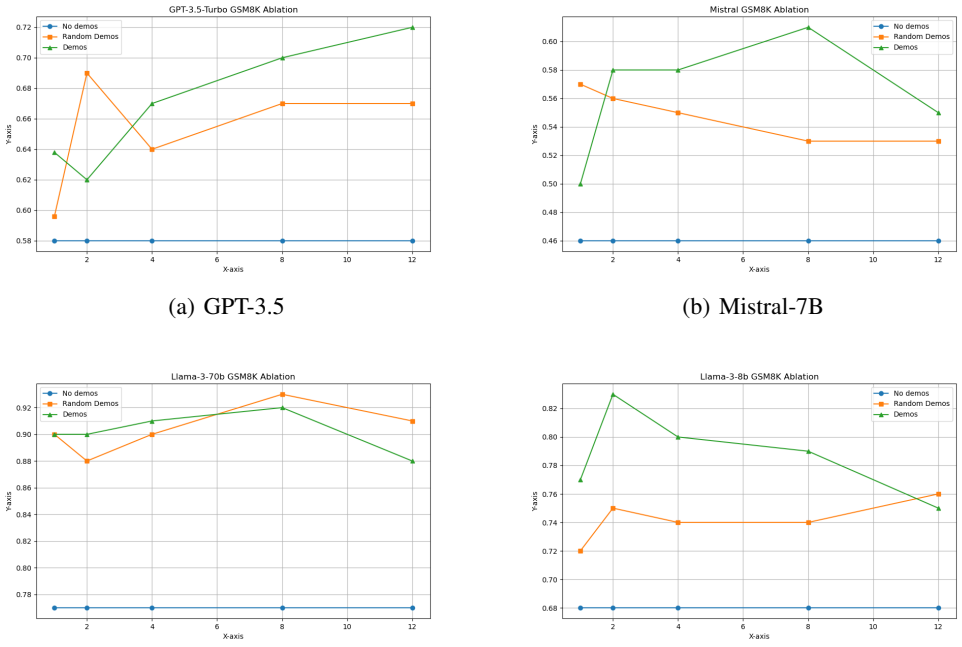
(j) news (k) winograd

Figure 1: Dataset Specific Results

The results from the above figures highlight a couple of different stories here. First off, there are tasks where providing demonstrations doesn't necessarily help. In tasks like common sense QA, providing more examples doesn't really seem to help and it aligns well with our understanding of the

task. In common sense QA you are given a common sense question and asked to return a multiple choice response. Giving more examples of this doesn't necessarily help out the LLM and it reflects in these results. Second off, there are tasks where providing demonstrations help considerably like GSM, XNLI, and News. In these tasks they fall under two categories – tasks where there is little to no difference in whether or not the demonstrations Q/As are randomly matched (like GSM & XNLI) and tasks where there is some importance in if they are matched (like news).

From the above figures, we see that there doesn't seem to be a definite answer to the question of when and how to use in context examples. In some tasks, like the quora one, the result was the same regardless. In some tasks like paws, ground truth matters a lot. And in some tasks like GSM8K and xnli, the truthiness of the examples didn't matter; what mattered was simply the presence of examples. We were in particular surprised by the fact that the effect of demonstrations was so dependent on the task.



(a) GPT-3.5

(b) Mistral-7B

(c) Llama-70B

(d) Llama-8B

Figure 2: Gpt 3.5 vs Llama-3-8b vs Llama-3-70b vs Mistral-7b on GSM8K Demonstration Count ABlation

We also run an ablations over the number of in-context demonstrations for GSM8K. We range the demonstrations from 1 to 2 to 4 to 8 to 12. We were unable to do 16 due to a context window size error. First, it's good to note that for GPT-3.5 as the demonstration count increases, we are seeing slightly improving results (the green line). It's also encouraging to see that the randomization of the demos here does not significantly reduce accuracy when compared to the "no demos" blue line. This same general upward trend is observed – in other words, increasing the number of incorrectly paired demonstrations helps for GPT-3.5! It's fascinating to observe this with a model as new as GPT-3.5 (relative to the publication of the original paper). To summarize the results of this ablation, the similarity between green & orange lines vs the disparity with the blue line is an encouraging result that in essence highlights the results from the original in-context paper we consider for this final report still hold for today's state of the art models!

6 Analysis

The analysis of results spans a couple different aspects of in-context learning. The first being that providing demonstrations for certain tasks does not always help. In hindsight, we should have avoided running as many experiments as we did on top of multiple-choice & true/false tasks like BoolQ, Race,

and common sense QA where giving more examples intuitively would not help. If we consider the initial hypothesis of our paper, it is that the distribution of output answer is more important than the actual pairing of question to answer. With this in mind, it makes sense the that tasks like true/false where the distribution of answer is very small to begin with would not do well. It seems to be that the adding of examples here only increases the number of noisy/useless tokens for the question itself. It was important to run these experiments to understand this, but we could have run less of the true/false ones in hindsight. It's also important to factor in the noise that could be produced by our evaluation criteria not being perfect for some of the tasks. Exactly figuring out what the model is outputting for multiple-choice can be tricky with just regexes: did the model output a), a., A, etc? Especially when the character A is frequently occurring in other parts of the answer string. As aforementioned, the ideal way of dealing with this is to feed it through another LLM and get the exact output, but due to financial constraints paired with the scale of our experiments this was not possible.

We now focus on the results for tasks where demonstration data significantly helped: GSM, XNLI, and News. It was very encouraging to see that in all of these cases the random demonstrations improved upon the "no demonstration" experiments. This aligns with the paper and, like the original paper, also makes us re-consider our intuition of how in-context learning actually works. One would assume that providing wrong demonstrations should make results worse than giving no demonstrations at all – making these results completely counterintuitive (yet supportive of Min's conclusions).

After running these experiments we wanted to focus our analysis on just GSM since it's a perfect dataset for analyzing in-context learning. We ran an ablation over the number of demonstrations and observed that the random in-context demonstrations results are consistently closer to the correct demonstration results than the no demo results – regardless of the number of demonstrations. It's interesting to note that the only model with an upward trend with regards to the number of demos was GPT-3.5. This aligns with our understanding of LLMs – smaller models with significantly less parameters like Mistral-7B and Llama-8B make use of more demonstrations less than larger models. Regardless, in both cases the no-demonstration models did significantly worse than the other two prompt configurations which indeed supports the hypothesis put forth by Min.

7 Conclusion

We have investigated the importance of truth in question answer demonstrations during inference. Our findings show that the importance of the presence and correctness of demonstrations varies based on the task. Some tasks, like BoolQ and Race that had few choices and a simple output format, benefited little from demonstrations. Some tasks, like GSM and XNLI, benefited very much from demonstrations yet very little from correct demonstrations as compared to incorrect ones. And some tasks, like PAWS, benefited only from correct demonstrations, and incorrect demonstrations were as helpful as no demonstrations. We also show the relationship between number of demonstrations and accuracy for a variety of state of the art models and show that larger models can take advantage of an increase in number of demonstrations.

We feel that our experiments were quite comprehensive. We both tested many models on the GSM8K datset and many datasets with a single model(llama 8b). A limitation was that we couldn't test all 12 datasets with many models due to a lack of Modal and Together.ai credits, but this is a minor limitation as the rest of our experiments were relatively comprehensive across models and datasets. Future work would include testing the state of the art(GPT-4, Claude Opus, Llama 400b), more datasets, and conducting detailed analyses on potential bias that incorrect in-context examples can introduce.

8 Ethics Statement

It is important to understand the limitations of this study. In particular, we only analyzed the accuracy scores that were affected by in-context learning. We did not look at, for example, whether incorrect question, answer pairs introduce bias. Although recent studies, including this one, shed light on how in context learning helps and the specific features that are necessary, there still remains a lot to be learned about in context learning. As a result, without comprehensive ablations analyzing bias, we cannot say whether or not incorrect examples introduce bias. Thus, the reality is that, while we show that incorrect question answer demonstrations can be valuable and thus are a good technique

for LLMs, they should not be used in real-life settings unless the potential for bias is and potential harmful effects are mitigated. Whenever in context learning with incorrect demonstrations is used in the real world, it is essential that for *each task*, a comprehensive analysis of the bias in context learning introduces is conducted, as we earlier demonstrated that the effects of in context learning vary per task.

Another potential ethical concern is the usage of this technique could lead to a decrease in trust of models. Both this paper and the original paper show that providing a LLM with incorrect information can help improve its accuracy. Assuming the aforementioned due dilligence is performed, this is a technique that has many practical applications, as it is much easier to generate incorrect question answer demonstrations than it is to generate correct demonstrations. However, this is a counterintuitive technique, and most users of large language models are not well versed in the research behind them. Upon hearing that incorrect demonstrations were used, users might conceivably lose trust in the models. This trust is crucial for the adoption and effective use of AI systems, particularly in sensitive applications such as healthcare, education, and legal contexts, and a lack of trust could undermine the potential benefits of AI that could otherwise, for example, be used to save lives by making the healthcare system more efficient. A potential mitigation for this is to always accompany a mention of this technique with a justification and referral to the research that backs it up. Even if the user doesn't understand the paper, they will know it is a research backed technique and this will increase the perceived legitimacy of this.

References

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.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2022. Surface form competition: Why the highest probability answer isn't always right.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022a. Noisy channel language model prompting for few-shot text classification.

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

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. Ethos: an online hate speech detection dataset.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. 2019. Natural language understanding with the quora question pairs dataset.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. of NAACL*.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models.