# Man, I Just Love Writing Research Papers

**Jacob Badolato**
University of Texas at Austin
jacobbadolato@utexas.edu

**Andrew McGehee**
University of Texas at Austin
andrewmcgehee@utexas.edu

## Abstract

State of the art language models have been shown to perform well across a broad spectrum of language modeling tasks like natural language inference (NLI), question answering (QA), and next token generation tasks as seen in applications like ChatGPT [8]. However, it is still quite simple to find examples that confuse language models and cause significant degradation in performance metrics. In this work, we propose one such set of "tricky" examples. We analyze the performance of the ELECTRA-SMALL model [3] on the NLI task [1] across three subsets of tricky examples: sarcasm, grammatically incorrect dialect, and idiomatic speech. We show that the model struggles to accurately classify such examples. While the model achieves a strong 85.8% accuracy on the Stanford NLI (SNLI) dataset [1], it is only able to achieve 50.9% accuracy on the holdout portion of our curated tricky dataset. We also show that simply fine tuning the model on a small training portion of these examples improves the models ability to accurately classify new tricky examples while incurring almost no additional loss on the original SNLI dataset. Hence, we provide some evidence that providing higher quality, more diverse examples of real, albeit complex language patterns in training data can improve a model's ability to generalize across tricky examples.

## 1 Introduction

Models like GPT 3 [2], BERT [4], and [9] GPT 4 have left the world speechless at the awe-inspiring capabilities of large language models (LLMs). In contrast, LLM failures like those in the early versions of Bing + ChatGPT have served as comedic inspiration for memes as well as the inspiration of nightmares for Microsoft stakeholders. This raises the question: given the strong results of LLMs in some domains, which problem classes cause LLMs to experience performance degradations? This work begins to explore this question.

Drawing inspiration from some of the well-known LLM failures mentioned earlier, we devised a set of "tricky" examples to test LLMs against, and we observed their behavior. These tricky examples are all examples of real language patterns that humans can decipher with relative ease. The examples are subdivided into three groups: sarcasm, grammatically incorrect dialect, and idiomatic speech. We provide examples from each of these groups in a later section. A criticism the reader may levy against the use of sarcasm in text is that the notions of tone, body language, rate of speech, inflection, etc. aren't signals which are included in the text. The authors encourage those with such criticisms to consider times where they have been able to decipher sarcasm in text messages without any additional context. While the authors acknowledge the limitation that even humans cannot perfectly decipher sarcasm via text alone, it is certainly an achievable task and therefore still within the scope of our research question.

We interrogate this question by training an instance of the ELECTRA-SMALL model. Despite its proven capabilities on other datasets, the model struggles with our tricky examples. Sarcasm, dialect, and idioms are often context dependent, non-literal, and difficult to parse grammatically. The authors expected when devising these examples that the model would struggle with these examples for the

Preprint. Under review.

following reason: the "default mode" of speech in most language training data is literal, mostly grammatically correct, and directly interpretable with respect to its meaning (not sarcastic).

Given this expectation, we also hypothesize that simply fine tuning a well-performing model on a carefully curated set of difficult examples can dramatically improve its ability to decipher these complex language patterns. To test this, we conducted an in-depth analysis of the ELECTRA-SMALL model, initially trained on the SNLI dataset and later fine tuned on our curated tricky examples.

## 2 Background

### 2.1 Challenges in Natural Language Processing

Modern LLMs perform exceedingly well across a broad range of language modeling tasks. Particularly massive, multi-hundred billion parameter models are approaching near superhuman language capabilities in some domains. However, smaller more practical models (like those which may realistically fit on a mobile device for example) still struggle with more complicated constructs of language like sarcasm and idioms. To emphasize this point, if an NLI model were to classify the hypothesis "the author loves research papers" given the title of this work as the premise, it would inaccurately label the example as supportive rather than a contradiction. In contrast, the subtleties of using the words "Man," and "just" are enough for most humans to accurately decipher the sarcastic tone.

### 2.2 The ELECTRA-SMALL Model

The ELECTRA-SMALL model was selected simply because it is a small, quickly trained model. Despite its small memory footprint, it is a robust model that outperforms many larger models. Furthermore, it was one of few models which were accessible to train given the authors' limited access to hardware accelerators.

### 2.3 Adversarial Examples

We drew inspiration for curating our tricky dataset from previous works exploring the concept of adversarial examples [5, 6, 7]. An adversarial example is one which is devised with the sole intent of causing the model to become confused or to navigate areas of its search space which cause instability in the model's outputs. Some works have shown that including adversarial examples in training data (as we have done via fine tuning) improve a models ability to be resilient to adversarial data. In our case, this improves the model's ability to generalize to more nuanced and complicated language patterns.

### 2.4 The SNLI Dataset

The Stanford Natural Language Inference (SNLI) dataset has been pivotal in training numerous NLP models, including ELECTRA-SMALL. While it provides a broad spectrum of linguistic examples, its coverage of intricate language patterns is limited [1]. This limitation prompted our exploration of supplemental training data to bridge these gaps.

## 3 Examples of Errors

### 3.1 Grammatically Incorrect Dialect

- Premise: He doesn't hardly speak to anyone.
- Hypothesis: He is very outgoing.
- Gold label: 2
- Actual: 0

### 3.2 Sarcasm

- Premise: What a pleasant surprise, another bill in the mail.
- Hypothesis: I do not enjoy paying my bills.
- Gold label: 2
- Actual: 0

### 3.3 Idiomatic Speech

- Premise: Come on, spill the beans about the party plans!
- Hypothesis: Someone wants them to spoil the party plans.
- Gold label: 0
- Actual: 1

## 4 Discussion on Tuning the Model

The categories of mistakes that we tested were: sarcasm, grammatically incorrect dialect, idiomatic speech. In the sarcasm category, the meaning is often the opposite of the literal text and is often signaled only by small nuances in the text. In the grammatically incorrect dialect category, the text often includes a double negation where the meaning is actually meant to be interpreted as a single negation. For example, in some parts of the southern United States the phrase "That ain't nothing to be scared of." simply means "That isn't something to fear." In the idiomatic speech category, the text simply contains some idiom or figure of speech which has a non-literal meaning. These categories are difficult for the model to interpret because each have unorthodox or non-literal usages. Our findings suggest that the model struggles most with context-dependent or counter-intuitive examples. We show that providing more context through fine tuning data allows the model to perform significantly better. Figure 1 shows the prediction error rates for the three categories of tricky examples.

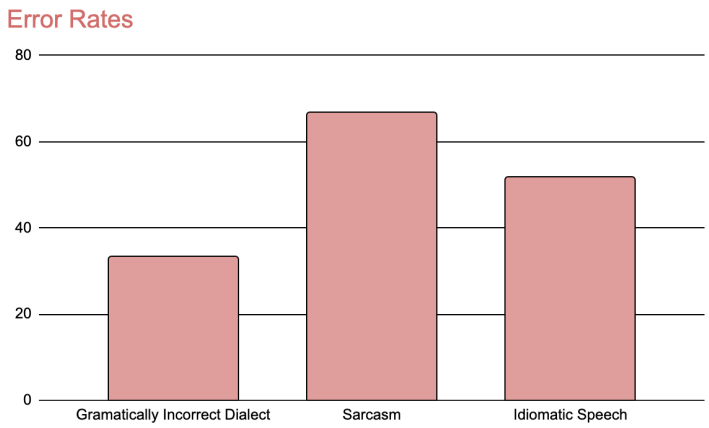


Figure 1: prediction error rates $\in [0, 100]$ for each category of tricky example

Using a confusion matrix, we can more deeply understand what kind of misclassifications the model is making. We can see from the data that the model has a very difficult time classifying examples where the true label is contradiction (2), as compared to those with a true label of entailment (0) or neutral (1). This is expected, because many of our tricky examples are designed to seem like entailments, when in reality they are contradictions. Figures 2 and 3 show the confusion matrices on our tricky dataset and the SNLI dataset for an untuned ELECTRA-SMALL model. When we evaluated the ELECTRA-SMALL model on the SNLI dataset it achieved an accuracy of 85.8%.

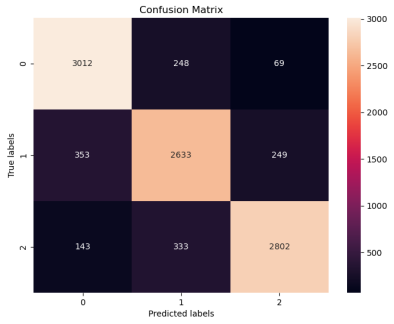


Figure 2: SNLI dataset before fine tuning. lighter regions are better. notice that a clear diagonal emerges.

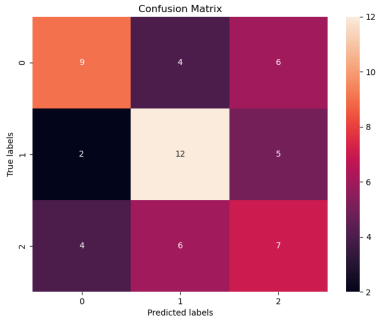


Figure 3: our tricky dataset before fine tuning. lighter regions are better.

## 5 Tuning on Adversarial Examples

To increase the model's resilience against tricky examples, we fine tune the model on a training portion of our tricky example dataset. The dataset is comprised of 60 hypotheses for each of the three categories discussed earlier (sarcasm, grammatically incorrect dialect, and idiomatic speech). Each hypothesis is paired with three premises where one is an entailment, one is neutral, and one is a contradiction. This results in a total of 180 examples. These 180 are then split into a training portion and an evaluation portion. The train/eval split was roughly 70/30 resulting in 125 training examples and 55 evaluation examples. The seed used to randomly split examples was 123. After fine tuning the model on our examples, the accuracy achieved improved to 60.0% on the "tricky examples", up from 50.9%. The confusion matrices for our tricky dataset before and after fine tuning are shown in Figures 4 and 5 respectively.

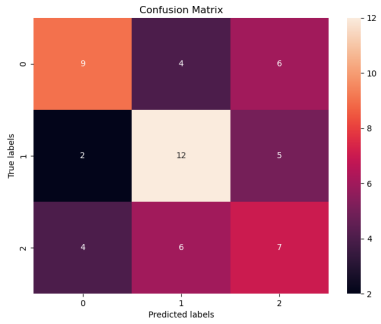


Figure 4: our tricky dataset before fine tuning. lighter regions are better. note this figure is identical to Figure 3.

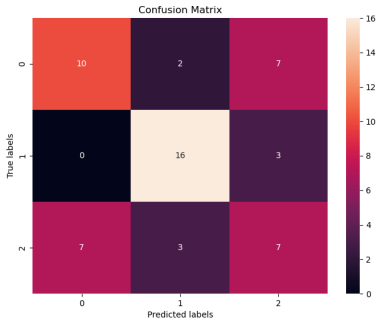


Figure 5: our tricky dataset after fine tuning. lighter regions are better. note the emergence of a diagonal as in Figure 2.

When re-evaluating the tuned model on the SNLI dataset, the model still achieved commensurate performance as the untuned model. Only a small decrease in accuracy from 85.8% to 84.6% was observed. This drop in accuracy is expected behavior since we fine tuned the model on examples that were significantly different from the original SNLI examples. The authors expect that given a larger curated dataset of tricky examples, the trends we observed would continue. The models ability to accurately pick up sarcasm, dialect, and idioms would continue to improve, while its ability to perform the original SNLI task would continue to degrade as the fine tuning dataset increases in size. At some point, the fine tuning dataset would become large enough as to induce catastrophic forgetting, where the model will begin to unlearn the patterns of the original dataset altogether and start to overfit to sarcastic, dialectic, and idiomatic speech. The confusion matrices for the untuned model and the tuned model on the original SNLI dataset are shown in Figures 6 and 7 respectively.

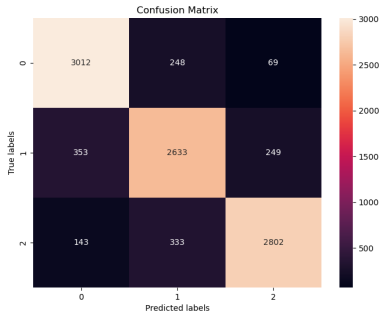


Figure 6: SNLI dataset before fine tuning. lighter regions are better. note that this is identical to Figure 2.

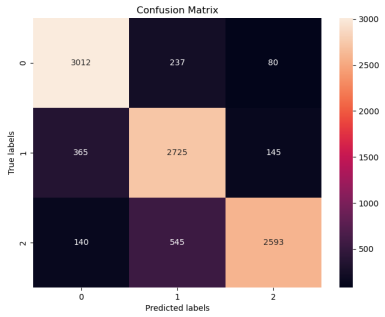


Figure 7: SNLI dataset after fine tuning. lighter regions are better.

We also report the accuracy, precision, recall, and f1 scores for each of our treatments. All treatments achieve tight clusters overall with each metric being within 1% of each other. This suggests that fine tuning on the tricky examples improves the models general ability to correctly classify tricky examples, rather than only improving precision or recall alone. Figure 8 provides a table of these metrics.

| Dataset | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| SNLI | 85.8 | 85.8 | 85.8 | 85.8 |
| SNLI (tuned) | 84.6 | 85.1 | 84.6 | 84.6 |
| Tricky | 50.9 | 51.1 | 50.6 | 50.5 |
| Tricky (tuned) | 60 | 58.7 | 59.3 | 58.9 |

Figure 8: reported metrics for each model-dataset treatment.

## 6 Conclusion

This work strengthens existing evidence for methods that effectively enhance the performance of NLP models in processing complex language structures. Namely, we show the effectiveness of including adversarial examples in a fine tuning dataset on the NLI modeling task. We demonstrate that fine tuning a model with a thoughtfully selected set of hand-annotated examples specifically designed to challenge the model can lead to a more resilient, robust model. We also note that there is a trade-off point when fine tuning, and that it is critical to use caution to avoid catastrophic forgetting. Specifically consideration of the number and the nature of examples used is needed. This approach underscores the potential of targeted tuning as a powerful tool in refining state-of-the-art models.

## References

[1] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[2] 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. Language models are few-shot learners, 2020.

[3] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators, 2020.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[5] Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets, 2020.

[6] Max Glockner, Vered Shwartz, and Yoav Goldberg. Breaking nli systems with sentences that require simple lexical inferences, 2018.

[7] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems, 2017.

[8] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models, 2023.

[9] OpenAI. Gpt-4 technical report, 2023.