Do large language models use grammar to solve natural language tasks?

Ishan Shah [1] Isabel Papadimitriou [2] Richard Futrell [3] Kyle Mahowald [1]

[1]The University of Texas at Austin [2]Stanford University [3]University of California, Irvine

Introduction

Large language models are trained to predict the next most probable word in a sequence. Given the prompt *"I'm a student at the University of"*, OpenAI's GPT-3 found the following words most probable:

Word	Probability
Washington	6.42%
Michigan	5.74%
California	4.59%
Toronto	4.34%
Texas	3.01%
Maryland	2.34%

We are interested in exploring how models **learn grammar** while being trained on billions of pages of text [1].

Natural Language Inference

Natural Language Inference (NLI) is a task in which a model is given a premise and a hypothesis, and must predict whether the hypothesis is entailed by the premise, contradicts the premise, or is neutral. Our experiments involve two variations of this task.

MNLI — General Task

I feel like I'm letting them down, so to speak.

entails

It seems that I am disappointing them.

Figure 1. An example pulled from the MNLI dataset.

HANS — Subjecthood-Specific Task

The scientists were mentioned by the professor.

entails

The professor mentioned the scientists.

Figure 2. An example pulled from the HANS dataset.

Locating Grammatical Knowledge

We use **Iterative Nullspace Projection** (INLP) [2] to find a model's grammatical knowledge. Below is an example of INLP being used to locate and remove gender bias from a model:

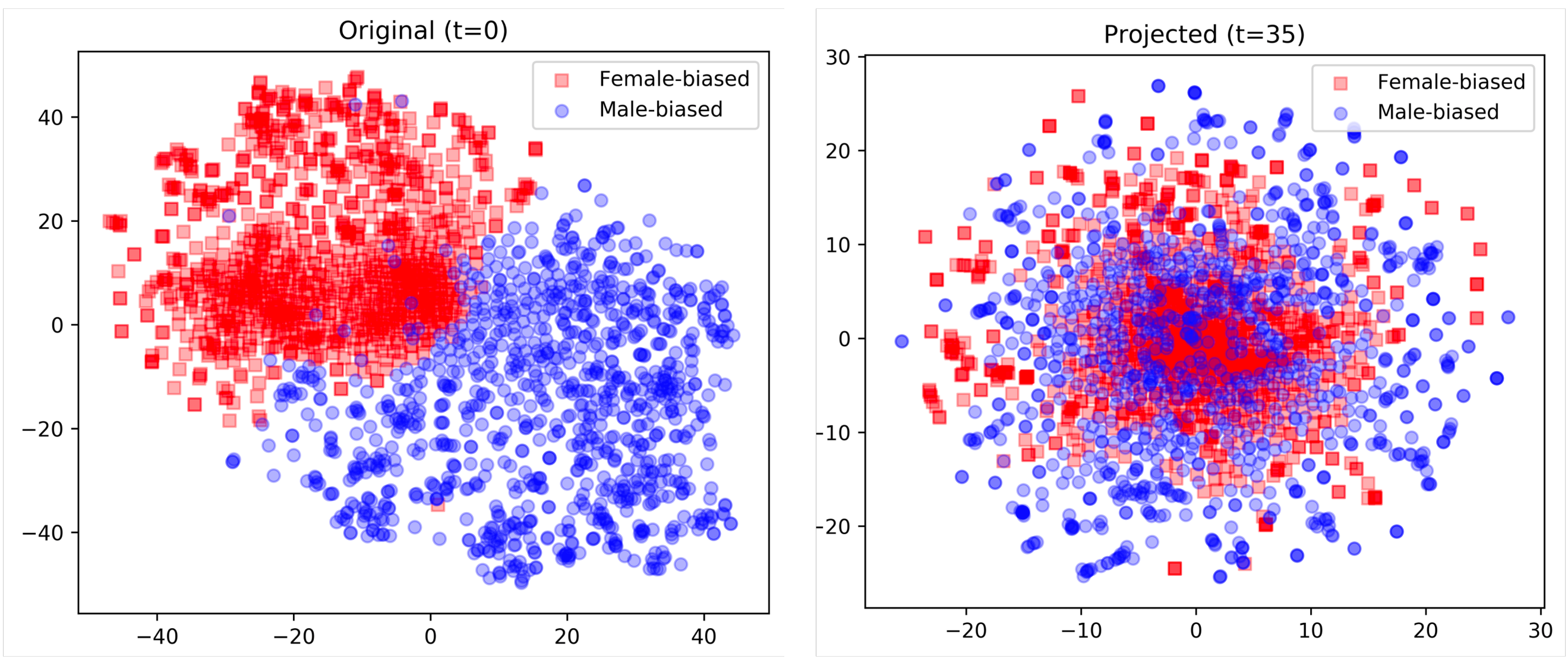

Figure 3. Gender-biased words in a language model before and after INLP (source: Ravfogel 2020).

Methods

We take a **baseline language model** and remove its grammatical knowledge using INLP to create a **nulled language model**. Both models are evaluated on two tasks:

1. **MNLI:** General language understanding task.
2. **HANS:** Subjecthood-specific language understanding task.

If language models understand grammar, performance of the nulled model should be worse in experiment 2 [3].

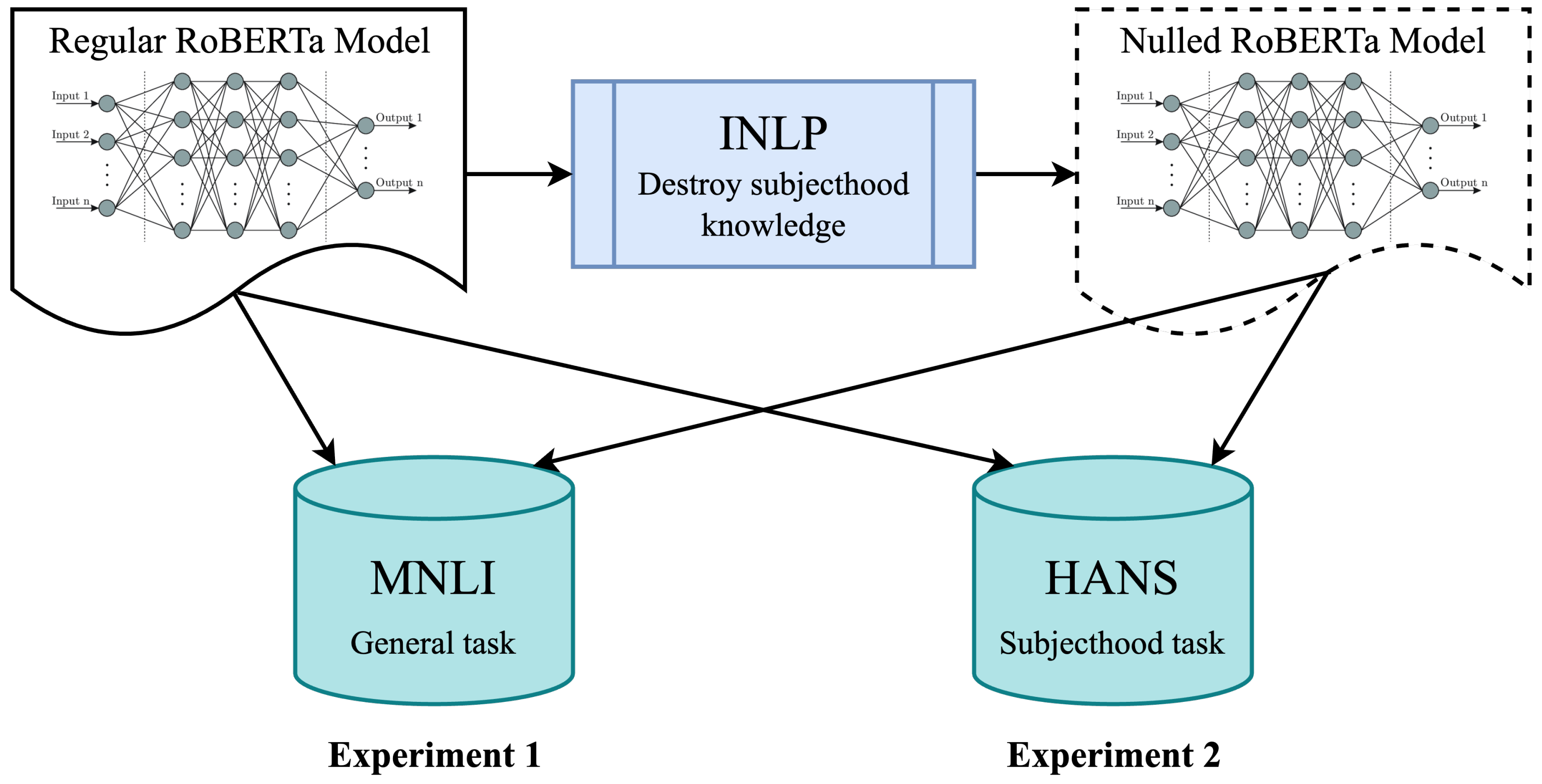

Figure 4. Overview of experimental procedure. The baseline and nulled models are tested against tasks to probe their grammatical knowledge.

It is critical to note that only ***some*** of the MNLI examples depend on subjecthood while ***all*** of the HANS examples depend on subjecthood.

Model Training

While performing INLP on a language model, we found that the model's performance on a basic subject-object classification task degraded. This gives us confidence that we are successfully removing grammatical knowledge from the model.

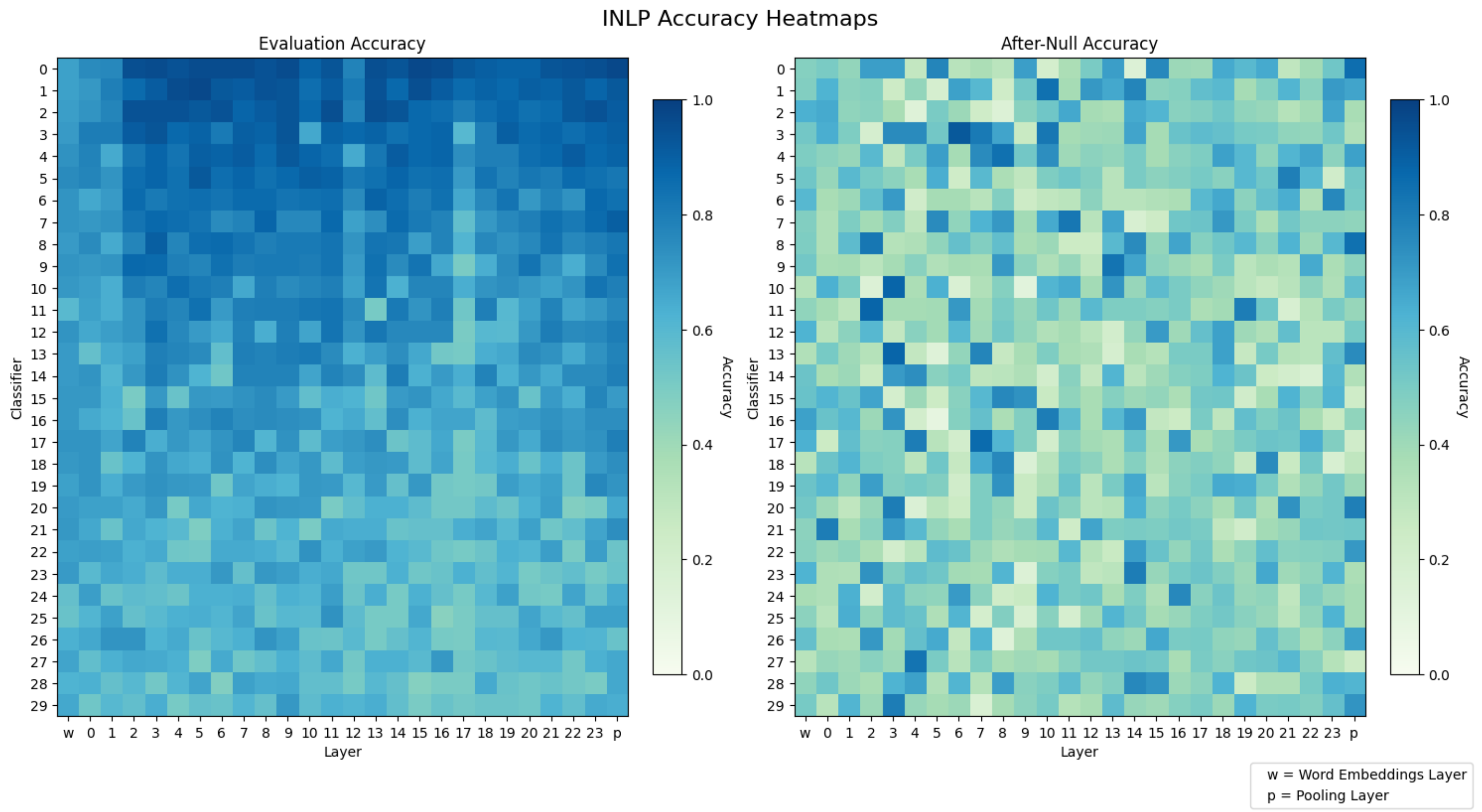

Figure 5. 30 iterations of INLP on the RoBERTa model. Accuracy drops after each iteration of nulling.

Results

We find that a language model *without* grammatical knowledge performs worse on a subjecthood-specific language task than a language model *with* grammatical knowledge. **This suggests that language models do use grammar to solve natural language tasks.**

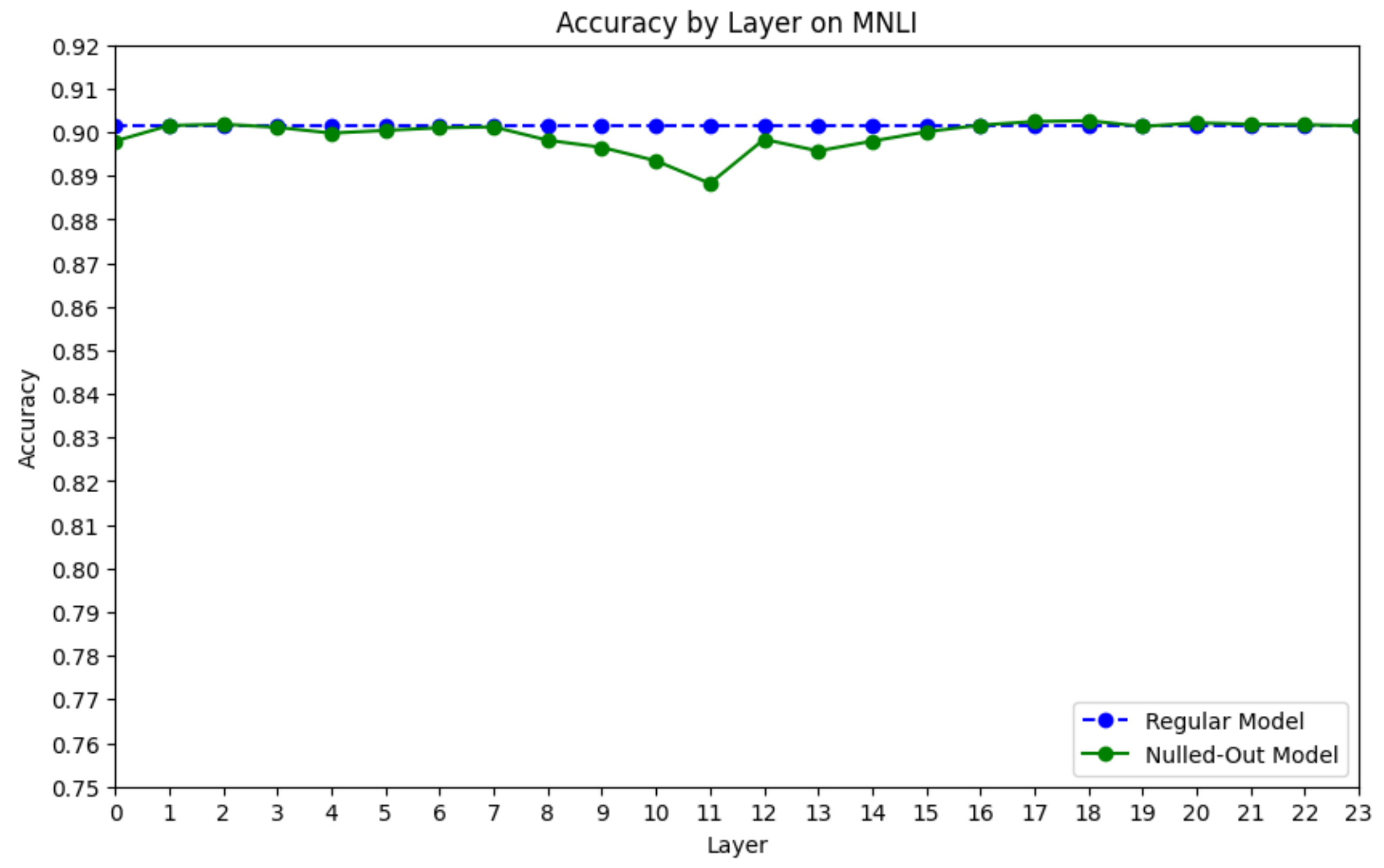

Figure 6. Accuracy on the general language task is almost the same for both models.

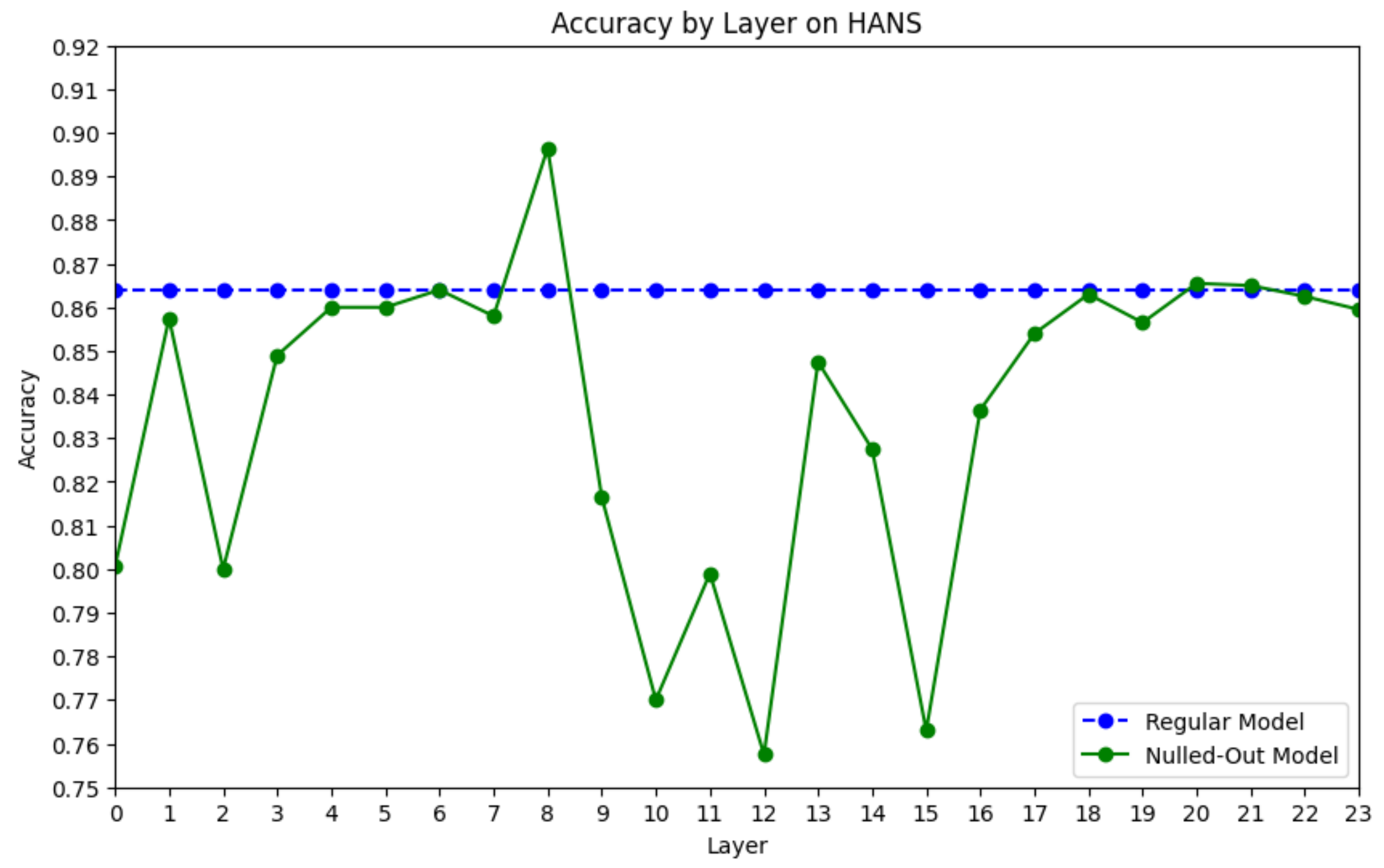

Figure 7. Accuracy on the subjecthood-specific language task is significantly lower for the nulled model.

Future Research

In the future, we are interested in exploring a number of research directions:

- Using AlterRep [4] to invert subjects and objects in a language model.
- Trying other methods of removing grammatical knowledge.
- Exploring whether these methods generalize to other language models.

Pursuing **interpretability** research is important because understanding how language models work is crucial to building safe and reliable AI systems.

References

[1] T. Linzen and M. Baroni, "Syntactic structure from deep learning," *CoRR*, vol. abs/2004.10827, 2020.

[2] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg, "Null it out: Guarding protected attributes by iterative nullspace projection," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 7237–7256, Association for Computational Linguistics, July 2020.

[3] Y. Elazar, S. Ravfogel, A. Jacovi, and Y. Goldberg, "Amnesic probing: Behavioral explanation with amnesic counterfactuals," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 160–175, 2021.

[4] S. Ravfogel, G. Prasad, T. Linzen, and Y. Goldberg, "Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction," in *Proceedings of the 25th Conference on Computational Natural Language Learning*, (Online), pp. 194–209, Association for Computational Linguistics, Nov. 2021.

https://www.ishanshah.me/ Linguistics Undergraduate Research Poster Session ishan0102@utexas.edu