



THE UNIVERSITY
OF ARIZONA



Computational
Language
Understanding

Making Language Models Robust Against Negation



MohammadHossein Rezaei
mhrezaei.com



Eduardo Blanco
eduardoblanco.github.io



NAACL 2025 (main)

Why study **negation** in LLMs?

Negation is present in **25%** of English sentences

LLMs *struggle* when **negation** is present

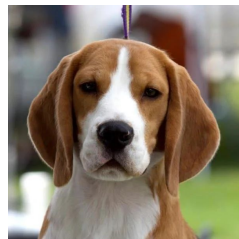
Example

A beagle is a type of [MASK]

A beagle is **not** a type of [MASK]



BERT



dog

Our Contributions

1. Introducing **two** novel **self-supervised tasks** for further pre-training LMs for **negation**
2. Creating a large-scale dataset ($\approx 6.4\text{M}$ samples) for these **tasks**
3. Experimental results with **BERT** and **RoBERTa** on **CondaQA** and **eight other** benchmarks

Pre-training Tasks



The computer screen stayed blank. It didn't display any images.

**Next Sentence
Prediction (NSP)**

The computer screen stayed blank.

+

It didn't display any images.



Reverse Polarity

The computer screen stayed blank.

+

It displayed some images.



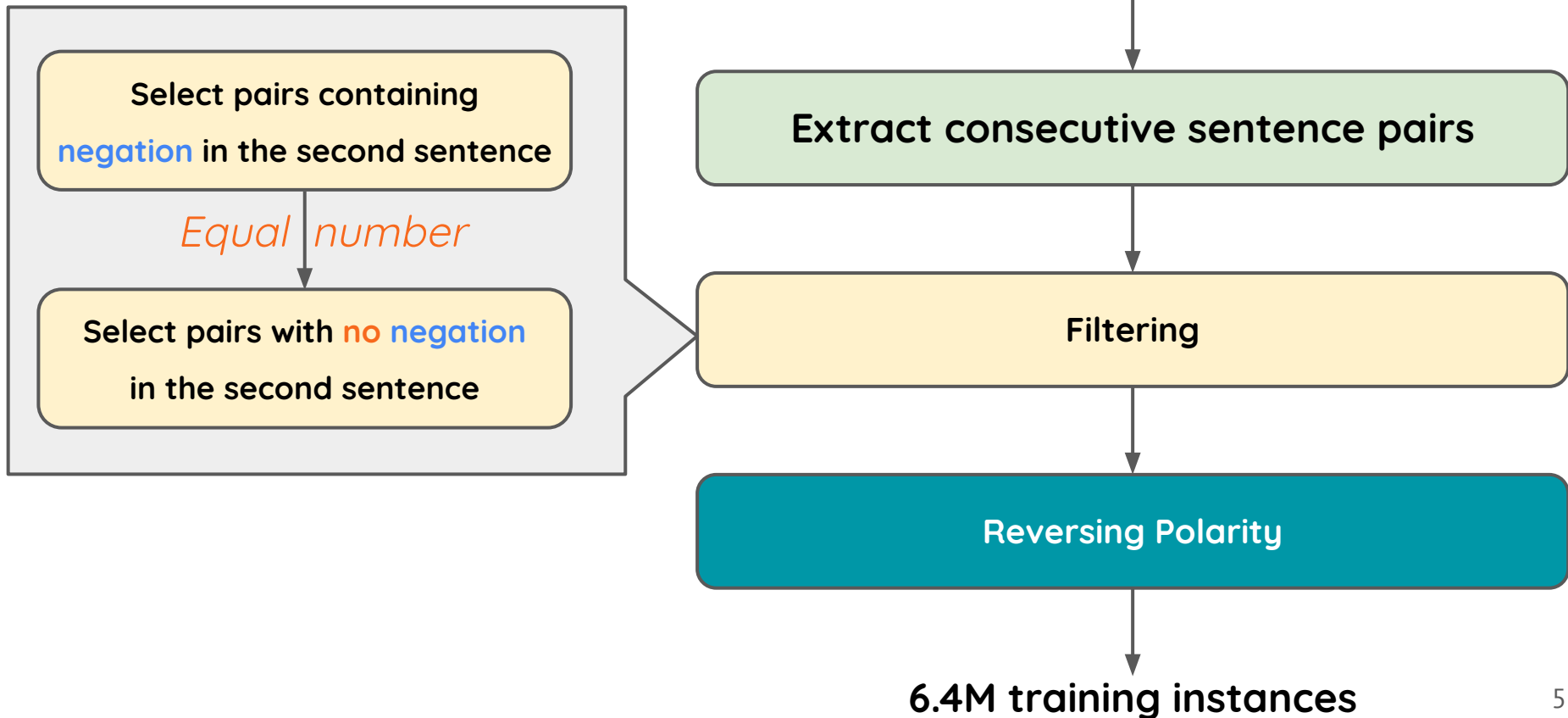
**Next Sentence Polarity
Prediction (NSPP)**

The computer screen stayed blank.

Does next sentence have negation?



Data Collection



Reversing Polarity

Definition: Remove negation if present in the sentence.
Add negation if not present.

We define **linguistic rules** to reverse polarity (§3.2.1)

🐼🧠 LLMs fail to reverse polarity due to safety guidelines!



TRUE

Large amounts of heat are wasted
when the boiler is **not** insulated

Reverse
Polarity

Large amounts of heat are wasted
when the boiler is insulated



FALSE

Experiments

1. Further pre-training **BERT** and **RoBERTa** on:
 - Next Sentence Prediction (NSP)
 - Next Sentence Polarity Prediction (NSPP)
 - Jointly on both tasks
2. **Evaluation** on:
 - CondaQA
 - NLI negation benchmarks (RTE, SNLI, MNLI)
 - NLU negation benchmarks (QNLI, WiC, WSC)
 - LAMA and LAMA-neg

CondaQA

The **largest** *question-answering* dataset requiring **reasoning** over **negation**

CondaQA contains **over 200** *unique* **negation** cues:

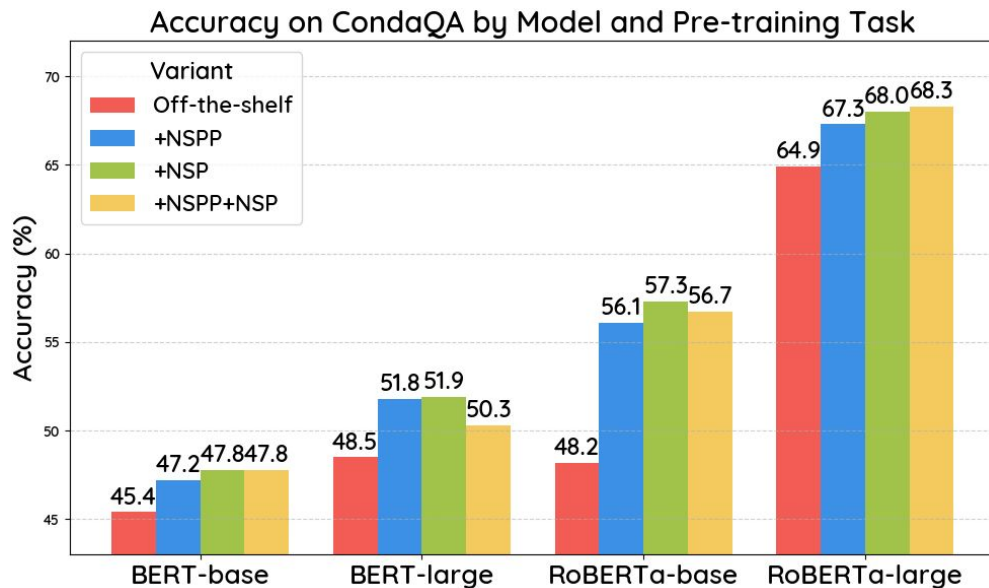
- Single-word (e.g., not, never)
- Affixal (e.g., unlucky, incorrect)
- Multi-word negation cues (e.g., a lack of, instead of)

CondaQA contains **three** types of *edits* made by crowdworkers:

- **Paraphrase**: Rewrite the **negated** sentence
- **Scope**: Change the scope of the **negation**
- **Affirmation**: Remove the **negation** from the sentence

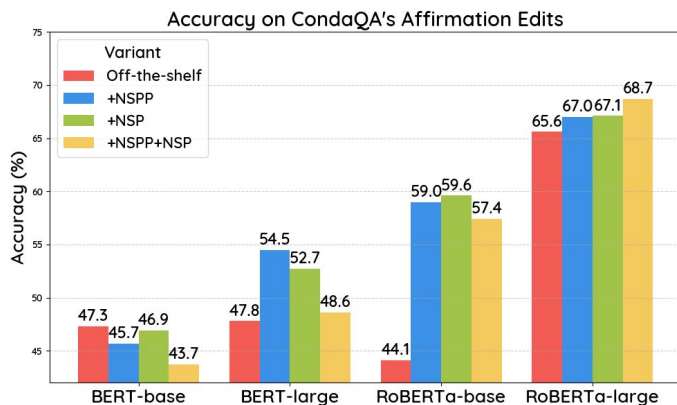
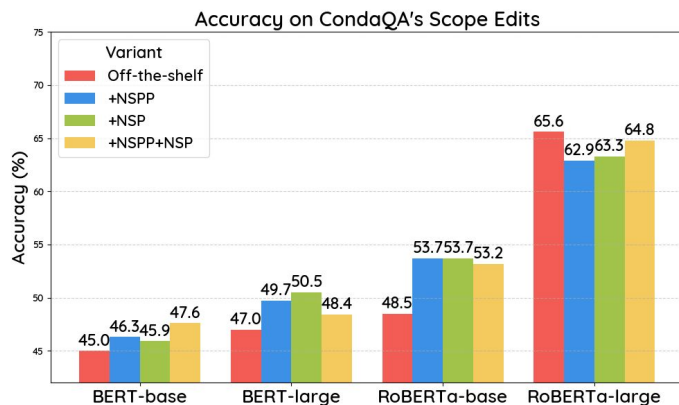
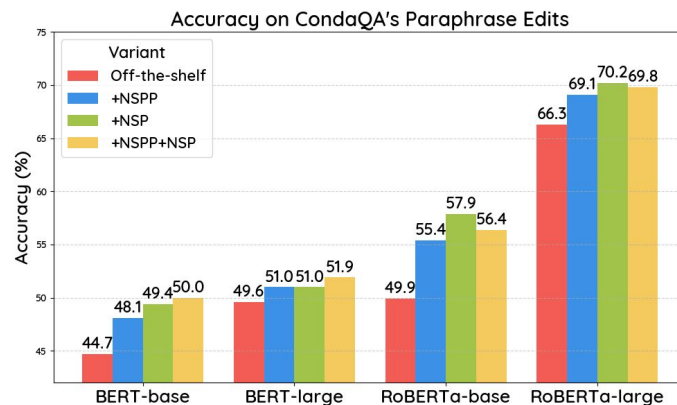
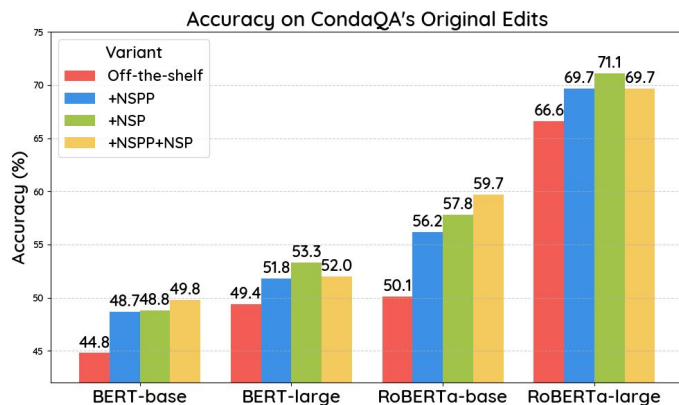
CondaQA **cannot** be solved by models relying **solely** on questions, edit types, or cues.

Results: CondaQA



Further pre-training on **any of our tasks** **statistically significantly** outperforms off-the-shelf LM

Results Breakdown: CondaQA



Key Takeaways

- Further pre-training on our tasks is beneficial
- The **NSP** task is more beneficial than the **NSPP** task
- Further pre-training on **both tasks** is *not beneficial*
- The methodology is task-agnostic

Future work

- **Expand** to *more* and *larger* models
- Use **more** *negation* cues in *pre-training*
- *Further pre-train* on **larger** corpora

Limitations

- Pre-training dataset is *only sourced* from Wikipedia
- We only use a subset of the data to pre-train
- Our linguistic rules only cover “not”, “n’t”, and “never”
- All the corpora we work with are in **English**

Thank You!

Questions:

mhrezaei@arizona.edu