MERGE: A TEST FOR NATURAL LANGUAGE INFERENCE

AUTHORS
MĂDĂLINA ZGREABĂN (PHD CANDIDATE)
TEJASWINI DEOSKAR
LASHA ABZIANIDZE (PI)

AFFILIATIONS
UTRECHT UNIVERSITY
INSTITUTE FOR LANGUAGE SCIENCES (ILS)

NATURAL LANGUAGE INFERENCE (NLI):
Does the premise entail the hypothesis?

P: A GIRL JUMPS IN THE AIR. H: A GIRL JUMPS HIGH.	A BOY JUMPS IN THE AIR. A BOY JUMPS HIGH.

Underlying reasoning: neutral

INTRODUCTION

GENERAZABILITY makes language models adapt their knowledge to new situations (Dutt et al., 2024, Hupkes et al., 2023, Yang et al., 2023).

Evaluating models on data slightly different than their training data (*out-of-distribution*, **OOD data**):

- proves important, as models might learn heuristics to get higher scores on in-distribution data (Dutt et al., 2024);
- results in decreased performance (Li et al., 2020; Petrov, 2025; Gardner et al., 2020; Kaushik et al., 2020);
- indicates a lack of generalization capacity.

However, testing on **OOD data:**

- **costly,** if OOD datasets are formed manually;
- **unfair**, if the data is not similar enough to the training data.

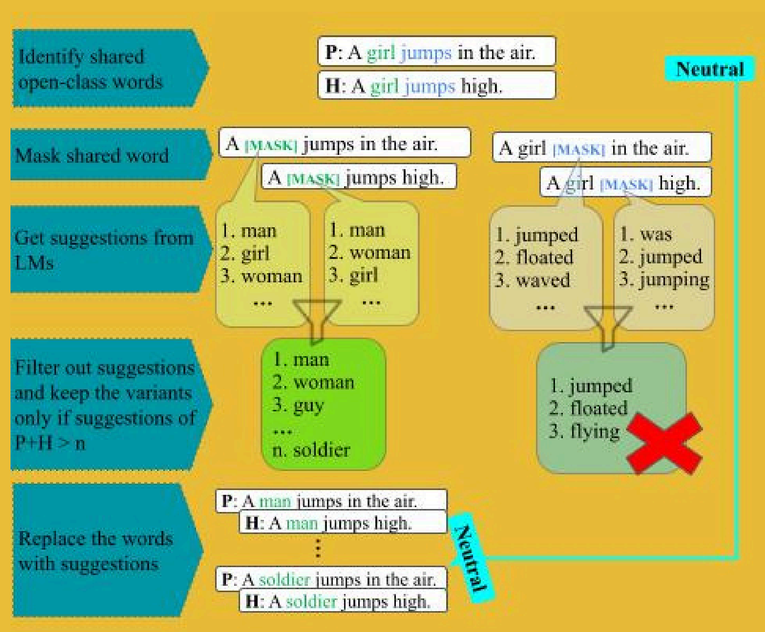
OBJECTIVE

The Minimal Expression Replacement GEneralization (**MERGE**) test **automatically** alters existing NLI datasets, while keeping their underlying reasoning, with **minimal** changes.

RESEARCH QUESTIONS:

- Do language models have decreased performance on evaluation tests obtained from MERGE?

DATASET CREATION

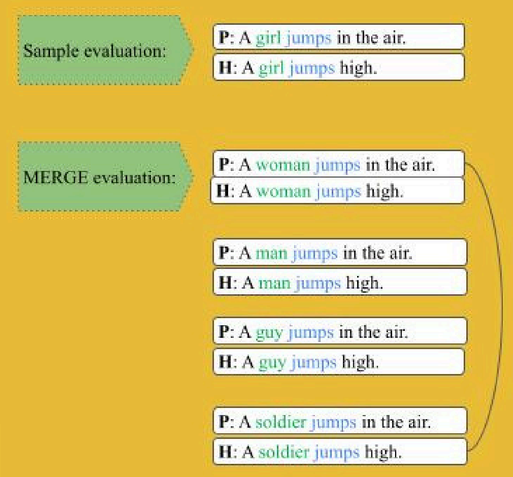


METHODOLOGY DATASET CREATION

- masked shared words and generated 200 suggestions for P and H occurrences with bert-base-cased and roberta-base;
- used en_core_web_sm to tag suggestions replaced in P and H;
- filtered out suggestions that were: punctuation signs, derivational morphemes, had a different pos tag or lower probability than the original replaced word;
- if filtered suggestions < 20, the original word was not replaced anymore.
- randomly sampled 20 suggestions for each replaced word, and replaced them in P and H, 10 times, creating a bigger ALL (R10) dataset with 10 mini-datasets.

Word	Seed	Var	N	C	E	R10
N	3706	534592	167244	120845	246503	74120
V	1130	126623	35591	21034	69998	22600
Adv	10	446	75	63	308	200
Adj	281	22429	7302	5043	10084	5620
All	**4474**	**684090**	**210212**	**147985**	**326893**	102540

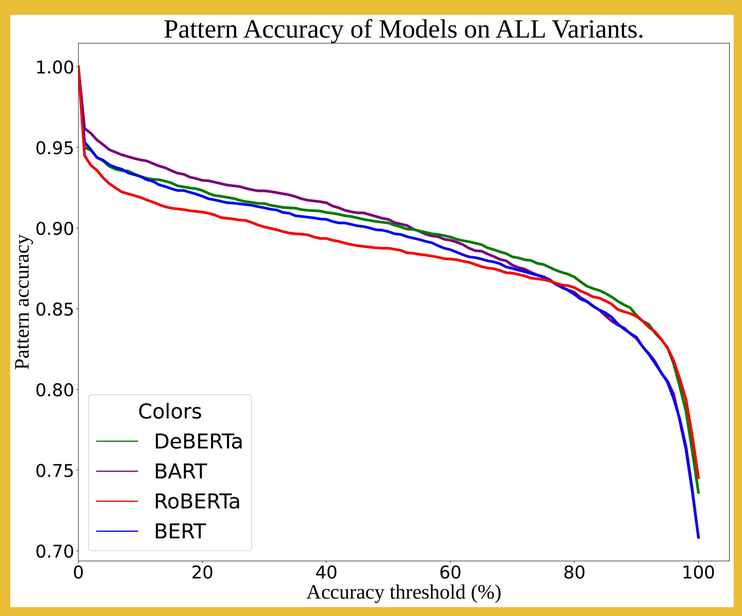
MODEL EVALUATION

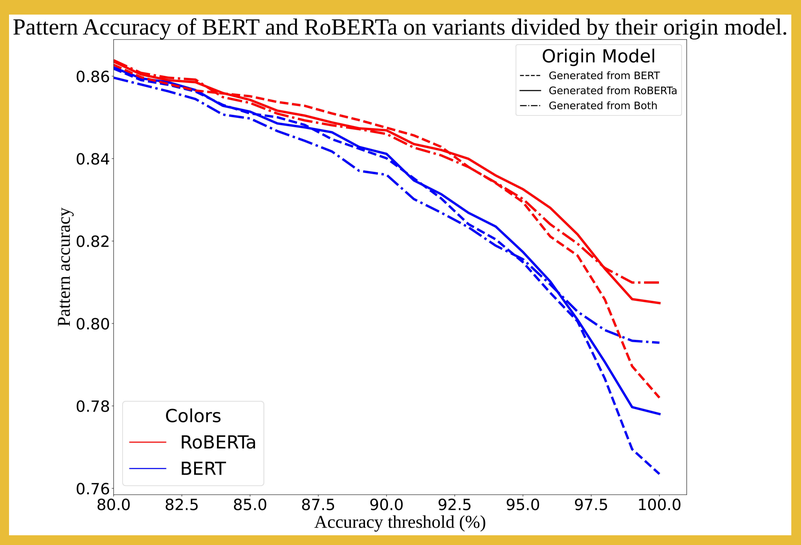


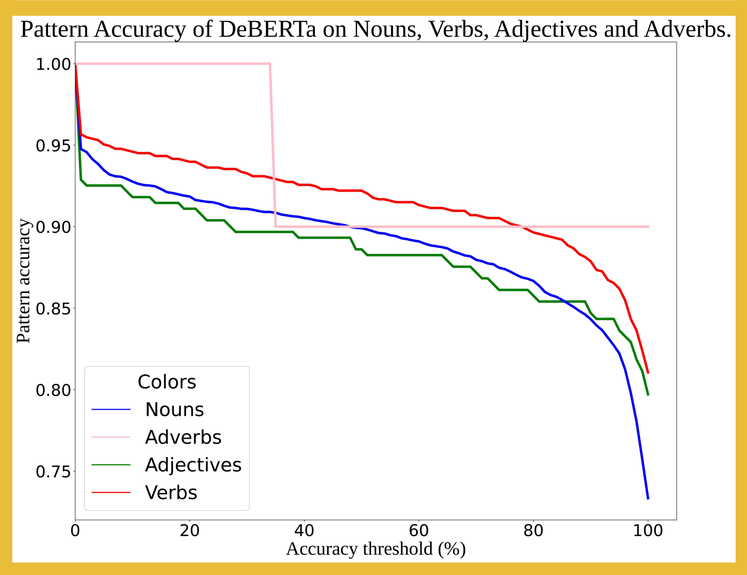
METHODOLOGY MODEL EVALUATION

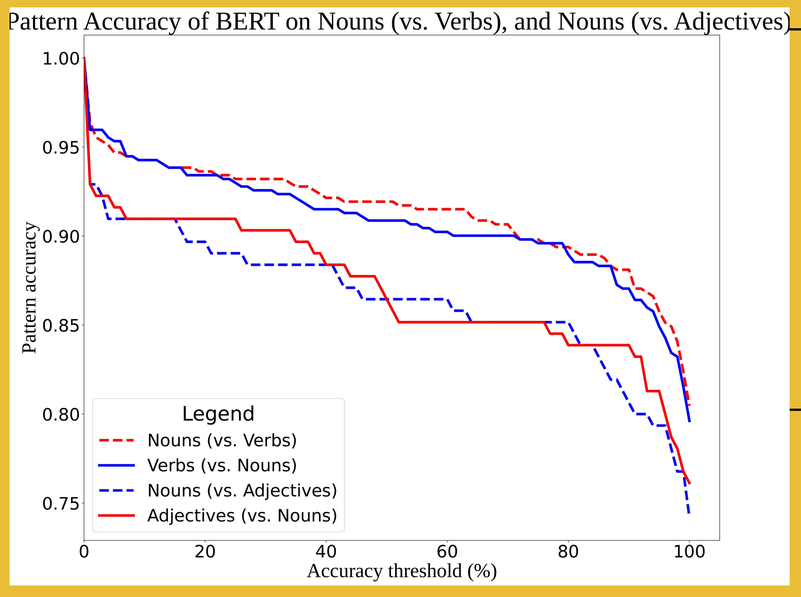
- evaluated textattack/bert-base-uncased-snli, ynie/bart-large-snli_mnli_fever_anli_R1_R2_R3-nli, pepa/deberta-base-snli and pepa/roberta-base-snli on ALL dataset, ALL dataset split into open-class categories (nouns, verbs, adjectives, adverbs); and ALL dataset divided by model used to generate suggestions (BERT, RoBERTa, or both).
- calculated Sample Accuracy (a correct prediction is when a variant is predicted correctly) and Pattern Accuracy (a correct prediction is when the model gets an x amount of variants correctly of the same seed problem).

RESULTS









Word	Dataset	BERT PA	RoBERTa PA	BART PA	DeBERTa PA
All	Seed	90.25	89.87	91.86	91.39

Table 3: PA scores of models on the seed dataset containing all problems. Note that the PA scores for Seed datasets equal their SA scores.

- Compared with PA scores on the seed dataset, PA scores of models on ALL variants indicate a decreased performance on variant datasets.
- Nouns are more difficult when we consider the ALL dataset, but when we take an equal amount of seed problems, verbs are more difficult, followed by nouns, and adjectives.
- Models perform best on suggestions that were common to both BERT and RoBERTa, followed by suggestions from RoBERTa, on higher threshold accuracies.

CONCLUSION

Our results suggested models lack the generalization ability to perform well on the task, in line with previous studies (Li et al., 2020; Petrov, 2025; Gardner et al., 2020; Kaushik et al., 2020).

Our results suggest that getting suggestions from more models might be beneficial for the evaluated models, and also that certain word classes might pose more difficulty for the models.

References