Minimal **E**xpression **R**eplacement **GE**neralization test for NLI

5th NALOMA@ESSLLI, 5 August 2025

Mădălina Zgreabăn, Tejaswini Deoskar, Lasha Abzianidze

Universiteit Utrecht

NLI task is popular, but…

We know the definition of the NLI task: $< P, H, l >$.

A very popular task: over 100 NLI datasets exist

One of the reasons of its popularity is being an easy task on reasoning:

- It is a three-way classification task.
- Simple/silly heuristics work due to annotation artifacts.
  - Hypothesis-only bias (Gururangan et al. 2018, Poliak et al. 2018, Tsuchiya 2018)
  - Word overlap bias (McCoy et al. 2019, Naik et al., 2018, Glockner et al. 2018)
  - Inverse word overlap bias (Rajaee et al, 2022)
  - Negation/antonymy bias (Lai&Hockenmaier 2014, Naik et al., 2018)

Not every **dog** without a collar is barking **loudly**	
E	Some **animal** without a red collar is not barking

Generalization & NLI

Breaking NLI (Glockner et al. 2018):

The man is holding a saxophone ⇨

	The man is holding a saxophone
C	The man is holding an electric guitar

HANS (McCoy et al. 2019):

	The lawyer near the actor ran
NE	The actor ran

IMPPRESS (Jeretic et al., 2020):

	Jo ate some of the cake
E	Jo didn’t eat all of the cake

PaRTE (Verma et al. 2023): $< P, H, l >$ ⇨ $< Para(P), Para(H), l >$

Generalization & NLI: but…

The datasets for generalization evaluation often have an adversarial nature:

small string edit with label change

Several elements involved in new tests, which makes it difficult to single out the reason of poor performance:

label, syntax, and sentence length change

Requires manual work:

writing templates

validating the generated NLI problems


MERGE
test

MERGE test

MERGE: Seed problem-based evaluation

Pattern accuracy (PA) with a threshold

$Acc_{th=0.5} = 1$

$Acc_{th=0.75} = 1$

$Acc_{th=0.95} = 0$

Sample-based evaluation

Sample/variant accuracy (SA)

$Acc_v = 0.75$

Original/seed NLI problem

P: A small girl carries a girl. — **E**

H: There is a small girl.

$\mathcal{M}_1, \ldots \mathcal{M}_n$ **Automatic generation of variants with MLMs**

NLI model's predictions

✓ **P**: A **small boy** carries a **boy**. **H**: There is a **small boy**. — **E** / **E**

⋮

✓ **P**: A **small dog** carries a **dog**. **H**: There is a **small dog**. — **E** / **E**

⋮

✗ **P**: A **little girl** carries a **girl**. **H**: There is a **little girl**. — **E** / **N**

⋮

✓ **P**: A **happy girl** carries a **girl**. **H**: There is a **happy girl**. — **E** / **E**

Precaution!

Certain minimal expression replacements can lead to unsound NLI problems:

P: Two dogs and three **boys** swim.
H: Only three **boys** swim.
E

boys/dogs →

P: Two dogs and three **dogs** swim.
H: Only three **dogs** swim.
E ✗

Don't replace original words with co-occurring words!

Are we good? Not really:

P: Two dogs and three **boys** swim.
H: Only three **boys** swim.
E

boys/animals →

P: Two dogs and three **animals** swim.
H: Only three **animals** swim.
E ✗

Don't replace with words being in semantic relation with co-occurring ones!

Are we good? Not really:

P: Two dogs and three **animals** swim.
H: Only three **animals** swim.
C

animals/boys →

P: Two dogs and three **boys** swim.
H: Only three **boys** swim.
C ✗

Minimality of MERGE

Variant problems require the exact same reasoning as the original/seed problems:

P: A small **girl** carries a **girl**.
H: There is a small **girl**.
E

P: A small **boy** carries a **boy**.
H: There is a small **boy**.
E

The sort of minimal string edits:

P: A **blond** **boy** carries a **boy**.
H: There is a **blond** **boy**.
E

Many biases are preserved:

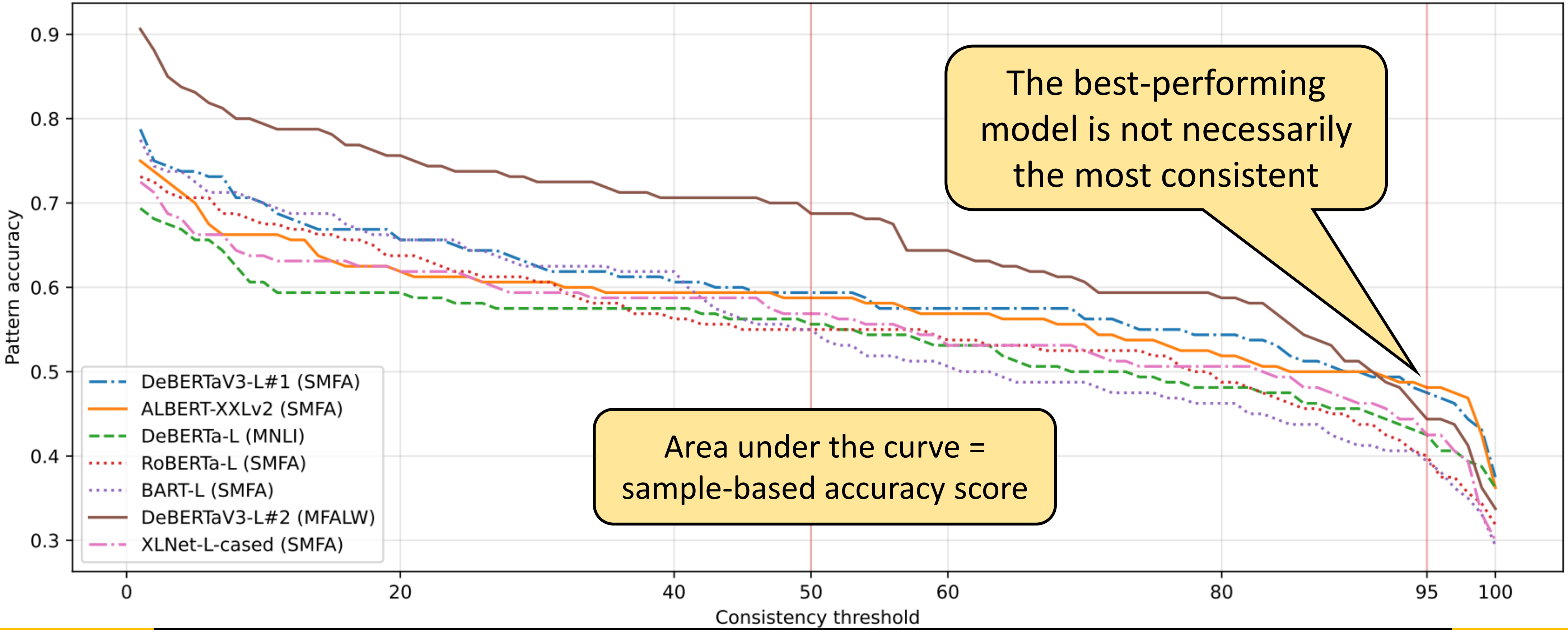
- The (reverse) word overlap — We replace single words with single words
- Negation/antonymy — Antonyms are different words; hence they remain
- Hypothesis only — Usually, give-away words only occurs in a hypothesis

Pattern/seed-based evaluation

Abzianidze et al. (2023)

Inspired by **SpaceNLI**: each pattern has n-number of samples

Similar to the idea behind ROC curve



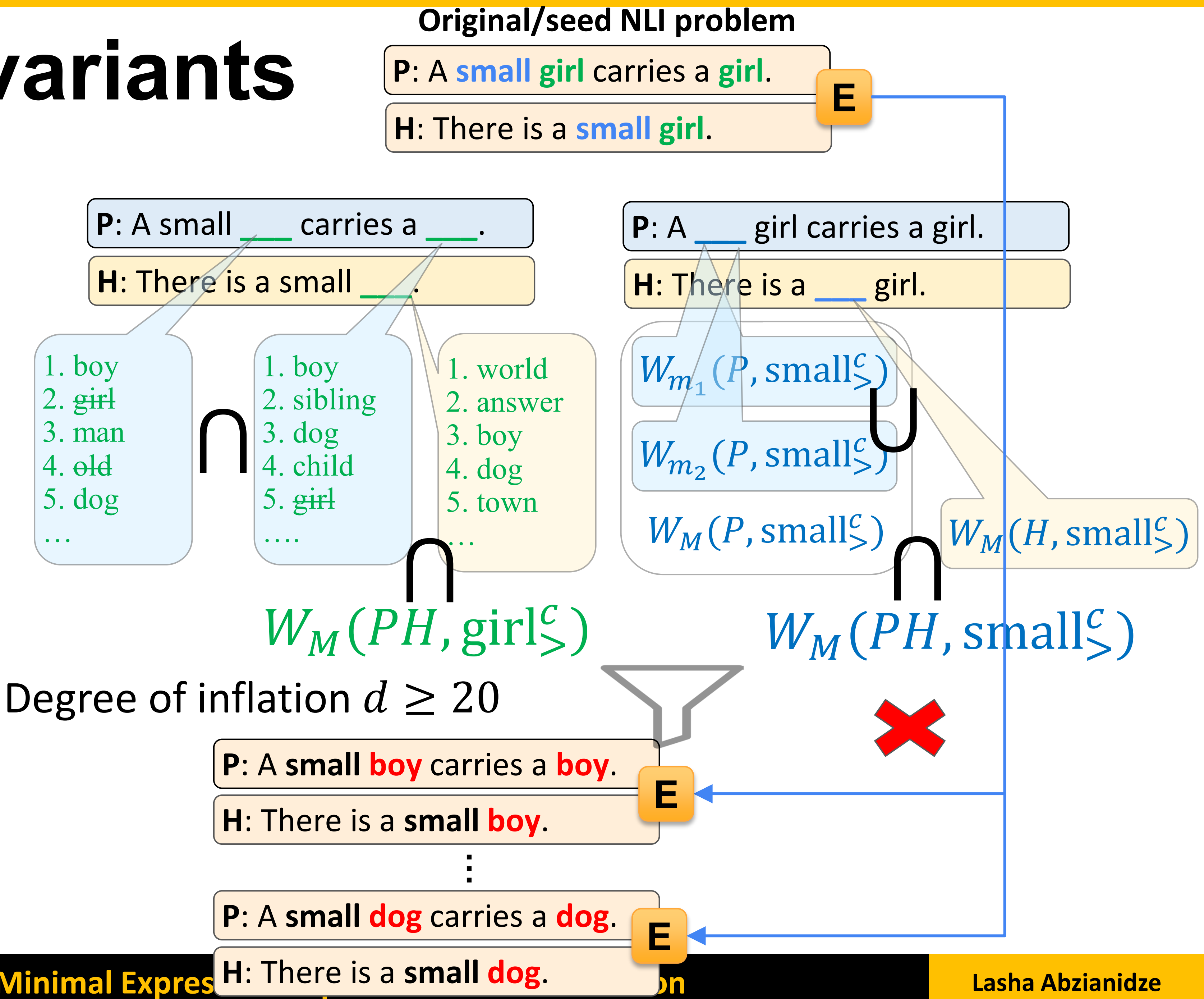
Generating variants

Identify and MASK shared words

Get fillers from MLMs $M = \{m_1, m_2\}$ and distill suggestions

Skip seed problems with insufficient inflation

Creating NLI problem variants



Generating variants (2)

Suggested words $W_M(PH, w^c_>)$ are such that:

- They differ from the co-occurring words in an NLI problem PH.
- At least one MLM from M suggests it and validates it, i.e., gives it a higher probability ($>$) than the original word.
- They get the same word class c tag as the original word.
- They are suggested for both premise P and hypothesis H.

If w is not in the tokenizer vocabulary of a MLM, then the suggestion set is empty, e.g., $W_M(PH, \mathrm{mentorship}^c_>) = \emptyset$

Setup of experiments

Masked Language Models (MLMs) used:

Roberta-base & Bert-base (both cased)

The test part of the Stanford NLI dataset:

~10K problems

Suffering from the hypothesis-only bias

For sufficient number of seeds: $W_M(PH, w^{C}_{>}) = W_M(P, w^{C}_{>}) \cup W_M(H, w^{C}_{>})$

Several NLI models:

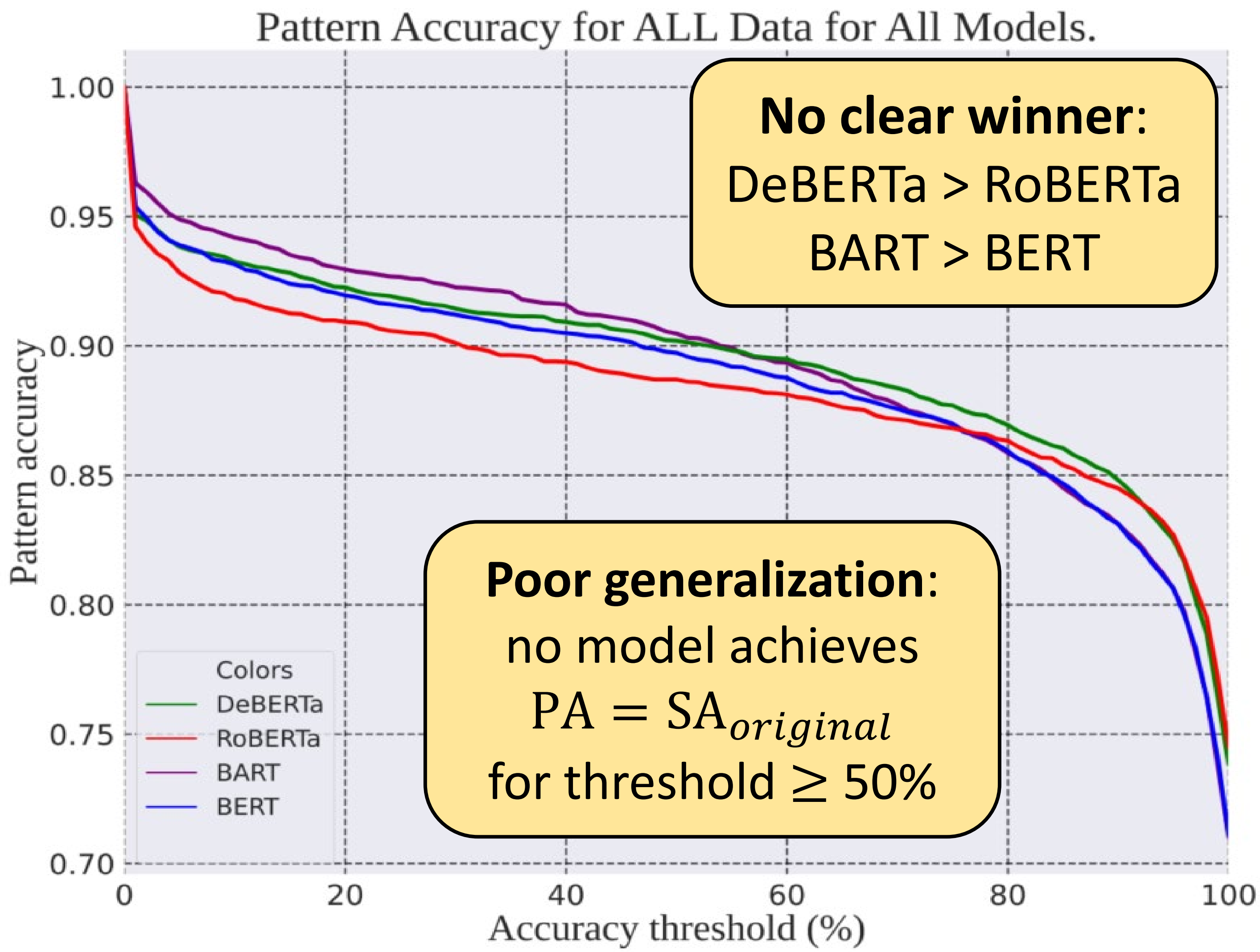
Model	Training set	SNLI test
BERT	SNLI train	90.48
RoBERTa	SNLI train	90.06
DeBERTa	SNLI train	91.70
BART	SNLI train + MNLI, FEVER-NLI, ANLI	**92.03**

Sample & pattern accuracy (PA) scores

Sample accuracy (SA) drops for the variants compared to the seed problems.

Model	Training set	SNLI test (~10K)	All seed (~4.5K)	Δ All variants (~102K)
BERT	S	90.48	90.24	-1.52
RoBERTa	S	90.06	89.86	-1.36
DeBERTa	S	91.70	91.38	**-1.97**
BART	SMFA	**92.03**	**91.85**	-2.74

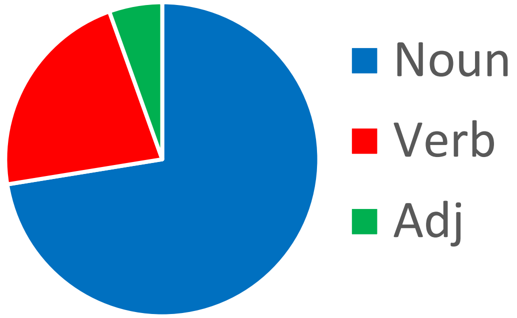
$SA_{original}$



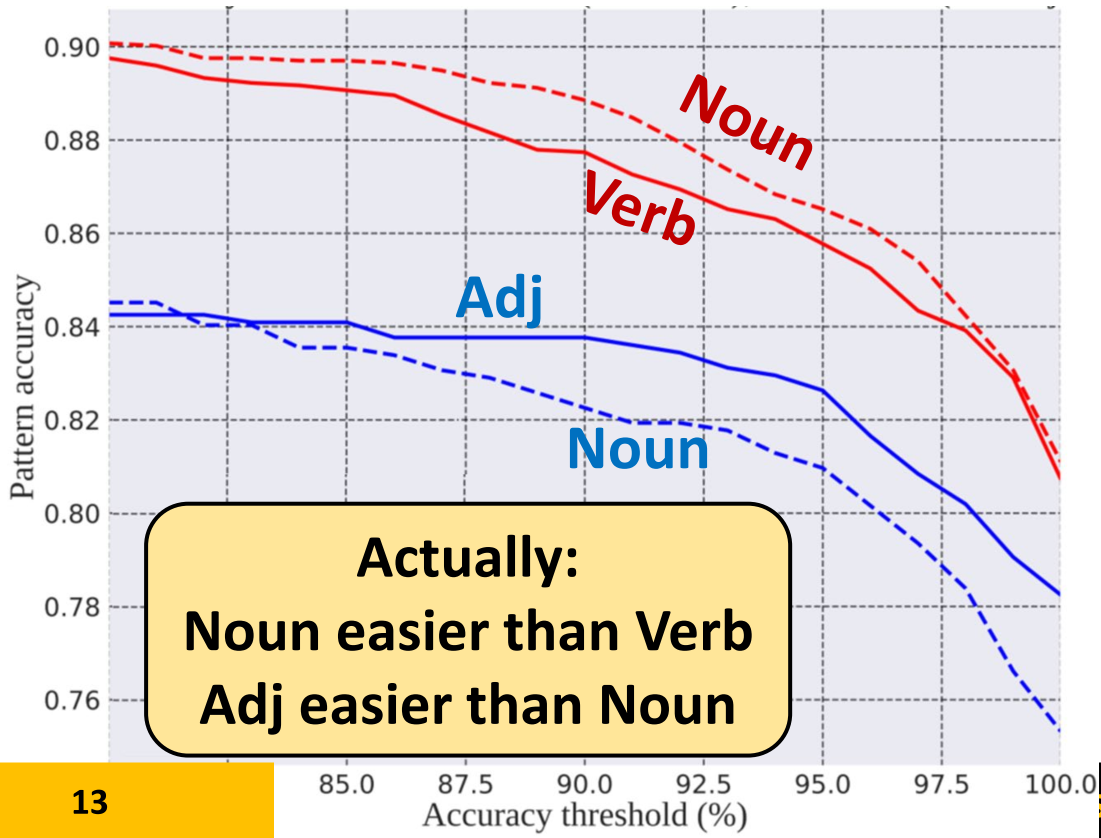
No clear winner: DeBERTa > RoBERTa BART > BERT

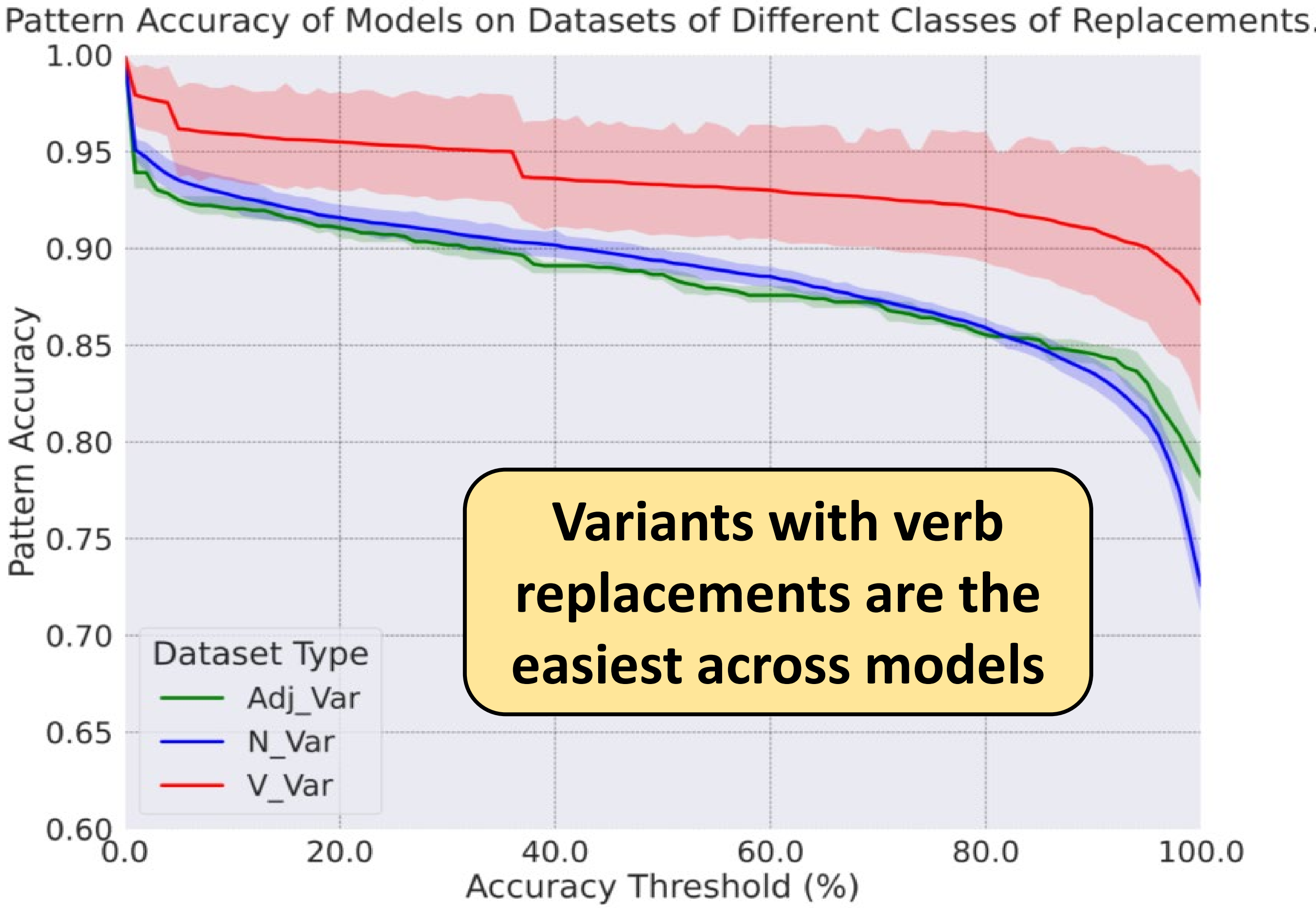
Poor generalization: no model achieves $PA = SA_{original}$ for threshold ≥ 50%

Easiest word classes



Removing the effect of different seed NLI problems, i.e. comparing on the same seed problems:





Do MLMs favor native NLI models?

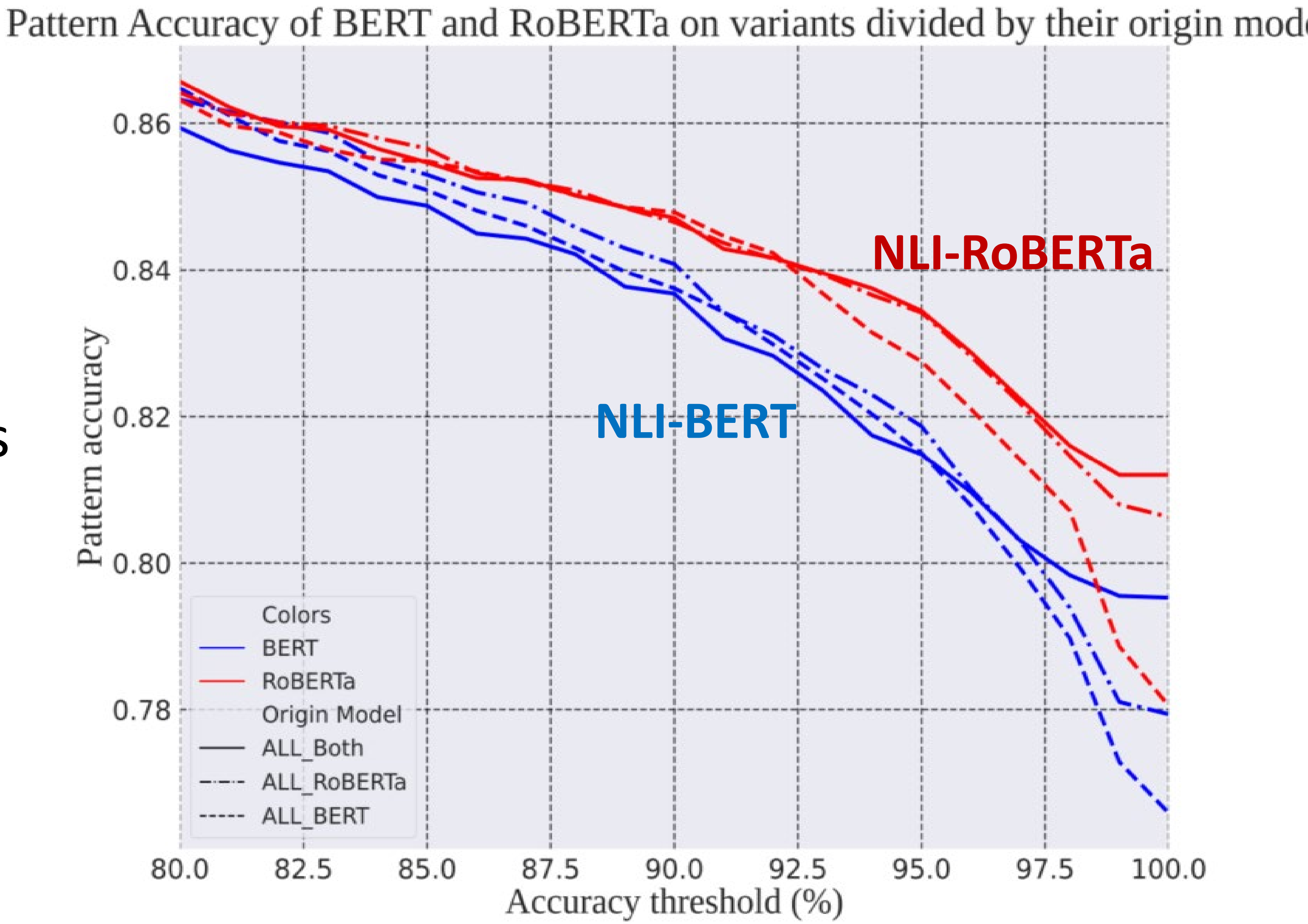
If there is any favoritism, it can be seen at the extreme th>97%.

However, MLMs do not favor native NLI models.

Easiest: shared suggestions

Easier: RoBERTa

Least easy: BERT



Conclusion

MERGE test:

- Auto generating sample variants with MLMs
- Most friendly generalization test
  - Maintains the underlying reasoning
  - Preserves the biases of the original samples

Increases to 60% when suggestion words are originating from both P and H

Models cannot maintain the same accuracy even for threshold of 50%.

Replacements with the easiest word classes: Adj, Noun, Verb.

No observable favoritism of NLI models from the native MLMs.

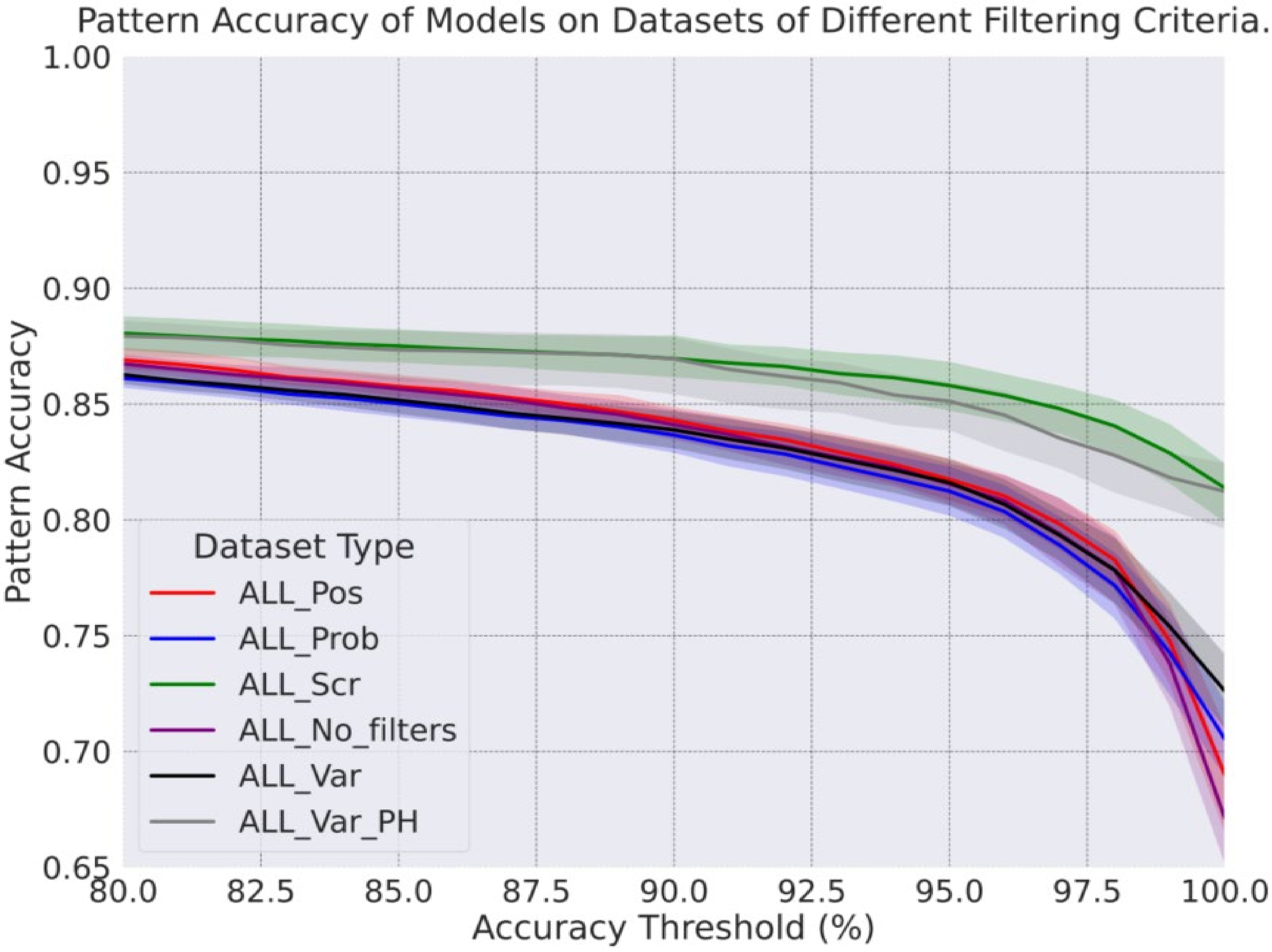
Future work might involve more NLI models, MLMs, and NLI datasets for stronger results.

Additional slides

Comparing various replacements

Variants obtained with cleaner replacements are easier.

However, variants obtained with replacements being non-existent words (e.g., scrambled characters) are also easier.



Contrasting PA across models