Utrecht University

MERGE: MINIMAL EXPRESSION REPLACEMENT GENERALIZATION TEST

MĂDĂLINA ZGREABĂN TEJASWINI DEOSKAR LASHA ABZIANIDZE

NLI BENCHMARKS

- Disturb lexical overlap heuristic of premise and hypothesis (*PH*);
- Have low lexical diversity;
- Costly, if formed manually;
- Syntax non-preserving;
- Unfair, if the data is not similar enough to the training data.

MERGE

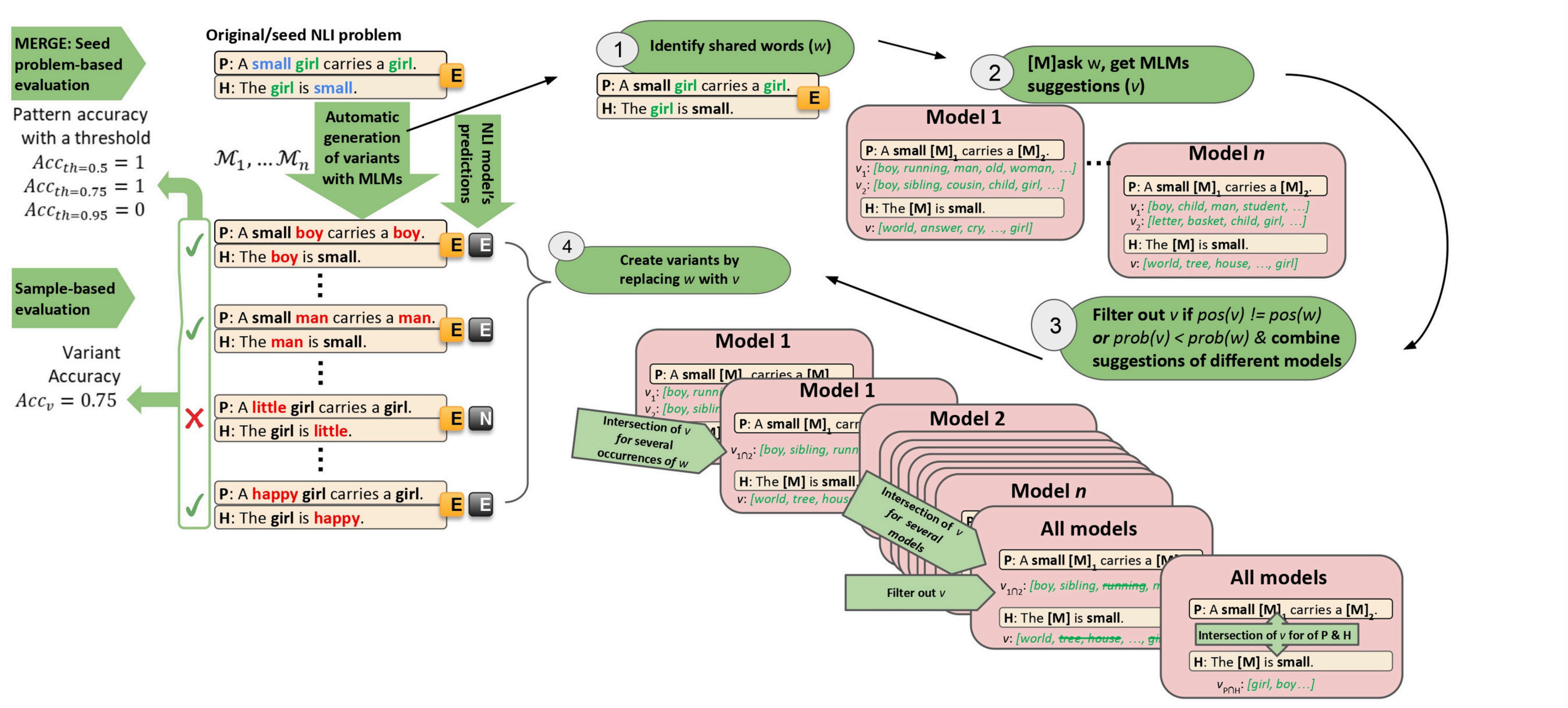
- Minimally alters already existing NLI datasets;
- Preserves underlying logical reasoning;
- Does not require human validation by strict minimal changing criteria;
- Preserves lexical overlap;
- Can add more lexical diversity by adding suggestions from other models;
- Automatic;
- Syntax preserving.

RESEARCH QUESTIONS

ARE LANGUAGE MODELS ROBUST AGAINST MINIMAL **VARIANTS** OF NLI PROBLEMS?

DO THE LIKELIHOOD, POS TAG, PLAUSIBILITY, OR MASKED MODELS MATTER?

FRAMEWORK



MERGE: Seed problem-based evaluation

Pattern accuracy with a threshold

$Acc_{th=0.5} = 1$

$Acc_{th=0.75} = 1$

$Acc_{th=0.95} = 0$

Sample-based evaluation

Variant Accuracy

$Acc_v = 0.75$

Original/seed NLI problem

P: A small girl carries a girl. E

H: The girl is small.

$\mathcal{M}_1, \dots \mathcal{M}_n$

Automatic generation of variants with MLMs

NLI model's predictions

✓ P: A small boy carries a boy. E E

H: The boy is small.

✓ P: A small man carries a man. E E

H: The man is small.

✗ P: A little girl carries a girl. E N

H: The girl is little.

✓ P: A happy girl carries a girl. E E

H: The girl is happy.

1 Identify shared words (*w*)

P: A small girl carries a girl. E

H: The girl is small.

2 [M]ask w, get MLMs suggestions (*v*)

Model 1

P: A small $[M]_1$ carries a $[M]_2$.

v_1: *[boy, running, man, old, woman, …]*

v_2: *[boy, sibling, cousin, child, girl, …]*

H: The [M] is small.

v: *[world, answer, cry, …, girl]*

Model *n*

P: A small $[M]_1$ carries a $[M]_2$.

v_1: *[boy, child, man, student, …]*

v_2: *[letter, basket, child, girl, …]*

H: The [M] is small.

v: *[world, tree, house, …, girl]*

3 Filter out *v* if *pos(v) != pos(w) or prob(v) < prob(w)* & combine suggestions of different models

4 Create variants by replacing *w* with *v*

Model 1

P: A small [M]₁ carries a [M]

v_1: [boy, runni…

v_2: [boy, siblin…

Intersection of *v* for several occurrences of *w*

Model 1

P: A small $[M]_1$ carr…

$v_{1\cap 2}$: *[boy, sibling, runn…*

H: The [M] is small.

v: *[world, tree, hous…*

Model 2

Model *n*

Intersection of *v* for several models

Filter out *v*

All models

P: A small $[M]_1$ carries a [M…

$v_{1\cap 2}$: *[boy, sibling, ~~running~~, m…*

H: The [M] is small.

v: *[world, ~~tree, house~~, …, gi…*

All models

P: A small $[M]_1$ carries a $[M]_2$.

Intersection of *v* for of P & H

H: The [M] is small.

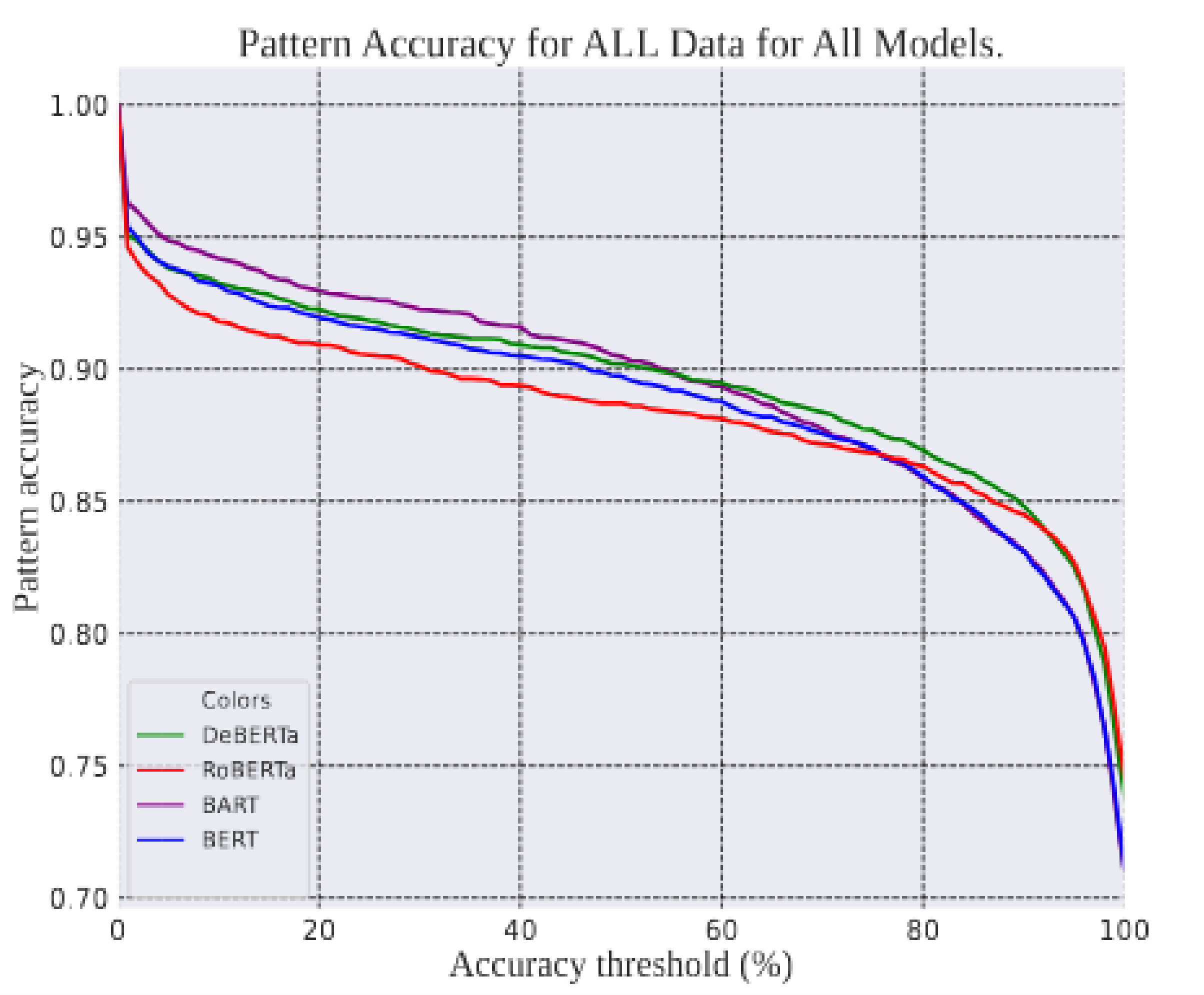
$v_{P\cap H}$: *[girl, boy…]*

METHODOLOGY

- 200 suggestions (v) with bert-base-cased and roberta-base;
- Suggestions tagged with en_core_web_sm;
- Exclude punctuation signs, derviational morephems, different POS(*v*), and lower probability(*v*).
- Required variants ==20;
- 10 random mini-datasets with 20 variants per problem (ALL_Var).
- Evaluate BERT, BART, DeBERTa, RoBERTa.

THE LIKELIHOOD, POS TAG, PLUASIBILITY & MLMS DO MATTER

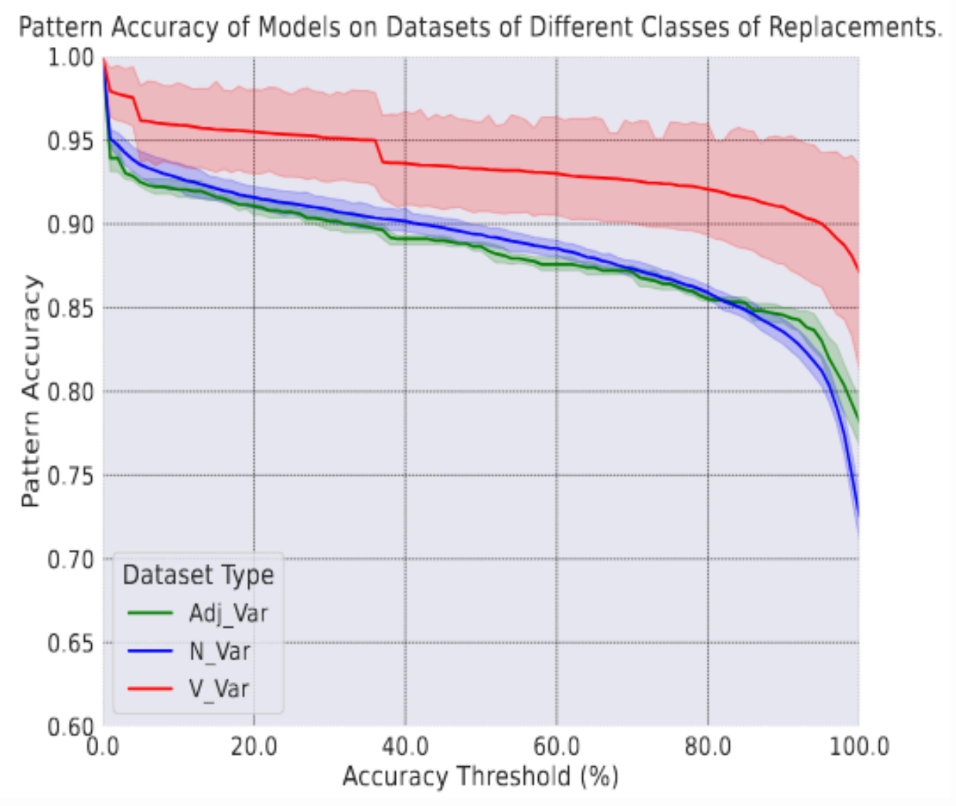
LMS ARE NOT ROBUST AGAINTS MINIMAL VARIANTS

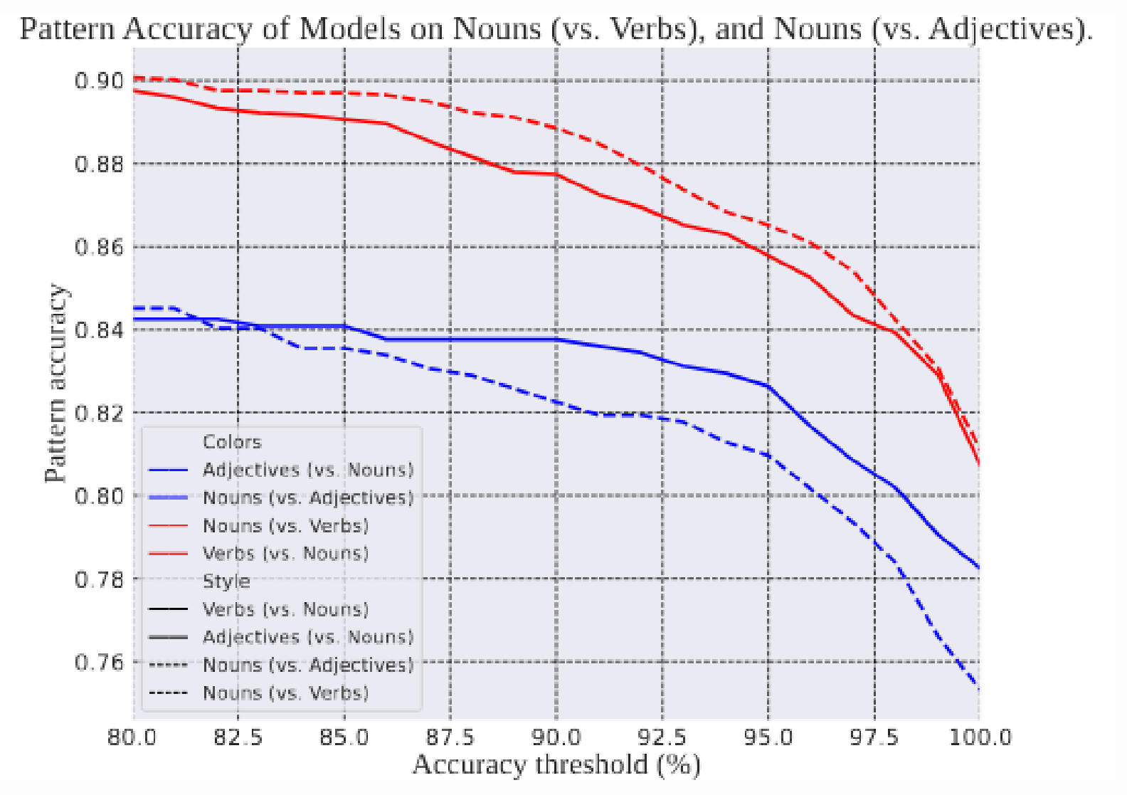


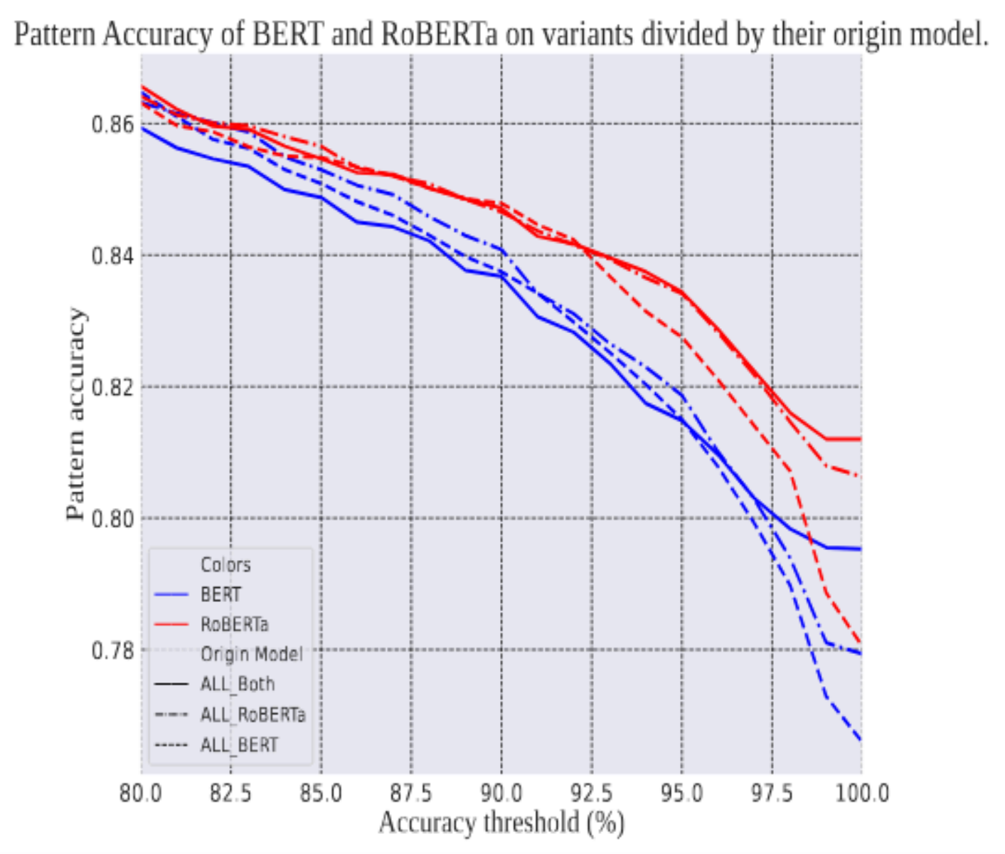
Word	Seed	Average	N(%)	C(%)	E(%)	Subs
N_{Var}	3704	144.2	12.5	22.6	46.1	74080
V_{Var}	1129	112	28.1	16.6	55.2	22580
Adj_{Var}	280	79.9	32.5	22.5	44.8	5620
ALL_{Var}	**4468**	**152.8**	**30.7**	**21.4**	**47.7**	102280

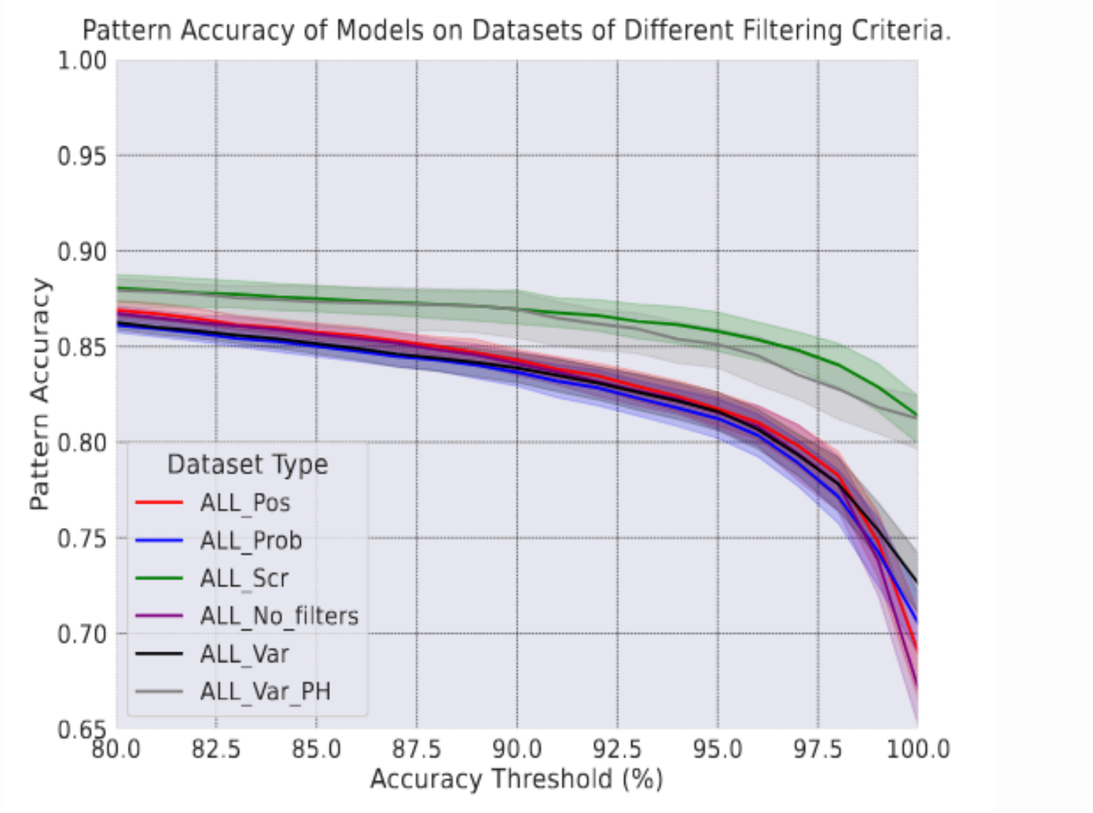
Model	Training	$SNLI_{test}$	ALL_{Seed}	$ALL_{Variants}$
BERT	S	90.48	90.24	88.72
RoBERTa	S	90.06	89.86	88.50
DeBERTa	S	91.70	91.38	89.41
BART	S, M, F, A	92.03	91.85	89.11

Model	Training	All_{BERT}	$All_{RoBERTa}$	All_{Both}
BERT	S	88.79	88.55	88.84
RoBERTa	S	88.58	88.33	88.56


Pattern Accuracy of Models on Datasets of Different Classes of Replacements.
Pattern Accuracy
Accuracy Threshold (%)
Dataset Type
Adj_Var
N_Var
V_Var


Pattern Accuracy of Models on Nouns (vs. Verbs), and Nouns (vs. Adjectives).
Pattern accuracy
Accuracy threshold (%)
Colors
Adjectives (vs. Nouns)
Nouns (vs. Adjectives)
Nouns (vs. Verbs)
Verbs (vs. Nouns)
Style
Verbs (vs. Nouns)
Adjectives (vs. Nouns)
Nouns (vs. Adjectives)
Nouns (vs. Verbs)


Pattern Accuracy of BERT and RoBERTa on variants divided by their origin model.
Pattern accuracy
Accuracy threshold (%)
Colors
BERT
RoBERTa
Origin Model
ALL_Both
ALL_RoBERTa
ALL_BERT


Pattern Accuracy of Models on Datasets of Different Filtering Criteria.
Pattern Accuracy
Accuracy Threshold (%)
Dataset Type
ALL_Pos
ALL_Prob
ALL_Scr
ALL_No_filters
ALL_Var
ALL_Var_PH