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

Mădălina Zgreabăn, Tejaswini Deoskar, Lasha Abzianidze

M
E
R
GE

A small girl carries a girl.
There is a girl.

from Stanford NLI (SNLI) dataset (Bowman et al., 2015)

<PREMISE>
<HYPOTHESIS>

typically what would a human reading (crowdworker) infer about the truth of **H** given **P**.

Dagan et al., 2005

ENTAILMENT
CONTRADICTION
NEUTRAL

typically what would a human reading (crowdworker) infer about the truth of **H** given **P**.

Dagan et al., 2005

NLI task

- Popular (100+).
- Easy task on reasoning.
- (Mostly) it is a three-way classification task.
- Simple/silly heuristics work due to annotation artifacts.
    - Hypothesis-only bias
    - Word overlap bias (WO)
    - Inverse WO bias
    - Negation bias

<HYPOTHESIS>

Gururangan et al. 2018, Poliak et al. 2018

A small girl carries a girl.
There is a girl.

McCoy et al. 2019, Naik et al., 2018

A small girl carries a girl.
There is a female.

A small girl is carrying a girl.
There is no girl is not true.

Generalization & NLI

HOW DO MODELS GENERALIZE?

HOW DO SPURIOUS CORRELATIONS AFFECT THEM?

Generalization & NLI

HOW DO MODELS $\boxed{\text{GENERALIZE}}$?

HOW DO SPURIOUS CORRELATIONS AFFECT THEM?

Generalization & NLI

HOW DO MODELS $\boxed{\text{GENERALIZE}}$ MODELS?

HOW DO $\boxed{\text{SPURIOUS CORRELATIONS}}$ AFFECT THEM?

Generalization & NLI & Contrasts sets

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

PaRTE (Verma et al. 2023):

$$< P, H, l > \Rightarrow < Para(P), Para(H), l >$$

Overview of contrast sets in NLI

Generalization & NLI & Contrasts sets

Study	Strategy	Creation	Val.	Sentence Mod.	M	R	S	WO	Evaluation	Dataset
Li et al. (2020)	Multiple	Auto.	HVal_p	P	Mix.	Mix.	N	N	Vs-G; Vs-O	SNLI; MNLI
Glockner et al. (2018)	Replace	Auto.	HVal_f	H	N	Mix.	Y	N	V-G	SNLI
Verma et al. (2023)	Paraphrase	Auto.	HVal_f	P/H; $P\&H$	Y	Y	N	N	Vs-O	Pascal RTE1-3 (Dagan et al., 2005)
Srikanth et al. (2024)	Paraphrase	Mix.	HVal_{pf}	H; U	Y	Y	N	N	Vs-Vs; Vs-G	α-NLI (Bhagavatula et al., 2019); δ-NLI (Rudinger et al., 2020)
Arakelyan et al. (2024)	Paraphrase	Auto.	HVal_p	H	Y	Y	N	N	V-O	SNLI; MNLI; ANLI
Petrov (2025)*	Multiple	Auto.	N/A	H	N	N	N	N	V-G	SNLI
Kaushik et al. (2020)	Multiple	Man.	HVal_f	P; H	N	N	N	N	V-G	SNLI
Srikanth and Rudinger (2025)	Decompose	Auto.	Mix.	H	Y	Y	N	N	V-G; Vs-O	SNLI; δ-NLI

Overview of contrast sets in NLI

Generalization & NLI & Contrasts sets

Study		Creation	Val.	Sentence Mod.	M	R	S	WO	Evaluation	Dataset
Li et al. (2020)	Multiple	Auto.	HVal_p	P	Mix.	Mix.	N	N	Vs-G; Vs-O	SNLI; MNLI
Glockner et al. (2018)	Replace	Auto.	HVal_f	H	N	Mix.	Y	N	V-G	SNLI
Verma et al. (2023)	Paraphrase	uto.	HVal_f	P/H; $P\&H$	Y	Y	N	N	Vs-O	Pascal RTE1-3 (Dagan et al., 2005)
Srikanth et al. (2024)	Paraphrase	ix.	HVal_{pf}	H; U	Y	Y	N	N	Vs-Vs; Vs-G	α-NLI (Bhagavatula et al., 2019); δ-NLI (Rudinger et al., 2020)
Arakelyan et al. (2024)	Paraphrase	uto.	HVal_p	H	Y	Y	N	N	V-O	SNLI; MNLI; ANLI
Petrov (2025)*	Multiple	uto.	N/A	H	N	N	N	N	V-G	SNLI
Kaushik et al. (2020)	Multiple	Man.	HVal_f	P; H	N	N	N	N	V-G	SNLI
Srikanth and Rudinger (2025)	Decompos	Auto.	Mix.	H	Y	Y	N	N	V-G; Vs-O	SNLI; δ-NLI

Overview of contrast sets in NLI

Generalization & NLI & Contrasts sets

Study	Strategy	n	Val.	Sentence Mod.	M	R	S	WO	Evaluation	Dataset
Li et al. (2020)	Multiple	Auto.	Val$_p$	P	Mix.	Mix.	N	N	Vs-G; Vs-O	SNLI; MNLI
Glockner et al. (2018)	Replace	Auto.	al$_f$	H	N	Mix.	Y	N	V-G	SNLI
Verma et al. (2023)	Paraph e	Auto.	al$_f$	P/H; $P\&H$	Y	Y	N	N	Vs-O	Pascal RTE1-3 (Dagan et al., 2005)
Srikanth et al. (2024)	Parapl se	Mix.	H l$_{pf}$	H; U	Y	Y	N	N	Vs-Vs; Vs-G	α-NLI (Bhagavatula et al., 2019); δ-NLI (Rudinger et al., 2020)
Arakelyan et al. (2024)	Parapl se	Auto.	H l$_p$	H	Y	Y	N	N	V-O	SNLI; MNLI; ANLI
Petrov (2025)*	Multip	Auto.	N	H	N	N	N	N	V-G	SNLI
Kaushik et al. (2020)	Multipl	Man.	al$_f$	P; H	N	N	N	N	V-G	SNLI
Srikanth and Rudinger (2025)	Decomp	Auto.	ix.	H	Y	Y	N	N	V-G; Vs-O	SNLI; δ-NLI

Overview of contrast sets in NLI

Generalization & NLI & Contrasts sets

Study	Strategy	Creation	Val.	Sentence Mod.	M	R	S	WO	Evaluation	Dataset
Li et al. (2020)	Multiple	Auto[illegible]	HVal$_p$		Mix.	Mix.	N	N	Vs-G; Vs-O	SNLI; MNLI
Glockner et al. (2018)	Replace	Aut[illegible]	HVal$_f$	[illegible]	N	Mix.	Y	N	V-G	SNLI
Verma et al. (2023)	Paraphrase	Au[illegible]	HVal$_f$	[illegible] H; [illegible] H	Y	Y	N	N	Vs-O	Pascal RTE1-3 (Dagan et al., 2005)
Srikanth et al. (2024)	Paraphrase	M[illegible]	HVal$_{pf}$	H [illegible] U	Y	Y	N	N	Vs-Vs; Vs-G	α-NLI (Bhagavatula et al., 2019); δ-NLI (Rudinger et al., 2020)
Arakelyan et al. (2024)	Paraphrase	Au[illegible]	HVal$_p$	[illegible]	Y	Y	N	N	V-O	SNLI; MNLI; ANLI
Petrov (2025)*	Multiple	Au[illegible]	N/A	[illegible]	N	N	N	N	V-G	SNLI
Kaushik et al. (2020)	Multiple	Ma[illegible]	HVal$_f$	[illegible] H	N	N	N	N	V-G	SNLI
Srikanth and Rudinger (2025)	Decompose	Auto[illegible]	Mix.		Y	Y	N	N	V-G; Vs-O	SNLI; δ-NLI

Overview of contrast sets in NLI

Generalization & NLI & Contrasts sets

Study	Strategy	Creation	Val.	...ter. Mod.	M	R	S	WO	Evaluation	Dataset
Li et al. (2020)	Multiple	Auto.	$HVal_p$	P	Mix.	Mix.	N	N	Vs-G; Vs-O	SNLI; MNLI
Glockner et al. (2018)	Replace	Auto.	HVal	H	N	Mix.	Y	N	V-G	SNLI
Verma et al. (2023)	Paraphrase	Auto.	HVa	P/H; $P\&H$		Y	N	N	Vs-O	Pascal RTE1-3 (Dagan et al., 2005)
Srikanth et al. (2024)	Paraphrase	Mix.	HV	H; U		Y	N	N	Vs-Vs; Vs-G	α-NLI (Bhagavatula et al., 2019); δ-NLI (Rudinger et al., 2020)
Arakelyan et al. (2024)	Paraphrase	Auto.	HVa	H		Y	N	N	V-O	SNLI; MNLI; ANLI
Petrov (2025)*	Multiple	Auto.	N/A	H		N	N	N	V-G	SNLI
Kaushik et al. (2020)	Multiple	Man.	HVal	P; H	N	N	N	N	V-G	SNLI
Srikanth and Rudinger (2025)	Decompose	Auto.	Mix.	H	Y	Y	N	N	V-G; Vs-O	SNLI; δ-NLI

Overview of contrast sets in NLI

Generalization & NLI & Contrasts sets

Study	Strategy	Creation	Val.	Sentence Mod.	M	R	S	WO	Evaluation	Dataset
Li et al. (2020)	Multiple	Auto.	HVal_p	P	Mix.	Mix.	N		Vs-G; Vs-O	SNLI; MNLI
Glockner et al. (2018)	Replace	Auto.	HVal_f	H	N	Mix.	Y		V-G	SNLI
Verma et al. (2023)	Paraphrase	Auto.	HVal_f	P/H; $P\&H$	Y	Y	N		Vs-O	Pascal RTE1-3 (Dagan et al., 2005)
Srikanth et al. (2024)	Paraphrase	Mix.	HVal_{pf}	H; U	Y	Y	N		Vs-Vs; Vs-G	α-NLI (Bhagavatula et al., 2019); δ-NLI (Rudinger et al., 2020)
Arakelyan et al. (2024)	Paraphrase	Auto.	HVal_p	H	Y	Y	N		V-O	SNLI; MNLI; ANLI
Petrov (2025)*	Multiple	Auto.	N/A	H	N	N	N	N	V-G	SNLI
Kaushik et al. (2020)	Multiple	Man.	HVal_f	P; H	N	N	N	N	V-G	SNLI
Srikanth and Rudinger (2025)	Decompose	Auto.	Mix.	H	[illegible]	Y	N	N	V-G; Vs-O	SNLI; δ-NLI

Overview of contrast sets in NLI

Generalization & NLI & Contrasts sets

Study	Strategy	Creation	Val.	Sentence Mod.	M	R	S	WC	Evaluation	Dataset
Li et al. (2020)	Multiple	Auto.	HVal_p	P	Mix.	Mix.	N	N	Vs-G; Vs-O	NLI; MNLI
Glockner et al. (2018)	Replace	Auto.	HVal_f	H	N	Mix.	Y		V-G	NLI
Verma et al. (2023)	Paraphrase	Auto.	HVal_f	P/H; $P\&H$	Y	Y	N		Vs-O	scal RTE1-3 (Dagan et al., 05)
Srikanth et al. (2024)	Paraphrase	Mix.	HVal_{pf}	H; U	Y	Y	N		Vs-Vs; Vs-G	NLI (Bhagavatula et al., 19); δ-NLI (Rudinger et al., 20)
Arakelyan et al. (2024)	Paraphrase	Auto.	HVal_p	H	Y	Y	N		V-O	LI; MNLI; ANLI
Petrov (2025)*	Multiple	Auto.	N/A	H	N	N	N		V-G	NLI
Kaushik et al. (2020)	Multiple	Man.	HVal_f	P; H	N	N	N	N	V-G	NLI
Srikanth and Rudinger (2025)	Decompose	Auto.	Mix.	H	Y	Y	N	N	V-G; Vs-O	SNLI; δ-NLI

CONSISTENCY

Overview of contrast sets in NLI

Generalization & NLI & Contrasts sets

Study	Strategy	Creation	Val.	Sentence Mod.	M	R	S	WO	Evaluation	Dataset
Li et al. (2020)	Multiple	Auto.	HVal_p	P	Mix.	Mix.	N	N	Vs-G; Vs-O	SNLI; MNLI
Glockner et al. (2018)	Replace	Auto.	HVal_f	H	N	Mix.	Y	N	V-G	SNLI
Verma et al. (2023)	Paraphrase	Auto.	HVal_f	P/H; $P\&H$	Y	Y	N	N	Vs-O	Pascal RTE1-3 (Dagan et al., 2005)
Srikanth et al. (2024)	Paraphrase	Mix.	HVal_{pf}	H; U	Y	Y	N	N	Vs-Vs; Vs-G	α-NLI (Bhagavatula et al., 2019); δ-NLI (Rudinger et al., 2020)
Arakelyan et al. (2024)	Paraphrase	Auto.	HVal_p	H	Y	Y	N	N	V-O	SNLI; MNLI; ANLI
Petrov (2025)*	Multiple	Auto.	N/A	H	N	N	N	N	V-G	SNLI
Kaushik et al. (2020)	Multiple	Man.	HVal_f	P; H	N	N	N	N	V-G	SNLI
Srikanth and Rudinger (2025)	Decompose	Auto.	Mix.	H	Y	Y	N	N	V-G; Vs-O	SNLI; δ-NLI

does not keep the WO;
introduces syntactic changes

Changes are not minimal

Generalization & NLI & Contrasts sets

Study	Strategy	Creation	Val.	Sentence Mod.	M	R	S	WO	Evaluation	Dataset
Li et al. (2020)	Multiple	Auto.	HVal_p	P	Mix.	Mix.	N	N	Vs-G; Vs-O	SNLI; MNLI
Glockner et al. (2018)	Replace	Auto.	HVal_f	H	N	Mix.	Y	N	V-G	SNLI
Verma et al. (2023)	Paraphrase	Auto.	HVal_f	P/H; $P\&H$	Y	Y	N	N	Vs-O	Pascal RTE1-3 (Dagan et al., 2005)
Srikanth et al. (2024)	Paraphrase	Mix.	HVal_{pf}	H; U	Y	Y	N	N	Vs-Vs; Vs-G	α-NLI (Bhagavatula et al., 2019); δ-NLI (Rudinger et al., 2020)
Arakelyan et al. (2024)	Paraphrase	Auto.	HVal_p	H	Y	Y	N	N	V-O	SNLI; MNLI; ANLI
Petrov (2025)*	Multiple	Auto.	N/A	H	N	N	N	N	V-G	SNLI
Kaushik et al. (2020)	Multiple	Man.	HVal_f	P; H	N	N	N	N	V-G	SNLI
Srikanth and Rudinger (2025)	Decompose	Auto.	Mix.	H	Y	Y	N	N	V-G; Vs-O	SNLI; δ-NLI

TEMPLATE BASED = plausible and correct, but…

- lexical diversity
- limited problems

Changes are not minimal

Generalization & NLI & Contrasts sets

Study	Strategy	Creation	Val.	Sentence Mod.	M	R	S	WO	Evaluation	Dataset
Li et al. (2020)	Multiple	Auto.	HVal_p	P	Mix.	Mix.	N	N	Vs-G; Vs-O	SNLI; MNLI
Glockner et al. (2018)	Replace	Auto.	HVal_f	H	N	Mix.	Y	N	V-G	SNLI
Verma et al. (2023)	Paraphrase	Auto.	HVal_f	P/H; $P\&H$	Y	Y	N	N	Vs-O	Pascal RTE1-3 (Dagan et al., 2005)
Srikanth et al. (2024)	Paraphrase	Mix.	HVal_{pf}	H; U	Y	Y	N	N	Vs-Vs; Vs-G	α-NLI (Bhagavatula et al., 2019); δ-NLI (Rudinger et al., 2020)
Arakelyan et al. (2024)	Paraphrase	Auto.	HVal_p	H	Y	Y	N	N	V-O	SNLI; MNLI; ANLI
Petrov (2025)*	Multiple	Auto.	N/A	H	N	N	N	N	V-G	SNLI
Kaushik et al. (2020)	Multiple	Man.	HVal_f	P; H	N	N	N	N	V-G	SNLI
Srikanth and Rudinger (2025)	Decompose	Auto.	Mix.	H	Y	Y	N	N	V-G; Vs-O	SNLI; δ-NLI

syntactic changes

Changes are not minimal

Generalization & NLI & Contrasts sets

Study	Strategy	Creation	Val.	Sentence Mod.	M	R	S	WO	Evaluation	Dataset
Li et al. (2020)	Multiple	Auto.	HVal$_p$	P	Mix.	Mix.	N	N	Vs-G; Vs-O	SNLI; MNLI
Glockner et al. (2018)	Replace	Auto.	HVal$_f$	H	N	Mix.	Y	N	V-G	SNLI
Verma et al. (2023)	Paraphrase	Auto.	HVal$_f$	P/H; $P\&H$	Y	Y	N	N	Vs-O	Pascal RTE1-3 (Dagan et al., 2005)
Srikanth et al. (2024)	Paraphrase	Mix.	HVal$_{pf}$	H; U	Y	Y	N	N	Vs-Vs; Vs-G	α-NLI (Bhagavatula et al., 2019); δ-NLI (Rudinger et al., 2020)
Arakelyan et al. (2024)	Paraphrase	Auto.	HVal$_p$	H	Y	Y	N	N	V-O	SNLI; MNLI; ANLI
Petrov (2025)*	Multiple	Auto.	N/A	H	N	N	N	N	V-G	SNLI
Kaushik et al. (2020)	Multiple	Man.	HVal$_f$	P; H	N	N	N	N	V-G	SNLI
Srikanth and Rudinger (2025)	Decompose	Auto.	Mix.	H	Y	Y	N	N	V-G; Vs-O	SNLI; δ-NLI

does it bias the contrast set?

are they correct?

Changes are not minimal

Generalization & NLI & Contrasts sets

Study	Strategy	Creation	Val.	Sentence Mod.	M	R	S	WO	Evaluation	Dataset
Li et al. (2020)	Multiple	Auto.	HVal$_p$	P	Mix.	Mix.	N	N	Vs-G; Vs-O	SNLI; MNLI
Glockner et al. (2018)	Replace	Auto.	HVal$_f$	H	N	Mix.	Y	N	V-G	SNLI
Verma et al. (2023)	Paraphrase	Auto.	HVal$_f$	P/H; $P\&H$	Y	Y	N	N	Vs-O	Pascal RTE1-3 (Dagan et al., 2005)
Srikanth et al. (2024)	Paraphrase	Mix.	HVal$_{pf}$	H; U	Y	Y	N	N	Vs-Vs; Vs-G	α-NLI (Bhagavatula et al., 2019); δ-NLI (Rudinger et al., 2020)
Arakelyan et al. (2024)	Paraphrase	Auto.	HVal$_p$	H	Y	Y	N	N	V-O	SNLI; MNLI; ANLI
Petrov (2025)*	Multiple	Auto.	N/A	H	N	N	N	N	V-G	SNLI
Kaushik et al. (2020)	Multiple	Man.	HVal$_f$	P; H	N	N	N	N	V-G	SNLI
Srikanth and Rudinger (2025)	Decompose	Auto.	Mix.	H	Y	Y	N	N	V-G; Vs-O	SNLI; δ-NLI

but then manual…

Changes are not minimal

Generalization & NLI & Contrasts sets

Study	Strategy	Creation	Val.	Sentence Mod.	M	R	S	WO	Evaluation	Dataset
Li et al. (2020)	Multiple	Auto.	$HVal_p$	P	Mix.	Mix.	N	N	Vs-G; Vs-O	SNLI; MNLI
Glockner et al. (2018)	Replace	Auto.	$HVal_f$	H	N	Mix.	Y	N	V-G	SNLI
Verma et al. (2023)	Paraphrase	Auto.	$HVal_f$	P/H; $P\&H$	Y	Y	N	N	Vs-O	Pascal RTE 1-3 (Dagan et al., 2005)
Srikanth et al. (2024)	Paraphrase	Mix.	$HVal_{pf}$	H; U	Y	Y	N	N	Vs-Vs; Vs-G	α-NLI (Bhagavatula et al., 2019); δ-NLI (Rudinger et al., 2020)
Arakelyan et al. (2024)	Paraphrase	Auto.	$HVal_p$	H	Y	Y	N	N	V-O	SNLI; MNLI; ANLI
Petrov (2025)*	Multiple	Auto.	N/A	H	N	N	N	N	V-G	SNLI
Kaushik et al. (2020)	Multiple	Man.	$HVal_f$	P; H	N	N	N	N	V-G	SNLI
Srikanth and Rudinger (2025)	Decompose	Auto.	Mix.	H	Y	Y	N	N	V-G; Vs-O	SNLI; δ-NLI

what about these?

Changes are not minimal

no benchmark constructed by

automatic-logic-preserving-truly-minimal changes no benchmark constructed by

automatic-logic-preserving-truly-minimal changes

Mădălina Zgreabăn, Tejaswini Deoskar, Lasha Abzianidze

M
E
R
GE

MERGE

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. **E** **E** ✓

H: There is a small boy.

⋮

P: A small dog carries a dog. **E** **E** ✓

H: There is a small dog.

⋮

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

H: There is a little girl.

⋮

P: A happy girl carries a girl. **E** **E** ✓

H: There is a happy girl.

MERGE: Minimal Expression-Replacement GEneralization

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) WO
- Negation/antonymy
- Hypothesis only

We replace single words with single words

Antonyms are different words; hence they remain

Usually, give-away words only occurs in a hypothesis

MERGE: Minimal Expression-Replacement GEneralization

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

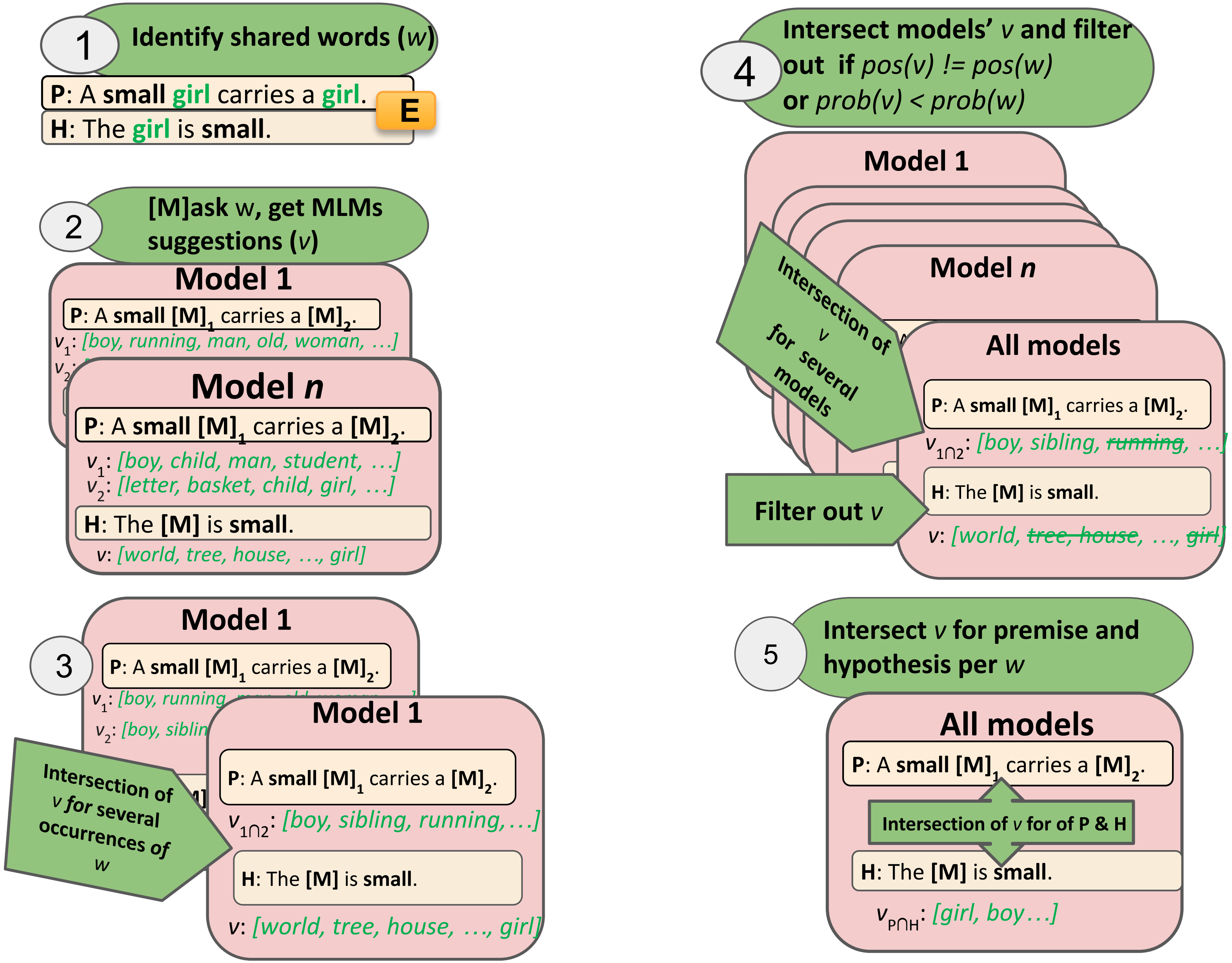
P: Two dogs and three **dogs** swim.
H: Only three **dogs** swim.
E

Which affect the WO! Don't replace original words with co-occurring words!

Are we good? WELL…

- Are the variants plausible?
- Do they keep the syntax?

Generating 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., $W_M(PH, \mathrm{mentorship}^c_>) = \emptyset$

AT LEAST ONE MODEL…

Setup of experiments

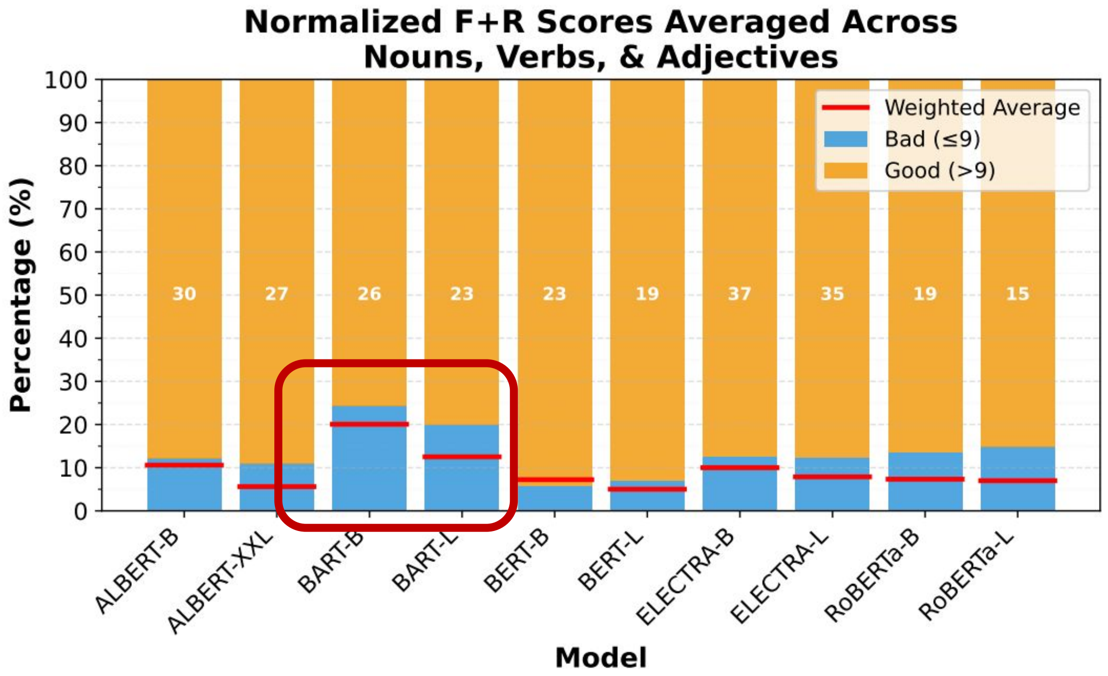
MLMs:

- BERT, RoBERTa, ALBERT, Electra, and BART, base and large, except ALBERT (b, xxl).
- 10k test SNLI, for nouns, verbs, adjectives.
- Manually annotated 100 examples * open-class category to evaluate efficiency of models.

Setup of experiments

MLMs (10):

- BERT, RoBERTa, ALBERT, Electra, and BART, base and large, except ALBERT (b, xxl).
- 10k test SNLI, for nouns, verbs, adjectives.
- Manually annotated 100 examples * open-class category to evaluate efficiency of models.
- After exclusion > annotate again.
- 91% plausible examples, but all logic-preserving.



no bart.

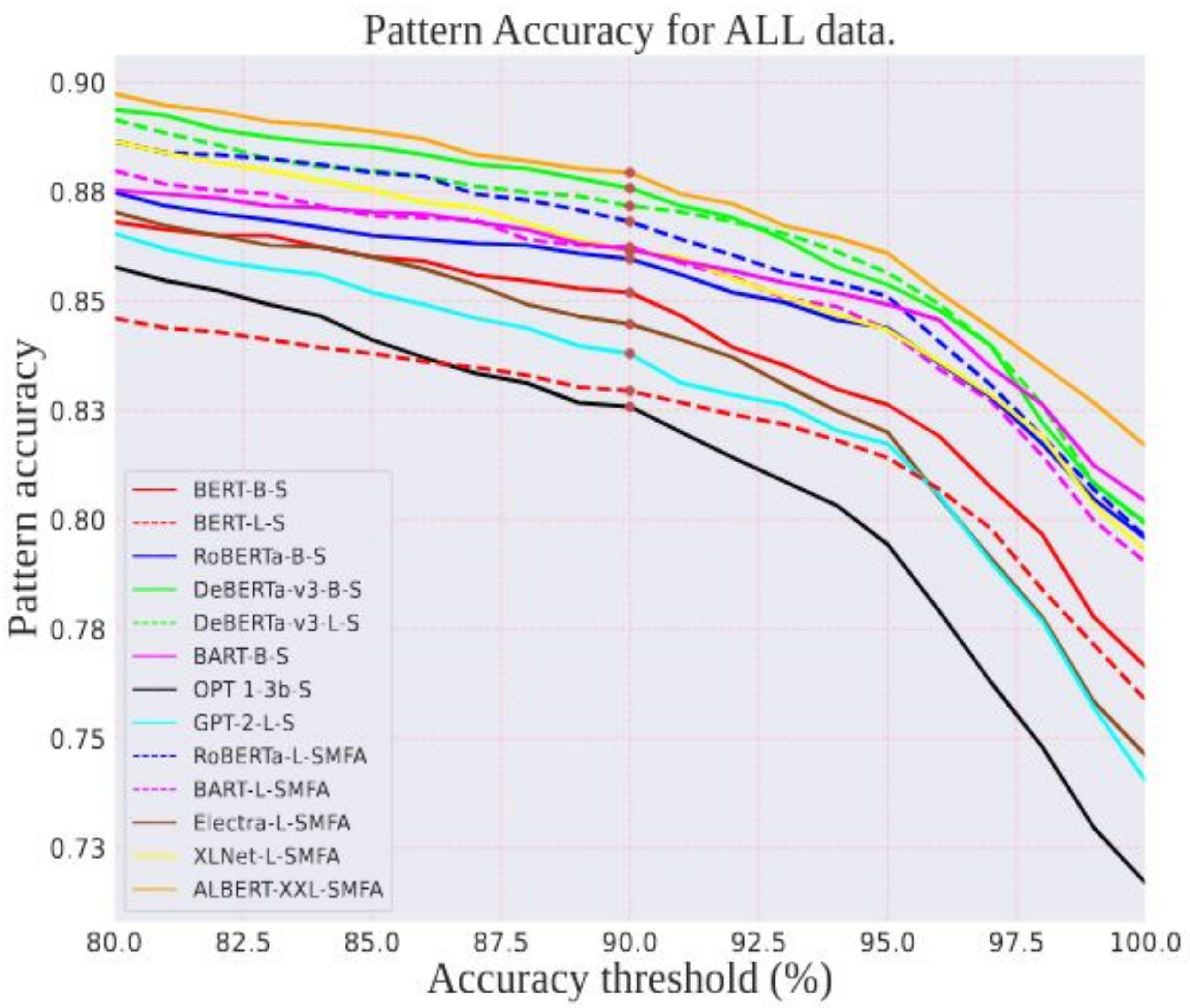
models are expected to get 90% variants correctly

Sample & pattern accuracy (PA) scores

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

~10K ~2.2K ~50K

Model	$SNLI_{Test}$	ALL_{Seed}	ALL_{Var}	90	MT
BERT-B-S	90.5	89.6	88.9	-4.9	59
BERT-L-S	87.1	87.2	87.4	-4.5	47
RoBERTa-B-S	90.1	90.1	89.2	-4.5	47
DeBERTa-v3-B-S	91.7	92.1	90.7	-4.9	58
DeBERTa-v3-L-S	91.7	91.9	91.0	-4.9	54
BART-B-S	90.6	90.2	89.4	-4.3	57
OPT-1-3b-S	91.0	90.5	89.1	-8.6	58
GPT-2-L-S	90.9	90.9	89.5	-7.7	55
RoBERTa-L-SMFA	91.8	91.4	90.5	-5.0	59
BART-L-SMFA	92.0	91.9	90.5	-6.0	55
Electra-L-SMFA	91.1	90.6	90.0	-6.5	56
XLNet-L-SMFA	91.7	91.4	90.6	-5.4	55
ALBERT-XXL-SMFA	91.9	92.2	91.2	-4.8	57



Error analysis

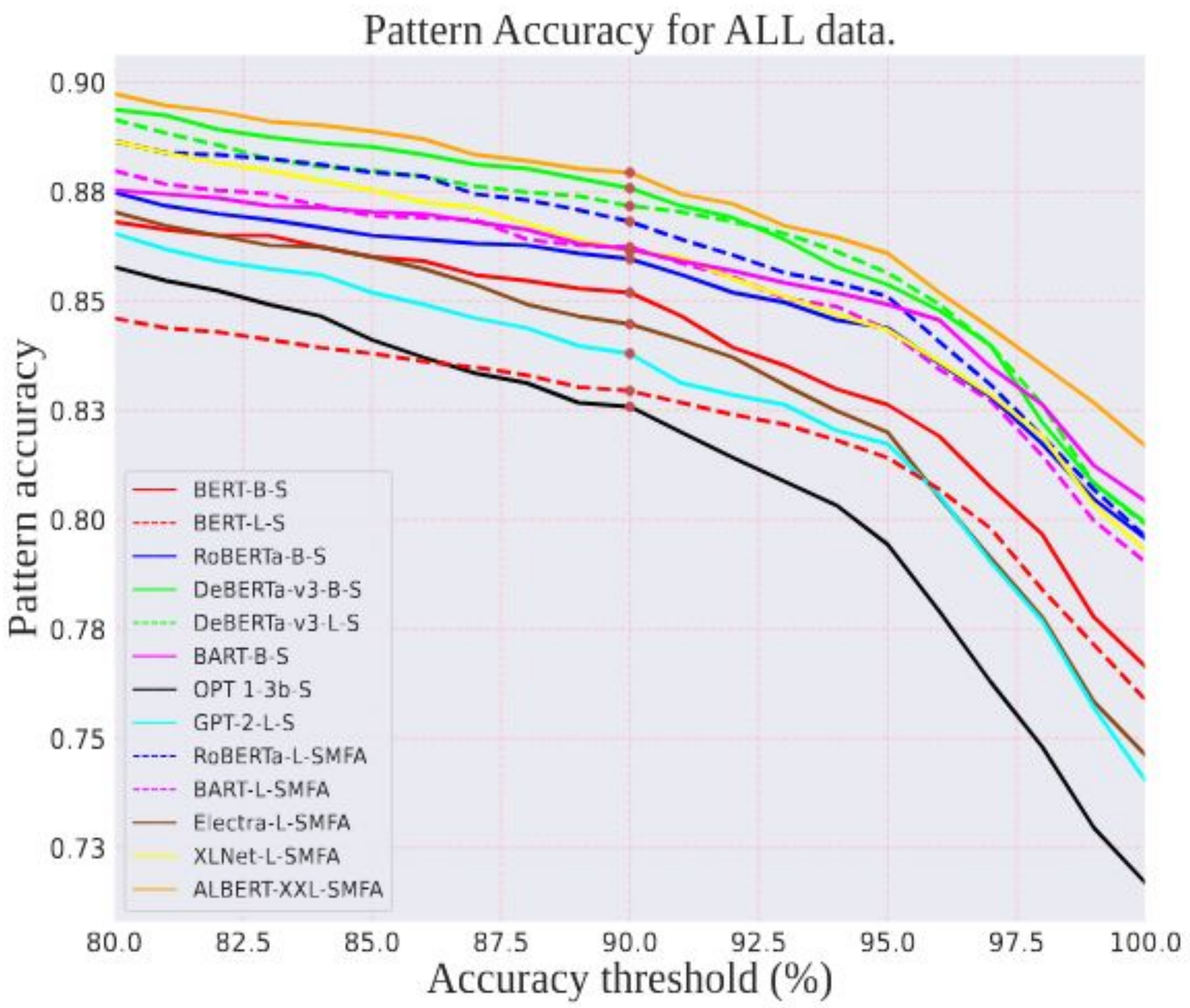
- Models make almost all mistakes on seed problems they initially got incorrectly
- Variants of 31 problems (31*20) are all predicted incorrectly across models:
  - Only 30% had a correct label assigned
  - In line with Maadan et al., 2024 which showed models' mistakes are in line with annotator variation
- No seed problems that were incorrectly classified, with any of their variants classified correctly

Sample & pattern accuracy (PA) scores

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

~10K ~2.2K ~50K

Model	$SNLI_{Test}$	ALL_{Seed}	ALL_{Var}	90	MT
BERT-B-S	90.5	89.6	88.9	-4.9	59
BERT-L-S	87.1	87.2	87.4	-4.5	47
RoBERTa-B-S	90.1	90.1	89.2	-4.5	47
DeBERTa-v3-B-S	91.7	92.1	90.7	-4.9	58
DeBERTa-v3-L-S	91.7	91.9	91.0	-4.9	54
BART-B-S	90.6	90.2	89.4	-4.3	57
OPT-1-3b-S	91.0	90.5	89.1	-8.6	58
GPT-2-L-S	90.9	90.9	89.5	-7.7	55
RoBERTa-L-SMFA	91.8	91.4	90.5	-5.0	59
BART-L-SMFA	92.0	91.9	90.5	-6.0	55
Electra-L-SMFA	91.1	90.6	90.0	-6.5	56
XLNet-L-SMFA	91.7	91.4	90.6	-5.4	55
ALBERT-XXL-SMFA	91.9	92.2	91.2	-4.8	57

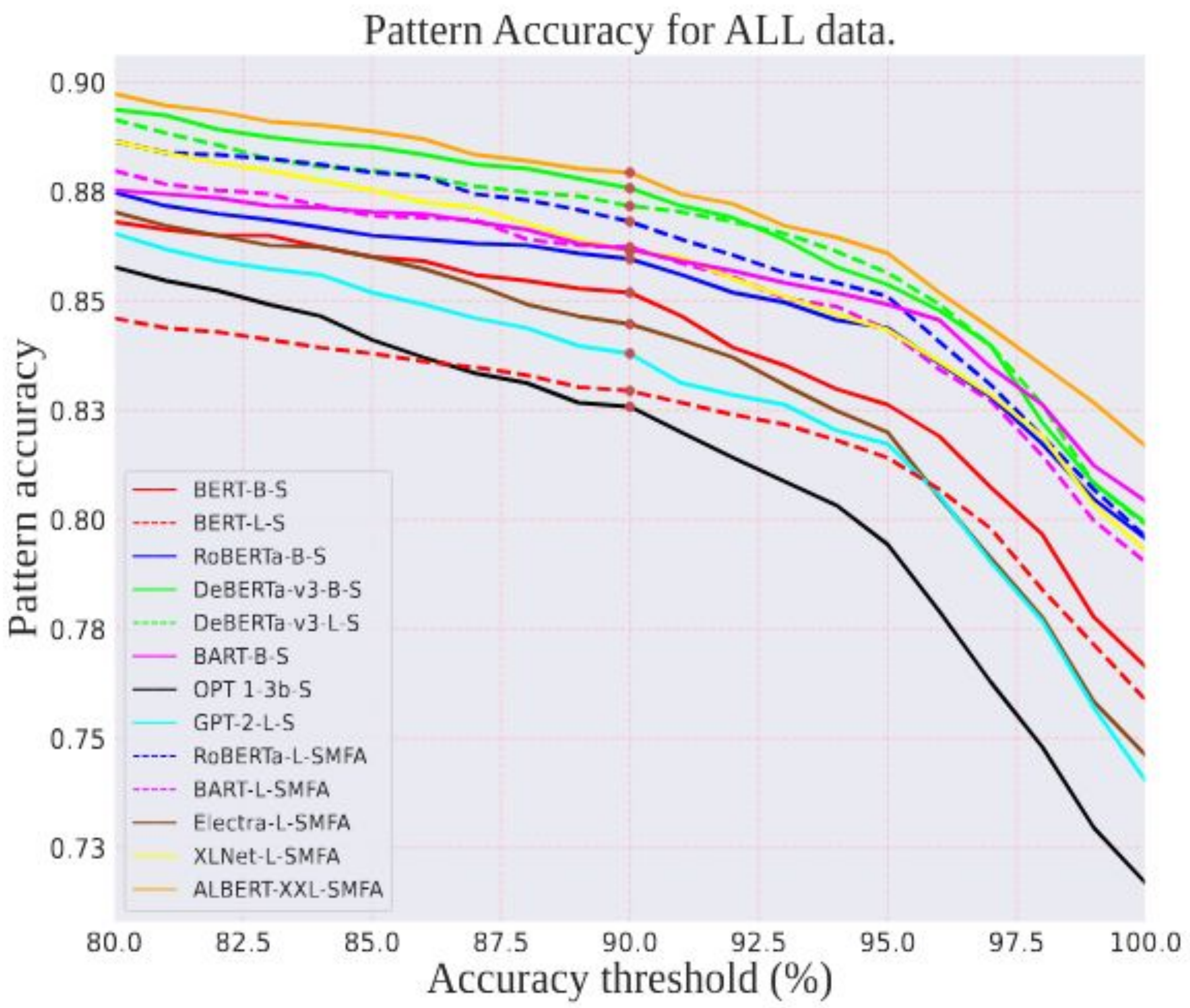


Sample & pattern accuracy (PA) scores

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

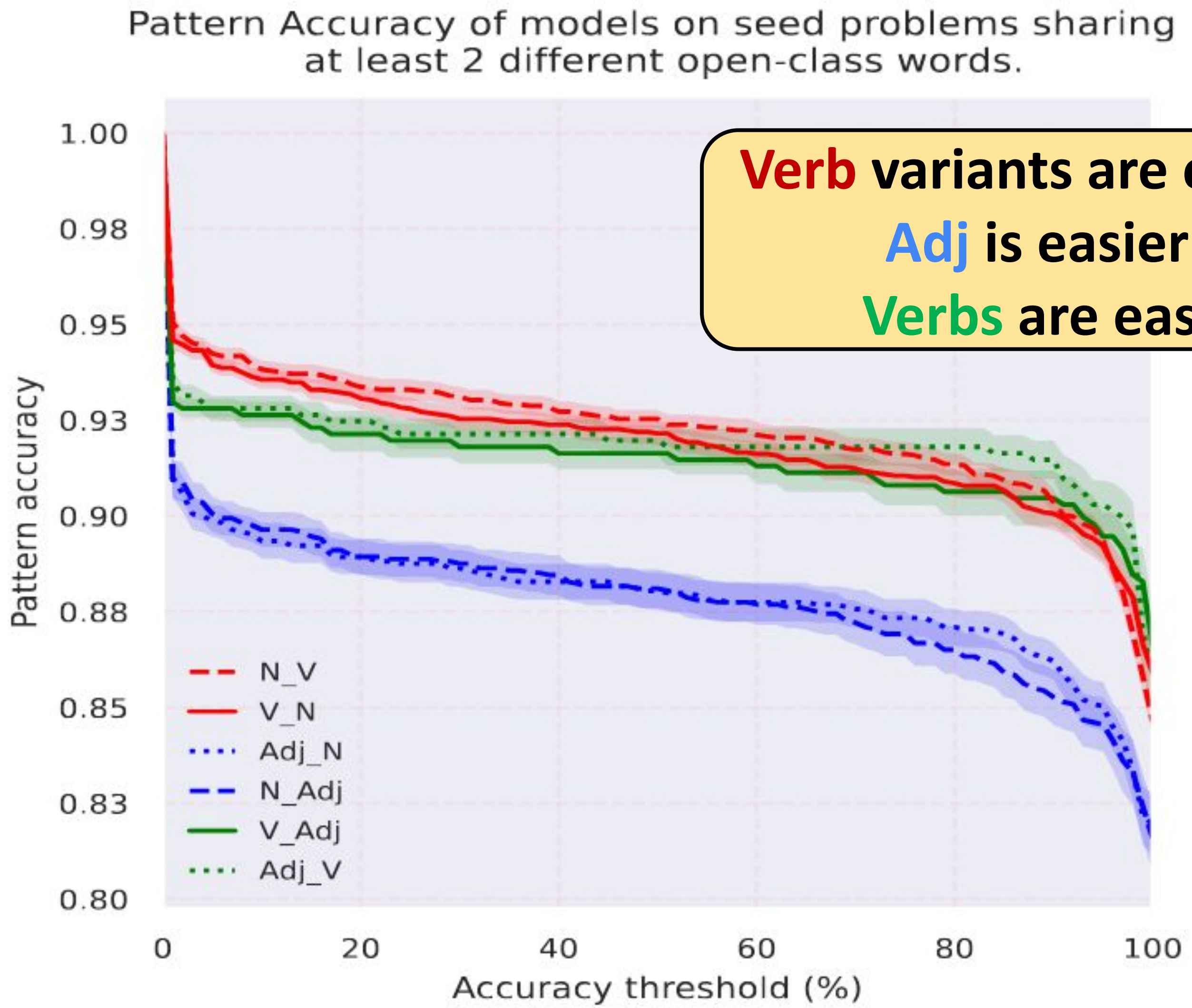
~10K ~2.2K ~50K

Model	$SNLI_{Test}$	ALL_{Seed}	ALL_{Var}	90	MT
BERT-B-S	90.5	89.6	88.9	-4.9	59
BERT-L-S	87.1	87.2	87.4	-4.5	47
RoBERTa-B-S	90.1	90.1	89.2	-4.5	47
DeBERTa-v3-B-S	91.7	92.1	90.7	-4.9	58
DeBERTa-v3-L-S	91.7	91.9	91.0	-4.9	54
BART-B-S	90.6	90.2	89.4	-4.3	57
OPT-1-3b-S	91.0	90.5	89.1	-8.6	58
GPT-2-L-S	90.9	90.9	89.5	-7.7	55
RoBERTa-L-SMFA	91.8	91.4	90.5	-5.0	59
BART-L-SMFA	92.0	91.9	90.5	-6.0	55
Electra-L-SMFA	91.1	90.6	90.0	-6.5	56
XLNet-L-SMFA	91.7	91.4	90.6	-5.4	55
ALBERT-XXL-SMFA	91.9	92.2	91.2	-4.8	57



*models also make more mistakes on original incorrect seed problems

Easiest word class variants



Verb variants are easier than Noun
Adj is easier than Noun
Verbs are easier than Adj

Comparisons are done on the same seed NLI problems, i.e., removing the difference in difficulty of NLI problems.

Do MLMs favor native NLI models?

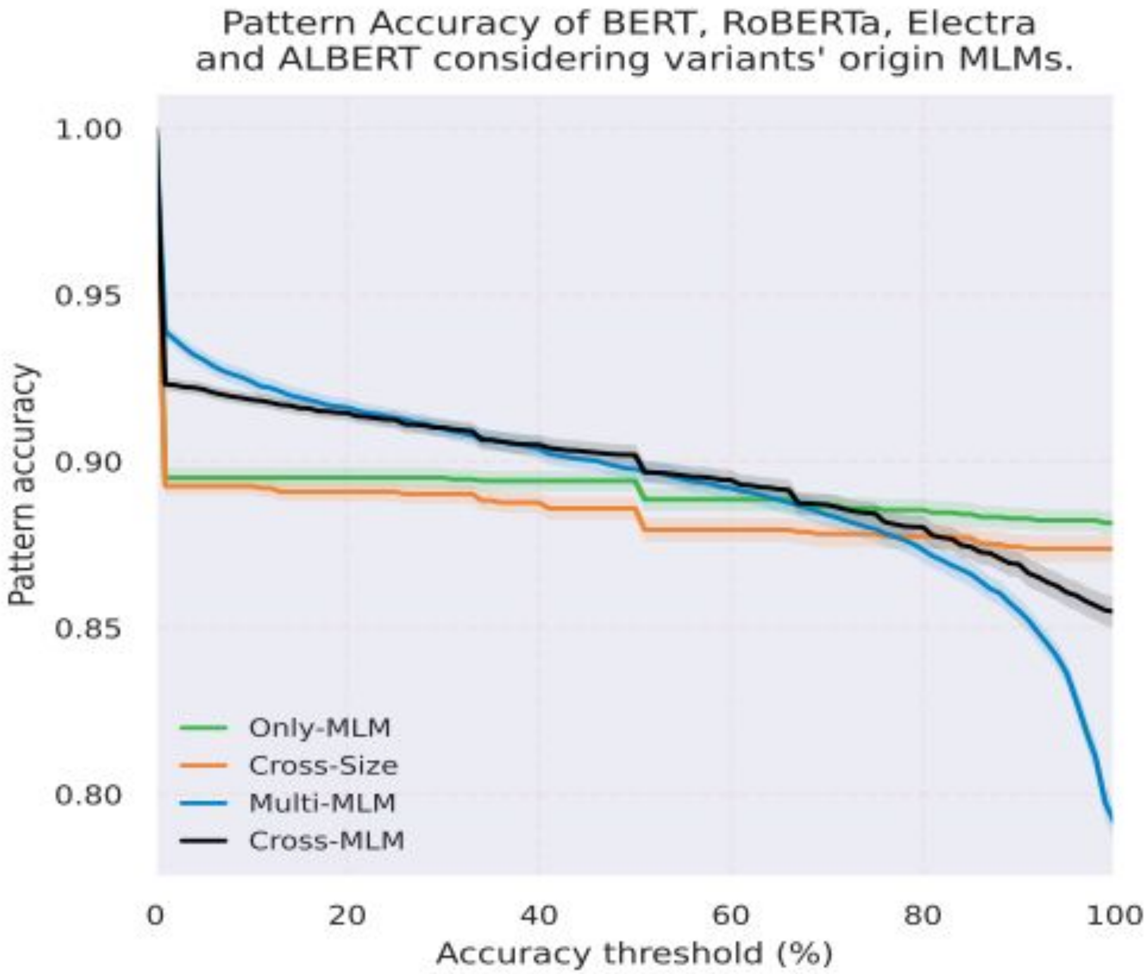
If there is ***any*** favoritism,
it can be seen
at the extreme th>90%.

However, MLMs ***do not favor native*** NLI models.

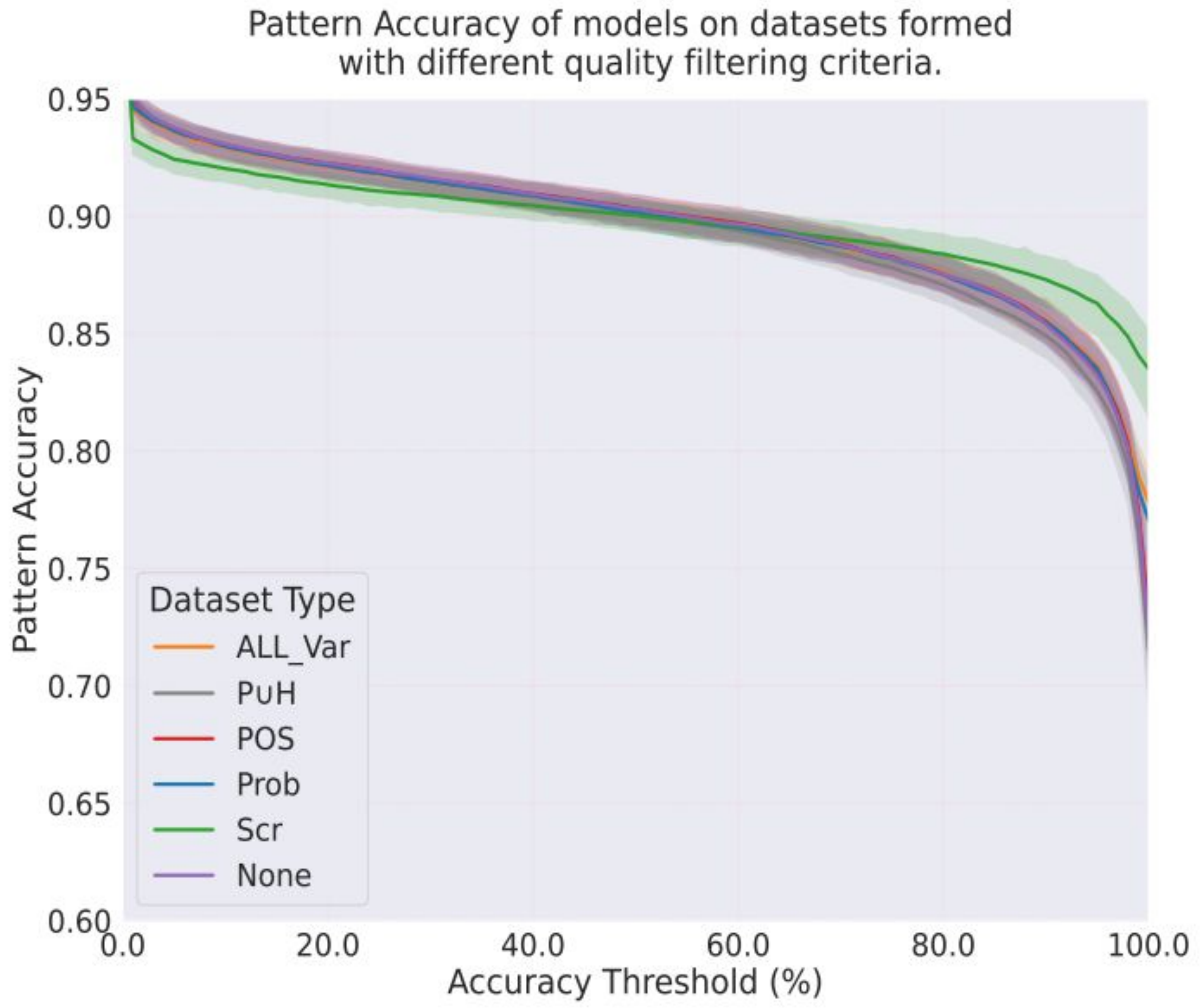
Easiest: same model

Easier: another model and same model, but diff size.

Least easy: multiple models.



FILTERING CRITERIA?



Conclusion

MERGE test:

- **Auto generating** sample variants with MLMs
- **Most friendly** generalization test: preserves reasoning & biases

Models **cannot** maintain the same accuracy even for threshold of 60%.

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

No favoritism between shared LLMs.

Future work will involve more NLI datasets and NLU tasks.