THE MINIMAL EXPRESSION REPLACEMENT GENERALIZATION TEST

MĂDĂLINA ZGREABĂN (PHD CANDIDATE)
UTRECHT UNIVERSITY
TEJASWINI DEOSKAR
LASHA ABZIANIDZE (PI)

GENERAZABILITY IN NLI

Out-of-distribution **(OOD data)** NLI benchmarks:
- are important, as in-distribution benchmarks are heuristics-prone [4, 3];
- result in decreased performance [6, 3, 1, 4, 8, 2, 7], indicating a lack of generalization capacity.

SHORTCOMINGS of previous OOD NLI benchmarks:
- disturb lexical overlap heuristic of premise and hypothesis (*PH*) > which can also cause a lower results [2, 7];
- have low lexical diversity [4, 1];
- are costly, if formed manually [3];
- are syntax non-preserving, which can also cause a decrease in models' scores [6];
- are unfair, if the data is not similar enough to the training data.

MERGE & OUR CONTRIBUTIONS

The Minimal Expression Replacement GEneralization (*MERGE*) test for NLI automatically & minimally alters existing NLI datasets, keeping their underlying reasoning, without requiring human validation by deploying strict minimal changes criteria.

Research questions:
- Are language models robust against variants of NLI problems?
- Do factors such as the likelihood, POS tag, plausibility, or masked models of the replacement influence models' performance?

DIAGRAM 1

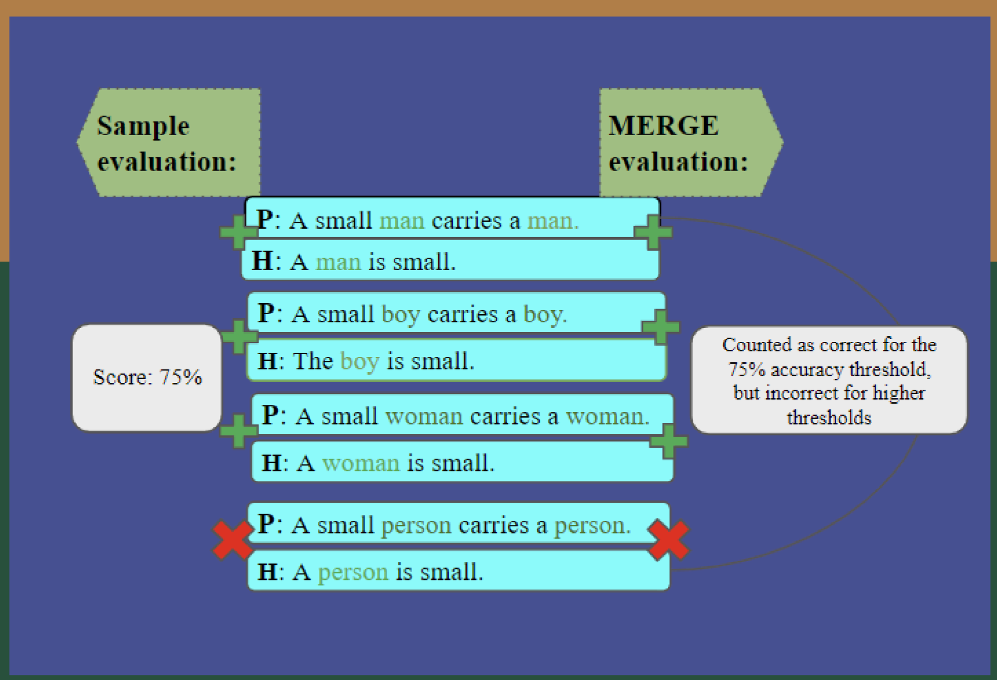
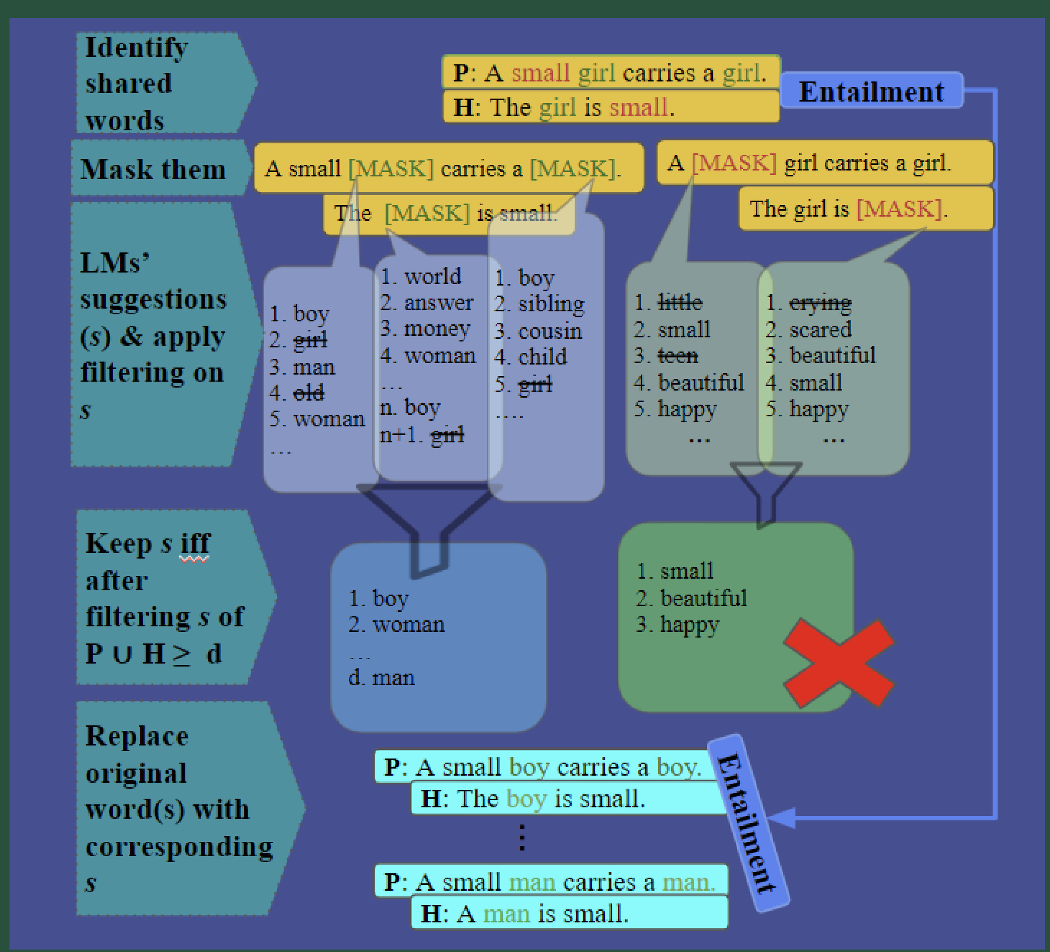


DIAGRAM 2



EXPERIMENTS (DIAGRAM 2)

- Mask shared open-class words *w* (nouns/verbs/adjectives) in SNLI test.
- Generate 200 suggestions (*s*) for all occurrences of *w* with bert-base-cased and roberta-base;
- Tag suggestions (en_core_web_sm);
- Exclude *s* if set(s) < 20 after filtering out: punctuation signs, derivational morphemes, s ≠ POS tag of *w;* probability(s) ⩽ probability(*w*);
- Variant dataset All_var: subsample 20 random suggestions for each open-class category for a NLI problem & replace them in <P,H>. Repeat 10 times. Statistics shown in Table 1.

MODELS & METRICS

- Evaluated BERT, BART, DeBERTa, RoBERTa on: ALL_Var, ALL_Var split by open-class categories; ALL_Var split by model used to generate suggestions (BERT, RoBERTa, or Both), ALL_Var with different filtering criteria for *s* (scrambled *s;* only *s* = POS tag of *w;* only with probability(*s*) ⩾ probability(w); all POS tags and probabilities).
- Metrics: Sample Accuracy (standard accuracy) and Pattern Accuracy (a correct prediction is when the model gets an x amount of variants correctly), see Diagram 1.

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

TABLE 1: STATISTICS OF ALL_VAR

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

TABLE 2

TABLE 3

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

RESULTS

- Low PA scores on high thresholds (Figure 1; 2), compared to SA scores in Table 1, further confirm a lack of generalization of models in line with previous studies [6; 3]. MERGE might dsitrub only-hyppthesis bias, or word associations between NLI problems and certain labels [5].
- Difficulty of open-class categories: verbs, followed by nouns and adjectives (Figure 3; 4).
- On higher PA thresholds, models do better on *s* from All_Both, and All_RoBERTa (Figure 6), compared to lower PA thresholds (Figure 5).
- No filtering criteria result in lower PA scores (Figure 7), but results could be influenced by other factors.

CONCLUSION

- Low PA scores on variants dataset > lack of generalization capacities.
- Models' scores influenced by the masked model source of the suggestions, the word category replaced, and by filtering criteria ⇒ strict quality control of suggestions is needed.

FUTURE RESEARCH

- Only one dataset modified; more masked models and evaluated models are needed.
- Potential confounds: disagreement the article and the noun, strategy used for scrambled words.

FIGURE 1

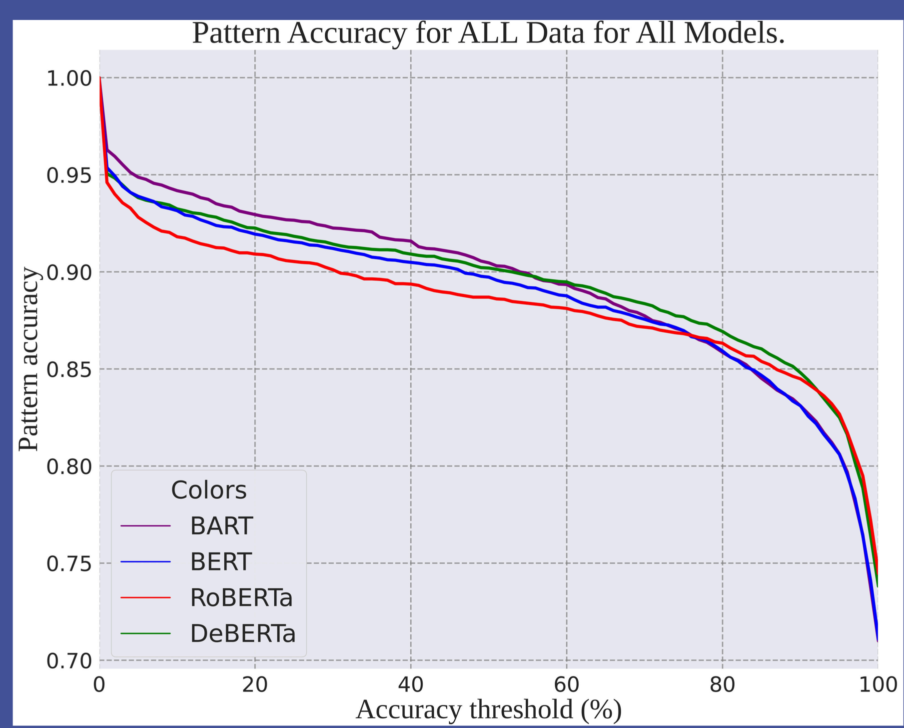


FIGURE 2

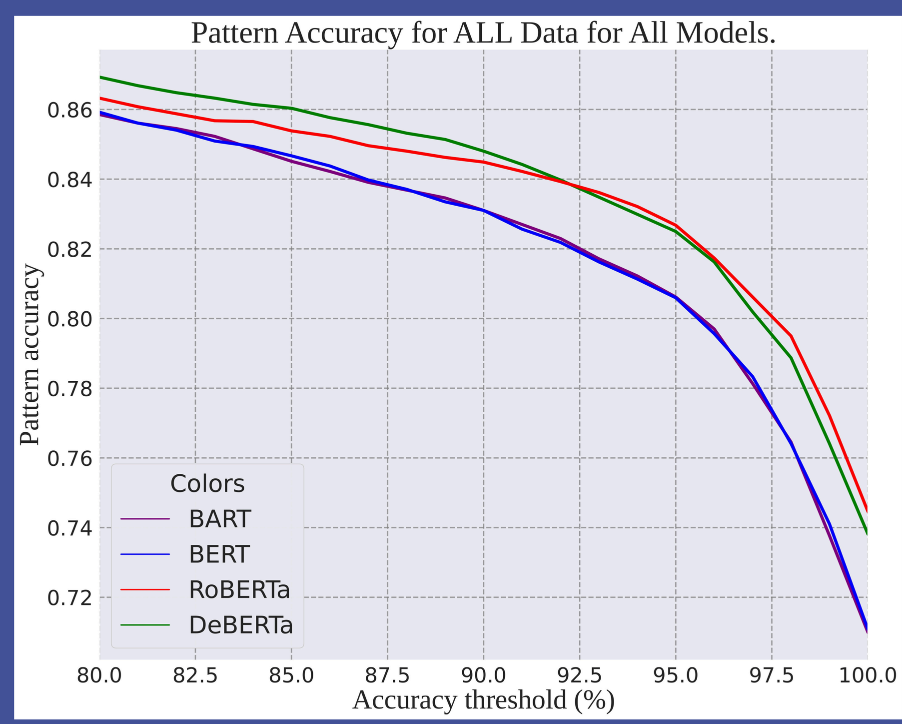


FIGURE 3

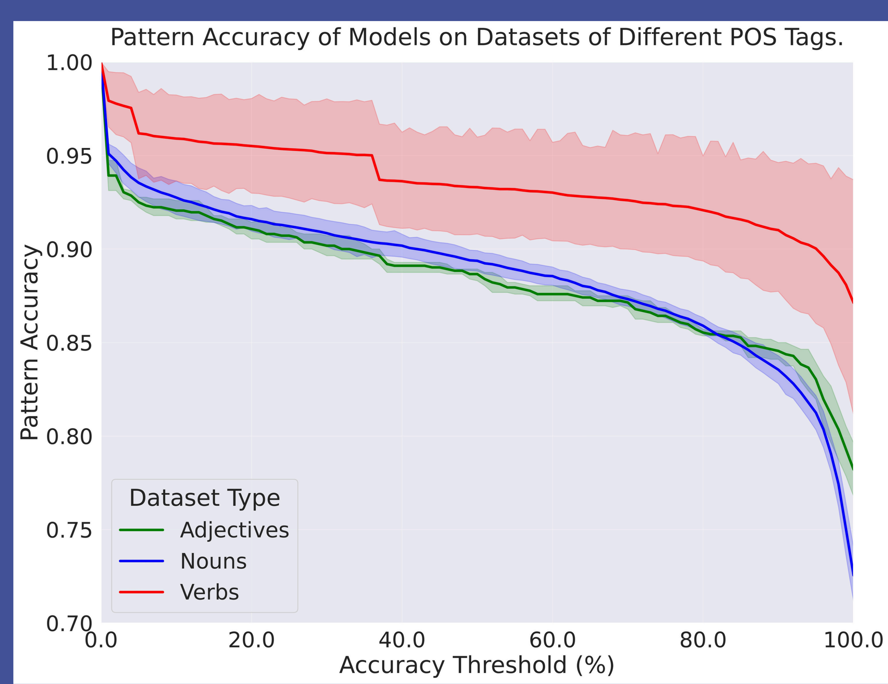


FIGURE 4

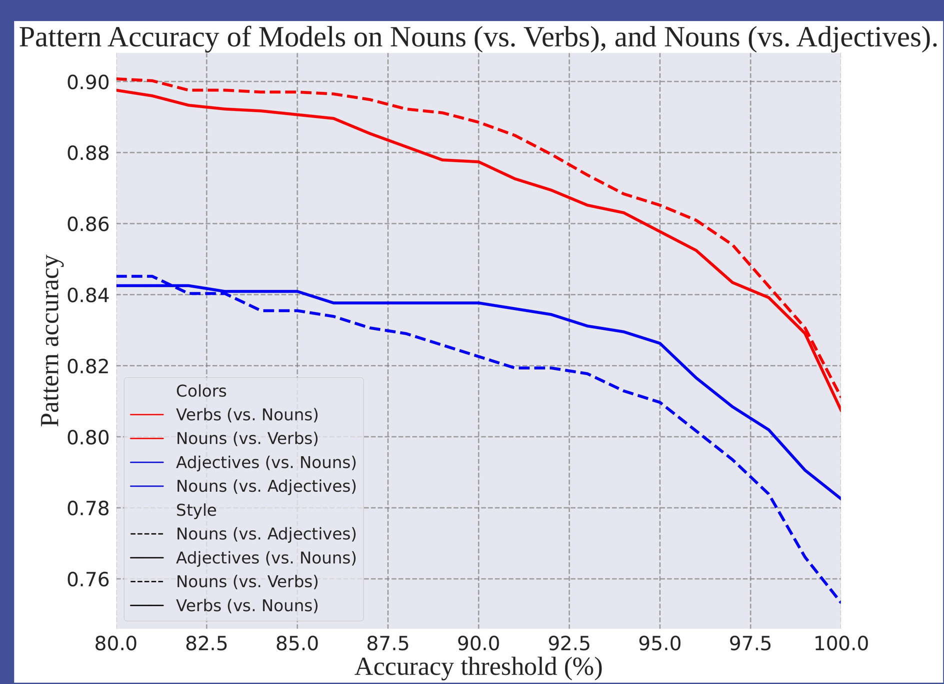


FIGURE 5

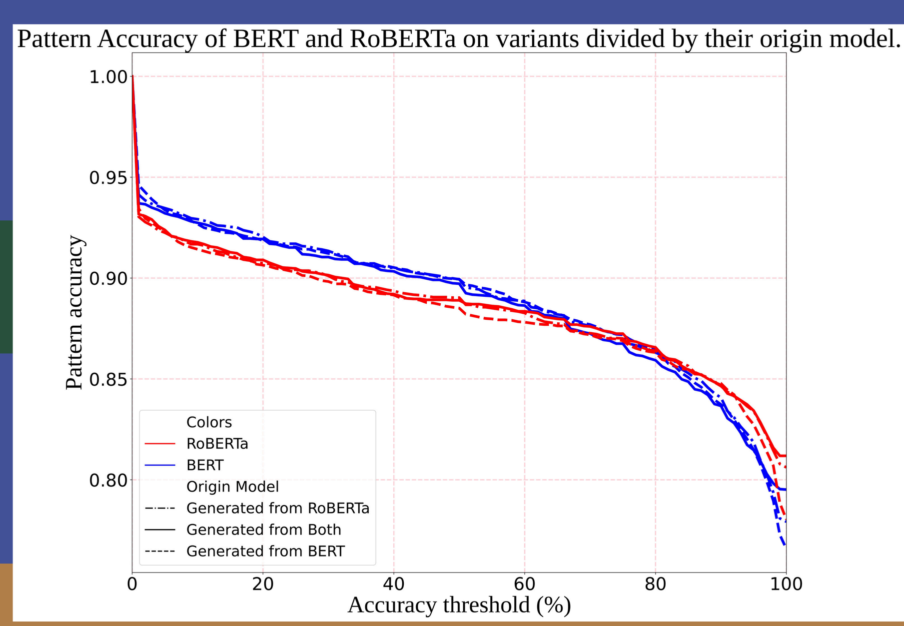


FIGURE 6

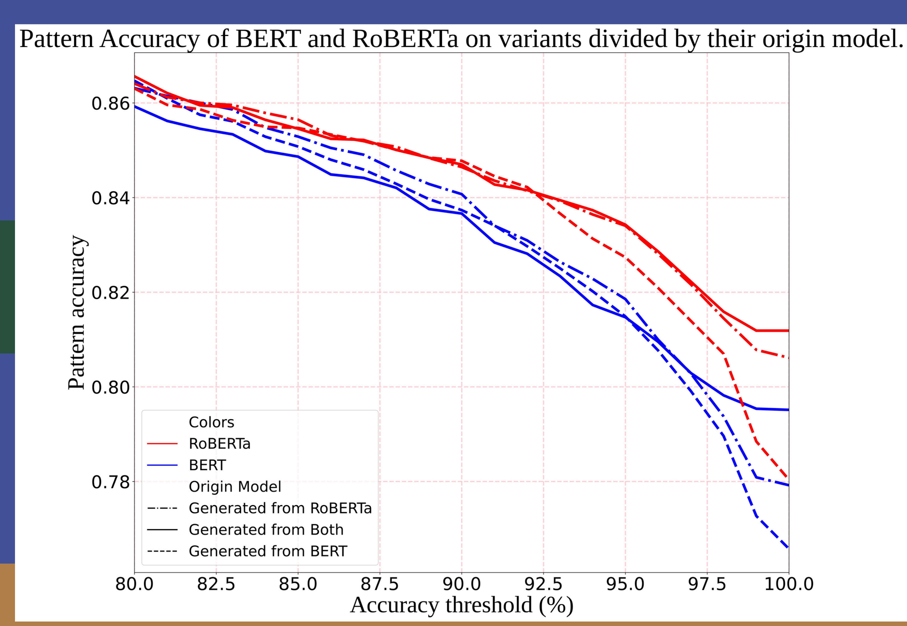
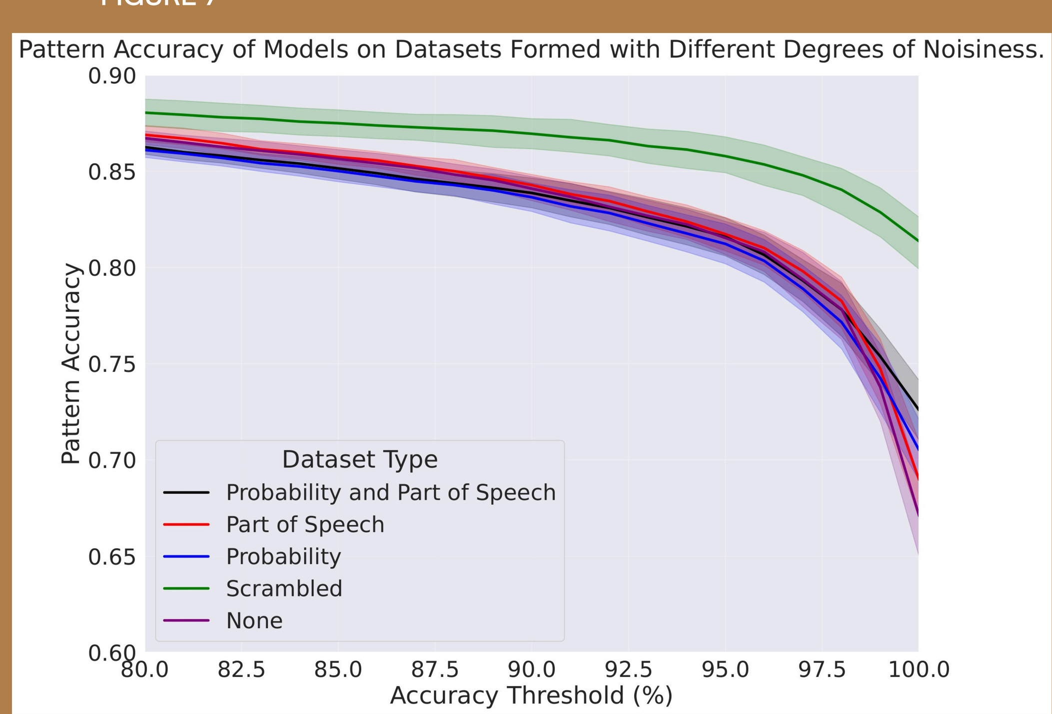


FIGURE 7



REFERENCES

[1] Abzianidze, L., Zwarts, J., & Winter, Y. (2023). SpaceNLI: Evaluating the consistency of predicting inferences in space. *arXiv preprint arXiv:2307.02269*.
[2] Arakelyan, E., Liu, Z., & Augenstein, I. (2024). Semantic sensitivities and inconsistent predictions: Measuring the fragility of NLI models. *arXiv preprint arXiv:2401.14440*.
[3] Gardner, M., Artzi, Y., Basmova, V., Berant, J., Bogin, B., Chen, S., ... & Zhou, B. (2020). Evaluating models' local decision boundaries via contrast sets. arXiv preprint arXiv:2004.02709.
[4] Glockner, M., Shwartz, V., & Goldberg, Y. (2018). Breaking NLI systems with sentences that require simple lexical inferences. *arXiv preprint arXiv:1805.02266*.
[5] Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.
that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.
[6] Li, C., Shengshuo, L., Liu, L. Z., Wu, X., Zhou, X., & Steinert-Threlkeld, S. (2020). Linguistically-informed transformations (LIT): A method for automatically generating contrast sets. *arXiv preprint arXiv:2010.08580*.
[7] Srikanth, N., & Rudinger, R. (2025). NLI under the Microscope: What Atomic Hypothesis Decomposition Reveals. *arXiv preprint arXiv:2502.08080*.
[8] Verma, D., Lal, Y. K., Sinha, S., Van Durme, B., & Poliak, A. (2023). Evaluating paraphrastic robustness in textual entailment models. *arXiv preprint arXiv:2306.16722*.