On the Zero-Shot Generalization of Machine-Generated Text Detectors

Xiao Pu, Jingyu Zhang, Xiaochuang Han, Yulia Tsvetkov, Tianxing He

1.Motivation

- Due to the wide adoption of large language models (LLMs), machine-generated misinformation has become a urgent societal threat, giving unprecedented importance to the detection of machine-generated text.
- This work is motivated by an important research question: **How will the detectors of machine-generated text perform on outputs of a new generator, that the detectors were not trained on?**

2.Experiment Setup

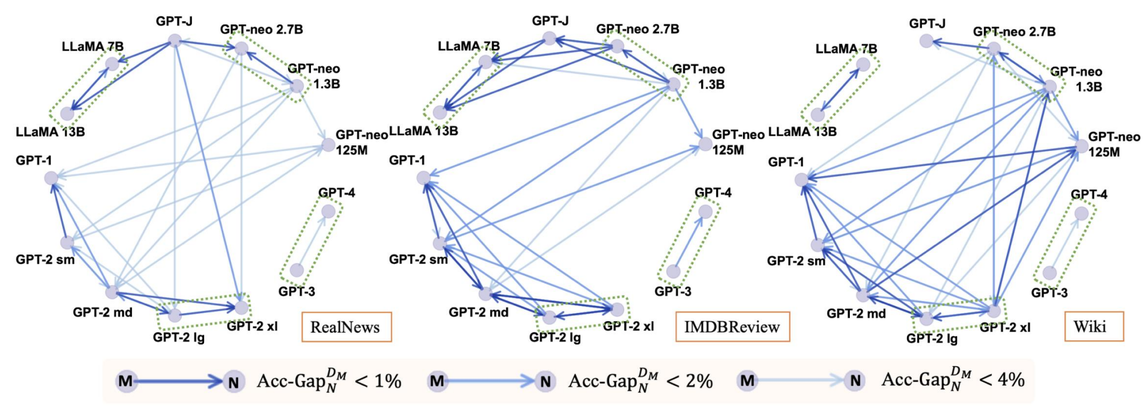
- We collect generation data from a wide range of LLMs, and train neural detectors on data from each generator and test its performance on held-out generators.
- Generators: GPT1, GPT-2 models (small, medium, large, and xl), GPT3, GPT-4, GPT-Neo models (125M, 1.3B and 2.7B), GPT-J and LLaMAs (7B and 13B) .
- Detectors: for data from each generator, we train an ELECTRA-large model as a binary classifier.
- Datasets: RealNews, IMDBreview and Wikipedia.

3. Measuring Generalization Ability of Detectors

- We use Acc−Gap to measure the drop of performance when the detector is trained on generator M instead of N itself:

$$\text{Acc-Gap}_N^{D_M} = \text{Acc}_N(D_N) - \text{Acc}_N(D_M).$$

- We expect Acc-Gap to be larger than zero in general, and a large Acc-Gap means D_M has poor generalization on generator N.
- Acc-Gap of each detector/generator pair. We link from node M to node N if $\text{Acc-Gap}_N^{D_M} < T$ (good generalization), where the threshold T is set to a small number from {1%, 2%, 4%} :

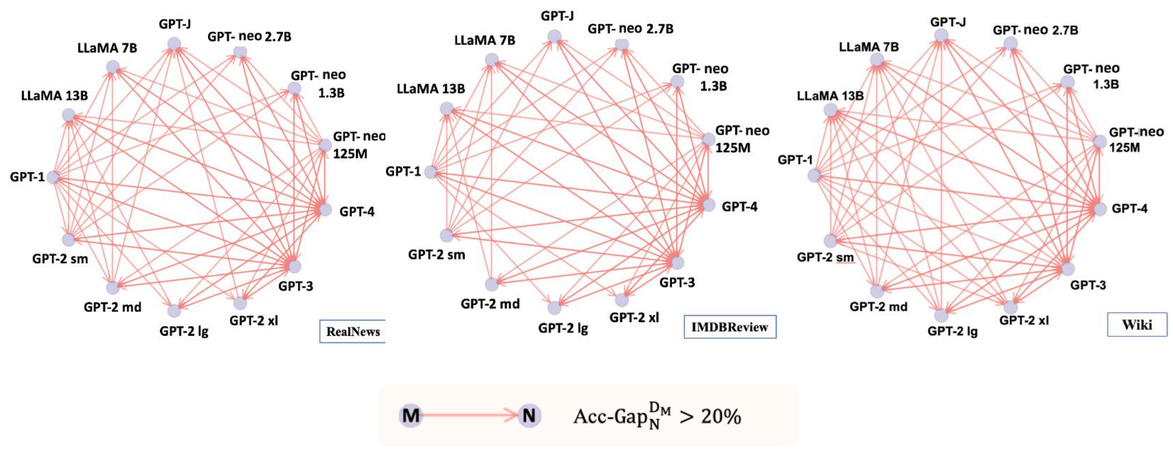


4. Two Interesting Generalization Patterns

- **The detectors for the medium-version LMs can generalize to the large-version models**, while the generalization of the reverse direction is weaker.

M	N	RealNews Gap$_N^{D_M}$	RealNews Gap$_M^{D_N}$	IMDBReview Gap$_N^{D_M}$	IMDBReview Gap$_M^{D_N}$
GPT3	**GPT4**	3.64%	5.46%	1.47%	5.48%
LLa7B	**LLa13B**	-1.11%	1.18%	-1.50%	1.47%
Neo1.3B	**Neo2.7B**	0.04%	2.16%	-2.31%	3.41%
GPT2lg	**GPT2xl**	-4.40%	4.94%	-0.02%	0.32%

- We plot detector-generator pairs with large (>20%) Acc-Gap on the three datasets, and find that **none of the detectors, on its own, can generalize to all generators**.



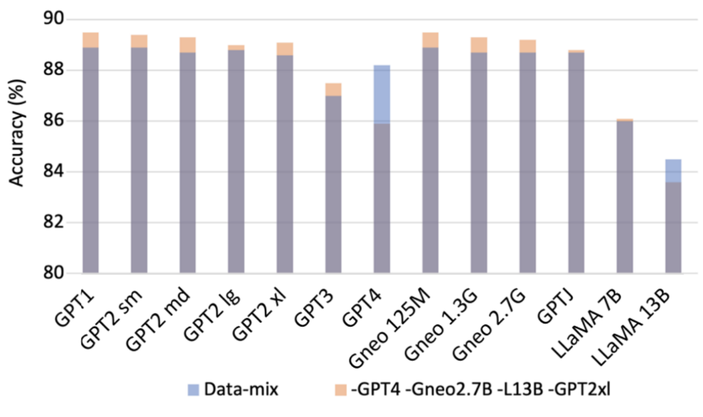
If we want an “universal” detector which can cover all generators, an ensemble of detectors/data is necessary.

5. Pruning out Large-Version models in a Mixed Training Dataset

- We now demonstrate a concrete application of our findings, and the following realistic threat scenario is considered: The task is still binary classification but the machine text is composed of generations from a range of models.
- We propose to prune out data from the large-version models when training a detector by mixing up training data from generators. Here is how it compares to baselines:

	Baseline			Pruned (Proposed)		Pruned (Comparison)	
RealNews Acc(%)	**Ensemble Vote**	**Ensemble Prob-avg**	**Data-mix**	**-GPT4 -L13B**	**-GPT4-GNeo2.7B -L13B-GPT2xl**	**-GPT3 -GPT4**	**-L13B -L7B**
Average	81.1	81.1	**88.0**	88.6 (+0.6)	88.2 (+0.2)	79.5 (-8.5)	91.6 (+3.6)
Worst-case	50.6	50.5	**84.5**	84.6 (+0.1)	83.6 (-0.9)	42.7 (-41.8)	80.8 (-3.7)
GPT4	50.6	50.5	**88.2**	84.6 (-3.6)	85.9 (-2.3)	42.7 (-45.5)	93.2 (+5.0)
GPT3	52.9	52.8	**87.0**	87.2 (+0.2)	87.5 (+0.5)	52.3 (-34.7)	91.4 (+4.4)
L13B	58.8	58.4	**84.5**	85.0 (+0.5)	83.6 (-0.9)	83.5 (-1.0)	80.8 (-3.7)
L7B	61.6	61.3	**86.0**	86.1 (+0.1)	86.1 (+0.1)	83.7 (-2.3)	82.6 (-3.4)
IMDBReview Acc(%)	**Ensemble Vote**	**Ensemble Prob-avg**	**Data-mix**	**-GPT4 -L13B**	**-GPT4-GNeo2.7B -L13B-GPT2xl**	**-GPT3 -GPT4**	**-L13B -L7B**
Average	85.2	85.1	**94.3**	93.2 (-1.1)	94.1 (-0.2)	89.0 (-5.3)	93.7 (-0.6)
Worst-case	52.0	51.8	**93.7**	92.2 (-1.5)	92.7 (-1.0)	62.6 (-31.1)	89.8 (-3.9)
GPT4	52.0	51.8	**94.4**	93.4 (-1.0)	94.4 (0)	62.8 (-31.6)	94.4 (0)
GPT3	54.1	54.2	**93.7**	93.1 (-0.6)	93.5 (-0.2)	62.6 (-31.1)	93.7 (0)
L13B	70.2	70.1	**94.0**	92.2 (-1.8)	92.7 (-1.3)	93.5 (-0.5)	89.8 (-4.2)
L7B	72.1	71.9	**94.1**	93.0 (-1.1)	93.9 (-0.2)	93.7 (-0.4)	91.7 (-2.4)

- In the case of limited budget or computing, **data from the medium version LM can decently approximate the large version in an ensembled data collection**:



6. Conclusion and Discussion

- We observe a generalization relationship among detectors trained on different generators, where detectors for medium version models demonstrate the ability to effectively generalize to the larger-version.
- Building upon this finding, we prune out data from large version generators in an ensembled training dataset and demonstrate that the performance loss is minimal.
- With the rapid release of various LLMs and generation APIs, a detector needs to cover a wide range of generators. Our experiments show that the detection of an unseen (or non-public) generator is still a difficult and open question.

Our dataset is available at https://github.com/SophiaPx/detectors-generalization.

Please scan QR code: