

TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation

Adaku Uchendu Zeyu Ma[†] Thai Le Rui Zhang Dongwon Lee

The Pennsylvania State University, University Park, PA, USA

{azu5030, thaile, rmz5227, dongwon}@psu.edu

Carnegie Mellon University, Pittsburgh, PA, USA[†]

mazeyuwhu@gmail.com[†]

Abstract

Recent progress in generative language models has enabled machines to generate astonishingly realistic texts. While there are many legitimate applications of such models, there is also a rising need to distinguish machine-generated texts from human-written ones (e.g., fake news detection). However, to our best knowledge, there is currently no benchmark environment with datasets and tasks to systematically study the so-called “Turing Test” problem for neural text generation methods. In this work, we present the TURINGBENCH benchmark environment, which is comprised of (1) a dataset with 200K human- or machine-generated samples across 20 labels {Human, GPT-1, GPT-2_small, GPT-2_medium, GPT-2_large, GPT-2_xl, GPT-2_PyTorch, GPT-3, GROVER_base, GROVER_large, GROVER_mega, CTRL, XLM, XLNET_base, XLNET_large, FAIR_wmt19, FAIR_wmt20, TRANSFORMER_XL, PPLM_distil, PPLM_gpt2}, (2) two benchmark tasks—i.e., *Turing Test* (TT) and *Authorship Attribution* (AA), and (3) a website with leaderboards. Our preliminary experimental results using TURINGBENCH show that FAIR_wmt20 and GPT-3 are the current winners, among all language models tested, in generating the most human-like indistinguishable texts with the lowest F1 score by five state-of-the-art TT detection models. The TURINGBENCH is available at: <https://turingbench.ist.psu.edu/>

1 Introduction

Recently, the field of Natural Language Generation (NLG) has seen a massive improvement. While the field of NLG has existed for some time since even before the onset of the first chatbot ELIZA (Weizenbaum, 1966), the recent neural architecture Transformers (Vaswani et al., 2017) has led to speedy improvement in the generation of long coherent texts. GPT-1 (Radford et al., 2018) created by OpenAI is the first installment of these

PROMPT: Some investigators struggle to reach Comey

HUMAN: Some congressional investigators want to bring former FBI Director James Comey to the Hill to tell them about his meetings with President Donald Trump but they can't find him.

GPT-2: Some investigators struggle to reach Comey and his team because the FBI's headquarters is in Northern Virginia. But the FBI has a lot of other people who can help.

FAIR: We have a duty to be vigilant. James Comey's testimony today should make him think twice about it. The White House said Trump had spoken with Comey during the hour-long meeting and that all the senior U.S. officials were given a "heads up" on the situation.

GPT-3: “Like Louis XVI, he doesn’t see the storm growing in the distance,” says the Democratic operative. The lack of specifics, even from surrogates on Trump’s behalf, forces well-known Democrats to point out the obvious.

Figure 1: Snippets of **PROMPT** (a news article title), **HUMAN** vs. three examples of machine-generated texts (**GPT-2**, **FAIR** and **GPT-3**).

new waves of text-generators. In fact, GPT-1 was built with 117 million parameters, however, in less than 3 years, Google’s Switch Transformer (Fedus et al., 2021) was the largest language model with 1.6 trillion parameters as of January-June 2021. Currently, the largest language model is Beijing Academy of Artificial Intelligence’s (BAAI) Wu-Dao 2.0 with 1.75 trillion parameters. Even more alarming, since the birth of GPT-1, the field of NLG has grown exponentially such that Hugging Face’s model repo houses more than 9K English and non-English language models (of which over 2K are text-generators). See Figure 2 for evolution of neural text-generators. Naturally, these newer language models are able to generate texts that can be easily misconstrued as human-written. Thus, due to the superior quality of recent generated texts and how easily such text-generators can be used, the potential for misuse is great. This misuse includes but is not limited to the spread of *misinformation* (Zellers et al., 2019) and *political propaganda*.

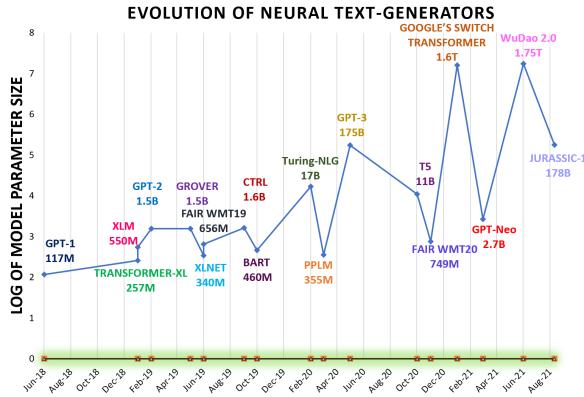


Figure 2: Evolution of neural text generators (Y-axis depicts model parameters in millions in log plot).

ganda (Varol et al., 2017). Therefore, it is urgent that we tackle ways to automatically distinguish machine-generated texts from human-written ones accurately.

To build accurate detectors of machine-generated texts, sufficient data is required but lacking. Therefore, we create a benchmark environment, **TURINGBENCH**, to combat the obvious security issue language models could pose. Just in line with benchmark environments such as SQuAD (Rajpurkar et al., 2016) and GLUE (Wang et al., 2018) that tremendously facilitate the progress of Natural Language Understanding, we build the first benchmark for Authorship Attribution in the form of the Turing Test by including humans and neural language models.

The **TURINGBENCH** Environment comprises benchmark datasets, benchmark tasks, and a website to host leaderboards. This benchmark dataset is created by collecting 10K news articles (mostly in politics) written by journalists in media outlets such as CNN, Washington Post, etc. Using the **Title** of each article, we **Prompt** 19 selected neural text-generators to generate an article similar to the human-written one. This creates 200K articles with 20 labels (or authors). Next, we have two benchmark tasks - *Turing Test* and *Authorship Attribution*. The *Turing Test* task is modeled after the *Turing Test* concept (Turing, 2009), where if a machine shows intelligent behavior or characteristics usually attributed to a human, then the machine has passed the test. In this scenario, the goal is to cause the machine to fail the *Turing Test*. Thus, we define this benchmark task as a binary classification problem with *human* and *machine* labels. Given 19 neural text-generators, there are 19 *Turing Test*

subtasks with 19 human-machine pairs.

Furthermore, we understand that due to the ubiquitous nature of these neural language models, distinguishing machine-generated texts from human-written ones is no longer sufficient. It is now also important we inquire as to which particular neural text-generator authored a piece of text. To this end, the *Authorship Attribution* task aims to assign authorship to one of the many text-generators. We study 20 authors for this task, however, as we have observed, this can easily become 2K authors very soon which will grossly exacerbate the difficulty of this task. Finally, to host all these tasks and datasets, we build a **TURINGBENCH** website with leaderboards for each benchmark task and call for participation in tackling this very relevant and non-trivial problem.

Lastly, we compare State-of-the-art (SOTA) and baseline *Turing Test* and *Authorship Attribution* models. From the experimental results, we observe that we need more complex models to accurately distinguish machine-generated texts from human-written ones, including text-generators that are yet to be created.

2 Related Work

Neural Text Generation Recent advances in neural network-based language modeling have demonstrated promising results in text generation (Garbacea and Mei, 2020). Current state-of-the-art neural text generation models can produce texts approaching the quality of human-written ones, especially in terms of grammar, fluency, coherency, and usage of real world knowledge (Radford et al., 2018, 2019; Keskar et al., 2019; Zellers et al., 2019; Deng et al., 2019; Brown et al., 2020). The progress in neural text generation has facilitated a wide range of applications: dialog response generation (Zhang et al., 2020), storytelling (Fan et al., 2018; See et al., 2019), table-to-text generation (Lebret et al., 2016), code comment generation (Alon et al., 2018), medical report generation (Liu et al., 2019a).

However, as these language models can generate text indistinguishable from human-written text, they can also be misused by adversaries to generate fake news (Shu et al., 2017; Wang, 2017; Zellers et al., 2019; Mosallanezhad et al., 2020; Shu et al., 2021), fake product reviews (Fornaciari and Poesio, 2014; Adelani et al., 2020), spam emails (Das and Verma, 2018).

Automatic Detection of Generated Text Given the potential malicious applications of text generation (Solaiman et al., 2019), it is thus vital to build detectors to distinguish text generated by machines from humans (Gehrman et al., 2019; Bakhtin et al., 2019; Jawahar et al., 2020; Varshney et al., 2020; Çano and Bojar, 2020). Most current work focus on fake news detection (Rashkin et al., 2017; Zhou et al., 2019; Bhat and Parthasarathy, 2020; Zhong et al., 2020; Schuster et al., 2020; Ippolito et al., 2020). Despite this progress, it remains a challenging task to build generalizable, interpretable, and robust detectors (Jawahar et al., 2020).

Authorship Attribution Authorship Attribution (AA) aims to decide the author of a given text from a set of candidates (Houvardas and Stamatatos, 2006; Stamatatos, 2009b; Zhang et al., 2014). AA has a broad range of applications including author profiling (López-Monroy et al., 2020), computer forensics (Lambers and Veenman, 2009), and plagiarism detection (Stamatatos, 2009a). Previous work on AA has explored and combined various features and representations at different levels including n-grams (Escalante et al., 2011; Sapkota et al., 2015, 2016), POS-tags (Ferracane et al., 2017; Halvani et al., 2020) psycholinguistics features (Li et al., 2014; Uchendu et al., 2019), while recent approaches also build deep neural network based classifiers such as feed-forward NNLMs (Ge et al., 2016), CNNs (Hitschler et al., 2017; Shrestha et al., 2017), LSTMs (Jafariakinabad and Hua, 2019, 2020), and BERT-based models (Uchendu et al., 2020).

However, previous AA work largely focuses on authorship attribution among humans, while only a few papers (Manjavacas et al., 2017; Uchendu et al., 2020; Munir et al., 2021) study neural generated text. Our work aims to provide the first benchmark for Authorship Attribution in the form of the Turing Test by including humans and neural language models.

3 The TURINGBENCH Environment

Figure 3 overviews the framework of the TURINGBENCH Environment.

3.1 Chosen Language Models

We generated texts using 10 language model architectures - *GPT-1* (Radford et al., 2018), *GPT-2* (Radford et al., 2019), *GPT-3* (Brown et al., 2020), *GROVER* (Zellers et al., 2019),

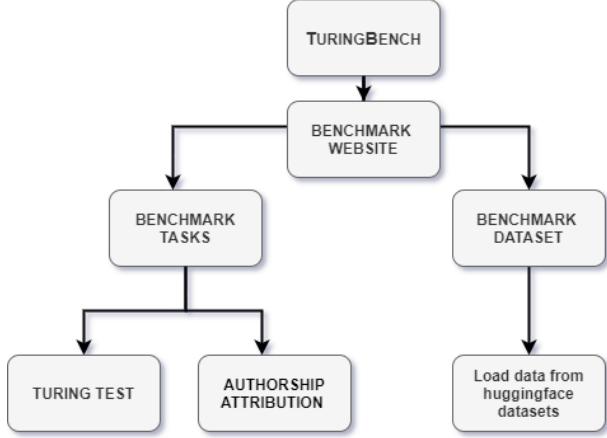


Figure 3: The TURINGBENCH Environment.

CTRL (Keskar et al., 2019), *XLM* (Lample and Conneau, 2019), *XLNET* (Yang et al., 2019), *FAIR* (Ng et al., 2019; Chen et al., 2020), *TRANSFORMER-XL* (Dai et al., 2019), and *PPLM* (Dathathri et al., 2020). In addition, some of these language models have multiple pre-trained models and thus, we were able to generate texts with 19 neural machine text-generators. We choose these 10 language model architectures because they are currently considered as the SOTA text-generators, many of the text-generators on Hugging Face’s model repo are variants of these language models, and both their pre-trained models and codes were publicly available.

To generate texts, all 19 neural generators require a short prompt and a specified number of words to generate texts. Table 1 (and Appendix) describes each language model in detail. Figure 4 illustrates the data creation process. Table 2 summarizes the stats of dataset and the model sizes.

3.2 TURINGBENCH Benchmark Tasks

The Turing Test (TT) Task Our proposed Turing Test task aims to answer the question: *Can we determine if a piece of text is human-written or machine-generated?* This task is formulated as a binary classification problem with two labels – *human* and *machine* – modeled after the classical *Turing Test* problem. The *Turing Test* examines the ability of a machine text-generator to exhibit intelligible behavior ascribed to humans. The goal is to build a model that causes the machine-generated texts to fail the *Turing Test*. Lastly, the TT task contains 19 subtasks with 19 human-machine pairs (e.g. GPT-2 XL vs. Human, GROVER_base vs. Human, etc.).

Text Generator	Description
Human	We collected news titles (mostly Politics) and contents from CNN, Washington Post, and Kaggle. The Kaggle datasets had news articles from 2014–2020, and 2019–2020 for the CNN and Washington Post news articles. Next, we removed articles that did not have the desired word length (i.e., 200–500). This resulted in 130K articles, but only 10K was used for the article generations. See data generation process in Figure 4.
GPT-1	Texts are generated with the Hugging Face github repo (Wolf et al., 2019).
GPT-2	We use 5 GPT-2 pre-trained models - PyTorch model, small (124 million parameters), medium (355 million parameters), large (774 million parameters), and x-large (1558 million parameters) to generate texts.
GPT-3	Texts are generated with the OpenAI GPT-3 API using the <i>davinci</i> engine.
GROVER	We use code from repo to generate from Grover’s 3 pre-trained models: GROVER-base , GROVER-large , GROVER-mega .
CTRL	Conditional Transformer Language Model For Controllable Generation uses control codes to guide generation. We use <i>News</i> control code to generate long articles.
XLM	We generated texts using Hugging Face repo (Wolf et al., 2019).
XLNET	We generated texts with: 2 XLNET pre-trained models: XLNET-base , and XLNET-large using Hugging Face.
FAIR_wmt	We use two Facebook’s FAIR English models - wmt19 (Ng et al., 2019) and wmt20 (Chen et al., 2020) to generate texts with FAIRSEQ sequence modeling toolkit.
TRANSFORMER_XL	We generated texts with this language model’s setup on Hugging Face (Wolf et al., 2019).
PPLM	PPLM fuses GPT-2’s pre-trained model with bag of words to generate more specific texts. We used the <i>Politics</i> bag of words model to generate texts. Next, we fused PPLM with two pre-trained models (i.e., distilGPT-2, and GPT-2) and generated texts with them, forming: PPLM_distil , PPLM_gpt2 . These models are gotten from the Hugging Face model repository.

Table 1: Description of the text generators in the TURINGBENCH dataset.

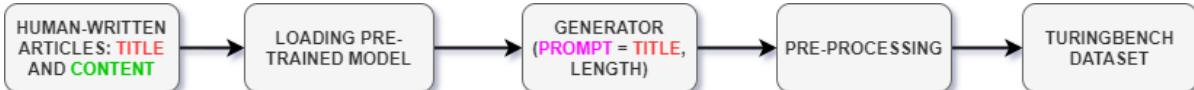


Figure 4: The TURINGBENCH Data Collection, Generation, and Building process.

Text Generator	# of words (AVG \pm Std. Dev.)	# of sentences (AVG \pm Std. Dev.)	Model Parameter Size
Human	232.7 \pm 42.0	15.0 \pm 6.6	N/A
GPT-1	316.7 \pm 12.9	10.5 \pm 3.7	117M
GPT-2_small	118.6 \pm 61.0	4.0 \pm 3.8	124M
GPT-2_medium	120.9 \pm 66.0	4.2 \pm 3.7	355M
GPT-2_large	119.7 \pm 62.1	4.1 \pm 3.8	774M
GPT-2_xl	117.8 \pm 63.3	4.1 \pm 3.8	1.5B
GPT-2_PyTorch	178.9 \pm 55.4	7.03 \pm 4.8	344M
GPT-3	129.5 \pm 54.9	5.0 \pm 3.7	175B
GROVER_base	299.2 \pm 108.6	9.4 \pm 6.9	124M
GROVER_large	286.3 \pm 101.3	8.7 \pm 5.9	355M
GROVER_mega	278.9 \pm 97.6	9.2 \pm 6.1	1.5B
CTRL	398.1 \pm 64.8	20.0 \pm 10.6	1.6B
XLM	387.8 \pm 30.3	4.2 \pm 1.7	550M
XLNET_base	226.1 \pm 97.5	11.6 \pm 7.9	110M
XLNET_large	415.8 \pm 53.2	4.3 \pm 2.1	340M
FAIR_wmt19	221.2 \pm 66.6	14.6 \pm 6.0	656M
FAIR_wmt20	100.6 \pm 28.1	5.1 \pm 3.0	749M
TRANSFORMER_XL	211.7 \pm 53.9	9.8 \pm 3.1	257M
PPLM_distil	156.9 \pm 40.1	10.7 \pm 3.6	82M
PPLM_gpt2	188.9 \pm 52.0	11.9 \pm 4.5	124M

Table 2: Summary statistics of the TURINGBENCH dataset.

The Authorship Attribution (AA) Task *Authorship Attribution* is the identification and proper assignment of the author of a piece of text (Coyotl-Morales et al., 2006). Our Authorship Attribution task aims to answer the question: *If we determine that an article is human-written or machine-generated, can we further determine which neural language model generated all the articles that are*

said to be machine-generated? This is a multi-class classification problem modeled after the traditional *Authorship Attribution* problem.

3.3 TURINGBENCH Benchmark Dataset

We keep 168,612 articles out of 200K after cleaning the text (see Appendix for data pre-processing details), and we build the benchmark dataset for

```

from datasets import load_dataset
import pandas as pd

# GPT-1 TT task
TT_gpt1 = load_dataset(
    'turingbench/TuringBench',
    name='TT_gpt1', split='train')
TT_gpt1 = pd.DataFrame.from_dict(
    TT_gpt1)

# AA task
AA = load_dataset(
    'turingbench/TuringBench',
    name='AA', split='train')
AA = pd.DataFrame.from_dict(AA)

```

Figure 5: Python code for loading the TURINGBENCH datasets using the Hugging Face API.

each benchmark task - *TT* and *AA*. For the *TT* task, there are 20 labels (i.e., 19 machine text-generators and 1 human), thus we can only have 19 pairs of human vs. machine. Therefore, we have 19 datasets for the *TT* task. To increase the difficulty of the *TT* task, we cut each article in the test set in half, using only 50% of the words. For the *AA* task, we have 1 dataset containing all the labels. All datasets have train/validation/test sets which were split using the 70:10:20 ratio, respectively. To avoid topic bias, these sets were carefully split, such that all articles in the sets were unique to each other. Therefore, all articles generated by a prompt belonged only to one set.

To make this dataset public, we added our datasets for each benchmark task and subtask to Hugging Face datasets¹. Figure 5 demonstrates how to load the TURINGBENCH dataset.

Evaluation Metrics We use the traditional evaluation metrics such as: Precision, Recall, F1 score, and Accuracy to evaluate Machine/Deep Learning models for the benchmark tasks. However, for the *TT* tasks, we only use F1 scores since it is a more robust measure for the imbalanced datasets.

3.4 The Web Environment

To create this TURINGBENCH environment, we built 2 versions of datasets - binary setting (i.e., *human vs. GROVER-large*, *human vs. GPT-1*, etc.) for the *TT* tasks, and multi-class setting (i.e., *human vs. GROVER-Large vs. GPT-1 vs. etc.*) for the *AA* task. To track progress, as shown in Figure 6, we create a website where each task and sub-task

¹<https://huggingface.co/datasets/turingbench/TuringBench/tree/main>

Leaderboard: Authorship Attribution

The TuringBench Datasets will assist researchers in building robust Machine learning and Deep learning models that can effectively distinguish machine-generated texts from human-written texts. This Leaderboard is for the Authorship Attribution scenario.

Rank	Model	Precision	Recall	F1	Accuracy
1 May 5, 2021	RoBERTa (Liu et al., '19)	0.8214	0.8126	0.8107	0.8173
2 May 5, 2021	BERT (Devlin et al., '18)	0.8031	0.8021	0.7996	0.8078
3 May 5, 2021	BertAA (Fabien et al., '20)	0.7796	0.7750	0.7758	0.7812
4 May 5, 2021	OpenAI detector	0.7810	0.7812	0.7741	0.7873
5 May 5, 2021	SVM (3-grams) (Sapkota et al. '15)	0.7124	0.7223	0.7149	0.7299
6 May 5, 2021	N-gram CNN (Shrestha et al., '17)	0.6909	0.6832	0.6665	0.6914
7 May 5, 2021	N-gram LSTM-LSTM (Jafariakaband, '19)	0.6694	0.6824	0.6646	0.6898
8 May 5, 2021	Syntax-CNN (Zhang et al. '18)	0.6520	0.6544	0.6480	0.6613
9 May 5, 2021	Random Forest	0.5893	0.6053	0.5847	0.6147
10 May 5, 2021	WriteprintsRFC (Mahmood et al. '19)	0.4578	0.4851	0.4651	0.4943

Figure 6: A screenshot of a leaderboard on the TURINGBENCH website.

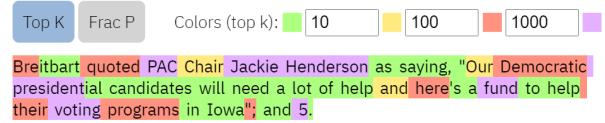


Figure 7: Using GLTR (Gehrmann et al., 2019) on a piece of text generated by GPT-3. Green represents the most probable words; yellow the 2nd most probable; Red the least probable; and purple the highest improbable words. Machine-generated texts are often populated with mostly Green and yellow words. However, we see that GPT-3-generated texts is very human-like.

has its own leaderboard that displays the evaluation metric scores of models. Furthermore, to ensure the integrity of the process, even though contributors can obtain the TURINGBENCH datasets from Hugging Face datasets, we still ask contributors to submit their code and/or trained model weights for private testing. After testing, we update the website with the new models’ scores. Lastly, we rank the model performance using the F1 score from best to worst.

4 Experiments

We experiment with several SOTA and baseline models as summarized in Table 3 for **Turing Test** and Table 4 for **Authorship Attribution**, and Ta-

TT Model	Description
GROVER detector	We use the GROVER-Large discriminator that is trained to detect GROVER-generated texts to predict the test labels.
GPT-2 detector	We use the trained weights of RoBERTa-large fine-tuned on GPT-2 XL outputs to predict the <i>human</i> and <i>machine</i> label of the test dataset.
GLTR	In the GLTR demo, the words are color coded to improve human detection of machine-generated texts. Top 0-10 probable words are green ; top 10-100 probable words are yellow ; top 100-1000 probable words are red and top greater than 1000 words are purple . See Figure 7 for an example of using GLTR and interpretation of its color schemes. Thus, we define human-written texts to be any article that 10% or more of the words belong in the top >1000 (i.e., purple words).
BERT RoBERTa	We fine-tune <i>bert-base-cased</i> on the train set and classify on the test set. We fine-tune RoBERTa-base, a variant of BERT with the train set.

Table 3: Description of the Turing Test (TT) models.

AA Model	Description
Random Forest	Using TF-IDF to represent the data, we classify the texts with Random Forest.
SVM (3-grams)	We represent the texts as 3-grams and classify the texts with SVM.
WriteprintsRFC	Writeprints features + Random Forest Classifier.
OpenAI detector	We re-purposed RoBERTa-base (<i>roberta-base-openai-detector</i>) model that was originally fine-tuned on GPT-2 XL outputs to detect machine-generated texts, by training the model as a multi-classifier for the AA task.
Syntax-CNN	Use Part-Of-Speech to capture the syntax of the texts and classify the texts with CNN
N-gram CNN	Represent the data with n-grams (uni-grams) and classify texts with CNN
N-gram LSTM-LSTM	Represent the data with n-grams (uni-grams) and classify texts with LSTM
BertAA	Using BERT + Style + Hybrid features to achieve automatic authorship attribution. Style features include: <i>length of text, number of words, average length of words</i> , etc. and Hybrid features include: <i>frequency of the 100 most frequent character-level bi-grams and the 100 most frequent character-level tri-grams</i> .
BERT-Multinomial	Using BERT for multi-class classification
RoBERTa-Multinomial	Using RoBERTa for multi-class classification

Table 4: Description of the Authorship Attribution (AA) models.

ble 5 and Table 6 show their results.

4.1 Results from Turing Test

The *Turing Test* task is formulated as a binary classification problem with *human* and *machine* labels. In order to make the TT task even more difficult, we train and validate on the full articles generated by the text-generators and test on only 50% of the words of each article in the test set. We intend to capture the differences that will exist between train and test data in the real world in this scenario.

We compare 3 SOTA TT models - GROVER detector (Zellers et al., 2019), GPT-2 detector (Sohaiman et al., 2019), and GLTR (Gehrmann et al., 2019). We observe in Table 5 that the average F1 scores are 0.56, 0.60, and 0.57, respectively. Next, using other text classifiers such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019b) brings a significant improvement in F1 scores (0.85 for both BERT and RoBERTa).

This performance improvement occurs mainly because BERT and RoBERTa are fine-tuned with the train set of each TT subtasks, while the TT models’ pre-trained models were used to classify the test set without any further training.

Additionally, averaging over all the 5 TT models, we find that FAIR_wmt20 and GPT-3, the most recent text-generators in the list, achieve the lowest average F1 score (0.49 and 0.55), thus making them the language models that produce the most indistinguishable texts, while XLNET_large has the highest average F1 score (0.87) using all TT models. XLNET has a high F1 score because it implements a text padding technique for generation which often negatively affects the generation quality.

We also run two human experiments using the Amazon Mechanical Turk (AMT) environment, recruiting workers with at least 95% approval rate of Human Intelligence Task (HIT). In the experiments, we randomly sampled 50 articles per each language model (across all 19 models) and performed two tests, where workers (1) vote if a given article is machine-generated or not, and (2) vote which of two given articles is machine-generated. These experiments yielded the AVG-accuracies of 0.535 and 0.513 (random-guess=0.5), respectively.

This part of experiments was reviewed and approved by the Institutional Review Board of our institution.

Human vs.	Human Test (machine)	Human Test (human vs. machine)	GROVER detector	GPT-2 detector	GLTR	BERT	RoBERTa	AVG
GPT-1	0.4000	0.5600	0.5792	0.9854	0.4743	0.9503	0.9783	0.7935
GPT-2_small	0.6200	0.4400	0.5685	0.5595	0.5083	0.7517	0.7104	0.6197
GPT-2_medium	0.5800	0.4800	0.5562	0.4652	0.4879	0.6491	0.7542	0.5825
GPT-2_large	0.7400	0.4400	0.5497	0.4507	0.4582	0.7291	0.7944	0.5964
GPT-2_xl	0.6000	0.4800	0.5549	0.4209	0.4501	0.7854	0.7842	0.5991
GPT-2_PyTorch	0.5000	0.5600	0.5679	0.5096	0.7183	0.9875	0.8444	0.7255
GPT-3	0.4400	0.5800	0.5746	0.5293	0.3476	0.7944	0.5209	<u>0.5534</u>
GROVER_base	0.3200	0.4200	0.5766	0.8400	0.3854	0.9831	0.9870	0.7544
GROVER_large	0.4800	0.5800	0.5442	0.5974	0.4090	0.9837	0.9875	0.7044
GROVER_mega	0.5400	0.4800	0.5138	0.4190	0.4203	0.9677	0.9416	0.6525
CTRL	0.5000	0.6900	0.4865	0.3830	0.8798	0.9960	0.9950	0.7481
XLM	0.6600	0.7000	0.5037	0.5100	0.8907	0.9997	0.5848	0.6978
XLNET_base	0.5200	0.5400	0.5813	0.7549	0.7541	0.9935	0.7941	0.7756
XLNET_large	0.5200	0.5200	0.5778	0.8952	0.8763	0.9997	0.9959	0.8690
FAIR_wmt19	0.5600	0.5600	0.5569	0.4616	0.5628	0.9329	0.8434	0.6715
FAIR_wmt20	0.5800	0.2800	0.5790	0.4775	0.4907	0.4701	0.4531	0.4941
TRANSFORMER_XL	0.5000	0.5000	0.5830	0.9234	0.3524	0.9721	0.9640	0.7590
PPLM_distil	0.5600	0.4400	0.5878	0.7178	0.6425	0.8828	0.8978	0.7457
PPLM_gpt2	0.5600	0.5000	0.5815	0.5602	0.6842	0.8890	0.9015	0.7233
AVG	0.5358	0.5132	0.5591	0.6032	0.5681	0.8799	<u>0.8280</u>	

Table 5: Compared Human Test vs. Test F1 scores of Turing Test models (**bold** and underlined are #1 and #2 performance, respectively). Human Test (machine) asked humans to decide if a given article is machine-generated or not, while Human Test (human vs. machine) asked humans which of the two given texts is machine-generated.

AA Model	P	R	F1	Accuracy
Random Forest	0.5893	0.6053	0.5847	0.6147
SVM (3-grams)	0.7124	0.7223	0.7149	0.7299
WriteprintsRFC	0.4578	0.4851	0.4651	0.4943
OpenAI detector	0.7810	0.7812	0.7741	0.7873
Syntax-CNN	0.6520	0.6544	0.6480	0.6613
N-gram CNN	0.6909	0.6832	0.6665	0.6914
N-gram LSTM-LSTM	0.6694	0.6824	0.6646	0.6898
BertAA	0.7796	0.7750	0.7758	0.7812
BERT-Multinomial	<u>0.8031</u>	<u>0.8021</u>	0.7996	0.8078
RoBERTa-Multinomial	0.8214	0.8126	0.8107	0.8173

Table 6: Performance of Authorship Attribution models (**bold** and underlined are #1 and #2 performance, respectively).

4.2 Results from Authorship Attribution

Since there are 20 labels in AA, the chance performance is at 0.05 (i.e., 5% in accuracy). Due to this difficulty, we use the full article contents in the test set. We compare different SOTA and popular techniques for automatic authorship attribution for our AA task including Random Forest, SVM (3-grams) (Sapkota et al., 2015), WriteprintsRFC (Mahmood et al., 2019), OpenAI detector², Syntax-CNN (Zhang et al., 2018), N-gram CNN (Shrestha et al., 2017), N-gram LSTM-LSTM (Jafariakinabad et al., 2019), BertAA (Fabien et al., 2020), BERT-Multinomial (Devlin et al., 2019), RoBERTa-Multinomial (Liu et al., 2019b). We find that BERT and RoBERTa outperform all the AA models, sometimes significantly, achieving the F1 scores of 0.80 and 0.81, respectively.

Interestingly, we observe that OpenAI detector, a RoBERTa-base model fine-tuned on GPT-2 XL outputs, does not outperform BERT-Multinomial and RoBERTa-Multinomial for this AA task although it performs comparatively, achieving a 0.77 as F1 score. BertAA achieves a slightly better F1 score (0.78).

5 Discussion

We present several observations from our experimental results.

1. Both TT and AA tasks are non-trivial: The average F1 score for each human vs. machine subtask and TT model is below 0.87, with FAIR_wmt20 achieving the lowest (0.49). FAIR_wmt20 is the newest text-generator in our list and before that we have GPT-3 which achieves the second lowest average F1 score (0.55). This suggests a trend that as newer text-generators get built, generated texts will become even more human-like, making the TT and AA tasks more difficult.

Additionally, the difficulty of the AA task is further demonstrated by the PCA plot of linguistic features LIWC of the TURINGBENCH dataset in Figure 8. Using LIWC to capture stylistic signatures of authors has been studied (Goldstein et al., 2009; Uchendu et al., 2020). However, we observe that there are quite a few overlaps in linguistic features across different authors (i.e.,

²<https://huggingface.co/roberta-base-openai-detector>

language models). This makes these authors' writing styles linearly inseparable.

2. No one size fits all: We observe in Table 5 that there is no one detection model that performs well across all 20 TT tasks. For instance, while BERT achieved the highest average F1 score, it still underperformed in detecting FAIR_wmt20. However, GROVER detector achieved a highest F1 score in detecting FAIR_wmt20.

3. Humans detect machine-generated texts at chance level: First two columns of Table 5 show the results of human detection test. In the first AMT-based tests, we randomly sampled 50 machine-generated texts and asked humans to decide if the given text is human-written or machine-generated (i.e., humans do not know whether they are shown only machine-generated texts in the test). In the second test, we showed two texts at random, one written by humans and the other generated by machines, and asked humans to decide which of the two are machine-generated (i.e., humans know that at least one of two is machine-generated).

Based on the average accuracies of two human tests, by and large, we observe that humans currently differentiate machine-generated texts from human-written ones, not much better (i.e., 0.535 and 0.513) than the level of random guessing (i.e., 0.5).

4. Not all text-generators are created equal: As shown in Table 5, the average F1 score for each human vs. machine subtask and TT model is below 0.87, with FAIR_wmt20 achieving the lowest (0.49). Consequently, this suggests that FAIR_wmt20 is the most sophisticated text-generator and thus, the hardest to detect. Other generators that are also hard to detect based on their < 0.62 F1 scores are: GPT-3, GPT-2_small, GPT-2_medium, GPT-2_large, and GPT-2_XL.

5. Sophisticated machine-generated texts often get detected as human-written: We observe an interesting phenomenon with these SOTA TT models. For instance, even though the labels in the binary classification task are approximately evenly split, GPT-2 detector and GLTR achieve below F1 score of 0.4 in some sub-tasks. This happens because TT models do not generalize well to those specific text-generators

(i.e., GROVER_base, CTRL, GPT-3, TRANSFORMER_XL) and mistakenly predicts the majority of the texts as *human-written*.

6. TT models do not always perform as expected: While both GROVER and GPT-2 detectors are trained to detect GROVER-generated and GPT-2-generated texts, respectively, they underperform in detecting those texts. For instance, GROVER detector performs the best in detecting PPLM_distil and PPLM_gpt2 texts, while GPT-2 detector performs significantly better at detecting GPT-1, TRANSFORMER_XL and XLNET_large texts.

7. Length of texts does not affect model performance: Due to the varying length of texts (i.e. 100-400) in Table 2, we plot the length of generated texts vs. the F1 scores of TT models in Figure 9. However, the figure suggests that there is no clear correlation between model performance and length of texts for all models except RoBERTa. This suggests that RoBERTa performance is text length-dependent.

8. Traditional AA models cannot fully capture an author's style “yet”: SOTA AA models cannot capture all of the stylistic features of human and machine text-generators. From Figure 8 we observe that the psycho-linguistic features of the 20 authors in the TURINGBENCH dataset are too similar, causing them to overlap in the plot. This suggests that machine-generated texts are becoming more similar to human-written texts in styles.

Therefore, traditional ways to capture an author's writing style will no longer be sufficient to achieve accurate automatic authorship attribution. This is further confirmed in the performance of classical AA models such as SVM and Random Forest. Similarly, we find that even deep learning based AA models are still unable to fully capture the distinct writing styles of all 20 authors. These results suggest that one needs to develop a model that can unearth more subtle yet distinct patterns that exist across 20 models.

9. Humans have a wide writing style range: In Figure 8, we observe that human-written features spread out all over the plot, while all machine-generated texts stay in little pockets of the plots. This suggests that humans may have a wider range of writing levels/styles, while

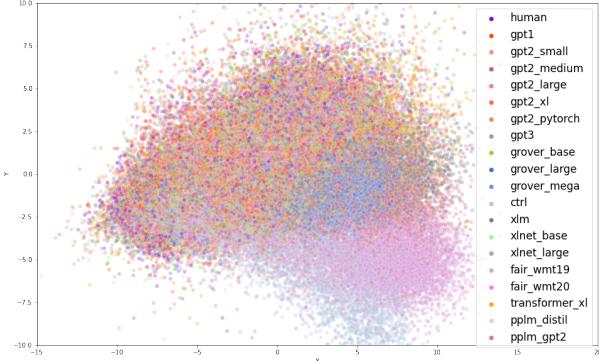


Figure 8: PCA plot of psycho-linguistics features of the TURINGBENCH dataset, using LIWC to attempt to capture the stylistic signatures of the dataset

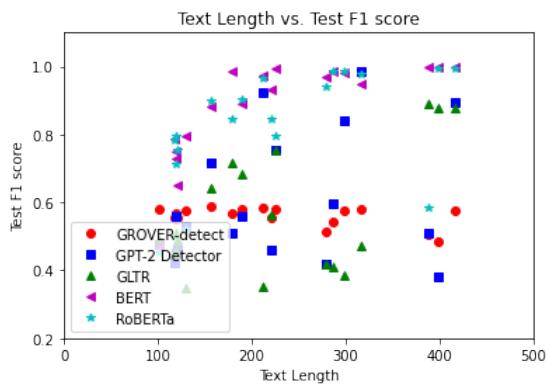


Figure 9: Despite the varying lengths of the generated texts (100–400) in Table 2, no correlation between text length and F1 score was found.

machines have a more limited range of writing levels/styles (e.g., high school to college).

6 Future Work

The experimental results suggest that we need better models to solve both TT and AA tasks as traditional AA models/features alone cannot solve the AA or TT problem. In addition, black-box detectors may be no longer sufficient for detection, as it cannot explain “why” a text is machine-generated or human-written yet. A research direction that GLTR-like framework points at may be able to help capture the nuanced nature of neural text-generators better. In addition, in future, a more complicated situation may emerge where a user may generate different parts of an article using different neural text-generators to intentionally mask the writing style of the generated text, thus confusing detectors—i.e., *Authorship Obfuscation*.

7 Conclusion

In this paper, we have introduced the TURINGBENCH environment and its preliminary results for both Turing Test (TT) and Authorship Attribution (AA) tasks. While varied, overall, (1) many contemporary language models can generate texts whose qualities are, to human eyes, indistinguishable from human-written texts, and (2) while some computational solutions for both TT and AA tasks can differentiate human-written texts from machine-generated ones much better than random guessing, overall, the community needs to research and develop better solutions for mission-critical applications. We hope that the TURINGBENCH environment will provide a platform on which insights into ways to tackle this urgent issue can be developed and shared.

8 Ethics Statement

We build TURINGBENCH by collecting public human-written news articles (mostly politics), and use the *Titles* of these articles as *Prompts* to generate similar news articles with neural text-generators. Some of these human-written articles were scraped from CNN and Washington Post news websites, and others from Kaggle. See Appendix for links to Kaggle datasets. However, while the purpose of the TURINGBENCH environment is to call attention to the urgent need for detectors of machine-generated texts, the potential negative uses of this research are not lost on us.

We understand that the insights we provide in this work can be used maliciously to thwart the performance of these detectors. Also, since we have released our dataset publicly, we understand that malicious users can copy the political articles generated by neural text-generators such as GPT-3, make minor changes, and post them online under the guise of real news. However, we believe that this work will lead to the creation of strong detectors of machine-generated texts, so that even human-edited machine-generated texts will still be detected in future.

Acknowledgments

This work was in part supported by NSF awards #1742702, #1820609, #1909702, #1915801, and #2114824.

References

- David Ifeoluwa Adelani, Haotian Mai, Fuming Fang, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. 2020. Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection. In *International Conference on Advanced Information Networking and Applications*, pages 1341–1354. Springer.
- Uri Alon, Shaked Brody, Omer Levy, and Eran Yahav. 2018. code2seq: Generating sequences from structured representations of code. In *International Conference on Learning Representations*.
- Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc’Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? learning to discriminate machine from human generated text. *arXiv preprint arXiv:1906.03351*.
- Meghana Moorthy Bhat and Srinivasan Parthasarathy. 2020. How effectively can machines defend against machine-generated fake news? an empirical study. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 48–53.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Erion Çano and Ondřej Bojar. 2020. Human or machine: Automating human likeliness evaluation of nlg texts. *arXiv preprint arXiv:2006.03189*.
- Peng-Jen Chen, Ann Lee, Changhan Wang, Naman Goyal, Angela Fan, Mary Williamson, and Jiatao Gu. 2020. Facebook ai’s wmt20 news translation task submission. In *Proceedings of the Fifth Conference on Machine Translation*, pages 113–125.
- Rosa María Coyotl-Morales, Luis Villaseñor-Pineda, Manuel Montes-y Gómez, and Paolo Rosso. 2006. Authorship attribution using word sequences. In *Iberoamerican Congress on Pattern Recognition*, pages 844–853. Springer.
- Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.
- Avisha Das and Rakesh Verma. 2018. Automated email generation for targeted attacks using natural language. In *TA-COS 2018: 2nd Workshop on Text Analytics for Cybersecurity and Online Safety*, page 23.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.
- Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc’Aurelio Ranzato. 2019. Residual energy-based models for text generation. In *International Conference on Learning Representations*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- H. Jair Escalante, Thamar Solorio, and Manuel Montes y Gómez. 2011. Local histograms of character n-grams for authorship attribution. In *Proceedings of the 49th Annual Meeting for the Association for Computational Linguistics: Human Language Technologies*, pages 288–298, Portland, Oregon. ACL.
- Maël Fabien, Esa ú Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. Bertaa: Bert fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing*, CONF. ACL.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.
- William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*.
- Elisa Ferracane, Su Wang, and Raymond Mooney. 2017. Leveraging discourse information effectively for authorship attribution. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 584–593.
- Tommaso Fornaciari and Massimo Poesio. 2014. Identifying fake amazon reviews as learning from crowds. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 279–287.
- Cristina Garbacea and Qiaozhu Mei. 2020. Neural language generation: Formulation, methods, and evaluation. *arXiv preprint arXiv:2007.15780*.

- Zhenhao Ge, Yufang Sun, and Mark Smith. 2016. Authorship attribution using a neural network language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116.
- Jade Goldstein, Ransom Winder, and Roberta Sabin. 2009. Person identification from text and speech genre samples. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 336–344.
- Oren Halvani, Lukas Graner, Roey Regev, and Philipp Marquardt. 2020. An improved topic masking technique for authorship analysis. *arXiv preprint arXiv:2005.06605*.
- Julian Hitschler, Esther Van Den Berg, and Ines Rehbein. 2017. Authorship attribution with convolutional neural networks and pos-eliding. In *Proceedings of the Workshop on Stylistic Variation*, pages 53–58.
- John Houvardas and Efstatios Stamatatos. 2006. N-gram feature selection for authorship identification. In *International conference on artificial intelligence: Methodology, systems, and applications*, pages 77–86. Springer.
- Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Fereshteh Jafariakinabad and Kien A Hua. 2019. Style-aware neural model with application in authorship attribution. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 325–328. IEEE.
- Fereshteh Jafariakinabad and Kien A Hua. 2020. A self-supervised representation learning of sentence structure for authorship attribution. *arXiv preprint arXiv:2010.06786*.
- Fereshteh Jafariakinabad, Sansiri Tarnpradab, and Kien A Hua. 2019. Syntactic recurrent neural network for authorship attribution. *arXiv preprint arXiv:1902.09723*.
- Ganesh Jawahar, Muhammad Abdul-Mageed, and VS Laks Lakshmanan. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309.
- Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.
- Maarten Lambers and Cor J Veenman. 2009. Forensic authorship attribution using compression distances to prototypes. In *International Workshop on Computational Forensics*, pages 13–24. Springer.
- Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.
- Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213.
- Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. 2014. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1566–1576.
- Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019a. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- A Pastor López-Monroy, Fabio A González, and Thamar Solorio. 2020. Early author profiling on twitter using profile features with multi-resolution. *Expert Systems with Applications*, 140:112909.
- Asad Mahmood, Faizan Ahmad, Zubair Shafiq, Padmini Srinivasan, and Fareed Zaffar. 2019. A girl has no name: Automated authorship obfuscation using mutant-x. *Proceedings on Privacy Enhancing Technologies*, 2019(4):54–71.
- Enrique Manjavacas, Jeroen De Gussem, Walter Daelemans, and Mike Kestemont. 2017. Assessing the stylistic properties of neurally generated text in authorship attribution. *arXiv preprint arXiv:1708.05536*.
- Ahmadreza Mosallanezhad, Kai Shu, and Huan Liu. 2020. Topic-preserving synthetic news generation: An adversarial deep reinforcement learning approach. *arXiv preprint arXiv:2010.16324*.
- Shaoor Munir, Brishna Batool, Zubair Shafiq, Padmini Srinivasan, and Fareed Zaffar. 2021. Through the looking glass: Learning to attribute synthetic text generated by language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1811–1822.

- Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair’s wmt19 news translation task submission. *arXiv preprint arXiv:1907.06616*.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/languageunderstandingpaper.pdf*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.
- Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.
- Upendra Sapkota, Steven Bethard, Manuel Montes y Gomez, and Thamar Solorio. 2015. Not all character n-grams are created equal: A study in authorship attribution. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–102, Denver, Colorado. ACL.
- Upendra Sapkota, Thamar Solorio, Manuel Montes y Gomez, and Steven Bethard. 2016. Domain adaptation for authorship attribution: Improved structural correspondence learning. In *Association for Computational Linguistics (ACL)*, Berlin, Germany. ACL.
- Tal Schuster, Roei Schuster, Darsh J Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2):499–510.
- Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. 2019. Do massively pretrained language models make better storytellers? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 843–861.
- Prasha Shrestha, Sebastian Sierra, Fabio Gonzalez, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*.
- Kai Shu, Yichuan Li, Kaize Ding, and Huan Liu. 2021. Fact-enhanced synthetic news generation. In *AAAI*.
- Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.
- Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Krepis, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.
- Efstathios Stamatatos. 2009a. Intrinsic plagiarism detection using character n-gram profiles. In *3rd PAN WORKSHOP. UNCOVERING PLAGIARISM, AUTHORSHIP AND SOCIAL SOFTWARE MISUSE*.
- Efstathios Stamatatos. 2009b. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.
- Alan M Turing. 2009. Computing machinery and intelligence. In *Parsing the turing test*, pages 23–65. Springer.
- Adaku Uchendu, Jeffery Cao, Qiaozhi Wang, Bo Luo, and Dongwon Lee. 2019. Characterizing man-made vs. machine-made chatbot dialogs. In *Proceedings of the Int’l Conf. on Truth and Trust Online (TTO)*.
- Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the International AAAI Conference on Web and Social Media*.
- Lav R Varshney, Nitish Shirish Keskar, and Richard Socher. 2020. Limits of detecting text generated by large-scale language models. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–5. IEEE.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

William Yang Wang. 2017. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, pages 9051–9062.

Chunxia Zhang, Xindong Wu, Zhendong Niu, and Wei Ding. 2014. Authorship identification from unstructured texts. *Knowledge-Based Systems*, 66:99–111.

Richong Zhang, Zhiyuan Hu, Hongyu Guo, and Yongyi Mao. 2018. Syntax encoding with application in authorship attribution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2742–2753.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Wanjun Zhong, Duyu Tang, Zenan Xu, Ruize Wang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Neural deepfake detection with factual structure of text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zhixuan Zhou, Huankang Guan, Meghana Moorthy Bhat, and Justin Hsu. 2019. Fake news detection via nlp is vulnerable to adversarial attacks. *arXiv preprint arXiv:1901.09657*.

A Appendices

A.1 Data Generation Implementation

Generating texts with these Language models is very computationally expensive. Some of the python code used to generate the texts were not written for large scale generation, so we had to repurpose it for our task. We mostly used Google Colab pro’s GPU - 12GB NVIDIA Tesla K80 to generate our texts. However, since *PPLM* was the heaviest language model computationally, we used a machine with more GPUs - NVIDIA Tesla K80s and P100s.

Most generators took 24 – 72 hours to generate 10K articles. However, *PPLM* took about 430 hours for *PPLM_distil* and about 600 hours for *PPLM_gpt2*. It is important to note that probably a few coding choices could reduce the computational cost of running *PPLM*, we just did not get to it. See the description of building the human dataset and 10 language model architectures used to generate the rest of the dataset. The table also contains the links to the dataset and github repo of some of the models.

A.2 Data Pre-processing

Some of the generated texts contain non-English tokens such as $\langle UNK \rangle$, $\langle eos \rangle$, $\langle eod \rangle$, $\langle eop \rangle$, $\langle |endoftext| \rangle$, etc. which we removed. Also, in an attempt to generate texts with the specified word count (i.e., 400), some of the generators had a tendency to repeat a particular word multiple times consecutively. This introduced bias into our Machine Learning models, making it easier to detect such generated texts. Therefore, we removed words that were repeated consecutively, leaving only one. Next, those same text-generators also had a tendency to generate texts where a random word would have the last character repeated multiple times. For instance, a word like “expressed”, could be spelt like “expressedddddddddd……ddddd”. This also made such generators easy to detect, so we removed words more than 20 characters to get rid of such words. Lastly, the word “CNN” was used heavily by a few generators, making it easier to detect such generators. Therefore, we removed the word, “CNN” from all the articles.

Before pre-processing of the data, we had 200K, and after the process, we have 168,612. See data distribution in Table 7 of the cleaned dataset. We can observe that the distribution of the dataset is

Text Generator	# of Data samples
Human	8,854
GPT-1	8,309
GPT-2_small	8,164
GPT-2_medium	8,164
GPT-2_large	8,164
GPT-2_xl	8,309
GPT-2_PyTorch	8,854
GPT-3	8,164
GROVER_base	8,854
GROVER_large	8,164
GROVER_mega	8,164
CTRL	8,121
XLM	8,852
XLNET_base	8,854
XLNET_large	8,134
FAIR_wmt19	8,164
FAIR_wmt20	8,309
TRANSFORMER_XL	8,306
PPLM_distil	8,854
PPLM_gpt2	8,854

Table 7: # of data samples in the TURINGBENCH dataset

still approximately the same.

A.3 TURINGBENCH Website

We create the TURINGBENCH website using the SQuAD website framework. The website contains a description of the benchmark datasets and benchmark tasks. Each benchmark task has a leaderboard that shows the models used to solve the tasks. These models are rated from best to worst. For the AA tasks, we use the standard Machine learning evaluation metrics such as: *Precision*, *Recall*, *F1 score*, and *Accuracy*. And we use only *F1 score* for the TT task because it is a binary classification problem and *F1 score* is sufficient for the problem. See website interface in Figure 10.

³<https://www.kaggle.com/snapcrack/all-the-news>,

⁴<https://www.kaggle.com/sunnysai12345/news-summary>

⁵<https://www.kaggle.com/ryanxjhan/cbc-news-coronavirus-articles-march-26>

⁶<https://www.kaggle.com/patjob/articlescrape>

⁷<https://github.com/huggingface/transformers>

⁸<https://github.com/graykode/gpt-2-Pytorch>

⁹<https://github.com/minimaxir/aitextgen>

¹⁰<https://github.com/rowanz/grover>

¹¹<https://github.com/salesforce/ctrl>

¹²<https://github.com/pytorch/fairseq/tree/master/examples/wmt19>

¹³<https://github.com/pytorch/fairseq/tree/master/examples/wmt20>

¹⁴<https://github.com/uber-research/PPLM>

¹⁵<https://huggingface.co/models>

TEXT-GENERATORS	DESCRIPTION
Human	We collected news titles (mostly Politics) and contents from CNN, Washington Post, and Kaggle ^{3 4 5 6} . Next, we removed articles that did not have the desired word length (i.e., 200–500). This resulted in 130K articles, but only 10K was used for the article generations.
GPT-1	Texts are generated with the huggingface github repo ⁷ .
GPT-2	We use 4 GPT-2 pre-trained models - PyTorch model ⁸ , small (124 million parameters), medium (355 million parameters), large (774 million parameters), and extra-large (1558 million parameters) ⁹ to generate texts.
GPT-3	Texts are generated with the OpenAI GPT-3 API using the <i>davinci</i> engine.
GROVER	We use code from repo ¹⁰ to generate from Grover’s 3 pre-trained models: GROVER-base , GROVER-large , GROVER-mega .
CTRL	Conditional Transformer Language Model For Controllable Generation ¹¹ uses control codes to guide generation. We use <i>News</i> control code to generate long articles.
XLM	We generated texts using huggingface repo.
XLNET	We generated texts with: 2 XLNET pre-trained models: XLNET-base , and XLNET-large using huggingface.
FAIR_wmt	We use two Facebook’s FAIR English models - wmt19 ¹² and wmt20 ¹³ to generate texts with FAIRSEQ sequence modeling toolkit.
TRANSFORMER_XL	We generated texts with this language model’s setup on huggingface.
PPLM	PPLM fuses GPT-2’s pre-trained model with bag of words to generate more specific texts. We used the <i>Politics</i> bag of words model to generate texts’, using the code ¹⁴ , and used the perturbed version. Next, we fused PPLM with two pre-trained models (i.e., distilGPT-2, and GPT-2) and generated texts with them, forming: PPLM_distil , PPLM_gpt2 . These models are gotten from the huggingface model repository ¹⁵ .

Table 8: Description of the Text-generators in the TURINGBENCH dataset.

Model	Run-time
GROVER detector	25 – 30 minutes
GPT-2 detector	5 – 10 minutes
GLTR	4 – 5 hours
BERT	25 – 40 minutes
RoBERTa	45 – 1 hour

Table 9: TT model Run-time per task

A.4 Experiments

All experiments, except *GLTR* and *GPT-2 detector* were done using the Google colab pro’s GPU stated above. Experiments with *GLTR* and *GPT-2 detector* were done in a machine with 4 GPUs - NVIDIA Quadro RTX 8000.

A.4.1 TT models

Each of the models used their default hyperparameters. There was no hyperparameter tuning performed. We used GROVER-Large discriminator for *GROVER detector*, the weights of roberta-large fine-tuned on GPT-2 XL outputs for *GPT-2 detector*, and GPT-2 117M model for *GLTR*. None of these models were trained on our dataset. We tested their performance on predicting on our test set. Next, we fine-tuned BERT and RoBERTa on our train set and validate these models on our validation set for each TT task. BERT was fine-tuned for 3 epochs and RoBERTa, 3–5 epochs with $2e^{-5}$ learning rate. See Table 9 for run-time of the mod-

els.

A.4.2 AA models

We used the default hyperparamters of the AA models for the AA task. Also, we did not perform any hyperparameter tuning on these models. *Random Forest* and *SVM* take about 30 minutes – 1 hour to converge. *WriteprintsRFC* took about 15 minutes to converge. *Syntax-CNN*, *N-gram CNN*, and *N-gram LSTM-LSTM* took about 30 minutes to converge. *OpenAI detector* took about an hour to converge. *BERT-Multinomial* and *RoBERTa-Multinomial* took about 1 – 2 hours to converge. *BertAA* took about 5 hours to converge.

TuringBench

Home

TuringBench

The Turing Test Benchmark Environment

What is TuringBench?

TuringBench is a benchmark environment that contains :

1. Benchmark tasks- Turing Test (i.e., human vs. machine) and Authorship Attribution: (i.e., who is the author of this texts?)
2. Datasets (Binary and Multi-class settings)
3. Website with leaderboard

The dataset has 20 labels (19 AI text-generators and human). We built this dataset by collecting 10K news articles (mostly Politics) from sources like CNN and only keeping articles with 200-400 words. Next, we used the Titles of these human-written articles to prompt the AI text-generators (ex: GPT-2, GROVER, etc.) to generate 10K articles each. This gives us a sum total of 200K articles and 20 labels. However, since there are two benchmark tasks - Turing Test and Authorship Attribution settings, we have all 20 labels in one dataset for the multi-class setting and only human vs. one AI text-generator, making 19 binary-class datasets.

How to get the Dataset

```
from datasets import load_dataset
import pandas as pd
...

```

Leaderboard: Authorship Attribution

The TuringBench Datasets will assist researchers in building robust Machine learning and Deep learning models that can effectively distinguish machine-generated texts from human-written texts. This Leaderboard is for the Authorship Attribution scenario.

Rank	Model	Precision	Recall	F1	Accuracy
1 May 5, 2021	RoBERTa (Liu et al., '19)	0.8214	0.8126	0.8107	0.8173
2 May 5, 2021	BERT (Devlin et al., '18)	0.8031	0.8021	0.7996	0.8078
3 May 5, 2021	BertAA (Fabien et al., '20)	0.7796	0.7750	0.7758	0.7812
4 May 5, 2021	OpenAI detector	0.7810	0.7812	0.7741	0.7873
5 May 5, 2021	SVM (3-grams) (Sapkota et al. '15)	0.7124	0.7223	0.7149	0.7299
6 May 5, 2021	N-gram CNN (Shreshta et al., '17)	0.6909	0.6832	0.6665	0.6914
7 May 5, 2021	N-gram LSTM-LSTM (Jafariakinabad, '19)	0.6694	0.6824	0.6646	0.6898

Figure 10: TURINGBENCH website interface