**Abstract**

**Introduction**

The increasingly articulate and realistic nature of texts generated by large-language models (LLM) have enabled their use in applications such as drafting essays, summarising information, as well as writing code proof-of-concepts. This also raises concerns about factual inaccuracies in journalism[[1]](#endnote-2), the potential corruption of technical knowledge[[2]](#endnote-3), misattribution of work extending to academic plagiarism, and the facilitating of phishing techniques and other scams. Researchers have hence recently been tackling the detection problem of identifying whether a candidate piece of text is machine or human-generated. However, this is complicated by the rapidly-evolving LLM landscape, with new models being frequently rolled out and the development of adversarial methods to circumvent detection methods.

**Related Work**

Current approaches to the detection problem make assumptions that the text was generated by a particular source model, to which the defender has different levels of access depending on the problem setting. We follow a similar classification approach to OpenAI[[3]](#endnote-4) when describing these approaches. The *black-box setting* presumes only access to the model outputs. Black-box approaches mainly consist of using sample text features such as TF-IDF (term frequency-inverse document frequency), skip-gram, continuous-bag-of-words (CBOW)[[4]](#endnote-5) and sometimes auxiliary features (such as account data in the setting of social media bot detection), and fitting a binary classifier to them, which can range from logistic regression models to convolutional neural networks[[5]](#endnote-6) or LSTM models[[6]](#endnote-7). However, high classification accuracy for these methods are reliant on sufficiently-long text length and a sufficiently-diverse corpus of training machine-generated samples in terms of stylometric and linguistic characteristics in order to prevent overfitting. As such, these classifiers need to be continually trained and updated, limiting their usefulness. In contrast, the *white-box* setting repurposes the generative model as a classifier, and provides the defender access to some or all of the model’s inner attributes, such as log-probabilities and scores. White-box approaches can be further classified into two subcategories: I) zero-shot models which aim to classify sample texts without any further inputs, and ii) fine-tuning based models, which further provide the model with domain-specific texts. Under the assumption that the discriminative model is also the generative model, white-box approaches tend to be highly effective even using simple metrics such as the average log probability of all tokens within the candidate text Solaiman et al. (2019).

In this paper, we modify the DetectGPT model developed in Mitchell et al. (2023), which uses perturbations of the existing text to estimate the local curvature of the sample’s log-probabilities rather than merely relying on the raw values. We extend DetectGPT by relaxing the constraint that the generative (or source) model and the discriminative (or scoring) model are the same, mirroring real-life scenarios where the model that generates the text (were it to be even machine-generated) is unknown. Mitchell et al show that detection performance is expectedly lower when this assumption is lifted, with significant variations in accuracy (as measured by area under the Receiver-Operating Curve, or AUROC). To tackle this question, we leverage on the ensembling of the log-probability outputs of multiple DetectGPT models (each assuming a separate source model) using classifiers in the machine-learning literature such as logistic regression and random forests. Our paper hence combines the black-box approach with the white-box approach, to attempt to achieve an improvement in detection accuracy without any source model assumptions.

**Methodology**

The key observation underpinning DetectGPT is the hypothesis that LLMs implicitly choose samples to maximise the log-probability of successive tokens when generating text, and hence tend to sample from regions of negative curvature of the log-probability function of p\_theta. Thus, by perturbing an original text x into a rewritten text x\_tilde which is semantically similar to x, we can compute the perturbation discrepancy d, defined to be log(p\_theta(x)/p\_theta(x\_tilde)). The perturbation discrepancy is then expected to be positive and larger for text if the text were indeed generated by the source model. In practice, normalising the perturbation discrepancy is more effective. The text perturbation can be performed consistently and at scale using an off-the-shelf mask-filling model such as T5.

We adopt a similar training pipeline by adapting the code developed by Mitchell et al (2023). e use news articles from the XSum dataset and generate samples from the raw conditional distribution of the corresponding source model with a temperature of 1, following Mitchell et al. We use smaller models than current state-of-the-art LLMs for the mask-filling and scoring models to evaluate the accuracy of our approach on less-complex models, to decrease the relatively long compute time already required by the vanilla DetectGPT method. We use the following source, mask-filling, scoring models and datasets during our experiments; however, our methodology can be easily extended to encompass additional models. We use the same set of hyperparameters for all experiments, such as masked span length, fraction of words masked for rewriting and number of perturbations.

We attempted various methods to aggregate the outputs of the individual classifiers. The first category consists of simple summary statistics which serve as our baseline methods, in particular the mean, median and maximum of the scores of the three classifiers, which mimic a soft-voting approach. However, we note that each individual classifier is evaluating a slightly different question, as each one is assessing the likelihood that the sample text is generated by their model. The second category consists of more sophisticated approaches such as logistic regression, random forest and Naive-Bayes methods. This approach requires prior training, and thus work best with a fine-tuning dataset similar to the fine-tuning approaches mentioned earlier, although it is used to train our additional classifier and not the LLMs themselves. We also attempt a “multi-stage” classifier that orders the DetectGPT models by decreasing complexity and successively considers each model. If the perturbation discrepancy d is below a threshold, which we take to be mu – 2\*SD where mu and SD are taken over the training dataset, we halt and take the mean of all the models we have considered so far. This approach aims to reduce the likelihood of false positives by combining the idea of taking the maximum of all DetectGPT models with the hypothesis that more complex scoring models, by virtue of having more representational power to “encompass” sub-models, are more accurate at evaluating if a text is machine-generated.

**Results**

Our results are shown in Table 1. We observe that our first category of ensembling methods perform better(?) than the individual DetectGPT classifiers in isolation when the scoring model is not the source model. In addition, our second category of ensembling methods performs better than the first category, reaching an accuracy of about 0.99. We observe that even if the source model is used as one of the scoring models, combining the outputs of the additional classifiers still yields an overall improvement in accuracy.

**Discussion**

**Limitations**

In addition to the known limitations of DetectGPT, which are its relatively long compute time and requirement of access to the log-probabilities of the base and scoring models, our current ensembling methods are performed on a relatively small dataset and using a limited range of models. As such, more work can be done to validate the approach on other datasets as well as to create a ‘minimal set’ of LLMs that when combined with our method can reliably and accurately classify any sample text. Furthermore, while simply taking the mean and maximum of individual DetectGPT classifiers to observe marginal increases in accuracy, a fine-tuning dataset as training data for the random forest classifier is required for more significant accuracy improvements.

**Conclusion**

We apply ensembling methods to the zero-shot, white-box DetectGPT approach developed by Mitchell et al., which estimates local probability curvature of the sampled text, by combining the outputs from individual DetectGPT classifiers that are each trained on a different source model. We find that a random forest classifier is highly accurate at detection as compared to the individual classifiers.

1. CNET secretly used AI on articles that didn’t disclose that fact, staff say [↑](#endnote-ref-2)
2. Cross-Domain Detection of GPT2-Generated Technical Text [↑](#endnote-ref-3)
3. Release strategies and the social impacts of language model [↑](#endnote-ref-4)
4. Text-mining based fake news detection using ensemble methods [↑](#endnote-ref-5)
5. Identifying Russian Trolls on Reddit with Deep LLearning and BERT Word Embeddings [↑](#endnote-ref-6)
6. Deep Neural Networks for Bot Detection [↑](#endnote-ref-7)