From Text to Source: Results in Detecting Large Language Model-Generated Content

Wissam Antoun **Benoît Sagot** **Djamé Seddah**
Inria, Paris
{firstname,lastname}@inria.fr

Abstract

The widespread use of Large Language Models (LLMs), celebrated for their ability to generate human-like text, has raised concerns about misinformation and ethical implications. Addressing these concerns necessitates the development of robust methods to detect and attribute text generated by LLMs. This paper investigates "Cross-Model Detection," evaluating whether a classifier trained to distinguish between source LLM-generated and human-written text can also detect text from a target LLM without further training. The study comprehensively explores various LLM sizes and families, and assesses the impact of conversational fine-tuning techniques on classifier generalization. The research also delves into Model Attribution, encompassing source model identification, model family classification, and model size classification. Our results reveal several key findings: a clear inverse relationship between classifier effectiveness and model size, with larger LLMs being more challenging to detect, especially when the classifier is trained on data from smaller models. Training on data from similarly sized LLMs can improve detection performance from larger models but may lead to decreased performance when dealing with smaller models. Additionally, model attribution experiments show promising results in identifying source models and model families, highlighting detectable signatures in LLM-generated text. Overall, our study contributes valuable insights into the interplay of model size, family, and training data in LLM detection and attribution.

1 Introduction

Large Language Models (LLMs), characterized by their ability to generate human-like text (Dou et al., 2022), have found applications in various domains, including content generation, chatbots, and language translation. However, as the use of LLMs becomes more widespread, concerns about their misuse, misinformation, and ethical implications have surfaced (McGuffie and Newhouse, 2020; Bender et al., 2021; Chiesurin et al., 2023). One of the ways to address these concerns is with robust methods that are able to detect and attribute text generated by LLMs (Jawahar et al., 2020), allowing us to differentiate between human-authored and machine-generated content, identify the source model, or even the model creator. Such capabilities are crucial for maintaining trust in online communication platforms, content moderation, and ensuring responsible AI deployment.

Our motivation for this research stems from real-life scenarios where we often lack knowledge of the specific model used to generate a piece of text. These scenarios can be formulated as a "Cross-Model Detection", where we investigate whether a classifier originally trained to distinguish between text generated by one LM and human-written text, can also identify text generated by a different LM without requiring fine-tuning or training on the text it produces.

Our contribution to this area is characterized by the comprehensiveness of our study. While previous works in the literature have been limited in their exploration of a few model sizes and families, we take a more expansive approach. We systematically examine a wide range of LLM sizes, spanning from base models to exceptionally large ones, and encompassing diverse model families such as GPT-2, LLaMA, Pythia, OPT and others (Zhao et al., 2023). Additionally, we explore the impact of conversational fine-tuning techniques, including Chat, Instruct (Mishra et al., 2022; Wei et al., 2022), and Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ziegler et al., 2020), on the generalization and transferability of the classifier across this wide array of models. This comprehensive investigation enables us to gain a deeper understanding of the generalization and transferability of the classifier across a

diverse array of models, thus eliminating a potential source of bias in our results. It also allows us to identify how factors like model size and family impact the detection and attribution of generated text.

Our contributions in this study can be summarized as follows:

- A comprehensive investigation into cross-model detection, evaluating the classifier's ability to detect text generated by different LLMs, and in model attribution, encompassing a broad range of sizes and model families.
- We highlight the role of both model size and family in the detection of text generated by Language Model Models (LLMs). We observed an inverse relationship between classifier effectiveness and LLM size. Detecting larger models can be challenging, but training on similarly sized LLMs can improve performance.
- Our experiments in model attribution reveal the potential for identifying the source model of generated text. While human-generated text is distinguishable, confusion primarily occurs between models from the same family or with adjacent sizes. This suggests that LLMs leave distinct signatures, enabling source model identification and model family classification, further enhancing our understanding of how different LLMs generate text.

In the subsequent sections, we present a summary of relevant related works followed by the details of our methodology, experiments, and results, shedding light on the interplay between model size, family, and training data in the context of LLM detection and attribution in the ever-evolving landscape of Large Language Models.

2 Related Works

Detecting AI-generated text is a recent and rapidly growing area of research (Jawahar et al., 2020). Although Sadasivan et al. (2023) demonstrated a theoretical impossibility of distinguishing between human-written and machine-generated when the total variation (TV) norm between the two is low, a more recent study by Chakraborty et al. (2023) showed that detection is still possible given enough samples.

Popular methods to detect AI-generated text can be grouped into three categories: 1) Using statistical features of text such as perplexity, n-grams, entropy, etc. (Gehrmann et al., 2019; Mitchell et al., 2023). 2) Watermarking generated text which was first demonstrated by Atallah et al. (2001) who embedded a watermark bit in the syntactic structure of the text. More recently, Kirchenbauer et al. (2023) used the LLM's output log probability at each generation step to embed a watermark based on "green/red" toke list where an LLM will have an artificially increased likelihood of selecting tokens from the "green" list. Other work on watermarking include (Fernandez et al., 2023; Christ et al., 2023). 3) Classifier-based approaches which use a classifier trained on a dataset containing both human-written and machine-generated text to detect LM-generated text (Zellers et al., 2019; Solaiman et al., 2019; Uchendu et al., 2020; Fagni et al., 2021; Antoun et al., 2021; Guo et al., 2023; Mitrović et al., 2023). This approach is vulnerable against adversarial text mimicking, among others, Wikipedia style and informative (Antoun et al., 2023).

We highlight recent work by Mireshghallah et al. (2023) that studies cross-model detection and detector transferability by examining the effect of using classifier models other than the generator itself to detect machine-generated text. The authors studied training LMs from 5 different model families with sizes ranging from 70M to 6.7B and trained the generator LMs to detect machine-generated text generated by other LMs. They demonstrated that using smaller language models for detection can lead to a higher performance. Our work differs from Mireshghallah et al. (2023) in that we assume we don't have access to the underlying model but only to a set of text generated by the model. We hence use an separate encoder classifier to detect generated text instead of using the generator. We also extend the study to more model families, sizes and further finetunings, while also studying model attribution.

3 Methodology

Cross-Model Detection Our objective is to evaluate whether a classifier, initially trained to distinguish text produced by a source LLM from human-written text, can also detect text generated by a target LLM.

We conduct a comprehensive evaluation, by using LLMs with a range of sizes (base models to up to very large LLMs) from different families. We consider a model's family as a proxy for pretraining

dataset variation, since apart from slight changes in model architecture, namely positional embeddings type, layer-norm order, or activation type, the only difference between the models from different families is the dataset used for pretraining.

We also investigate the effect of Chat, Instruct and Reinforcement Learning from Human Feedback (RLHF) which we refer to as conversational finetuning. This enables us to measure the generalization and transferability of the classifier across a diverse array of models.

Model Attribution We divide this task into three subtasks:

- Source Model Identification: We first investigate the ability to identify the source for a given piece of text, the source being either a human-written text or a text generated by an LLM.
- Model Family Classification: In the second investigation, we classify the source model into its corresponding family. This classification helps us understand how good a text can be attributed to a specific model family, and identifies instances where confusion arises between different model families. This task is a higher-level generalization of the Source Model Identification task.
- Model Size Classification: Lastly, we examine the ability to determine the model size responsible for generating a given text. This experiment aims to determine whether it is feasible to discern whether the text was generated by a small or large LLM. This information is valuable for understanding how model size impacts the generated content.

These investigations collectively contribute to a comprehensive understanding of model attribution in the context of our study.

Our research methodology for investigating cross-model detection and model attribution involves synthetic text generated using Large Language Models (LLMs) selected from diverse families, sizes, and architectures.

4 Experimental Protocol

4.1 LLM Choice

We chose the following model families and sizes for our experiments for a total of 55 models:

- **BLOOM** (Scao et al., 2022): 560M, 1.1B, 1.7B, 3B, 7.1B.
- **Cereberas-GPT** (Dey et al., 2023): 111M, 256M, 1.3B, 2.7B, 6.7B, 13B.
- **Falcon** (Almazrouei et al., 2023; Penedo et al., 2023): 7B, 40B.
- **GPT-2** (Radford et al., 2019): 124M, 355M, 774M, 1.5B.
- **LLaMA** (Touvron et al., 2023a): 7B, 13B, 30B, 65B.
- **LLaMA-v2** (Touvron et al., 2023b): 7B, 13B, 70B.
- **MPT** (MosaicML, 2023): 7B, 30B.
- **OPT** (Zhang et al., 2022): 125m, 350m, 1.3B, 2.7B, 6.7B, 13B, 30B, 66B.
- **OpenLLaMA** (Geng and Liu, 2023): 3B, 7B, 13B.
- **OpenLLaMA-v2** (Geng and Liu, 2023): 3B, 7B.
- **Pythia** (Biderman et al., 2023): 70m, 160m, 410m, 1B, 1.4B, 2.8B, 6.9B, 12B.

We select the following conversationally finetuned models to compare with their corresponding foundation models:

- **Falcon-Instruct** (Almazrouei et al., 2023; Penedo et al., 2023): 7B and 40B.
- **Alfred-0723**: 40B, an RLHF finetuned version of Falcon-40B.
- **LLaMA-v2-Chat** (Touvron et al., 2023b): 7B, 13B and 70B, an RLHF finetuned version of LLaMA-v2
- **MPT-Chat** (MosaicML, 2023): 7B, 30B, based on MPT finetuned on a large selection of chat datasets.
- **Vicuna-v1.3** (Zheng et al., 2023): 7B, 13B, 33B, based on LLaMA fine-tuned on user-shared conversations.

4.2 Data-Generation

We generate our data by prompting the different LLMs with the first 10 words of documents sampled from the OpenWebText dataset (Gokaslan et al., 2019). For conversational models, in addition to each model's specific prompt, we explicitly instruct the model to continue generation with the following prompt: *"Give the best continuation of the following text:"*, followed by the first 10 words from the sampled document.

We use the HuggingFace Text Generation Inference server [1] to load all models using up to 4 48GB NVIDIA GPUs for the largest models with float16

[1] https://github.com/huggingface/text-generation-inference

Figure 1: 5-seed averaged AUC scores for a classifier trained on text from a source model (*Side axis*) and tested on text from a target model (*Top axis*).

precision. The same set of hyper-parameters is used for all models, a maximum 256 tokens per generation, with beam-search of size 5, repetition penalty of 1.0, temperature of 1.0, top-k of 10, top-p of 0.9 and typical sampling with 0.9.

4.3 Data Splitting and Filtering

We first split our data into 80% for training and 20% for validation. Then we filter each split to remove bad generations, by filtering (i) generations that are too short, (ii) generations that are repetitive and (iii) generations that contain apologies or sentences similar to "As an AI language model". To ensure a fair comparison between classifier trainings, we sample 800 samples for training and 200 samples for validation from all models, except both *pythia-70m* models, *Cereberas-GPT-110m & 256m* and *OPT-350m*.

For negative human-generated samples, we sample a new set of texts (800 samples for training and 200 for validation) from the same OpenWebText dataset.

4.3.1 Cross-Model Detection Training Data

For each LLM we merge its own train and test sets with the negative examples sets for a total of 1600 training samples and 400 validation samples. To quantify classifier performance, and following (Sadasivan et al., 2023; Chakraborty, 2023), we utilize the Area Under the Receiver Operating Characteristic Curve (AUC score). The AUC score provides a robust measure of the classifier's ability to distinguish between different models, taking into account both true positive and false positive rates.

4.3.2 Model Attribution Training Data

We conduct three distinct investigations as mentioned earlier. We use F1 score to evaluate classifier performance in all three tasks due to the imbalanced nature of our data.

Source Model Identification This task involves classifying the text into one of the 50 LLMs used in our experiments, spanning various families and sizes. We also include a class for human written text for a total of 51 classes.

Family Classification We group models into 12 classes including one for human written text, and then sub-sample the data to have a balanced 1600 training samples and 400 validation samples for each class.

Model Size Identification We bin the models into 6 bins: <1B, 1-5B, 5-10B, 10-20B, 20-50B, >50B. Refer to Appendix A for the class distribution.

4.4 Classifier

Our classifier of choice for all experiments uses the transformer encoder architecture namely

Figure 2: Average target AUC scores vs model size. OLS Trend lines are drawn for each set of model family

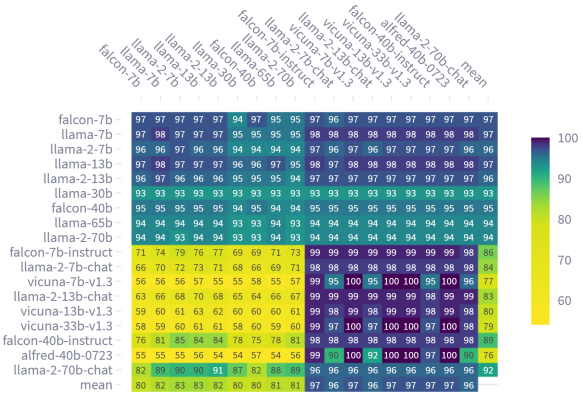

Figure 3: Conversational models cross-model detection with their foundation LLM.AUC scores (5-seed averaged) for a classifier trained on text from a source model (*Side axis*) and tested on text from a target model (*Top axis*).

DeBERTaV3-base (He et al., 2023, 2021). All models are trained with a batch size of 32, learning rate of 2e-5 for 5 epochs. The classification experiments were conducted using five different random seeds, and the resultant scores were averaged to enhance the robustness of our findings, as this approach helps mitigate the potential impact of seed-specific variations on the results.

5 Results

5.1 Cross-Model Detection Results

Figure 1 presents a heatmap of the AUC scores for the cross-model detection experiments. The side axis represents the classifier's source model, and the top axis represents the classifier's target model. We sort the models by their size (from left to right, top to bottom). From Figure 1, we observe several interesting patterns in the cross-model detection results:

Model Size Influence In general, our findings suggest a clear inverse relationship between the classifier's effectiveness and the size of the test models. The pattern is showcased better in Figure 2, which plots the self-detection and average AUC scores trend lines against the model size. This pattern indicates that larger LLMs tend to pose a greater challenge for the classifier, particularly when the classifier is trained on data from a smaller source model. Notably, the detection performance on very large Language Models (LMs) tends to improve when the model is trained on data sourced from similarly sized large LMs. However, it is essential to highlight the trade-off that training only on very large LMs leads to, results in decreased performance in detecting smaller-sized models.

Model Family Influence We observe that performance on detecting GPT2 and LLaMA generated text tends to be slightly lower than other model families *(Refer to the corresponding heatmap columns and their means in Figure 1 to observe the corresponding data patterns)*. This pattern suggests that the two model families are harder to detect relative to their similar-sized counterparts due to their superior language modeling capabilities and hence "closer" to human written text. We can also observe that the performance of a classifier trained on text sourced from *pythia-160m* and *opt-2.7b* tends worse overall, while a classifier trained on text sourced from *Cereberas-GPT-6.7B* is performing better than similarly sized models *(Refer to the corresponding heatmap rows in Figure 1)*. **The lack of a discernible pattern in the cross-model detection performance across different model**

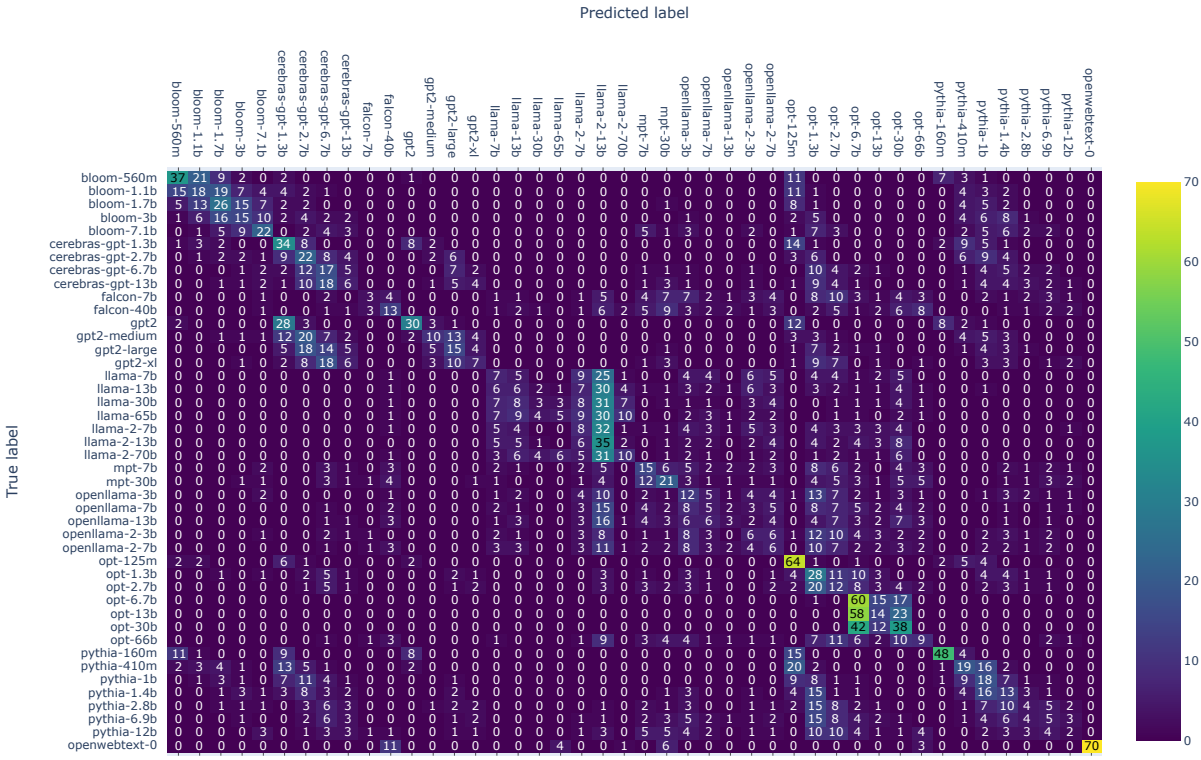

Figure 4: Normalized confusion matrix for Source Model Identification. 5-seed averaged and normalized by the predicted class support.

families may be attributed to the extensive overlap in their pretraining data, with a predominant reliance on ThePile (Gao et al., 2020) dataset or its subsets across most models, supplemented by Common Crawl as the primary data source. Consequently, the primary distinguishing factor among these models lies in their respective data cleaning pipelines.

Influence of Conversational Finetuning Our experiments reveal a clear pattern in the cross-model detection results, as shown in Figure 3. Specifically, a classifier trained on text generated by chat models exhibits limited capability in detecting normal language models (LMs). However, it demonstrates improved performance when tasked with detecting other chat models. Notably, when trained on LLaMA 2 70b chat data, the classifier achieves the highest scores, albeit with a slight decline in detection accuracy when tested on chat models. This observation suggests that the LLaMA 2 70b chat model effectively follows instructions to continue input text. Surprisingly, training the classifier on vanilla LM output also yields commendable results in detecting these distinct model categories. These findings underscore the nuanced relationship between chat models and traditional language models in the context of detection.

5.2 Model Attribution Results

Source Model Identification In the Model Attribution experiments, our objective was to investigate the ability of our classifier to identify the source model of generated text accurately. Figure 4 displays the confusion matrix for the Model Attribution experiments, where rows represent the true source models, and columns represent the predicted source models. We can draw the following conclusions from our results:

- Human-generated text proved to be the most easily distinguishable source, as it exhibited minimal confusion, primarily with a few 30B+ Large Language Models (LLMs).
- The majority of confusions occurred between models from the same family. We also notice that within a model family, the confusions tend to happen between models with adjacent sizes.
- An interesting case was the confusion between GPT-2 models and Cereberas-GPT models. It's worth noting that both models share the same GPT-2 architecture but differ in their pretraining data, with Cereberas-GPT being

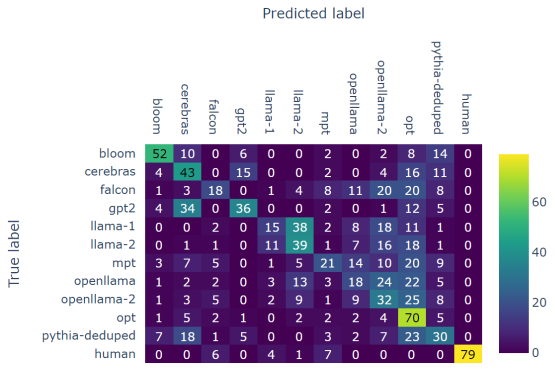

Figure 5: Normalized confusion matrix for model family classification. 5-seed averaged and normalized by the predicted class support.

trained on ThePile, which includes an open replication of the OpenWebText dataset.

Overall, our classifier achieved an F1-score of 17.7% across 44 distinct labels, indicating that LLMs leave detectable signatures, thus enabling source model identification.

Model Family Classification In the Model Family Classification experiments, our primary aim was to evaluate the classifier's efficacy in identifying the model family responsible for generating a given text. This assessment allows us to determine if distinct signatures emerge from the diverse pretraining data and architectural choices utilized by different language models. Figure 5 provides an overview of the Model Family Classification results. Notably, we observe that human-generated text exhibits the highest distinguishability from other model families, followed by the OPT model family. It's worth noting that this distinction might be partly influenced by the subpar generation quality of the OPT-125m model, which stands out and can be easily identified among the models as seen in Section 5.2. Furthermore, we notice a consistent confusion pattern between GPT-2 models and Cereberas-GPT models. These two model families, sharing the GPT-2 architecture but differing in their pretraining data sources, appear to exhibit a higher degree of similarity in their generated text, leading to increased misclassifications. The overall F1-score across 12 distinct model family labels was 37%, underscoring the potential for detecting model family signatures.

5.3 Model Size Classification

In the Model Size Classification experiments, we aimed to assess the classifier's ability to determine

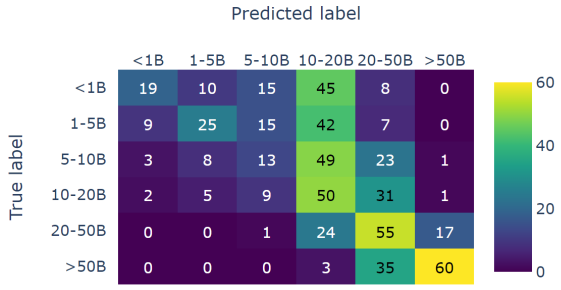

Figure 6: Normalized confusion matrix for model size classification. 5-seed averaged and normalized by the predicted class support.

the size category of the model responsible for generating a given text. This evaluation allows us to discern whether the differences in model sizes translate into detectable signatures in the generated text. As depicted in Figure 6, the results of the Model Size Classification experiment reveal a discernible pattern. Larger models consistently exhibit the least amount of confusion, while models with sizes that are closely related tend to be more frequently misclassified. An interesting exception to this pattern is observed in the case of the 10-20B model, where the classifier tends to confuse other smaller models with it. In summary, the classifier achieves an overall F1-score of 38% across six distinct model size categories.

6 Discussion

The experiments and results presented in this study provide valuable insights into the challenges and nuances of detecting and attributing text generated by different LLMs.

In the cross-model detection experiments, we observed a clear inverse relationship between the effectiveness of the classifier and the size of the test models. Larger LLMs tend to be more challenging to detect, especially when the classifier is trained on data from smaller models. However, training on similarly sized LLMs can improve detection performance on larger models, although it may lead to decreased performance on smaller models. Interestingly, the performance varied across LLM families, with GPT2 and LLaMA-generated text proving harder to detect due to their advanced language modeling capabilities. These findings emphasize the importance of considering both model size and family when developing detection strategies. In addition to the observations made in the cross-model detection experiments, we also conducted experiments to assess the influence of finetuned chat

models, shedding light on the relationship between chat models and traditional language models in the context of detection.

In the Model Attribution experiments, our classifier demonstrated the ability to identify the source model of generated text to a certain extent. Human-generated text was the most distinguishable, while confusions mainly occurred between models from the same family and between models with adjacent sizes. Furthermore, in Model Family Classification, the classifier showed promise in identifying the model family responsible for generating text, highlighting the potential for detecting distinct signatures arising from diverse pretraining data and architectural choices. This indicates that LLMs leave detectable signatures, enabling source model identification, and model family classification. In Model Size Classification, we observed that larger models were less frequently misclassified, emphasizing the influence of model size on detection.

Building upon the findings of (Antoun et al., 2023) which demonstrated the challenging nature of identifying adversarial text composed in an academic, pedagogic, or encyclopedic style for state-of-the-art classifiers trained on a mixture of text generated by LLMs and human content, we also investigated how the detection of adversarial content text could influence the trends we exposed earlier. As shown in Figure 7, in the Adversarial column, the results are massively inferior to the ones reported in our main experimental setting. The inherent out-of-domain distribution of this content[2] compared to our main experiment setting may have indeed contributed significantly to this performance degradation. Nevertheless, it is worth noting that the top-five detection models, with F1-scores ranging from 80 to 72, mostly consist of models trained on text generated by smaller models and primarily from the BLOOM family.

This observation suggests that these detectors are likely taking advantage of relevant textual features to distinguish between automatically generated text of lower quality and human-produced content. However, it is important to acknowledge that the results exhibit variability across models, with models of similar size encountering difficulties in this task, while larger model-trained classifiers also face challenges in this specific context.

Further work is required to investigate the precise factors at play in this scenario. Our key takeaway is that our study was conducted within a controlled environment, aiming to single-out variable influences. Therefore, the level of performance we demonstrated should not be interpreted as indicative of real-world expectations for this task. Overall, our results underscore the complex interplay between model size, family, and training data in the context of LLM detection and attribution. We provide all our experiments results in an interactive online repository https://huggingface.co/spaces/wissamantoun/LLM_Detection_Attribution.

[2]We translated the original data set from French to English using google translate. As shown by Antoun et al. (2023), using translated text from English to French has no effect in the detectability of automatically generated content. We believe this result holds in the French to English direction.

Acknowledgements

This work was partly funded by Benoît Sagot's chair in the PRAIRIE institute funded by the French national research agency (ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001. This work also received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 101021607. The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

We would also like to thank Francis Kulumba, Arij Riabi, and Roman Castagné for the productive discussions.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Maitha Alhammadi, Mazzotta Daniele, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of language models: Towards open frontier models.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraGPT2: Pre-trained transformer for Arabic language generation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 196–207, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Wissam Antoun, Virginie Mouilleron, Benoît Sagot, and Djamé Seddah. 2023. Towards a Robust Detection of Language Model-Generated Text: Is ChatGPT that easy to detect? In *18e Conférence en Recherche d'Information et Applications – 16e Rencontres Jeunes Chercheurs en RI – 30e Conférence sur le Traitement Automatique des Langues Naturelles – 25e Rencontre des Étudiants Chercheurs en Informatique*

pour le Traitement Automatique des Langues, pages 14–27, Paris, France. ATALA.

Mikhail J Atallah, Victor Raskin, Michael Crogan, Christian Hempelmann, Florian Kerschbaum, Dina Mohamed, and Sanket Naik. 2001. Natural language watermarking: Design, analysis, and a proof-of-concept implementation. In *Information Hiding: 4th International Workshop, IH 2001 Pittsburgh, PA, USA, April 25–27, 2001 Proceedings 4*, pages 185–200. Springer.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling.

Abir Chakraborty. 2023. RGAT at SemEval-2023 task 2: Named entity recognition using graph attention network. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 163–170, Toronto, Canada. Association for Computational Linguistics.

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2023. On the possibilities of ai-generated text detection. *arXiv preprint arXiv:2304.04736*.

Sabrina Chiesurin, Dimitris Dimakopoulos, Marco Antonio Sobrevilla Cabezudo, Arash Eshghi, Ioannis Papaioannou, Verena Rieser, and Ioannis Konstas. 2023. The dangers of trusting stochastic parrots: Faithfulness and trust in open-domain conversational question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 947–959, Toronto, Canada. Association for Computational Linguistics.

Miranda Christ, Sam Gunn, and Or Zamir. 2023. Undetectable watermarks for language models. *arXiv preprint arXiv:2306.09194*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Nolan Dey, Gurpreet Gosal, Zhiming, Chen, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, and Joel Hestness. 2023. Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274, Dublin, Ireland. Association for Computational Linguistics.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415.

Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. 2023. Three bricks to consolidate watermarks for large language models.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander 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, Florence, Italy. Association for Computational Linguistics.

Xinyang Geng and Hao Liu. 2023. Openllama: An open reproduction of llama.

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. 2019. Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, Barcelona, Spain (Online). International Committee on Computational Linguistics.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. *arXiv preprint arXiv:2301.10226*.

Kris McGuffie and Alex Newhouse. 2020. The radicalization risks of gpt-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807*.

Fatemehsadat Mireshghallah, Justus Mattern, Sicun Gao, Reza Shokri, and Taylor Berg-Kirkpatrick. 2023. Smaller language models are better black-box machine-generated text detectors.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv preprint arXiv:2301.13852*.

NLP Team MosaicML. 2023. Introducing mpt-30b: Raising the bar for open-source foundation models. Accessed: 2023-06-22.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

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)*, pages 8384–8395, Online. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9054–9065. Curran Associates, Inc.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences.

A Model Size Distribution

Size Bin	Train	Test
<1B	9600	2400
1-5B	11200	2800
5-10B	12000	3000
10-20B	7200	1800
20-50B	6400	1600
>50B	3200	800

B Adversarial Results

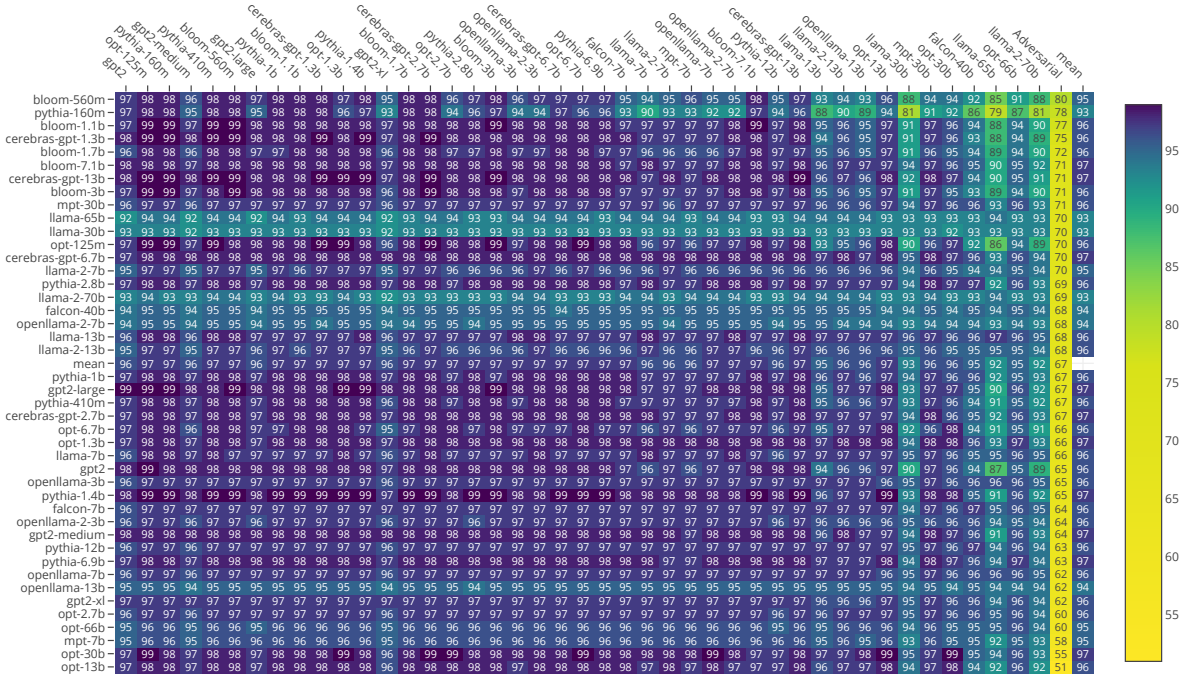

Figure 7: Figure 1 sorted by Adversarial results