Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Towards a Robust Detection of Language Model-Generated Text: Is ChatGPT that Easy to Detect?

Wissam Antoun Virginie Mouilleron Benoît Sagot Djamé Seddah
Inria, Paris, France
firstname.lastname@inria.fr

ABSTRACT

Recent advances in natural language processing (NLP) have led to the development of large language models (LLMs) such as ChatGPT. This paper proposes a methodology for developing and evaluating ChatGPT detectors for French text, with a focus on investigating their robustness on out-of-domain data and against common attack schemes. The proposed method involves translating an English dataset into French and training a classifier on the translated data. Results show that the detectors can effectively detect ChatGPT-generated text, with a degree of robustness against basic attack techniques in in-domain settings. However, vulnerabilities are evident in out-of-domain contexts, highlighting the challenge of detecting adversarial text. The study emphasizes caution when applying in-domain testing results to a wider variety of content. We provide our translated datasets and models as open-source resources.[1]

RÉSUMÉ

Vers une détection robuste de texte généré par un modèle de langue : ChatGPT est-il si facile à détecter ?

Les récents progrès en traitement automatique des langues (TAL) ont conduit au développement de grands modèles de langage (*Large Language Models*, LLM) tels que ChatGPT. Cet article propose une méthodologie pour développer et évaluer des modèles de détection de contenus en français produits par ChatGPT, en mettant l'accent sur leur robustesse face à des données hors domaine et face à des attaques classiques. La méthode proposée consiste à traduire un ensemble de données anglaises en français et à entraîner un classificateur sur les données traduites. Les résultats montrent que les détecteurs peuvent efficacement détecter le texte généré par ChatGPT, avec un bon niveau de robustesse contre les techniques usuelles d'attaque sans changement de domaine. Cependant, les résultats sont moins bons dans les contextes hors domaine, soulignant le défi que constitue toujours de contenus adversariaux. Notre étude souligne l'importance de rester prudent lorsque l'on cherche à généraliser des résultats obtenus sans changement de domaine à une plus grande variété de contenus. Tous nos jeux de données et nos modèles sont distribués librement.

KEYWORDS: ChatGPT, text generation, detection of machine-generated text, robustness.

MOTS-CLÉS : ChatGPT, génération de texte, détection de texte généré par machine, robustesse.

[1] https://gitlab.inria.fr/wantoun/robust-chatgpt-detection

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs, pages 14–27.

1 Introduction

Advances in natural language processing (NLP) have been driven mainly by scaling up the size of pre-trained language models, along with the amount of data and compute required for training (Raffel *et al.*, 2020; Radford *et al.*, 2019; Rae *et al.*, 2021; Fedus *et al.*, 2021; Hoffmann *et al.*, 2022). OpenAI recently released ChatGPT, a text generation model with conversational capabilities. The model is based on GPT3.5 which is a version of GPT3 (Brown *et al.*, 2020) first fine-tuned on code then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) (Christiano *et al.*, 2017; Stiennon *et al.*, 2020), a method previously demonstrated by OpenAI with InstructGPT (Ouyang *et al.*, 2022). This fine-tuning process contributes not only to the model's knowledge but also simplifies the model's interface compared to GPT3, which necessitated substantial *prompt engineering* to achieve satisfactory outcomes, and hence facilitating the extraction and application of that built-in knowledge.

As a result of these significant performance improvements, ChatGPT and other large language models have gained much popularity in the media and in the social context, often without fully understanding the underlying limitations of the models – e.g., the possibility of generating hateful, hateful, toxic, or disrespectful content (Bender *et al.*, 2021; McGuffie & Newhouse, 2020; Weidinger *et al.*, 2021). Another potential misuse of LLMs or ChatGPT is industrializing radicalization and harmful propaganda which poses a significant and unconventional threat to civil society.

In response to the mounting concerns surrounding potential misuse, numerous researchers are now exploring various strategies to mitigate associated risks. For example, some have proposed watermarking techniques to trace the origin of generated text[2], while others are developing methods to detect and flag text generated by these models. Of particular interest, a recent study by Guo *et al.* (2023) investigated the text generation capabilities of ChatGPT and its proximity to human-generated text. To create a dataset of ChatGPT-generated text, the authors leveraged pre-existing question-answering datasets in both English and Chinese, using the questions as prompts to generate responses from the model. In addition, the authors conducted a linguistic analysis to compare the output generated by ChatGPT with human-written text, and they also developed a detector to distinguish between ChatGPT-generated text and human-written text by fine-tuning a separate language model on a dataset containing both types of text.

The aim of this research is to explore the development of ChatGPT detectors in multiple languages, along with evaluating their robustness on out-of-domain text, we selected French as the language of interest. Therefore, we propose a methodology that involves translating the English dataset into French and subsequently training a classifier on the translated data. We conducted a series of evaluations in a monolingual and multilingual setting on both in-domain and out-of-domain data. The in-domain data consisted of text generated by ChatGPT using prompts related to the topics covered in the training dataset. The out-of-domain data included text generated in French by ChatGPT and Bing, a search engine powered by ChatGPT[3], which has access to a broader range of internet content and may generate text on a wider range of topics than ChatGPT. Given that Wolff & Wolff (2020) demonstrated the vulnerability of BERT-based detectors for GPT-2 against basic attack schemes, such as substituting characters with homoglyphs or misspelled words, we also evaluated the robustness of our models against these types of attacks. Furthermore, we hypothesize that the detector models we trained rely

[2]At the time of writing, OpenAI was working on a tool to statistically watermark text generated by GPT-like models according to Scott Aaronson, a guest researcher at OpenAI https://scottaaronson.blog/?p=6823

[3]https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/

heavily on the didactic response style of ChatGPT to distinguish between human-generated content.

The contributions of this study can be summarized as follows:

- We build upon the work of Guo *et al.* (2023) and propose a methodology to develop ChatGPT detectors in multiple languages, focusing on French as a case study.
- We evaluated the performance of the ChatGPT detectors in both a monolingual and multilingual setting. Specifically, we trained and evaluated our models on both the English and French datasets, as well as on a combined dataset containing both languages.
- We investigate the generalizability of our detector by testing its performance on out-of-domain data.
- We evaluate the robustness of our models against common attack schemes, such as substituting characters with homoglyphs or misspelled words. This is an important aspect to consider in the deployment of ChatGPT detectors, as attackers may attempt to evade detection by modifying the generated text in subtle ways.
- We investigated the reliance of detector models on ChatGPT's didactic response style for distinguishing between human-generated content.
- We release all translated datasets and models as open source to encourage further research in this area and to enable others to replicate our experiments.

Overall, our work contributes to the growing body of research on developing and evaluating ChatGPT detectors, with a focus on multilinguality, generalizability, and robustness to attacks. Our findings have practical implications for the use of ChatGPT detectors in various settings, including social media platforms, online forums, and chatbots, where the detection of harmful content is critical for maintaining a safe and respectful online environment.

2 Related Works

2.1 Large Language Models

The race to scale up language models to new heights has been a hot topic in recent years. Researchers and tech companies have been competing to develop larger and more powerful models, often breaking records for model size and performance. The trend began with OpenAI's GPT-2 (Radford *et al.*, 2018), which was released in 2019 and featured 1.5 billion parameters. This was quickly followed by Megatron (Shoeybi *et al.*, 2019), a 8.3 billion parameter model, displaying steadily increasing superior zero-shot language model performance, and T5 (Raffel *et al.*, 2020), an 11 billion parameter encoder-decoder model which advanced transfer learning and performance on several closed-book question answering tasks. The release of GPT-3 (Brown *et al.*, 2020), and PaLM (Chowdhery *et al.*, 2022) represented a major milestone in the race to scale up language models, with their unprecedented 175 and 540B billion parameters. Scaling models to such massive scales "unlocks" new emergent capabilities as shown in Chowdhery *et al.* (2022). In November 2022, OpenAI released ChatGPT, a conversational language model based on GPT-3.5 fine-tuned using Reinforcement Learning from Human Feedback (RLHF) (Christiano *et al.*, 2017; Stiennon *et al.*, 2020), a method previously demonstrated by OpenAI with InstructGPT (Ouyang *et al.*, 2022).

2.2 Detecting Synthetic Text

Detecting synthetically generated text is one of the defense mechanisms against harm caused by LLMs. One of the first major explorations of this topic was conducted following the release of GROVER (Zellers *et al.*, 2019) a fake news generator and detector. Since this approach has been shown to work quite well, and as part of their model release strategies (Solaiman *et al.*, 2019), OpenAI also released a GPT2 detector based on a fine-tuned RoBERTa model (Liu *et al.*, 2019), later Fagni *et al.* (2021) demonstrated the performance of another RoBERTa-based detector on machine-generated tweets, Uchendu *et al.* (2020) also used RoBERTa to spot news generated by several language models, while Antoun *et al.* (2021b) created an ELECTRA-based model (Antoun *et al.*, 2021a) to spot articles generated by their AraGPT2. The authors stated that the success of their method was due to the model being pre-trained on the exact same dataset as AraGPT2, and also due to the replaced-token detection pre-training objective (RTD) (Clark *et al.*, 2020), which bears a resemblance to the synthetic text detection objective. Nguyen-Son *et al.* (2021) proposed a detector that uses the text similarity with round-trip translation (TSRT), to detect a machine-translated text from a never before seen translator, and achieved 86.9% detection accuracy. On the other hand, Wolff & Wolff (2020) showed that the RoBERTa GPT-2 detector is vulnerable against simple attack schemes such as substituting characters with homoglyphs or misspelled words. In these cases, the detector's recall went down from 97% to 0.26% and 22.68% respectively. In order to further enhance the accuracy of detecting manipulated news articles that may deceive readers, Jawahar *et al.* (2022) proposed a neural network-based detector that uses factual knowledge via graph convolutional neural network to distinguish between human-written news articles and manipulated news articles that mislead readers.

Following the release of ChatGPT, and with the recognition of the potential risks posed by this highly-capable model, there has been a surge of investigation into methods for detecting ChatGPT-generated text, which led to the commercial release of multiple detector products. However, as the methods used in these products are often not publicly verifiable, this study focuses solely on the academic literature surrounding detection methods. Mitchell *et al.* (2023) proposes a new method called DetectGPT, which leverages negative curvature regions of the model's log probability function and does not require training a separate classifier or watermarking generated text, resulting in a more discriminative approach than existing zero-shot methods for model sample detection. Notebly, Guo *et al.* (2023) created a dataset of ChatGPT responses to queries from diverse sources in English and Chinese. The authors investigated the linguistic and stylistic differences between human and ChatGPT-generated text, in addition to training a variety of classifiers of which a finetuned pretrained language model turned out to be the best.

3 Methodology

3.1 Data Collection

To train and evaluate our ChatGPT detectors, we leveraged the Human ChatGPT Comparison Corpus (HC3) created by Guo *et al.* (2023) which contains both human-written and ChatGPT-generated text in English and Chinese. We primarily focus on the English portion of the dataset as machine translation performs optimally on it. The dataset consists of 24,322 human-written questions and 58546 answers sourced primarily from ELI5 (Fan *et al.*, 2019), WikiQA (Yang *et al.*, 2015), Crawled Wikipedia,

Medical Dialog dataset (Chen *et al.*, 2020), and FiQA (Maia *et al.*, 2018). The authors generated ChatGPT responses using OpenAI's web application,[4] automating the input of questions and scraping the answers with the help of automation testing tools, for a total of 26903 ChatGPT-generated answers.

To create a French dataset, we translated the English dataset using the Google Cloud Translation API. We then split the dataset into three splits train, validation, and test, by first selecting 710 balanced question and answer pairs to be validated, manually annotated[5] and to serve as our test set. We split the rest in an 80/20 split to get the training and validation set.

Furthermore, to assess the ChatGPT detectors' ability to generalize, we manually compiled out-of-domain test data by means of:

- Manually collecting 113 ChatGPT French responses to high-quality translated questions from the test set, referred to as the **ChatGPT-Native** set.
- Using Bing, we manually collect 106 French responses to high-quality translated questions from the test set which we refer to as the **BingGPT**. Given that BinGPT includes source citations in its output, we remove these artifacts from the data (as well as all of its self-referring mentions).
- Randomly sampling 4454 French question-answer pairs from the French subset of the Multilingual FAQ Dataset (MFAQ) (De Bruyn *et al.*, 2021), known as the **FAQ-Rand** set.
- Since the French FAQ data featured in the MFAQ dataset could be machine translated, we create a smaller set from the French FAQ data featured by filtering for .gouv domains, named the **FAQ-Gouv** set.
- 1235 sentences from The French Treebank test set, corpus from Le Monde (Abeillé *et al.*, 2000) articles, which we denote as the **FTB** set.
- Moreover, in order to investigate our hypothesis that the detector relies heavily on the style of ChatGPT and Bing answers to distinguish between human-generated content, we created an additional set of responses to 61 questions. These responses were crafted as "open-book" answers with the same style as those provided by ChatGPT and Bing, resulting in a set of responses that we refer to as the **Adversarial** set.

3.2 Detector Architecture

Our approach fine-tunes pre-trained transformer-based models on our binary classification dataset.

For English, we used two pre-trained transformer models: RoBERTa (Liu *et al.*, 2019) and ELECTRA (Clark *et al.*, 2020). RoBERTa is a variant of BERT (Devlin *et al.*, 2019), trained using masked language modeling. ELECTRA, on the other hand, introduced a new training objective, Replaced Token Detection (RTD), that replaces tokens in the input sequence with tokens generated by another model and then requires the discriminator to distinguish between the replaced and original tokens. We hypothesize that this objective should improve performance since the RTD objective greatly resembles the machine-generated text detection objective.

For French, we used two pre-trained transformer models: CamemBERT (Martin *et al.*, 2020) and CamemBERTa.[6] CamemBERT is a RoBERTa model trained from scratch on French text, while CamemBERTa is based on the DeBERTaV3 (He *et al.*, 2021) architecture and trained from scratch on

[4] https://chat.openai.com/chat

[5] The detailed annotation guideline will be publicly released with our dataset.

[6] The model paper is currently under review and will be released soon.

French text using RTD.

For the multilingual setting, we only fine-tune XLM-R (Conneau *et al.*, 2020), a multilingual RoBERTa model with supports for 100+ languages.[7]

4 Experimental Methodology and Results

4.1 Experiment Design

Motivated by Guo *et al.* (2023), and given that the HC3 dataset comprises question/answer pairs, we investigated three distinct methods for generating dataset examples:

- Jointly incorporating the question and answer into the model input, which we refer to as the **qa** subset.
- Using only the full answer text, which we refer to as the **full** subset.
- Splitting the answer text into sentences, resulting in shorter text segments and producing 455,320 training examples and 114,117 validation examples. We refer to this subset as the **sentence** subset.

To test the robustness of our approach against adversarial attacks, we add misspellings and simulate homoglyph substitution on the **full** subset of the test sets, using the *nlaug* (Ma, 2019) library.

Regarding our choices of training hyperparameters, we maintain a fixed batch size of 32, adopt a linear scheduler with a warmup ratio of 0.1%, and restrict our learning rate tuning to a range between 10^{-5} and 5.10^{-5} with a step size of 10^{-5}. Our model is trained for 5 epochs, and we report the results averaged over 5 distinct random seeds for all in-domains results. For the out-of-domains experiments, we used the best models.

4.2 Results

4.2.1 In Domain

Table 1 presents the results obtained from hyperparameter tuning. Notably, both evaluated French models demonstrated exceptional performance, and consistent stability evidenced by the low standard deviation scores. However, the scores for French models were comparatively lower than the English models, indicating the impact of translation on model performance. Our findings suggest that the performance of models trained on the QA and Full subset significantly deteriorates when assessed on short-length or sentence data, indicated also by the high standard deviation scores. Conversely, models trained on sentences exhibit a relatively consistent performance across all subsets. Considering the overall highest performance on the **Full** subset, we opted to conduct subsequent experiments with the CamemBERTa and RoBERTa models trained on the Full subset.

Furthermore, Table 2 displays a detailed breakdown of the scores obtained from the **Full** subset. Notably, the models consistently achieve high recall scores in identifying ChatGPT across all tested languages. Additionally, the inclusion of misspellings and homoglyph substitutions improves the

[7]We also tested mDeBERTa (He *et al.*, 2021) but it wasn't converging in any of our hyper-parameter tuning experiments.

models' ability to detect human-written text while slightly reducing their performance for machine-generated text. The multilingual model XLM-R demonstrates superior and more resilient performance on both the French and English test sets, exhibiting increased robustness against adversarial attacks. When compared to a native French-speaking human linguist, the trained model accurately identifies ChatGPT-generated content with higher accuracy, while the human linguist achieves a similar human detection score.

Train	QA			Full			Sentence		
Test	QA	Full	Sentence	QA	Full	Sentence	QA	Full	Sentence
				French					
CamemBERT	**98.37±0.5**	97.79±0.4	**40.20±8.6**	**92.43±1.2**	**98.44±0.4**	25.08±4.7	**93.48±5.2**	96.41±0.6	90.27±0.3
CamemBERTa	98.23±0.3	**98.48±0.3**	32.00±6.3	90.13±1.0	**98.49±0.4**	**29.11±3.6**	81.82±3.4	**96.71±0.1**	**91.18±0.2**
				English					
RoBERTa	**99.88±0.03**	**98.91±0.2**	51.23±7.6	**98.58±0.7**	**99.86±0.03**	**66.93±5.4**	71.10±19.4	**99.39±0.07**	**98.17±0.1**
ELECTRA	99.27±0.2	**99.07±0.2**	**65.23±8.3**	96.24±0.7	99.35±0.1	43.82±9.1	**93.57±1.9**	97.05±0.4	93.60±0.1

Table 1: Average and standard deviation of F1 scores for the best model on the validation set with adversarial perturbations.

Evalution set		French			English			Multilingual						Human Expert		
								French-Test			English-Test					
		Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Full subset	ChatGPT	0.95	1	0.97	0.99	1	0.99	0.99	1	0.99	0.98	1	0.99	0.98	0.87	0.92
	Human	1	0.94	0.97	1	0.99	0.99	1	0.99	0.99	1	0.98	0.99	0.88	0.98	0.93
+misspelling	ChatGPT	1	0.95	0.98	0.99	0.79	0.88	1	0.96	0.98	0.99	0.99	0.99	-	-	-
	Human	0.95	1	0.98	0.82	0.99	0.9	0.96	1	0.98	0.99	0.99	0.99	-	-	-
+homoglyphs	ChatGPT	1	0.94	0.97	0.99	0.87	0.93	1	0.97	0.99	0.99	0.99	0.99	-	-	-
	Human	0.94	1	0.97	0.88	0.99	0.93	0.99	0.99	0.99	0.97	1	0.99	-	-	-

Table 2: Detailed test set scores (full subset) breakdown of CamemBERTa (French), RoBERTa (English), XLM-R (Multilingual) trained on the full subset.

4.2.2 Out-of-Domain

To assess the potential for overfitting to our in-domain data, we evaluated the performance of our French detector on the out-of-domain test sets described previously. The results, shown in Table 3, reveal the detector's exceptional performance on the FTB and FAQ-Gouv test sets, with a drop in accuracy to 88.75 on the FAQ-Rand subset. This suggests the model may be detecting translation artifacts that remain in some FAQ web pages after automatic translation. Remarkably, our detector correctly identified French text generated natively by ChatGPT, suggesting that it may be possible to develop detectors for other languages by translating existing datasets. Similarly, the detector models displayed surprising performance in detecting content generated by BingGPT.

The multilingual detector model consistently outperformed the monolingual model only in detecting human-generated text but fell behind in detecting ChatGPT or BingGPT-generated text, this behavior might be due to the significantly larger pre-training dataset of XLM-R compared to CamemBERTa.

The detector models also exhibited clear weaknesses against misspelling and homoglyph-based attacks. For instance, the performance of CamemBERTa and XLM-R dropped to 44.81 and 28.18, respectively, when detecting BingGPT-generated text with misspellings added.

Finally, the low scores obtained by the detector on the adversarial response dataset we developed in the style of ChatGPT and BingGPT serve to validate our detector's heavy reliance on the writing style utilized in generating responses.

True label	Human												ChatGPT					
Model	FTB			FAQ-Rand			FAQ-Gouv			Adversarial			Native			BingGPT		
	raw	*+ms*	*+hg*	raw	*+ms*	*+hg*	*raw*	*+ms*	*+hg*	*raw*	*+ms*	*+hg*	raw	*+ms*	*+hg*	raw	*+ms*	*+hg*
CamemBERTa	99.19	**99.92**	**100**	88.75	99.01	99.10	96.17	**100**	**99.57**	33.57	87.61	**85.49**	99.19	81.42	84.96	**92.45**	44.81	48.37
XLM-R	**99.43**	99.59	99.76	**95.35**	**99.39**	**99.55**	**96.59**	**100**	**99.57**	**59.12**	**89.05**	82.67	94.69	60.18	62.83	77.46	28.18	35.72
*Trained on a mix of raw, misspellings and homoglyphs**																		
CamemBERTa	98.98	98.54	98.79	80.56	84.51	84.73	90.64	91.49	90.21	45.90	42.62	44.26	**100**	**99.12**	**99.95**	91.51	**91.51**	**90.57**
XLM-R	98.54	98.78	98.79	85.20	88.84	95.32	92.34	96.17	95.32	62.26	60.66	62.30	100	97.34	99.16	62.26	53.77	56.60

Table 3: Accuracy scores of CamemBERTa and XLM-R on the French out-of-domain test sets. *ms*: misspelling, *hg*: homoglyphs. **Dataset mix was 100% raw, 50% misspellings and 50% homoglyphs.*

5 Discussion

About the link between translation quality and the model detectability As part of our study to assess the possibility of differentiating between texts written by humans and those generated by LLMs, following the work of Guo *et al.* (2023), we analyzed and re-evaluated the responses in the translated French dataset. The purpose was to confirm the hypothesis that a human expert can generally distinguish between a ChatGPT-generated text and one written by a human. We initially rated the translation quality on a scale of 1 to 5, with 5 indicating a good translation. Translations with scores exceeding 3 were retained even though ChatGPT managed to interpret badly translated questions extremely well.

Additionally, we assessed the correlation between our detector's performance and translation quality scores, and found it to be weak.[8]

About discriminating linguistics clues We identified several visible characteristics in the generated texts. ChatGPT uses an impersonal and didactic style, characterized by extensive use of the impersonal form, conditionals statements, as shown here:

- "Cela **pourrait** également nuire à la réputation de l'entreprise (...)"
- "Cela **pourrait** entraîner une baisse des dépenses des consommateurs (...)"
- "**Si** vous êtes allergique aux chats, cela signifie que votre corps a une réaction anormale aux protéines présentes dans leur peau, leur urine ou leur salive. **Si** vous deviez manger de la viande de chat, il est possible que (...)"

The language model structures its responses to create an impression of coherence and clarity. It often reformulates the question in its answer, resulting in a didactic response that aligns with the question.

[8]With three different correlations measures showing the same trend: Spearman's τ of -0.25, Pearson's R of -0.26 and Kendall's τ of -0.24.

- Question: “Pourquoi **mon signal wifi semble se dégrader avec le temps** ? Je réinitialise/redémarre constamment mon routeur et/ou mon modem. Je dois noter que je vis dans un petit appartement et que j’ai utilisé 2 routeurs haut de gamme. Explique comme si j’avais cinq ans.”
  GPT : “Il peut y avoir plusieurs raisons pour lesquelles **votre signal Wi-Fi se dégrade avec le temps**. Voici quelques explications possibles : (...)”

ChatGPT’s responses are often general, and it redefines the subject on which the question is asked.

When asked “How does nature solve for Pi ?” or “*Comment la nature résout-elle pour Pi ?”, it started by stating the definition of Pi:

- “Pi, ou le nombre 3,14, est une constante mathématique qui représente le rapport de la circonférence d’un cercle à son diamètre. La valeur de Pi est d’environ 3,14, mais c’est un nombre irrationnel (...)”

Additionally, ChatGPT is characterized by the absence of some human markers, such as errors in punctuation, spelling, or grammar. The language model does not use any tone, judgment, or personal touch, such as (“je pense que” / “je juge que”) , which creates a neutral impression. While its responses lack a human touch, it provides a specific recommendation when discussing technical or sensitive issues, such as consulting a specialist or seeking medical attention. Also, It does not ask any questions except towards the end of the response.

- “(...) **il est important de consulter un médecin** dès que possible. **N’hésitez pas à appeler le** 911 ou votre numéro d’urgence local si vous ressentez des douleurs à la poitrine ou d’autres symptômes d’une condition médicale grave.”
- “(...) **Il est important de** vérifier les instructions de votre four à micro-ondes pour voir si le support en métal peut être utilisé en toute sécurité.”
- “(...) Encore une fois, **je vous recommande de** parler avec un dermatologue ou un autre professionnel de la santé pour déterminer le plan de traitement le plus approprié pour votre cas spécifique.”
- “(...) J’espère que cela vous aidera à l’expliquer ! **Y a-t-il autre chose que vous aimeriez savoir sur Vénus ou sur la façon dont elle se déplace dans l’espace ?**”

Our study suggests that these visible differences could be used to differentiate between human-written texts and AI-generated texts automatically. It shall be noted that the ChatGPT tendency to produce didactic text can lead any detector trained on its content to be easily fooled assuming the text follows the same patterns. This is what showed our results in Table 3 (“Adversarial” column results).

About the character-level perturbations Interestingly, the introduction of character-level perturbations increased the model’s capability of detecting the adversarial human content, albeit at the expense of its capacity to detect Bing automatically generated content. This finding suggests that the addition of perturbations to content renders it more comparable to human-generated content, confirming, for French, previous work on the subject (Wolff & Wolff, 2020). These effects were much more difficult to notice in the in-domain scenarios because of the high-accuracy of the model.

Enhancing the robustness to noise of our models Although not the focus of this work, one obvious path of improvement is to add the same kind of perturbations to the training data in order to make the model more robust. To this end, we performed a quick set of experiments where we added to the training set, 50% of its content perturbed by mispellings and 50% with homoglyphs leading to a training set twice as big.[9] These results, presented in the lower half of Table 3, demonstrate that both models exhibit a minor decrease in human detection accuracy. However, they achieve substantial enhancements and improved robustness, particularly when utilizing CamemBERTa, for detecting ChatGPT-generated text in the presence of noisy data. Consequently, the detector models are now less inclined to attribute writing errors to human authors and instead focus more on writing style. This is evident from the scores obtained on the Adversarial set, where the performance on noisy data aligns more closely with that on the original set. However, this does not make the model less sensitive to other kinds of noises but it is an interesting path of improvement. As always with noisy adversarial user-generated content, the question is to find a more general approach that will avoid a constant *cat and mouse* game when it comes to processing productive content.

Take home message The key takeaway from our study is that detecting adversarial text, which is designed to evade detection by language models, presents a significant challenge. OpenAI has reported[10] a success rate of 26% in their own supervised settings when identifying adversarial content in a challenge set of English text.[11] Furthermore, OpenAI has stated that their detection methods are unreliable for text shorter than 1000 characters. We would like to emphasize that our study does not claim to have produced an universally accurate detector. Our strong results are based on in-domain testing and, unsurprisingly, do not generalize in out-of-domain scenarios. This is even more so when used on text specifically designed to fool language model detectors and on text intentionally stylistically similar to ChatGPT-generated text, especially instructional text.
We are currently extending the adversarial dataset using much more various sources as we believe that understanding the shortcoming of these models is of crucial importance.

6 Conclusion

In conclusion, this paper proposed a methodology for developing and evaluating ChatGPT detectors in multiple languages, focusing on French as a case study. The proposed method involved translating an English dataset into French and training a classifier on the translated data. The results demonstrate that the proposed method can effectively detect ChatGPT-generated text, with a certain degree of robustness against basic attack techniques, albeit exclusively within the in-domain setting. However, the detectors display evident vulnerabilities in out-of-domain contexts, emphasizing the importance of considering different writing styles in training language models. Additionally, the study highlights the significant challenge of detecting adversarial text, which even OpenAI's detection methods have difficulties with. The key takeaway is that caution should be exercised when applying in-domain testing results to a wider variety of content. We provide open-source resources to further advance research in this and are currently working to extend the adversarial dataset to better understand the limitations of these models.

[9] We also tested a 50% original training set + 25% misspelling + 25% homoglyphs pertubations model that led to slightly inferior performance, less than one percentage point of difference.

[10] https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

[11] Not released as the time of writing.

Acknowledgments

We thank the reviewers for their insightful comments. 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" program 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.

References

ABEILLÉ A., CLÉMENT L. & KINYON A. (2000). Building a treebank for French. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece: European Language Resources Association (ELRA).

ANTOUN W., BALY F. & HAJJ H. (2021a). AraELECTRA: Pre-training text discriminators for Arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, p. 191–195, Kyiv, Ukraine (Virtual): Association for Computational Linguistics.

ANTOUN W., BALY F. & HAJJ H. (2021b). AraGPT2: Pre-trained transformer for Arabic language generation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, p. 196–207, Kyiv, Ukraine (Virtual): Association for Computational Linguistics.

BENDER E. M., GEBRU T., MCMILLAN-MAJOR A. & SHMITCHELL S. (2021). On the dangers of stochastic parrots: Can language models be too big? DOI : 10.1145/3442188.3445922.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A. *et al.* (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901.

CHEN S., JU Z., DONG X., FANG H., WANG S., YANG Y., ZENG J., ZHANG R., ZHANG R., ZHOU M., ZHU P. & XIE P. (2020). Meddialog: a large-scale medical dialogue dataset. *arXiv preprint arXiv:2004.03329*.

CHOWDHERY A., NARANG S., DEVLIN J., BOSMA M., MISHRA G., ROBERTS A., BARHAM P., CHUNG H. W., SUTTON C., GEHRMANN S. *et al.* (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

CHRISTIANO P. F., LEIKE J., BROWN T., MARTIC M., LEGG S. & AMODEI D. (2017). Deep reinforcement learning from human preferences. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 30: Curran Associates, Inc.

CLARK K., LUONG M.-T., LE Q. V. & MANNING C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics, p. 8440–8451, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.747.

DE BRUYN M., LOTFI E., BUHMANN J. & DAELEMANS W. (2021). MFAQ: a multilingual FAQ dataset. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, p. 1–13, Punta Cana, Dominican Republic: Association for Computational Linguistics. DOI : 10.18653/v1/2021.mrqa-1.1.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (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)*, p. 4171–4186, Minneapolis, Minnesota: Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

FAGNI T., FALCHI F., GAMBINI M., MARTELLA A. & TESCONI M. (2021). Tweepfake: About detecting deepfake tweets. *Plos one*, **16**(5), e0251415.

FAN A., JERNITE Y., PEREZ E., GRANGIER D., WESTON J. & AULI M. (2019). ELI5: long form question answering. In A. KORHONEN, D. R. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, p. 3558–3567: Association for Computational Linguistics. DOI : 10.18653/v1/p19-1346.

FEDUS W., ZOPH B. & SHAZEER N. (2021). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*.

GUO B., ZHANG X., WANG Z., JIANG M., NIE J., DING Y., YUE J. & WU Y. (2023). How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

HE P., GAO J. & CHEN W. (2021). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

HOFFMANN J., BORGEAUD S., MENSCH A., BUCHATSKAYA E., CAI T., RUTHERFORD E., CASAS D. D. L., HENDRICKS L. A., WELBL J., CLARK A. *et al.* (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

JAWAHAR G., ABDUL-MAGEED M. & LAKSHMANAN L. (2022). Automatic detection of entity-manipulated text using factual knowledge. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 86–93, Dublin, Ireland: Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-short.10.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

MA E. (2019). Nlp augmentation. https://github.com/makcedward/nlpaug.

MAIA M., HANDSCHUH S., FREITAS A., DAVIS B., MCDERMOTT R., ZARROUK M. & BALAHUR A. (2018). Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of The Web Conference 2018*, WWW '18, p. 1941–1942, Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. DOI : 10.1145/3184558.3192301.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

MCGUFFIE K. & NEWHOUSE A. (2020). The radicalization risks of gpt-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807*.

MITCHELL E., LEE Y., KHAZATSKY A., MANNING C. D. & FINN C. (2023). Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

NGUYEN-SON H.-Q., THAO T., HIDANO S., GUPTA I. & KIYOMOTO S. (2021). Machine translated text detection through text similarity with round-trip translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 5792–5797.

OUYANG L., WU J., JIANG X., ALMEIDA D., WAINWRIGHT C., MISHKIN P., ZHANG C., AGARWAL S., SLAMA K., GRAY A., SCHULMAN J., HILTON J., KELTON F., MILLER L., SIMENS M., ASKELL A., WELINDER P., CHRISTIANO P., LEIKE J. & LOWE R. (2022). Training language models to follow instructions with human feedback. In A. H. OH, A. AGARWAL, D. BELGRAVE & K. CHO, Éds., *Advances in Neural Information Processing Systems*.

RADFORD A., NARASIMHAN K., SALIMANS T. & SUTSKEVER I. (2018). Improving language understanding by generative pre-training.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D. & SUTSKEVER I. (2019). Language models are unsupervised multitask learners.

RAE J. W., BORGEAUD S., CAI T., MILLICAN K., HOFFMANN J., SONG F., ASLANIDES J., HENDERSON S., RING R., YOUNG S. *et al.* (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, **21**(140), 1–67.

SHOEYBI M., PATWARY M., PURI R., LEGRESLEY P., CASPER J. & CATANZARO B. (2019). Megatron-LM: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

SOLAIMAN I., BRUNDAGE M., CLARK J., ASKELL A., HERBERT-VOSS A., WU J., RADFORD A., KRUEGER G., KIM J. W., KREPS S. *et al.* (2019). Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

STIENNON N., OUYANG L., WU J., ZIEGLER D., LOWE R., VOSS C., RADFORD A., AMODEI D. & CHRISTIANO P. F. (2020). Learning to summarize with human feedback. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 3008–3021: Curran Associates, Inc.

UCHENDU A., LE T., SHU K. & LEE D. (2020). Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 8384–8395, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.673.

WEIDINGER L., MELLOR J., RAUH M., GRIFFIN C., UESATO J., HUANG P.-S., CHENG M., GLAESE M., BALLE B., KASIRZADEH A. *et al.* (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

WOLFF M. & WOLFF S. (2020). Attacking neural text detectors. *arXiv preprint arXiv:2002.11768*.

YANG Y., YIH S. W.-T. & MEEK C. (2015). Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*: ACL - Association for Computational Linguistics.

ZELLERS R., HOLTZMAN A., RASHKIN H., BISK Y., FARHADI A., ROESNER F. & CHOI Y. (2019). Defending against neural fake news. *Advances in neural information processing systems*, **32**.