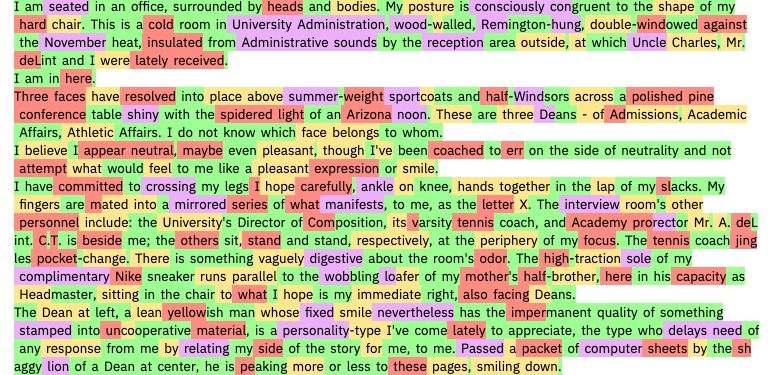
Ghostbuster: How to Spot Fake Texts Written by AI

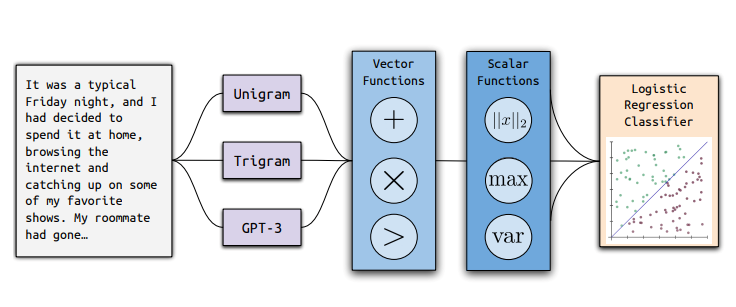


Have you ever wondered if the text you are reading online is written by a human or a machine? With the rapid advances in natural language generation (NLG) technologies, it is becoming harder and harder to tell the difference. NLG models, such as GPT-3, can produce fluent and coherent texts on various topics and styles, often fooling human readers and evaluators. However, these models also pose serious ethical and social challenges, such as spreading misinformation, plagiarism, and manipulation.

In this blog post, I will introduce you to **Ghostbuster**, a state-of-the-art system for detecting text ghostwritten by large language models, developed by researchers at the University of California, Berkeley. Ghostbuster is a novel and powerful tool that can help you identify and verify the authenticity and origin of the texts you encounter online. I will explain the main features and advantages of Ghostbuster, and show you some examples of how it works in practice.

Feature 1: Model-agnostic Detection

One of the key features of Ghostbuster is that it does not require access to token probabilities from the target model, making it useful for detecting text generated by **black-box models** or **unknown model versions**. This means that Ghostbuster can detect text generated by any NLG model, regardless of its architecture, parameters, or training data. This is a significant impro vement over previous methods, which often relied on specific model features or assumptions, and could be easily bypassed by changing the model or its settings.

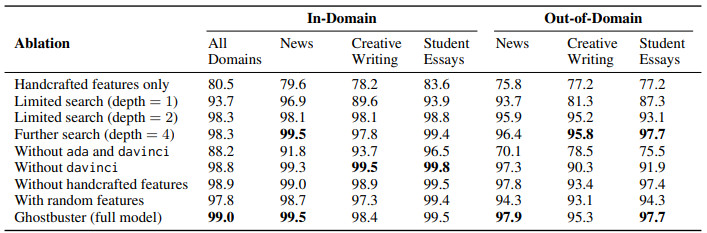


**Outline of the model training procedure of Ghostbuster.**

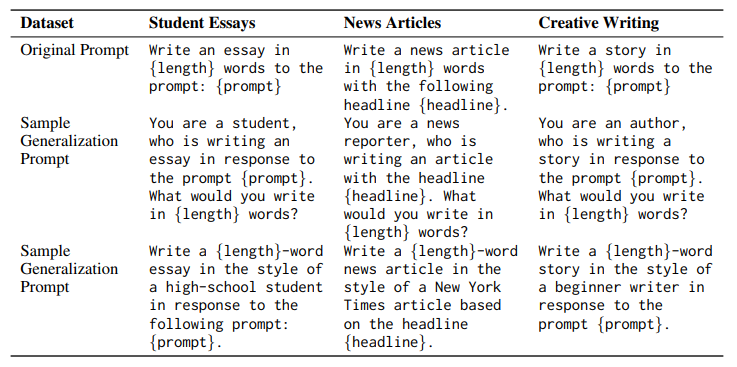
For example, suppose you want to detect text generated by ChatGPT, a commercial NLG model that powers several chatbots and conversational agents online. ChatGPT is a black-box model, meaning that you do not have access to its internal workings or outputs. How can you tell if the text you are chatting with is written by ChatGPT or a human? Ghostbuster can help you with that, by analyzing the text and comparing it with a reference corpus of human-written texts. Ghostbuster can detect subtle differences and patterns that distinguish ChatGPT-generated texts from human-written texts, such as word choices, sentence structures, and semantic coherence.

Feature 2: Cross-domain Generalization

Another key feature of Ghostbuster is that it achieves high performance across different writing domains, such as student essays, creative writing, and news articles. This means that Ghostbuster can detect text generated by any NLG model, regardless of its domain, genre, or style. This is a significant improvement over previous methods, which often suffered from domain-specific biases or limitations, and could be easily fooled by changing the domain or the style of the text.

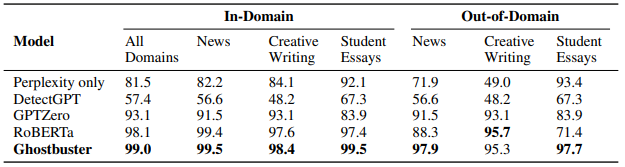


For example, suppose you want to detect text generated by GPT-3, which can produce texts on various domains and styles, such as fiction, poetry, and journalism. How can you tell if the text you are reading is written by GPT-3 or a human? Ghostbuster can help you with that, by analyzing the text and comparing it with a reference corpus of human-written texts from the same domain and style. Ghostbuster can detect subtle and nuanced differences and patterns that distinguish GPT-3-generated texts from human-written texts, such as tone, mood, and creativity.

To demonstrate Ghostbuster's cross-domain generalization ability, the researchers released three new datasets of human-written and AI-generated texts from different domains: student essays, creative writing, and news articles.

**Sample prompts used to produce paired ChatGPT-generated data.**

Ghostbuster achieved high accuracy and recall on all three datasets, outperforming previous methods and human baselines. The datasets are publicly available for anyone who wants to test their own methods or skills on detecting AI-generated texts.

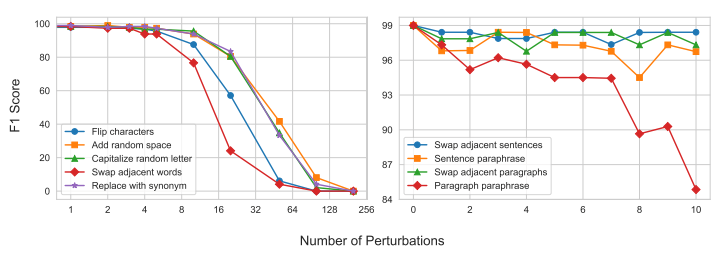


**Results of Ghostbuster model across a variety of text domains (F1 score)**

Feature 3: Robustness to Perturbations and Paraphrasing

A third key feature of Ghostbuster is that it is robust to various perturbations and paraphrasing attacks that could be used to evade detection, such as spelling errors, word substitutions, and sentence reordering. This means that Ghostbuster can detect text generated by any NLG model, regardless of how it is modified or altered. This is a significant improvement over previous methods, which often relied on specific text features or patterns, and could be easily tricked by changing or obscuring the text.

To demonstrate Ghostbuster's robustness to perturbations and paraphrasing, the researchers conducted several experiments where they applied different types of modifications to the AI-generated texts, such as adding spelling errors, replacing words with synonyms, and shuffling sentences. Ghostbuster was able to detect the AI-generated texts with high accuracy and recall, even when the texts were heavily modified or obscured. The results show that Ghostbuster is not easily fooled by superficial changes or manipulations, and can detect the underlying signals and features of AI-generated texts.



**Robustness experiments on Ghostbuster (F1 score). Character- and word-level perturbations (left), sentence -and paragraph- level perturbations (right).**

Conclusion

In this blog post, I have introduced you to Ghostbuster, a state-of-the-art system for detecting text ghostwritten by large language models. I have explained the main features and advantages of Ghostbuster, and showed you some examples of how it works in practice. I hope you have learned something new and useful about how to spot fake texts written by AI, and why it matters.

Ghostbuster is not a perfect or a final solution. It is a work in progress, and it has its own limitations and challenges. For instance, Ghostbuster requires a large and diverse reference corpus of human-written texts, which may not always be available or accessible. Ghostbuster also relies on statistical and linguistic features, which may not capture all the aspects and dimensions of text quality and credibility.

Therefore, Ghostbuster is not a substitute for human judgment or evaluation. It is a complement and a support. It is a tool that can help us detect and flag AI-generated texts, but it is up to us to verify and validate them.

What do you think of Ghostbuster? Have you ever encountered or used AI-generated texts? How do you tell if a text is written by a human or a machine? Share your thoughts and experiences in the comments below. And remember: **don't believe everything you read online. Be a ghostbuster.**