White paper

Turnitin's AI writing detection model architecture and testing protocol

Maximizing effectiveness and safety for academic use

This whitepaper presents the Turnitin AI writing detection system, focusing on its architecture and its testing protocol. This whitepaper also defines and discusses key concepts in generative AI and AI writing detection such as "transformers," "perplexity," "burstiness," "recall" and "false positive rate (FPR)." The Turnitin AI writing detection system has been independently shown to have high effectiveness in correctly identifying AI-generated content. Finally, potential future directions are presented and discussed.

Contents

Introduction and prior work in generative AI language models

Writing has long been a critical cornerstone of teaching and learning, helping to promote critical thinking, creativity, and idea and narrative development for students, among many other important skills. Recent developments in generative AI have created enormous disruptions across almost all sectors, with education and academic writing particularly affected.

The most impactful class of these generative AIs are called Large Language Models (LLMs). LLMs are deep learning models that can generate novel text based on simple writing "prompts." LLMs differ from previous natural language AIs such as sequence to sequence translation models (Sutskever et. al., 2014) in that the prompts are natural language requests, and the generated responses are both novel in nature and are typically remarkably cogent and human-like.

LLMs trace their origins to the invention of the transformer architecture (Vaswani et. al, 2014). Transformers are a particular deep learning architecture that enable the model to associate individual text tokens (words or subwords) with one another in highly nonlinear ways, thereby encoding significant inter-token association. At a high level, the training objective for a transformer language model is relatively simple; LLMs are trained to maximize accuracy on next-word prediction conditioned on a set of previously observed or generated words.

Breakthroughs in scaling computing infrastructure and model training pipelines by industrial research laboratories—most notably OpenAI, Anthropic and Google—have resulted in models with hundreds of billions of parameters (OpenAI, 2023; Chowdhery et. al., 2022; Askell et. al., 2021). The enormous parameterization of these models, combined with highly scaled data flow pipelines and training datasets spanning the breadth of the crawlable internet, allow the models to encode a massive amount of highly generalizable token patterns. At a certain parameter and training data scale, these collective sets of patterns begin to allow the LLM to perform remarkably complex reasoning and linguistic tasks. The existence of these emergent behaviors is the topic of much open research (Wei et. al., 2022; Hagendorff, 2023).

Overview of AI writing detection research

State of the art LLMs consistently exhibit human or near-human abilities on a wide variety of standardized assessments (Zellers et. al., 2019; Sakaguchi et. al., 2019; Chen et al., 2021). While the specific reasons and mechanisms for this performance remain open research questions, the impact of these models on our education and economic systems is undeniable. Within education, the use of these models present enormous opportunities, but there are clearly parts of a student's learning journey where an instructor would want to know about or limit the use of LLMs by the student to encourage critical thinking, learning, and growth.

Detecting writing generated by LLMs helps instructors gain visibility into when LLMs may have been used in the creation of a submitted assignment. While LLMs write in a very human-like manner, they exhibit noticeable statistical signals that are visible to specially trained AI systems. These signals originate from the fact that LLMs generate word tokens sequentially from a probability distribution. The sequences of tokens from LLMs tend to have much more consistent sequential probability than sequences of tokens on the same topic or concept written by a human—meaning LLMs select the most probable word tokens to continue the topic, giving it a more formulaic structure when compared to human writing. The simplest measure of these differences is in the concept of "perplexity" and "burstiness" (Gehrmann et. al., 2019). Perplexity measures the statistical "smoothness" of a sequence of words, while burstiness measures the deviation from norm of statistics such as sentence length. While perplexity and burstiness are useful measures of how AI writing deviates from human writing, in reality, there are an enormous number of long-range statistical dependencies that differentiate human writing and LLM writing.

Turnitin solution overview

Turnitin is a leading provider of academic integrity tools worldwide. These tools are built into popular learning management system (LMS) workflows across 16,000 institutions, 140 countries and used by over 40 million students. In April 2023, Turnitin launched its AI writing detection tool, which as of July 2023 has processed over 76 million paper submissions.

Turnitin's AI writing detection system is built around a state of the art transformer deep-learning architecture. It is trained on a representative sample of data that includes both AI-generated text and authentic academic writing across geographies and subject areas spanning roughly two decades. The AI written text was created by Turnitin to mirror known human writing. Care was taken during dataset construction to represent statistically under-represented groups like second-language learners, English users from non-English speaking countries, students at colleges and universities with diverse enrollments, and less common subject areas such as anthropology, geology, sociology, and others to minimize potential sources of bias when training the model. The system is currently tuned to work in English only to allow optimization in one language before moving onto additional languages. The training, validation and evaluation datasets were created to represent a broad spectrum of LLM prompt strategies, ranging from simple "write the whole essay for me" to more complex mixtures of human and AI writing. The complete testing and held-out evaluation datasets included a rich mixture of purely human, purely AI and mixed AI/human written text.

The use of the transformer architecture was chosen specifically for its flexibility and improved performance compared to a simpler model that relies primarily on hand-curated measures such as perplexity and burstiness that do not capture many higher order deviations. Transformers are designed to intricately model language and allow Turnitin's AI writing detection system to identify more subtle statistical patterns of AI generated writing. This in turn enables the Turnitin AI writing detection system to have high robustness, improved recall and - most importantly - safety (as measured by false positive rate) when compared to other AI writing detection systems. These concepts are defined and discussed in detail in the next section.

Turnitin's transformer model operates on a segment window of text that spans roughly a few hundred words (about five to ten sentences). Each document submission consists of one or more segment windows, with the segment windows being "slid" or "strided" across the document at one sentence stride lengths. This segment windowing allows the model to capture sufficient token statistics to make a reliable prediction on whether the text resembles the signature of AI writing.

The prediction output from the transformer classifier is a single real number between 0 and 1, with 0 meaning that the text in the segment window is highly unlikely to have been written by an AI, and 1 meaning that it is strongly plausible the text was written by an AI. To maintain prediction stability, Turnitin's AI writing detection system has a minimum document length limit of 300 words.

Sentence level AI writing predictions are achieved by a weighted average of the AI writing detection model predictions for the windows in which a sentence appears. This weighting results in a sentence level AI writing prediction score that is compared to a predetermined sentence level AI writing threshold chosen to maximize sentence level recall while minimizing sentence level FPR. The specific threshold for assigning the label of "AI-written" to a sentence varies depending on the specific transformer model, but is typically a value between 0.8 and 1.

A document is labeled as "AI-written" if more than 20% of the sentence level AI writing prediction scores are above a sentence level AI writing threshold described in the previous paragraph. Based on tests conducted by us, we've determined that in cases where we detect less than 20% of AI writing in a document, there is a higher incidence of false positives. Hence, the 20% document proportion cutoff as well as the predetermined model threshold were chosen to keep document level FPR below 0.01 (1%).

A sentence is labeled as "AI-written" if its weighted average AI writing score across all windows is greater than the model threshold.

Testing and evaluation protocol

Turnitin's AI writing detection system is rigorously tested using multiple datasets. Turnitin uses two main metrics to test its AI writing detection system: Recall and FPR.

Recall measures system efficacy. For example, consider a dataset of 100 pieces of writing, 40 of which are generated by a GPT style LLM. Recall would measure how many of the 40 AI written documents are "recalled" or correctly labeled by the AI writing detection system as being "AI-written." If the AI writing detection system in this example correctly labels 30 of the 40 AI written documents, then the recall is 30/40 = 0.75 or 75%.

The FPR measures system safety. In the above mentioned dataset of 100 pieces of writing, the FPR is computed as how many of the 60 human written documents were incorrectly labeled by the AI writing detection system as being "AI-written." If the AI writing detection system in this example flagged 3 of the 60 human written documents as "AI-written," the FPR is 3/60=0.05 or 5%.

Turnitin does not use accuracy as a metric as it is too easily manipulated and too dependent on the specific dataset upon which it is computed. For example, consider a dataset with 100 pieces of writing, 99 of which are human written. A simple, naive algorithm that identifies all pieces of writing as "human-written" would achieve 99% accuracy on this dataset, despite having no value as an AI writing detection system.

To measure FPR, Turnitin conducted a stress-test using 800,000 papers submitted before 2019 and therefore pre-dating GPT-3. All papers in this dataset are assumed to be human written. The production AI writing detection system and its attendant heuristics was run on this dataset, and achieved a document level FPR of 0.007 (0.7%) and a sentence level FPR of 0.002 (0.2%). Our findings are further supported by a recent comparison of popular AI writing detection solutions on the market, where Turnitin's AI writing detection system demonstrated zero false accusations (Weber-Wulff, 2023).

To measure Recall, a held out evaluation dataset of approximately 7,000 documents was used. The dataset comprises a mix of documents that are purely human, purely AI-written and a mix of AI-written and human written. This challenging dataset represents the complex use cases and textual features the Turnitin AI writing detection system may face in the real world. On this dataset, the system achieved a document level Recall of 0.842 (84.2%) and a sentence level Recall of 0.923 (92.3%). These numbers show that the Turnitin AI writing detection system is effective at identifying AI writing for a diversified and complex dataset that closely represents real-world submissions. Weber-Wulff (2023) showed that Turnitin's AI writing detection system outperformed all other AI writing detection solutions on the market in accurately detecting AI writing across a diverse set of evaluations.

Summary and future work

OpenAI's ChatGPT is creating significant disruption and challenges to academic integrity. Turnitin's AI writing detection system is designed to provide educators with valuable insight into the use of GPT style models in student writing, enabling instructors to have vital conversations with students on the appropriate use of these powerful new tools.

Turnitin's AI writing detection system is a safe and effective AI writing detection tool which has been trained and tested on a large collection of human-generated academic writing. Additional studies by Turnitin and by independent researchers further support the low FPR performance and show that Turnitin's AI writing detection system outperforms other AI writing detection systems in the market.

Future work in the AI writing detection space will seek to increase the breadth of generative AI sources that can be detected, including LLMs such as Bard and Claude, as well as generative AI paraphrasing and AI rewriting tools. This will ensure that the Turnitin AI writing system remains a vital tool for instructors in understanding and promoting academic integrity in an increasingly AI-centric world.

References

Askell A, Bai Y., Chen A., et al. (2021). A General Language Assistant as a Laboratory for Alignment.

Chen M., Tworek, J., Jun, H., et al. (2021). Evaluating Large Language Models Trained on Code.

Chowdhery A., Narang S., Devlin J., et al. (2022). PaLM: Scaling Language Modeling with Pathways.

Gehrmann, S., Strobelt, H., & Rush, A. (2019). GLTR: Statistical Detection and Visualization of Generated Text. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 111–116). Association for Computational Linguistics.

Hagendorff, T. (2023). Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods.

OpenAI. (2023). GPT-4 Technical Report.

Sakaguchi, K., Bras, R., Bhagavatula, C., & Choi, Y. (2021). WinoGrande: an adversarial winograd schema challenge at scale. *Communications of the ACM*, 64, 99-106.

Sutskever, I., Vinyals, O., & Le, Q. (2014). Sequence to Sequence Learning with Neural Networks.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Sigut, P., & Waddington, L. (2023). Testing of Detection Tools for AI-Generated Text. *European Network for Academic Integrity.*

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models.

Zellers, R., Holtzman, A. Bisk, Y., Farhadi, A., & Choi, Y. (2019). HellaSwag: Can a Machine Really Finish Your Sentence?. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4791–4800). Association for Computational Linguistics.

Turnitin

Turnitin is a global company dedicated to ensuring the integrity of education and meaningfully improving learning outcomes. For more than 20 years, Turnitin has partnered with educational institutions to promote honesty, consistency, and fairness across all subject areas and assessment types. Our products are used by educational institutions and certification and licensing programs to uphold integrity and increase learning performance, and by students and professionals to do their best, original work.

turnitin™

www.turnitin.com