White paper

Turnitin’s AI Writing Detection Model Architecture and Testing Protocol

Turnitin AI Technical Staff

This white paper presents the Turnitin AI writing detection system, focusing on its architecture and its testing protocol, recent enhancements to the core AI writing detection model, and a new AI paraphrase detection model. This white paper 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 shown by independent researchers to be highly effective in correctly identifying AI-generated content. Finally, potential future developments are presented and discussed.

www.turnitin.com

Updated: August 2024

Contents

Introduction and Prior Work in Generative AI Language Models

Writing is 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 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 et al.).

Overview of AI Writing Detection Research

State of the art LLMs consistently perform at or near the levels of human performance 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 presents 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.

Identifying 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

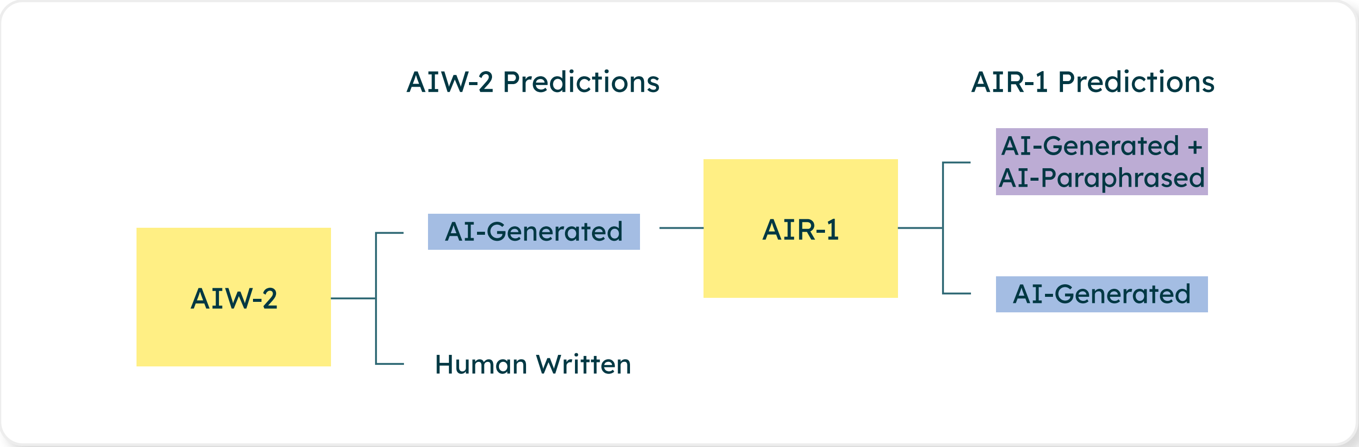
Turnitin is a leading provider of academic integrity tools worldwide. These tools are built into popular learning management system (LMS) workflows across 16,000 institutions, 185 countries and used by more than 70 million students. In April 2023, Turnitin launched its AI writing detection tool, powered by the AIW-1 (AIW stands for AI writing detection) model, which as of June 2024 has processed more than 250 million paper submissions.

In December 2023, Turnitin launched AIW-2, an updated and improved model to replace the AIW-1 model. Compared to AIW-1, AIW-2 improves LLM detection performance, especially detecting AI generated text that has been modified by AI paraphrasing tools (otherwise known as “text spinners”), while improving on AIW-1’s below 1% false positive rate for documents with 20% or more predicted AI text. We discuss these improvements in the Testing and Evaluation Protocol section of this white paper.

In July 2024, Turnitin launched the AI paraphrase highlighting feature in the AI writing report. This feature is powered by the AIR-1 model (AIR stands for AI rewriting detection). AIR-1 is designed to find the statistical signatures of AI rewriting and AI paraphrasing models. These models use a different generative architecture framework from a standard LLM, and therefore leave a distinct and detectable signature.

The system architecture for the Turnitin AI writing detection system is shown in Figure 1, showing how the AIR-1 is integrated into the detection process. When a submission is made, the AIW-2 model determines whether the text is human-written or AI-generated. If the document is predicted to have 20% or more AI-generated text, the AIR-1 model tags sentences predicted to be AI-generated as also being “AI paraphrased” if the statistical signature of an AI paraphraser is detected. In the AI writing report, sentences that are predicted to be both AI-generated and AI-paraphrased are highlighted as purple. Note that the AIR-1 model does not make any predictions for documents for which the AIW-2 model has predicted to be less than 20% AI-generated writing. Additionally, in documents in which the AIW-2 model predicts at least 20% of the text to be AI-generated, the AIR-1 model does not make AI-paraphrasing predictions for sentences the AIW-2 model predicts to be human-written. We discuss the overall system level statistics and performance in the Testing and Evaluation Protocol section of this white paper.

Figure 1. Turnitin AI Writing Detection System with both AIW-2 and AIR-1 models



The Deep Learning Model Framework of AIW-2 and AIR-1

The AIW-2 and AIR-1 models in Turnitin's AI writing detection system are both built around a state-of-the-art transformer deep-learning architecture. AIW-2 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.

A key difference between the AIW-2 and AIW-1 training datasets is the inclusion of "AI+AI paraphrased" text. This was a key development that significantly improved the AIW-2 ability to detect LLM-generated text, even after it had been manipulated by AI paraphrasers.

The AI-generated text was created by Turnitin to mirror text known to be human-written. 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 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. Complex "human-written+AI-paraphrased" and "AI-generated+AI-paraphrased" texts were also created to train and evaluate the AIR-1 model, further expanding the dataset. The complete testing and held-out evaluation datasets included a rich mixture of purely human, purely AI-generated, human-written+AI-paraphrased, AI-generated+AI-paraphrased, and various mixed AI-generated/human-written/AI-paraphrased 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. In multiple peer-reviewed studies, Turntin's AI writing detection system using the AIW-1 model compares favorably to other AI writing detection systems (Weber-Wulff et al., 2023; Walters, 2023; Liu et al., 2024), and the AIW-2 model continues to build upon this foundation with improved performance. Additionally, the new AIR-1 model exhibits high performance in identifying AI paraphrased text, as measured by two misidentification metrics. These concepts, metrics and evaluation results are defined and discussed in detail in the next section.

Turnitin’s transformer model architecture 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 striding 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 is AI-generated.

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 making a prediction of “AI-generated” (or in the case of AIR-1 “AI-paraphrased”) on a sentence varies depending on the specific transformer model and its final tuning.

A document is labeled as “AI-generated” if more than 20% of the sentence-level AIW-2 prediction scores are above a sentence-level AI writing threshold described in the previous paragraph. Based on tests conducted by Turnitin, it was determined that in cases where the system detects 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%). To maintain prediction stability, Turnitin’s AI writing detection system has a minimum document length limit of 300 words for the document to be processed.

The sentence aggregation logic for the AIR-1 model is identical to that of the AIW-2 model. However, the AIR-1 model does not have document-level aggregation logic since it has no impact on whether a document is labeled “AI-generated” within the product. The full aggregation logic of the AIW-2 and AIR-1 models was discussed in the previous section and shown in Figure 1.

AIW-2 Testing and Evaluation Definitions and Protocol

The AIW-2 and AIR-1 models that power Turnitin's AI writing detection system are tested using multiple datasets. Turnitin uses two main metrics to test AIW-2: 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-generated documents are "recalled" or correctly labeled by the AI writing detection system as being "AI-generated". If the AI writing detection system in this example correctly labels 30 of the 40 AI-generated documents, then the recall is 30/40 = 0.75 or 75%.

FPR measures system reliability. 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-generated". If the AI writing detection system in this example flagged 3 of the 60 human-written documents as "AI-generated", 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.

AIW-2 Testing and Evaluation Results

To measure FPR, Turnitin conducted a stress-test using over 700,000 papers submitted before 2019 and therefore pre-dating GPT-3. All papers in this dataset are human written. AIW-2 and its attendant heuristics was run on this dataset, and achieved a document-level FPR of 0.5% (compared to 0.7% for AIW-1, see Table 1) and a sentence-level FPR of 0.33% (compared to 0.42% for AIW-1, see Table 2). Our findings are further supported by multiple comparisons of popular AI writing detection solutions on the market, where Turnitin's AI writing detection system demonstrated zero false accusations (Weber-Wulff et al., 2023; Walters, 2023; Liu et al., 2024).

Recall was measured on a 2,970-document held-out evaluation dataset (See Table 3). The dataset comprises a mix of documents that are purely human, purely AI-generated and a mix of AI-generated and human-written text. 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, AIW-2 achieves a document recall of 91.18% (compared to 89.83% for AIW-1).

Table 1. AIW-1 vs AIW-2 Document-Level FPR

	Document Count (Human only)	AIW-1 Document FPR	AIW-2 Document FPR
Pre-2019 Student Writing	719,877	0.70%	0.51%

Table 2. AIW-1 vs AIW-2 Sentence-Level FPR

	Document Count (Human only)	AIW-1 Sentence FPR	AIW-2 Sentence FPR
Pre-2019 Student Writing	719,877	0.42%	0.33%

A key goal of AIW-2 is to improve its ability to detect AI writing even when it has been masked or modified by AI paraphrasers. On a dataset of 1,768 AI-generated documents (see Table 4) that have also been AI-paraphrased, AIW-1 achieved a document-level recall of 51.7%, while AIW-2 achieved a recall of 78.34%, an improvement of 26.64%.

Table 3. AIW-1 vs AIW-2 Document-Level Recall

	Document Count (AI and Mixed AI Human)	AIW-1 Document Recall	AIW-2 Document Recall
Standard Recall Dataset	2,970	89.83%	91.18%
AI-generated and AI-paraphrased documents	1,768	51.7%	78.34%

Table 4. AIW-1 vs AIW-2 Sentence-Level Recall

	Document Count (AI and Mixed AI Human)	AIW-1 Sentence Recall	AIW-2 Sentence Recall
Standard Recall Dataset	2,970	91.22%	95.06%
AI-generated and AI-paraphrased documents	1,768	45.56%	85.22%

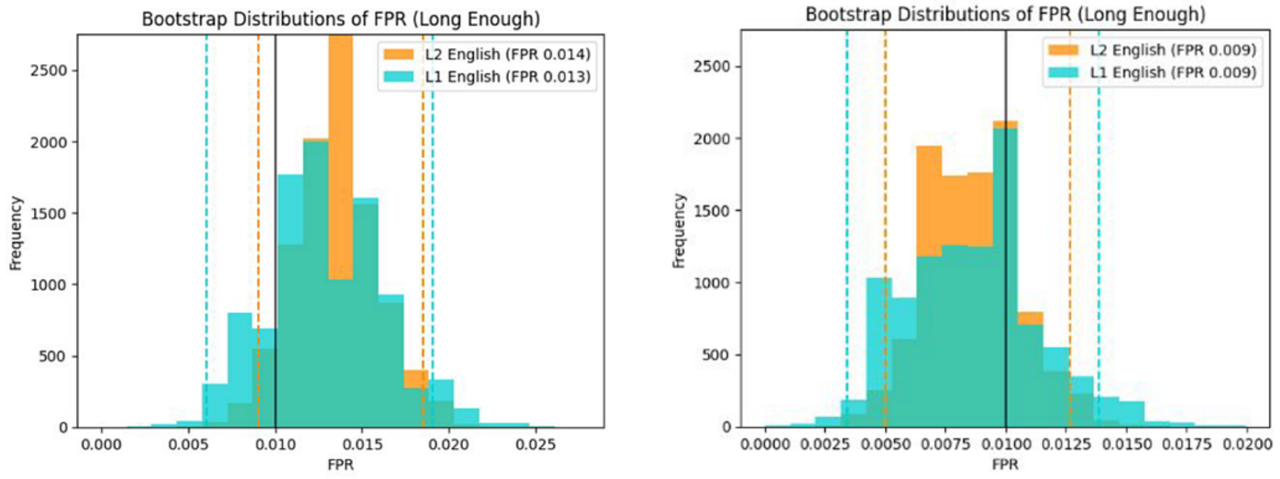
Sentence-level recall was measured on the same 2,970 document dataset. Here, AIW-2 achieved a sentence-level recall of 95.06% (compared to 91.22% for AIW-1). On the AI-generated and AI-paraphrased dataset mentioned above, AIW-1 achieved a sentence-level recall of 45.56% while AIW-2 achieves 85.22%. These numbers show that the Turnitin AI writing detection system is effective at identifying AI writing for a diversified and complex dataset.

English Language Learners (ELL) Bias Evaluation Results for AIW-2

Studies on potential bias by AIW-2 against ELL writers were conducted on an approximately 9,000 document dataset by combining the ASAP, ICNALE and PELIC datasets (Shermis et al. 2014; Ishikawa et al., 2023; Juffs et al., 2020). The goal of these tests was to validate that AIW-2 displays no tendency to over predict AI writing or AI paraphrasing for "L2" English writers (English language learners) compared to "L1" English writers (native English speakers). All data in these datasets are human written. This test was first conducted for the AIW-1 model (Adamson, 2023) and subsequently was repeated prior to the AIW-2 launch.

AIW-2's L2 English FPR is 0.86% (compared to 1.35% for AIW-1), and L1 English FPR is 0.87% (compared to 1.30% for AIW-1). Using bootstrap resampling, both AIW-1 and AIW-2 exhibit no statistically significant differences in FPR behavior between L1 and L2 English writers as shown by the highly overlapping distributions for L1 and L2 English FPR on this dataset (see Figure 2).

Figure 2. Bootstrap Sampling Distributions for FPR on L1 (green) and L2 (Orange) English Documents. AIW-1 is shown in Fig. 2(a) and AIW-2 is shown in Fig. 2(b).



(a)

(b)

Improved AIW-2 LLM Detection Efficacy

AIW-2 is designed to detect AI writing even when it has been modified by masking by AI paraphrasers. A byproduct of this improvement is that AIW-2 is able to detect more LLM signatures than AIW-1, which predominantly detects the signature of GPT-3.5.

Turnitin ran tests to determine the level of performance of AIW-2 compared to AIW-1 on GPT-4, Llama-2, and Gemini Pro 1.0. All data was generated using our automated complex prompt chain system. At the time of writing this white paper, AIW-1 has not been tested on Gemini-Pro 1.0. Additionally, Gemini-Pro 1.0 has multiple settings for text generation, so we chose to report our lowest (worst performing) recall value. Table 5 below shows the performance improvement of AIW-2 over AIW-1.

Table 5. AIW-2 Detection Efficacy of GPT-4, Llama-2 and Gemini-Pro 1.0

	Document Count	AIW-1 Recall	AIW-2 Recall
GPT-4	7,610	39.7%	76.8%
Llama-2	6,649	99.4%	99.2%
Gemini-Pro 1.0	3,550	N/A	87.25%

In summary, AIW-2 demonstrates consistent performance on detecting GPT-4, Llama-2 and Gemini Pro 1.0, even though by design it is still predominantly trained to detect writing outputs from GPT-3.5. We attribute this improvement to the addition of the AI-paraphrased LLM-generated data, which increases the complexity and covered LLM footprint of the training dataset.

AIR-1 Testing and Evaluation Definitions and Protocol

For AIR-1, the primary metrics are misidentification “Type 1” and misidentification “Type 2” errors. The AIR-1 model does not impact document recall or FPR since it operates at the sentence level and only on sentences that have already been identified as AI-generated.

Misidentification Type 1 measures the rate at which AIR-1 labels AI-generated sentences that are not AI-paraphrased as AI-generated and AI-paraphrased. For example, consider a set of 100 sentences that are known to be AI-generated but not AI-paraphrased. If the AIR-1 model labels 5 of these sentences as AI-generated and AI-paraphrased, the misidentification Type 1 rate is 5/100 = 0.05 or 5%.

Misidentification Type 2 measures the rate at which AIR-1 labels AI-generated and AI-paraphrased sentences as AI-generated only. For example, consider a set of 100 sentences that are known to be AI-generated and AI-paraphrased. If the AIR-1 model labels 10 of these sentences as AI-generated but not AI-paraphrased, the misidentification Type 2 rate is 10/100 = 0.1 or 10%.

In tuning AIR-1, we prioritized reducing misidentification Type 1 over misidentification Type 2 due to the potential that Type 1 might be interpreted as an indicator of academic misconduct.

AIR-1 Testing and Evaluation Results

AIR-1 is used to determine whether sentences identified as AI-generated should be further classified as being also AI-paraphrased. AIR-1's performance is measured through Type 1 and Type 2 misidentification rates, which are defined in the section "Testing and Evaluation Definitions and Protocol".

The counts for the various types of documents in the AIR-1 evaluation dataset is shown below in Table 6. This dataset consists of documents ranging from fully human written with no AI paraphrased text to documents that are fully AI-generated and AI-paraphrased. We also consider some complex mixed AI-generated and AI-paraphrased cases.

Table 6. AIR-1 Evaluation Dataset Composition

		Document Counts		
		Document AI-Paraphrased?		
		None	Fully	Mixed
Document AI-Generated?	None	762	748	975
	Fully	798	796	977
	Mixed	972	972	-

Table 7 shows the results of the AIR-1 sentence-level evaluation. When a sentence is AI generated and AI paraphrased, the combined AIW-2/AIR-1 system correctly identifies the sentence 81.68% of the time. The system labels the sentence purely human 9.31% of the time and 9.01% of the time the system labels the sentence AI-generated but not AI-paraphrased.

When a sentence is AI-generated and not AI-paraphrased, the combined AIW-2/AIR-1 system (see Figure 1) correctly identifies the sentence 86.25% of the time. The system labels the sentence purely human in origin 10.84% of the time, and 2.91% of the time the system labels the sentence AI-generated and AI-paraphrased.

Table 7. Sentence-Level Performance of the AIW-2 + AIR-1 Models on the AI-Generated and AI-Generated/AI-Paraphrased dataset.

		AIW-2 + AIR-1 System Prediction		
		Human	AI-Generated and NOT AI-Paraphrased	AI-Generated and AI-Paraphrased
Ground Truth	AI-Generated and NOT AI-Paraphrased	10.84%	86.25%	2.91% (Misidentification Type 1)
	AI-Generated and AI-Paraphrased	9.31%	9.01% (Misidentification Type 2)	81.68%

Note that AIR-1 cannot add a label to a sentence that is marked by AIW-2 as human, and operates only on sentences that AIW-2 has marked as AI-generated.

Summary and Future Work

The explosion in popularity of LLM tools 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 potential 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 reliable and effective AI writing detection tool that has been trained and tested on a large collection of human-generated academic writing.

Turnitin has continued to improve its AI writing detection system, most recently with the launch of AIW-2 and AIR-1. These models improve the AI writing detection system's ability to detect likely AI-generated text that may have been AI-paraphrased, as well as further improving the system's ability to detect text from additional LLMs, including GPT-4, Llama-2 and Gemini Pro 1.0, with a low false positive rate. These changes represent ongoing research and development to continuously improve the performance of the AI writing detection system and demonstrate Turnitin's commitment to give educators state of the art solutions to navigate academic integrity in the age of AI.

References

Adamson, D. (2023). New research: Turnitin's AI detector shows no statistically significant bias against English Language Learners [Blog]. https://www.turnitin.com/blog/new-research-turnitin-s-ai-detector-shows-no-statistically-significant-bias-against-english-language-learners

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.

Ishikawa, S. (2023). The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English (Routledge).

Juffs, A., Han, N-R., & Naismith, B. (2020). The University of Pittsburgh English Language Corpus (PELIC) [Data set]. http://doi.org/10.5281/zenodo.3991977 This dataset was funded with a grant via the Pittsburgh Science of Learning Center, award number SBE-0836012. (Previously NSF award number SBE-0354420.)

Liu, J.Q.J., Hui, K.T.K., Al Zoubi, F. et al. (2024). The great detectives: humans versus AI detectors in catching large language model-generated medical writing. *Int J Educ Integr* 20, 8. https://doi.org/10.1007/s40979-024-00155-6

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

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

Shermis, M. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a united states demonstration. Assessing Writing, 20:53–76.

Sutskever, I., Vinyals, O., & Le, Q. (2014). Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*.

Vaswani, A., Shazeer, N., Parmar, N. et al.(2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*.

Walters, W. H. (2023) "The Effectiveness of Software Designed to Detect AI-Generated Writing: A Comparison of 16 AI Text Detectors" *Open Information Science*, vol. 7, no. 1, pp. 20220158. https://doi.org/10.1515/opis-2022-0158

Weber-Wulff, D., Anohina-Naumeca A., Bjelobaba S. et al.(2023). Testing of Detection Tools for AI-Generated Text. In *European Network for Academic Integrity*.

Wei, J., Tay, Y., Bommasani, R., et al. (2022). Emergent Abilities of Large Language Models.

Zellers, R., Holtzman, A., Bisk, et al. (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.

Non-English AIW Detector Addendum

This addendum clarifies how Turnitin builds the models for non-English AIW detectors. Non-English AIW detectors are built using the same general transformer architecture as AIW-2. Currently, Turnitin does not offer a non-English equivalent of AIR-1.

Compared to AIW-2, building a non-English AIW model typically involves a new base transformer model, sentence splitter and a new tokenizer tuned for the specific base model, as per industry best practices. Additionally, new training and evaluation datasets are created and vetted with the help of in-house and external speakers of the target language.

When a document is submitted to Turnitin and is intended for AIW detection, a language classifier selects the appropriate AIW model to use. For documents with multiple languages, the language with the largest proportion of text within the document will determine the language. The minimum document length to be processed by an AIW Detector may vary based on language, but is in general 300 words unless otherwise noted. The AIW indicator will only display a number if more than 20% of the qualified content in the document is predicted as likely AI generated.

Turnitin plans to only release non-English AI detectors that exhibit a document false positive rate of below 1% on pre-2019 human-written documents in the target language, while also achieving a sufficiently high document recall rate on challenging test datasets of documents that are AI-generated or contain a mix of human-written and AI-generated text.

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