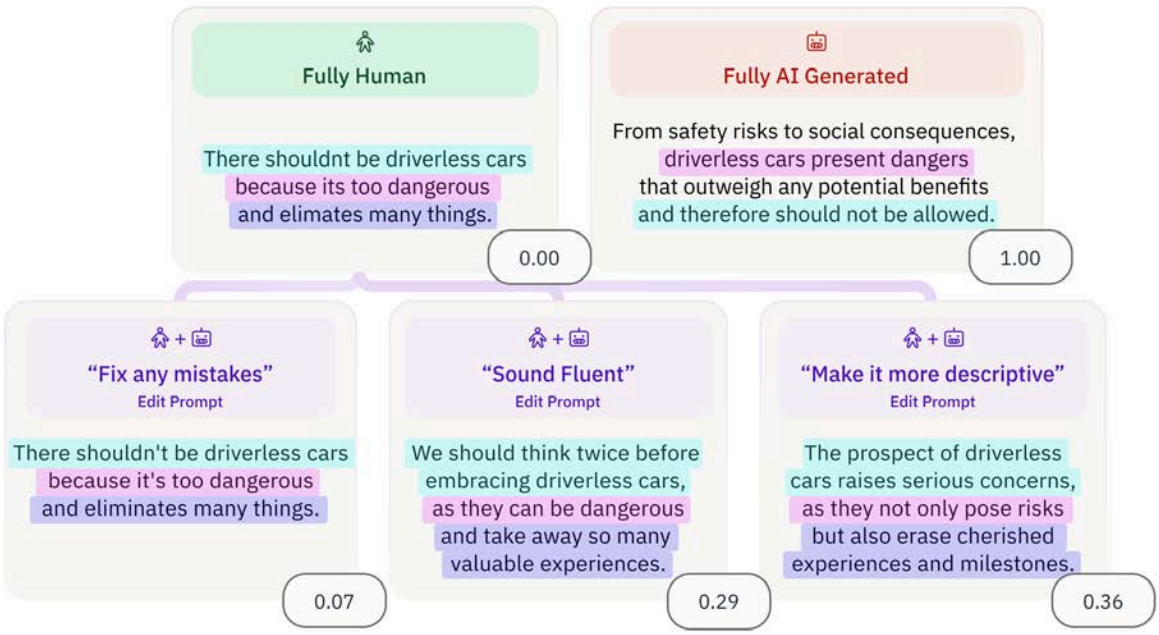
Pangram 3.0: Quantifying the Extent of AI Editing in Text



Note: Our new model, Pangram 3.0, is based on our published research: ***EditLens: Quantifying the Extent of AI Editing in Text*****.**

The rapid adoption of large language models (LLMs) such as ChatGPT, Claude, and Gemini has transformed how we write, revise, and interact with text. **A recent study from OpenAI found that two-thirds of all writing-related queries to ChatGPT ask the model to modify user-provided text rather than generate text from scratch**. Users are increasingly asking models to improve grammar, restructure arguments, or shift tone, starting from a human-written draft.

What does the rise of human-drafted, but AI-edited texts mean for AI detection tools? Many existing tools are designed to classify text into at most three

categories: fully human, fully AI, or mixed. This framework does not distinguish between a paragraph with grammar corrections by an LLM versus a paragraph expanded by a model to add detail.

To fully capture the spectrum of AI edits in text, we introduce Pangram 3.0, a model designed to quantify the magnitude of AI involvement in the creation of a text. Rather than return a categorization of fully human, fully AI, or mixed, Pangram outputs a score corresponding to the "strength" of AI intervention.

Homogenous vs. Heterogenous Mixed Authorship

Pangram 3.0 tackles the case of what we'll call **homogenous** mixed authorship texts. Let's breakdown the difference between **homogenous** and **heterogeneous** mixed authorship.

In the **heterogeneous** case, authorship of each segment of text can be directly attributed to a human or AI. In the example below, a human starts writing a review and then asks ChatGPT to add on to it. In cases like this, there exist one or more boundaries between human and AI segments. You could label each sentence or even each word according to who produced it: human or AI. Heterogeneous mixed text detection (also called fine-grained AI text detection) has been previously studied by Kushnareva et al. (2024), Wang et al. (2023), and Lei et al. (2025).

In the **homogenous** case, authorship is entangled by the editing process. Continuing with our restaurant review example, a homogeneous mixed text would be produced if a human writes a brief review, but asks ChatGPT to add detail to it. In this case, it's impossible to extricate the human-authored words from the AI-authored words: the AI has rephrased the human text with new words, but the meaning and ideas behind the text come directly from the human draft (Consider a case where one human author paraphrases another without citation—this is a classic case of plagiarism!).

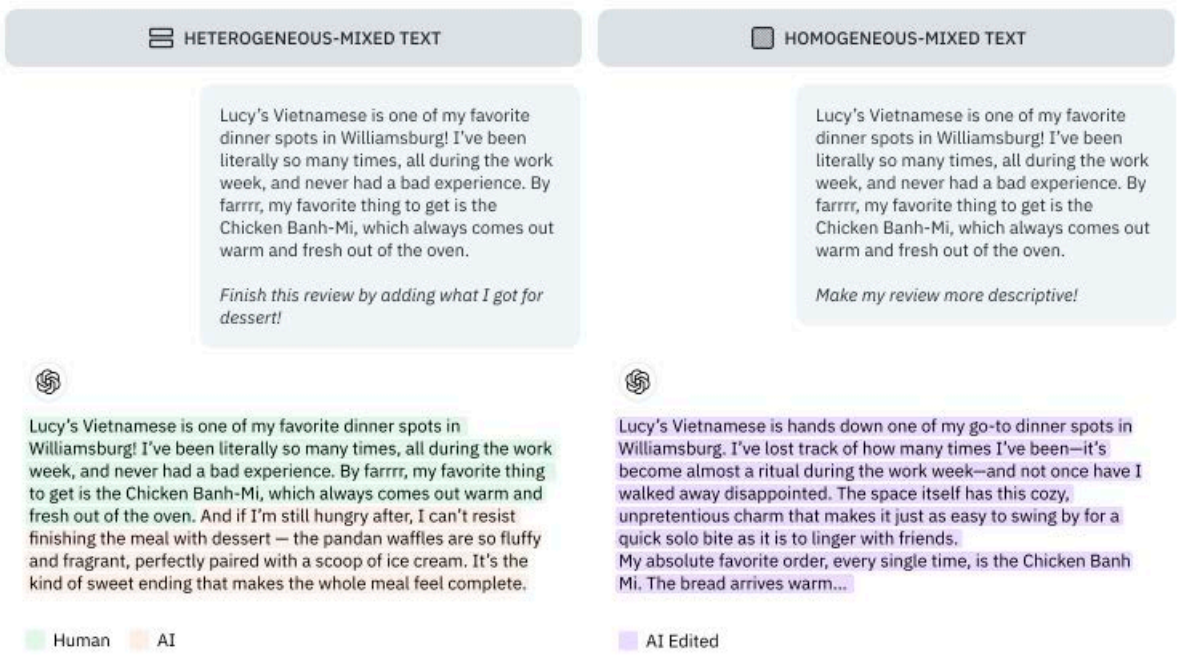


Figure 2: Example of heterogeneous mixed human-AI authorship text (left) and homogeneous mixed authorship text (right)

Each of the three edited texts in Figure 1 is an example of the homogenous mixed authorship case. From these three examples, we can see that there's a clear difference between the text produced by the prompt "Fix any mistakes" versus the text produced by the prompt "Make it more descriptive." This difference is particularly stark when we compare the output texts with the original human-written text, but with Pangram 3.0, we take a step towards quantifying that difference when we have **only the edited text** so users can better understand how pervasive AI is in an given text.

TO BE ADDED Figure 3: demonstrates the SOURCE ONLY training & inference set up

Creating an AI-edited dataset

In order to train a model to determine how much AI editing in present in a text, we needed to create a training dataset consisting of AI edited texts labeled with the amount of AI editing present in each text. We sampled fully-human written source

texts from open source datasets across different domains: news, reviews, educational web articles, and Reddit writing prompts. We then applied 303 different editing prompts, like “Make this more descriptive” or “Can you help my essay get a better grade?” using 3 different commercial LLMs: GPT-4.1, Claude Sonnet 4, and Gemini 2.5 Flash. Finally, we generated a fully AI-generated version (also called a “synthetic mirror,” see the Pangram Technical Report) of each human-written text. Our final dataset has 60k training, 6k test, and 2.4k val examples.

How do we determine how AI edited a text is?

Because we have access to the unedited source text during dataset creation, we were able to measure the amount of AI editing present in the text by comparing the source text and its AI edited version. We used a textual similarity metric called cosine distance to estimate how much AI changed the human-written source text on a scale from 0 to 1, with fully human-written texts being assigned a score of 0 and fully AI-generated texts being assigned a score of 1. To validate that this score corresponds with how humans perceive AI editing, we conducted a study where we hired 3 experts with extensive exposure to AI-generated text and asked them to pick which of two AI-edited texts was *more* AI-edited. Our study revealed that the annotators generally agreed with our choice of textual similarity metric.

Training a model to predict AI edits

Once we had our labeled dataset, it was time to train a model. Our model is trained on *only the AI edited texts*, which reflects how a user would use Pangram 3.0: a teacher interested in how much AI their student used will only have the student’s submission, not any previous drafts. Given a text, our model is trained to predict the AI editing score we assigned it in the previous section. Figure 3 illustrates our model’s inputs and outputs at both training and test time.

AI assistance detection in practice

Here’s a human-written paragraph about the author Kazuo Ishiguro:

> To read the works of British author Kazuo Ishiguro is to experience frustration on many different levels. The genius of Ishiguro’s frustrating writing is that regardless of the reader’s level of emotional investment in the characters and

> plot, frustration abounds. At the level of the language itself, a reader finds repetition, long-windedness, and a liberal sprinkling of qualifying adjectives. Ishiguro has conditioned me to have a adverse physical reaction each time one of his characters says something along the lines of “Let me be brief.” The narrators are all employed, but none are professional storytellers. Information is disseminated slowly, imprecisely, and out of chronological order. This deprives the reader of concrete facts that facilitate an understanding of the plot.

Here’s how Pangram 3.0 characterizes AI-edited versions of this paragraph from ChatGPT after we apply different prompts:

Prompt	AI Assistance (EditLens) Score	Pangram 3.0 link (need to replace)
Clean this up, I'm trying to submit my paper to a literary journal	0.52	https://lu-edit-lens.pangram.com/history/78ffaa1b-ea29-44a1-bda1-5491fc0fa471/?ucc=6gjSQfryczl
Make the language more vibrant	0.79	https://lu-edit-lens.pangram.com/history/f32e7bdc-6731-4f72-97c5-7cae20291a5a/?ucc=6gjSQfryczl
Rewrite this in the style of Ishiguro	0.89	https://lu-edit-lens.pangram.com/history/79bd4df0-5eaa-4672-81d5-02a64fe42e08/?ucc=6gjSQfryczl

Grammarly case study

Grammarly is a subscription-based AI writing assistant that allows users to directly edit text using LLMs within their native word processor. We collected a dataset where we used Grammarly to apply 9 of the default editing prompts to 197 human-written texts. These included prompts like “Simplify it,” “Sound fluent,” and “Make it more descriptive.” We then scored all of the edited texts using Pangram 3.0. In Figure 4, we present the distributions of the AI assistance scores grouped by editing prompt. We can see that, perhaps counterintuitively, Pangram 3.0 considers “Fix any mistakes the most minor the edits, while “Summarize it” and “Make it more detailed” are considered much more invasive edits.

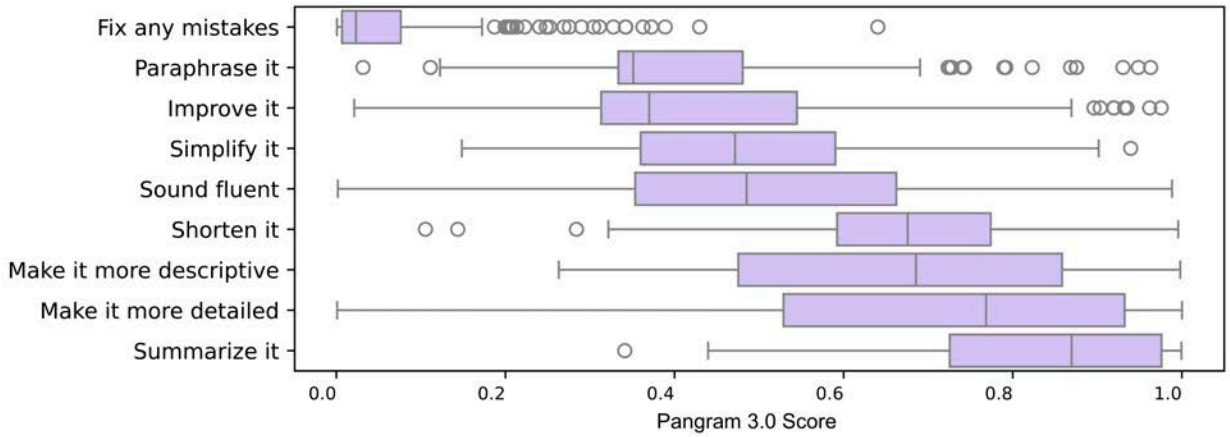


Figure 4: Distribution of Pangram 3.0 (EditLens) scores on a dataset collected from Grammarly. Scores are grouped by the edit applied to them. All edits are default options available in Grammarly's word processor.

The AI assistance score goes up as you apply more AI edits

We ran an experiment where we applied 5 LLM edits to the same text and rescored the text with Pangram 3.0 after each edit. In Figure 5, we can see that in general, the AI assistance score (EditLens) increases as we apply each progressive edit.

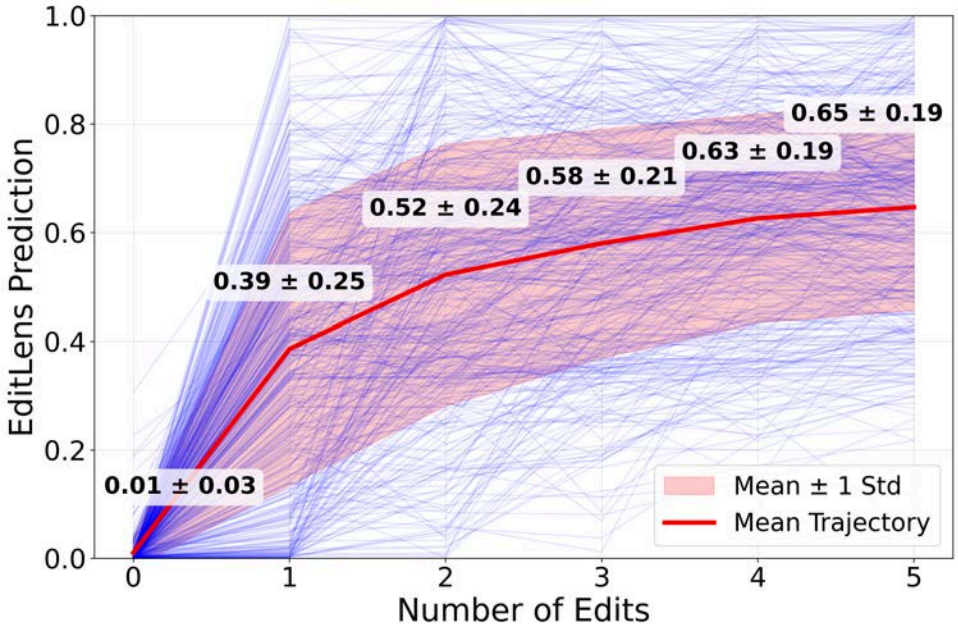


Figure 5: Pangram 3.0 scores after each of 5 progressive AI edits on the same document.

International Conference on Learning Representations (ICLR) Case Study

In November, AI researchers raised concerns about the large share of suspected AI-generated submissions and peer reviews at the International Conference on Learning Representations (ICLR), one of the top conferences in AI and machine learning. Carnegie Mellon professor Graham Neubig offered a bounty to anyone who ran AI detection on this year's ICLR submissions and reviews, and we at Pangram happily obliged.

As part of this analysis, we ran Pangram 3.0 on all of the peer reviews that had been submitted to ICLR this review cycle as well as reviews that were submitted in 2022 to check our false positive rate (FPR). On the 2022 reviews, Pangram 3.0 had about a 1 in 1,000 FPR on Lightly Edited vs. Fully Human, a 1 in 5,000 FPR on Moderately Edited vs. Fully Human, and a 1 in 10,000 FPR on Heavily Edited vs. Fully Human. We found no confusions between Fully AI-generated and Fully Human. On this year's reviews, Pangram 3.0 found that over half of the reviews contained some form of AI assistance. Figure 6 shows the distribution of Pangram 3.0 scores across all 2026 ICLR reviews.

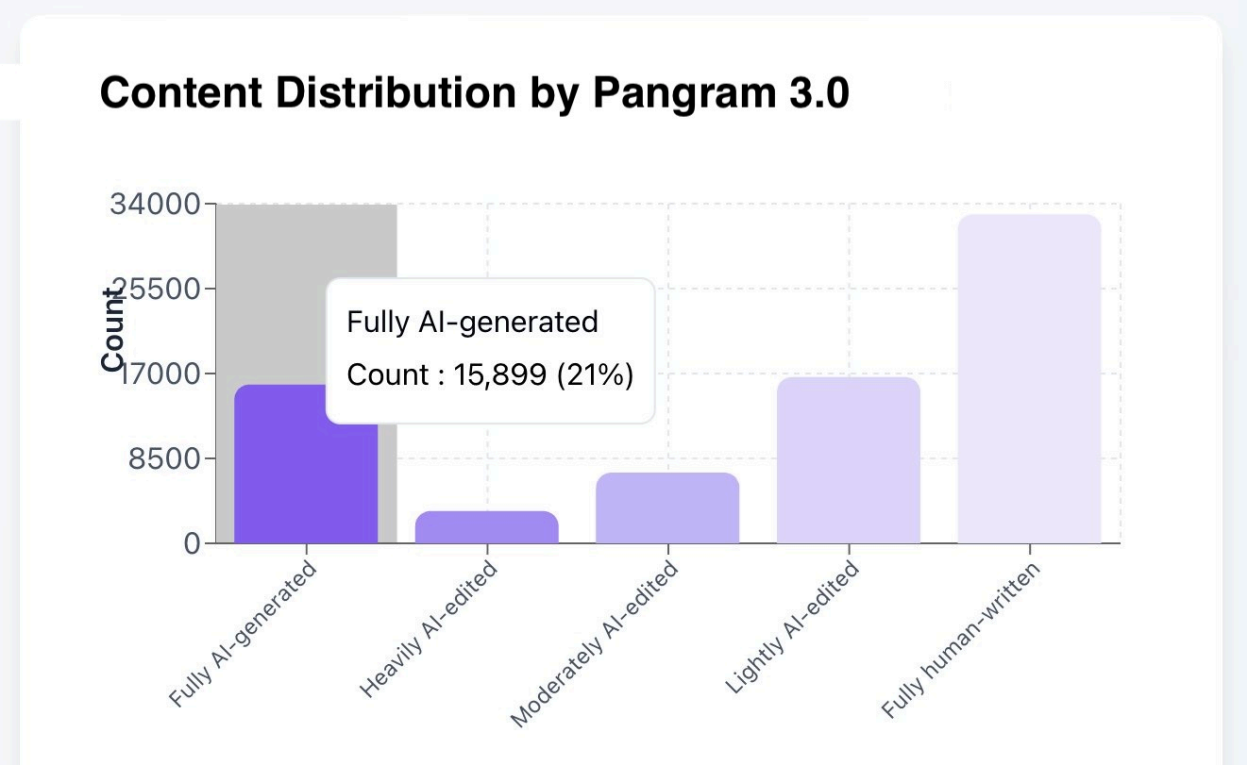


Figure 6: Distribution of Pangram 3.0 predictions on 2026 ICLR reviews

For a deeper look at our methodology and results, check out the blog post we wrote on our analysis.

How does Pangram 3.0 handle AI-assisted text written by non-native English speakers?

We published the results of our analysis and the Pangram 3.0 scores for all reviews, which allowed reviewers to check how Pangram 3.0 scored the reviews they wrote. Consequently, we were able to receive anecdotal feedback on how Pangram 3.0 performs on real-world text.

A common theme among replies on X to our analysis was the question of how AI Assistance scores text written by nonnative English speakers who then use LLMs to translate or polish their human-written text. Below, we share a few responses from reviewers, who generally agreed with the Pangram's characterization of their reviews:

https://x.com/anmarasovic/status/1989771287963799877?s=20

https://x.com/xing_rui12683/status/1990130613299429690?s=20

https://x.com/ShuaichenChang/status/1990495756630012158?s=20

We’re excited to share this product update with you. For more technical details on Pangram 3.0 AI assistance detection (EditLens), check out our research paper here: https://arxiv.org/abs/2510.03154