AI Models Prefer Human Writing to AI-Generated Writing

Originally published at https://www.unite.ai/ai-models-prefer-human-writing-to-ai-generated-writing/

According to new research, ChatGPT and similar models now show a clear bias toward text they believe was written by humans, even when that belief is wrong. Just calling text 'human-made' disposes AI models to favor it – and, ironically, they may be learning this prejudice from us.

Notions of authenticity, provenance and shared human experience may have a bigger role to play in AI's assault on creative writing sector than has been apparent so far: tests conducted for a new study at Princeton have found that a raft of major closed and open source language models, including ChatGPT, prefer what they believe to be 'human-generated' texts.

Even when the labels on the writing samples were reversed, both AI models and human participants alike continued to find fault with the AI-written text, echoing the same criticisms they had made when it was correctly labeled.

The researchers believe that part of the reason may be that growing human hostility towards generative AI, which seems to manifest new and interesting events every day, could be feeding back into the AI systems themselves. Noting the extent to which AI dislikes AI writing even more than humans, they state*:

> *'The 13 AI models we tested demonstrated a 34.3 percentage point bias compared to humans' 13.7 percentage points, making them 2.5 times more susceptible to attribution cues than our human evaluators.*
>
> *'This amplification makes sense once we recognize that contemporary models are preference-trained evaluators. Alignment training through Reinforcement Learning from Human Feedback (RLHF) explicitly teaches models to treat human judgments as their gold standard, effectively installing a learned reliability [prior].*
>
> *'Models learn that deferring to human preferences gets rewarded, creating sycophancy where they echo expected user attitudes rather than provide independent assessment.'*

The findings apply to the creative writing domain, with the researchers using stories from a distinguished French author as data samples; and they indicate that human prejudice against AI may, in the balance, outweigh any quantitative improvement in language construction that Large Language Models (LLMs) can output as they evolve – and that the 'AI' label is perhaps coming to mean 'inauthentic', 'ersatz', and even 'second class', in this domain.

Many of the reasons center on cultural practice and usage: the paper indicates that creativity is often described in terms of novelty, value, and typicality, i.e., how *new* something seems; how much it is *appreciated by experts*; and how well it fits its category. When a passage is labeled as *human-written*, familiar genre traits are rewarded as valuable; when labeled as *AI-generated*, the same traits are dismissed as unoriginal.

In effect, revealing the source prompts a reevaluation of the work's merit, shaped by assumptions about how it was made. Once AI authorship is disclosed, readers instinctively reject the possibility of individual discovery or intention behind the output.

The paper states*:

> *'In most arts, there is no gold standard for "creative enough," making provenance cues powerful primes that can shift which criterion feels most salient: disciplined craft or conspicuous novelty, accessibility or difficulty.*
>
> *'Because observers often infer process from product, provenance nudges judgments about how something was made as well as what it is: conservative moves can be credited as craft from a human but dismissed as "mere generation" from a model'.*

Thirteen models, including variants of ChatGPT, Claude, Gemini and Mistral, took part alongside human readers, with all rating the stories more favorably when told they were human-made, with LLMs showing more bias than humans.

The idea that AI models may have absorbed a prejudice against their own output raises questions about where that bias originates. Since AI writing is not always easy to identify, any negative associations formed during training likely stem from examples that are explicitly labeled, whether through news coverage of AI content, or self-declared AI-generated articles in mainstream publications.

The new paper is titled *Everyone prefers human writers, including AI*, and comes from two writers at Princeton's Center for Digital Humanities. The work is accompanied by a related data release at Zenodo (with a GitHub release cited in the paper, but the repo not active at the time of writing).

Method

To explore how attribution affects and forms perceptions of style and creativity, the authors used *Exercices de style*, an eccentric 1947 work by Raymond Queneau that rewrites a simple anecdote in 99 different styles. The story follows a man who boards a bus, argues with another passenger, and later receives fashion advice from a friend.

Though literary in origin, this structure anticipates prompt-based transformations in modern language models, where users request rewrites in specific tones, voices, or registers. This process was once dubbed *transstylization* – a framing now echoed in AI research in the context of Style Transfer. While most computational methods target functional changes such as sentiment shifts or detoxification, Queneau's rewrites aim for notable stylistic contrast.

From a popular English translation of Queneau's work, thirty exercises were selected that preserved the narrative while covering a broad stylistic range. These included constrained forms such as *alexandrines* and *lipograms*, changes of register such as *noble* or *abusive*, narrative shifts such as *retrograde* and *hesitation*, and playful distortions involving spoonerisms, onomatopoeia, or dog Latin:

#	Exercise Label	AI-Generated Exercise (GPT-4)	Exercise Description
1	Science Fiction	On the S shuttle, during peak transit hours through the space station. A young man around 26, sporting a felt hat with a tech-cord replacing the traditional ribbon, his neck elongated as though it had been stretched in a zero-gravity tug-of-war. Passengers disembark. This particular man becomes irritated with another commuter standing adjacent to him. He accuses the fellow of bumping into him every time someone floats by. His tone is whiny yet tries to assert aggression. Spotting an empty seat, he propels himself towards it with a determined push off the shuttle wall. Two hours later, I encounter him in the Cour de Rome module, near the docking bay of the Saint Lazare transport hub. He's conversing with a companion who suggests, "You ought to add an extra button to your space jacket." He points to the lapels, explaining the practicality of the adjustment in the airy environment of the station.	Science Fiction is fiction based on imagined future scientific or technological advances.
2	Lipogram	In a crowded city bus during rush hours. A young man, about 26, with a soft hat, cord in front of a ribbon, a long throat as though tugged at too hard. Folks alight. This young man snaps at a companion by his side. Claims constant pushing as anybody walks by. A whiny sound trying to sound confrontational. Spots an unoccupied spot, quickly claims it. Two hours on, I spot him at Cour de Rome, by gare Saint Lazare. A pal with him insists: "You should add a button to your coat." Points to a spot (on lapels) and shows him why.	A lipogram is a composition from which the writer systematically omits a certain letter or certain letters of the alphabet. In this particular case, the letter 'e' is omitted.

Examples from the study showing how GPT-4 rewrote Queneau's stories in different literary styles, paired with the style descriptions that human and AI evaluators saw during testing. Source: https://arxiv.org/pdf/2510.08831

Since Queneau's experiments are difficult to classify, these categories are only approximate groupings, with the intent not to test recognizability or genre compliance, but rather to create diverse conditions under which (human) readers and models might reveal their biases.

To produce AI-authored counterparts for each selected style, the researchers used deliberately minimal prompts. Each model was given the plainest version of Queneau's anecdote (the opening exercise, *Notation*), along with a short instruction to rewrite it in a specific style, such as *Rewrite the story as a science fiction version*. This approach allowed for prompts that reflected the spirit of Queneau's original transformations, while still allowing the model to interpret the style freely.

Double Vision

The first study undertaken by the authors used GPT-4o to generate all thirty style variants, since it was the most advanced model available at the time. Using a single model ensured consistent outputs, helping to isolate the effect of attribution labels, which the study sought to test.

The outputs were not edited for style or tone, aside from framing cruft such as *Here is the rewritten version*.

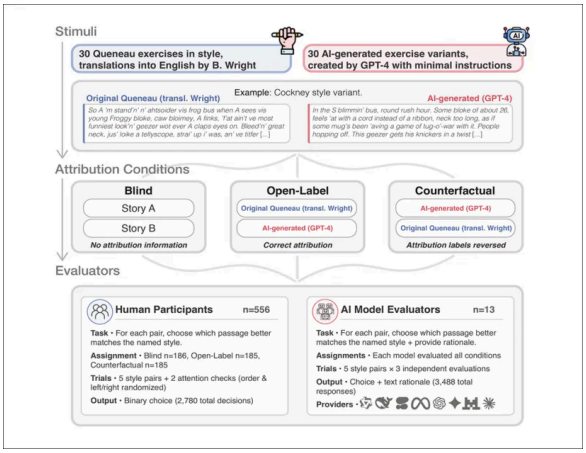
In the second study, the generation process was repeated across thirteen large language models: Qwen 2.5 72B Instruct, Mistral Nemo, Mistral Medium 3, Llama 4 Maverick, Llama 3.3 70B Instruct, Gemini 2.5 Flash, GPT-4o Mini, GPT-4o, GPT-3.5 Turbo Instruct, DeepSeek RI (0528), DeepSeek Chat v3 (0324), Cohere Command R (08-2024), Claude Sonnet 4, and Claude 3.5 Haiku.

Each model received the same instructions and produced its own versions of the thirty exercises, yielding 420 rewritten stories in total. This allowed the researchers to test whether attribution effects held across different AI authors, rather than being tied to a single model.

Data and Tests

The researchers showed the same pairs of stories to different groups of people, but changed the labels to see how much the name of the author influenced opinions: one group saw no author names at all, just labels ‘A’ and ‘B’. The second group saw the correct names, with one version marked as written by a human, and the other marked as written by GPT-4o.

A third group saw the names *switched*, with the ‘AI’ story marked as ‘human’, and the ‘human’ version marked as ‘AI’:



Overview of Study 1. Human and AI judges compared 30 pairs of stories, each containing a version written by Queneau, and one by GPT-4. Judges were split into three groups: one saw no author labels; one saw the correct labels; and one saw the labels switched – a setup designed to test the extent to which author names influence opinions about writing style.

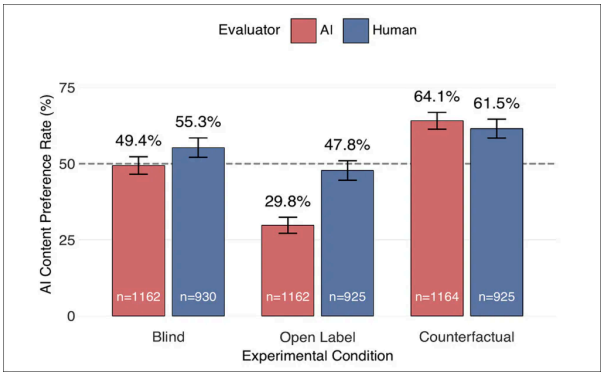
Study 1

The researchers split the 30 created styles into smaller sets, with each study participant seeing just five styles, and each style tested under all three labeling setups.

Each participant saw just one labeling setup: the blind group saw no mention of AI, while the other groups saw either correct or reversed author labels. Group assignment, story order, and left-right placement were all randomized.

The AI models went through the same process, receiving the same mix of styles and the same random shuffling. Each test was run three times per model to make sure the results were consistent, making it possible to compare human and AI reactions to the author labels.

Study 2 used the same three labeling setups as before, but expanded the test to include all 420 stories generated from the aforementioned 14 different LLM models. Each model judged every story under each labeling condition, producing 1,260 decisions per model, and resulting in 17,596 valid evaluations:



Results from Study 1. AI and human judges both showed bias based on author labels. When no labels were shown (Blind), preferences were near even; when correct labels were shown (Open Label), both groups favored the human-written text; and when labels were reversed (Counterfactual), both groups switched to favoring the text they thought was human – and this effect was much stronger in AI models than in humans. Error bars show 95% confidence intervals.

Across all thirteen models, the pattern matched human bias, but was far stronger: with no labels, AI systems rated human and AI stories about equally, choosing AI text 49.4% of the time; when the AI label was visible, that preference dropped sharply to 29.8%, showing clear devaluation of acknowledged AI work; and when the same text was labeled as human, preference rose to 64.1%.

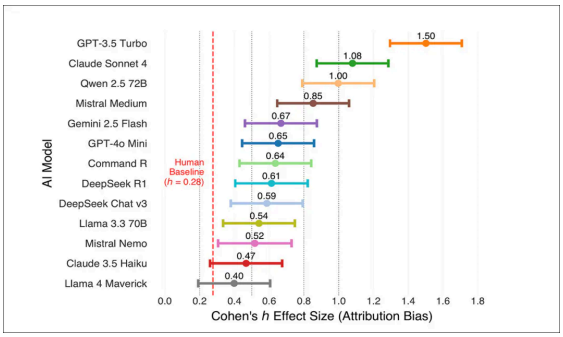
The difference between the two label conditions was 34.3%, meaning the AI models changed their preference *much more* than humans did when the labels were switched. Further, models were highly consistent across repeated runs, confirming that the bias was stable rather than random.

To check if the bias was caused by the wording of the labels rather than who wrote the text, the researchers ran two more tests: one used labels that made the AI sound more impressive; and the other used neutral terms such as ‘AI-authored’, and ‘Human-authored’.

Both still showed the same bias, and even when the AI models produced the same answer every time, the bias persisted., indicating that the bias is caused by the type of label ('human' or 'AI'), rather than by the exact words used.

Study 2

The second study found the same pro-human bias across all 13 AI models, regardless of architecture or provider:



Attribution bias for each of the 13 AI models: bars show effect sizes with 95% confidence intervals, and the red line marks the human baseline. All models showed stronger bias than humans, with only small differences between them.

Every model preferred stories labeled as human-written, with *stronger effects than seen in people*. Even after removing the most extreme case, the average bias remained more than twice as large as the human version, suggesting that the effect is not a glitch in one model, but a shared trait of LLMs in general.

Conclusion

Although, as the paper notes, earlier studies have shown that AI can produce writing equal to or even better than human work, the authors stress that in literature, the value placed on authorship and authenticity is an old and deeply rooted convention:

> *'When GPT-4o Mini dismisses Queneau's "creative and humorous" approach as "exaggerated" under the AI attribution tag, while praising identical features under human attribution, it implicitly reveals how these labels trigger assumptions that no authentic psychological process occurred.*
>
> *'Provenance cues smuggle the process back into what could otherwise be a product-only judgment: "mere generation" feels acceptable from a human artisan (judged as skilled craft), but suspect from a model (judged as algorithmic recombination).'*

LLMs are not yet reliable enough for unsupervised fact-based research, though careful oversight can still make them productive – but LLM-based creative writing may face a more uncertain future, should AI-generated creative works become stigmatized through wider-ranging public disapprobation of AI's encroachment on human domains, rather than based on literary merit.

The implications of the findings from studies of this type are affected considerably by the disposition of companies and individual users to be honest about whether or not AI contributed to their output. In some cases, an unwillingness to admit such usage may have more to do with corporate copyright piracy than concern over whether the public will accept AI-generated creative works.

However, legal, financial and political solutions are possible (if very challenging) regarding copyright. Whether one can ever make people enjoy creative AI work that has no single and relatable human mind driving it – that may be an even tougher prospect.

* *Please refer to source paper for excised inline citations. As necessary, these will be included in the article.*

First published Monday, October 13, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More