Readers Prefer Outputs of AI Trained on Copyrighted Books over Expert Human Writers

Tuhin Chakrabarty[1], Jane C. Ginsburg[2], Paramveer Dhillon[3,4]

[1]Department of Computer Science, Stony Brook University.

[2]Columbia Law School.

[3]School of Information Science, University of Michigan.

[4]MIT Initiative on the Digital Economy.

Corresponding authors: tchakrabarty@cs.stonybrook.edu, ginsburg@law.columbia.edu, dhillonp@umich.edu

The use of copyrighted books for training AI models has led to numerous lawsuits from authors concerned about AI's ability to generate derivative content. Yet it's unclear whether these models can generate high quality literary text while emulating authors' styles/voices. To answer this we conducted a preregistered study comparing MFA-trained expert writers with three frontier AI models: ChatGPT, Claude, and Gemini in writing up to 450-word excerpts emulating 50 award-winning authors' (including Nobel laureates, Booker Prize winners, and young emerging National Book Award finalists) diverse styles. In blind pairwise evaluations by 28 MFA-trained readers (writers from top U.S. writing programs) and 516 college-educated general readers (recruited via Prolific), AI-generated text from in-context prompting was strongly disfavored by MFA-trained readers for both stylistic fidelity (odds ratio [OR]=0.16, $p < 10^{-8}$) and writing quality (OR=0.13, $p < 10^{-7}$), while college-educated general readers showed no significant preference on stylistic fidelity (OR=1.06, $p = 0.50$) but favored AI-generated

text for writing quality (OR=1.82, $p < 10^{-11}$). However, fine-tuning ChatGPT on individual author's complete works reversed these findings: MFA-trained readers now favored AI-generated text for stylistic fidelity (OR=8.16, $p < 10^{-13}$) and writing quality (OR=1.87, $p = 0.010$), while college-educated general readers showed even stronger AI preference (stylistic fidelity OR=16.65, $p < 10^{-111}$; writing quality OR=5.42, $p < 10^{-57}$). Both reader groups preferred fine-tuned AI, but the writer-type × reader-type interaction remained significant after fine-tuning ($p = 0.021$ for fidelity; $p < 10^{-4}$ for quality), indicating that college-educated general readers favored AI-generated text by a wider margin than MFA-trained readers. These effects are robust under cluster-robust inference and generalize across authors and styles in author-level heterogeneity analyses. The fine-tuned outputs were rarely flagged as AI-generated (3% rate versus 97% for in-context prompting) by state-of-the-art AI detectors. Mediation analysis reveals this reversal occurs because fine-tuning eliminates detectable AI stylistic quirks (e.g., cliché density) that penalize in-context outputs, altering the relationship between AI detectability and reader preference. While we do not account for additional costs of human effort required to transform raw AI output into cohesive, publishable novel-length prose, the median fine-tuning and inference cost of $81 per author represents a dramatic 99.7% reduction compared to typical professional writer compensation. Author-specific fine-tuning thus enables non-verbatim AI writing that readers prefer to expert human writing, thereby providing empirical evidence directly relevant to copyright's fourth fair-use factor, the "effect upon the potential market or value" of the source works.

The U.S. publishing industry supports hundreds of thousands of jobs while generating $30 billion in yearly revenue, contributing to the larger American copyright sectors that account for $2.09 trillion in annual GDP contributions (*1*). Adult Fiction and Non-Fiction books alone accounted for $6.14 billion in 2024 (*2*). This economically important sector now faces an unprecedented challenge: its core products have become essential training data for generative-AI systems. Models trained on well-edited books produce more coherent, accurate responses—something crucial to creating the illusion of intelligence (*3*). Most technology companies building AI use massive datasets of

books, typically without permission or licensing (*4*), and frequently from illegal sources. In the recent copyright lawsuit *Bartz vs Anthropic* (*5*), Judge Alsup noted that Anthropic acquired at least five million books from LibGen and two million from Pirate Library Mirror (PiLiMi). Anthropic also used Books3 (a dataset of approximately 191,000 books also used by Meta and Bloomberg to train their language models), now removed after a legal complaint by anti-piracy group, the Rights Alliance. The Bartz v. Anthropic lawsuit also revealed how Anthropic cut millions of print books from their bindings, scanned them into digital files, and threw away the originals solely for the purpose of training Claude. This unauthorized use has sparked outrage among authors (*6*), triggering dozens of lawsuits against technology companies including OpenAI, Anthropic, Microsoft, Google, and Meta.

Generative AI systems such as ChatGPT that can be prompted to create new text at scale are qualitatively unlike most historical examples of automation technologies (*7*). They can now solve Olympiad-level Geometry (*8*), achieve an impressive rating of 2700 on Codeforces, one of the most challenging coding competition platforms (*9*), and deliver medical guidance that meets professional healthcare standards (*10, 11*). Recent findings from Microsoft's Occupational Implications of Generative AI (*12*), Anthropic's Economic Index (*13*) and OpenAI (*14*) reveal AI usage primarily concentrates in writing tasks. This concentration threatens creative writing professionals in particular—novelists, poets, screenwriters, and content creators who shape cultural narratives and human expression. Based on U.S. Bureau of Labor Statistics May 2023 national estimates, creative writing constitutes almost 50% of writing jobs (*15*), making these positions especially vulnerable to GenAI-based automation, as the writing community has warned (*16*).

While it is widely established that most state-of-the-art Large Language Models (LLMs) have been trained on copyrighted books, it remains unclear whether such training can produce expert-level creative writing. Past research has shown that AI cannot produce high-brow literary fiction or creative nonfiction through prompting alone when compared to professionally trained writers (*17*). More recent work demonstrates that AI-generated creative writing still remains characterized by clichés, purple prose, and unnecessary exposition (*18*). Additionally, relying on Generative AI for creative writing reduces the collective diversity of novel content (*19*). AI often produces formulaic, mediocre creative writing because it lacks the distinctive personal voice that typically distinguishes one author from another (*20, 21*). As Pulitzer Fiction finalist Vauhini Vara observes, *"ChatGPT's*

voice is polite, predictable, inoffensive, upbeat. Great characters, on the other hand, aren't polite; great plots aren't predictable; great style isn't inoffensive; and great endings aren't upbeat" (*22*). To address this limitation, practitioners now increasingly prompt AI systems to perform style/voice mimicry by emulating specific writers' choices (*23*). This practice has become so common that a fantasy author recently published a novel containing an accidentally included AI prompt requesting emulation of another writer's style (*24*). While the effectiveness of such stylistic emulation remains contested, the more pressing question concerns whether style/voice mimicry genuinely improves AI-generated text quality and whether MFA-trained readers and college-educated general readers perceive these improvements as meaningful.

To address this question, we conducted a preregistered behavioral study comparing a representative sample of MFA-trained expert writers with state-of-the-art LLMs.[1] MFA-trained writers are broadly representative of the universe of expert writers. Not all professional writers attended MFA programs, but enough did to stand in for the wider group. Historically, several award-winning authors who have written great American Novels (*25*) have graduated with an MFA degree in Creative Writing (See Figure 9 in SI) and especially from the programs we recruited our writers for this study (See Figure 10 in SI). Eminent literary agent Gail Hochman of Brandt & Hochman said, *"We look favorably on anyone who has an M.F.A., simply because it shows they're serious about their writing" (26).* In June 2010, the editors of The New Yorker Magazine announced to widespread media coverage their selection of 20 under 40[2]—the young fiction writers who are, or will be, central to their generation; 17 out of those 20 writers had an MFA degree (See Fig 11 in SI). This elite sample of MFA trained expert writers provides a conservative test—if AI can compete with the best emerging talent, the disruption to average writers is likely even greater.

We selected closed-source LLMs readily accessible to users without technical expertise: GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro. We also tried open-weights Llama 3.1 model but its empirical performance was not good at following long context instructions at the time of our study. Both human experts and LLMs were given the same task: write an excerpt of up to 450 words emulating the style and voice of one of 50 internationally acclaimed authors, including Nobel

[1] In the United States, a Master of Fine Arts (MFA) is typically recognized as a terminal degree for practitioners of creative writing (equivalent to a Ph.D.)—meaning that it is considered the highest degree in its field, qualifying an individual to become a professor at the university level in these disciplines.

[2] `https://www.newyorker.com/magazine/20-under-40-fiction`

laureates, Booker Prize winners, and Pulitzer Prize winners, spanning multiple continents and cultures. Among others, our study includes Nobel laureates Han Kang and Annie Ernaux; Booker Prize winners Salman Rushdie, Margaret Atwood, and George Saunders; and Pulitzer Prize winners Junot Díaz and Marilynne Robinson. Beyond style emulation, our writing task requires originality under explicit content constraints. Although both AI and human writers received prompts with detailed content specifications, composing literary fiction/creative non-fiction still requires agency especially in deliberate choices of words, syntax, voice, and narrative framing to produce novel, coherent prose. Because literary fiction/creative non-fiction is inherently non-formulaic, the writing task extends beyond style imitation to creative composition, making it different from mere parody or pastiche.

Our study examines shorter (up to 450-word) excerpts for two reasons. First, an excerpt of this length isolates the voice, style, and quality that writers must sustain across any longer work. Second, in its current form, an LLM cannot write long-form narratives that automatically balance coherence, quality, thematic consistency, and plot development across thousands of words. But it is merely a matter of time before LLMs reach that capacity. It should also be noted that human novel-writing is itself iterative and compositional, making excerpt-level quality a meaningful building block. With human steering and iterative prompting, it is feasible to produce long-form fiction and non-fiction using shorter model-generated excerpts. This is already happening in real life with startups such as Sudowrite or Genre Fiction publisher Inkitt (*27*) specifically focusing on helping consumers write full books using AI. In self-publishing marketplaces such as Kindle, these AI-generated books have already gained popularity (*28, 29, 30*).

We tested two AI conditions: (1) in-context prompting, where models received the same instructions as human experts, and (2) fine-tuning, where models were additionally trained on each author's complete oeuvre. For authors writing in non-English languages (Han Kang, Yoko Ogawa, Annie Ernaux, Haruki Murakami), we used the same translator's work across all books to maintain voice consistency. Our MFA-trained expert writers who wrote excerpts also evaluated submissions in our study (though they never evaluated their own excerpts). We selected these writers as expert readers because their degrees serve as a strong proxy for quality in both creation and evaluation: a good writer is also likely to be a good reader, capable of discerning quality in others' writings. In addition to MFA-trained readers we also recruited college-educated general readers from Prolific.

Both groups of readers performed blind pairwise evaluations (*31, 32, 33*) of ⟨Human-AI⟩ excerpts on writing quality and stylistic fidelity. This design addresses three preregistered research questions: (1) Can AI match expert performance in writing quality and stylistic fidelity across both conditions? (2) Do both groups of readers show similar preference patterns? (3) Does AI detectability correlate with human quality judgments, and does fine-tuning remove this correlation? Our full experimental setup is shown in Figure 1.

Results

Overall Performance Comparisons

Our final data consist of 10,920 reader-level pairwise judgments/evaluations, with judgments from 28 MFA-trained readers and 516 college-educated general readers. We fit a logit model for each outcome and condition and include dummies for writer type, reader group, and their interaction. We further employ CR2 cluster robust standard errors (*34*) clustered at the reader-level to account for within-reader correlation in ratings. Our hypotheses, outcomes, design, and analysis closely follow our OSF pre-registration (SI Sections S4-S8); minor deviations are detailed in SI Section S9.

Figure 2A–B presents the odds ratios, with corresponding predicted probabilities shown in Figure 2C–D. Under in-context prompting, MFA-trained readers demonstrated strong preference for human-written text. Odds ratios were 0.16 (95% CI: 0.08–0.29, $p < 10^{-8}$) for stylistic fidelity and 0.13 (95% CI: 0.06–0.28, $p < 10^{-7}$) for writing quality, indicating six- to eight-fold preferences for human excerpts (Fig. 2A–B). College-educated general readers showed no significant preference regarding stylistic fidelity (OR = 1.06, 95% CI: 0.90–1.25, $p = 0.50$) but clearly favored AI-generated text for writing quality (OR = 1.82, 95% CI: 1.53–2.15, $p < 10^{-11}$), selecting AI excerpts in 50–62% of writing quality trials depending on the model (Fig. 2C–D). Inter-rater agreement reflected this divergence: MFA-trained readers achieved $\kappa = 0.58$ for stylistic fidelity and $\kappa = 0.41$ for writing quality, while college-educated readers showed minimal agreement among themselves ($\kappa = 0.10$ and $\kappa = 0.08$, respectively). Given that college-educated general readers' assessments of literary quality and style reflect inherently diverse tastes, we anticipated lower inter-rater agreement compared to MFA-trained evaluators. The writer-type × reader-type interaction was significant for

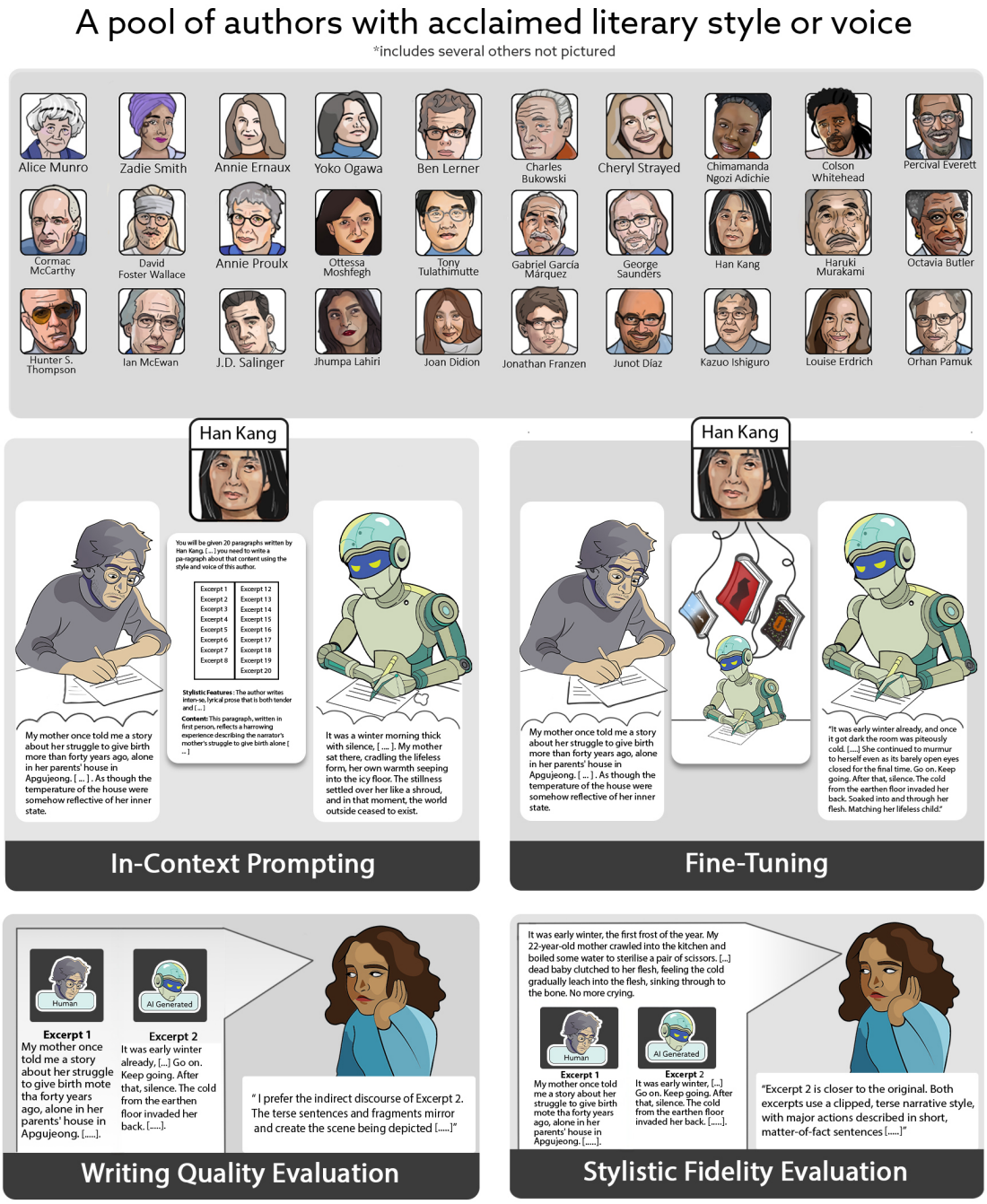


Figure 1: Figure showing our study design. (1) Select a target author and prompt. (2) Generate up to 450-word candidate excerpts from MFA experts and from LLMs under two settings: in-context prompting (instructions + few-shot examples) and author-specific fine-tuning (model fine-tuned on that author's works). (3) Readers (MFA-trained and college-educated) perform blinded, pairwise forced-choice evaluations on two outcomes: stylistic fidelity to the target author and overall writing quality. Pair order and left/right placement are randomized on every trial.

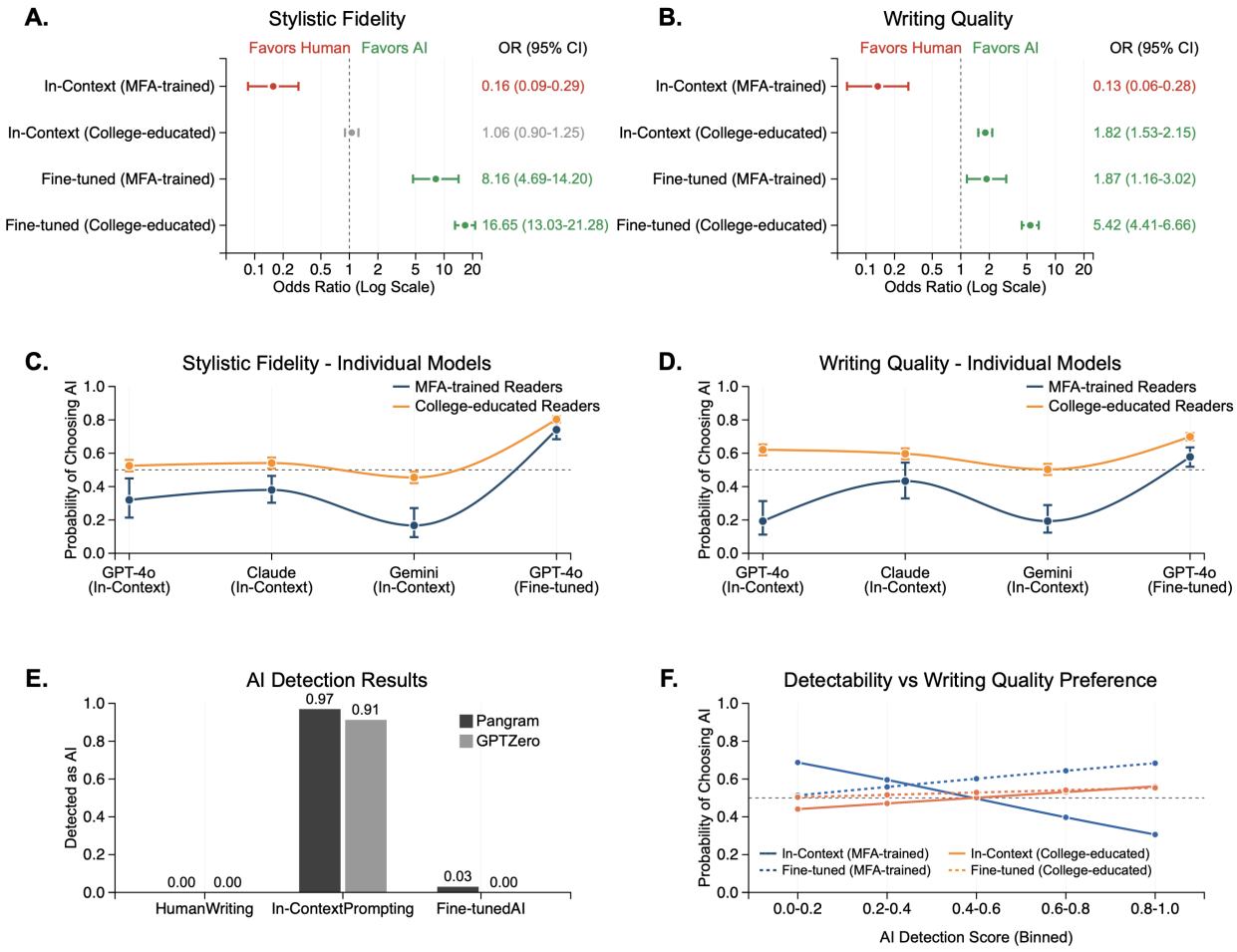


Figure 2: (A-B) Forest plots showing odds ratios (OR) and 95% confidence intervals comparing AI-generated and human-written excerpts in pairwise evaluations of stylistic fidelity (A) and writing quality (B) where values >1 favor AI and values <1 favor humans. MFA-trained readers show preference for human writing when prompted in an in-context setting (OR = 0.16 and 0.13) but that changes when AI is fine-tuned (OR = 8.16 and 1.87). College-educated general readers show a distinct preference pattern rather than mere noise: under in-context prompting they are at parity with humans on stylistic fidelity (OR = 1.06) but prefer AI-generated text on writing quality (OR = 1.82). (C-D) Probability of choosing AI excerpts across individual language models for stylistic fidelity (C) and writing quality (D). Error bars represent 95% confidence intervals. Dashed line indicates chance performance (50%). (E) AI detection accuracy with chosen threshold of τ=0.9 using two state-of-the-art AI detectors (Pangram and GPTZero). Human written text was never misclassified (0.00), in-context AI was detected with 97% accuracy by Pangram and 91% by GPTZero, but fine-tuned AI evaded detection 97% of the time (0.03) for Pangram and 100% of the time in case of GPTZero. (F) Relationship between AI detectability (Pangram) and preference for writing quality across detection score bins. For the in-context prompting setup, higher detection scores predict lower AI preference mainly among MFA-trained readers. After fine-tuning, that negative detectability penalty is strongly attenuated, and the model-implied slopes are flat-to-positive. Fine-tuning on an author's complete oeuvre gets rid of AI quirks while achieving expert-level performance. n = 28 MFA-trained readers, 516 college-educated general readers; 10,920 reader-level pairwise judgments/evaluations.

both outcomes ($\chi^2_{(3)} = 36.6$, $p = 5.6 \times 10^{-8}$ for fidelity; $\chi^2_{(3)} = 46.9$, $p = 3.7 \times 10^{-10}$ for quality).

Fine-tuning on authors' complete works reversed these preferences. For MFA-trained readers, the *odds* of selecting the AI excerpt were 8.16 times the odds of selecting the human excerpt for stylistic fidelity (OR = 8.16, 95% CI: 4.69–14.2, $p < 10^{-13}$) and 1.87 times the odds for writing quality (OR = 1.87, 95% CI: 1.16–3.02, $p = 0.010$). College-educated general readers showed even larger shifts in their preferences (stylistic fidelity OR = 16.65, 95% CI: 13.03–21.28, $p < 10^{-15}$; writing quality OR = 5.42, 95% CI: 4.41–6.66, $p < 10^{-15}$). Unlike the in-context condition, both reader groups now favored fine-tuned AI, but the magnitude of this preference differed: model-based predicted AI win probabilities were 0.74 for MFA-trained readers and 0.80 for college-educated general readers on stylistic fidelity, and 0.58 versus 0.70 on writing quality (Fig. 2C–D). The writer-type × reader-type interaction remained significant in the fine-tuned models (stylistic fidelity $\chi^2_{(1)} = 5.33$, $p = 0.021$; writing quality $\chi^2_{(1)} = 15.98$, $p = 6.4 \times 10^{-5}$), indicating that college-educated general readers favored AI-generated text by a wider margin than MFA-trained readers even after fine-tuning. Inter-rater agreement among MFA-trained readers increased ($\kappa = 0.67$ for writing quality; $\kappa = 0.54$ for stylistic fidelity), while agreement among college-educated general readers remained low ($\kappa = 0.12$ and $\kappa = 0.26$, respectively).

AI Detection and Stylometric Analysis

We probe whether differences in AI detectability can account for these preference reversals. Pangram, a state-of-the-art AI detection tool (*35,36,37*), correctly classified 97% of in-context prompted texts as machine-generated but only 3% of fine-tuned texts were classified as AI-generated (Fig. 2E). GPTZero, another state-of-the-art AI detection, tool showed comparable performance but had higher false-positive rate, so we used Pangram for our subsequent analyses.

Higher AI-detection scores strongly predicted lower preference rates among MFA-trained readers in the in-context prompting condition. For stylistic fidelity, each unit increase in detection score reduced the odds of selecting an excerpt by a factor of 6.3 ($\beta = -1.85 \pm 0.29$, $p < 10^{-9}$); a similar pattern held for writing quality ($\beta = -2.01 \pm 0.33$, $p < 10^{-9}$). College-educated general readers were substantially less sensitive to detectability even before fine-tuning (Pangram × Reader Type for style: $\beta = 1.88 \pm 0.30$, $p < 10^{-9}$; for quality: $\beta = 2.61 \pm 0.34$, $p < 10^{-13}$). Fine-tuning

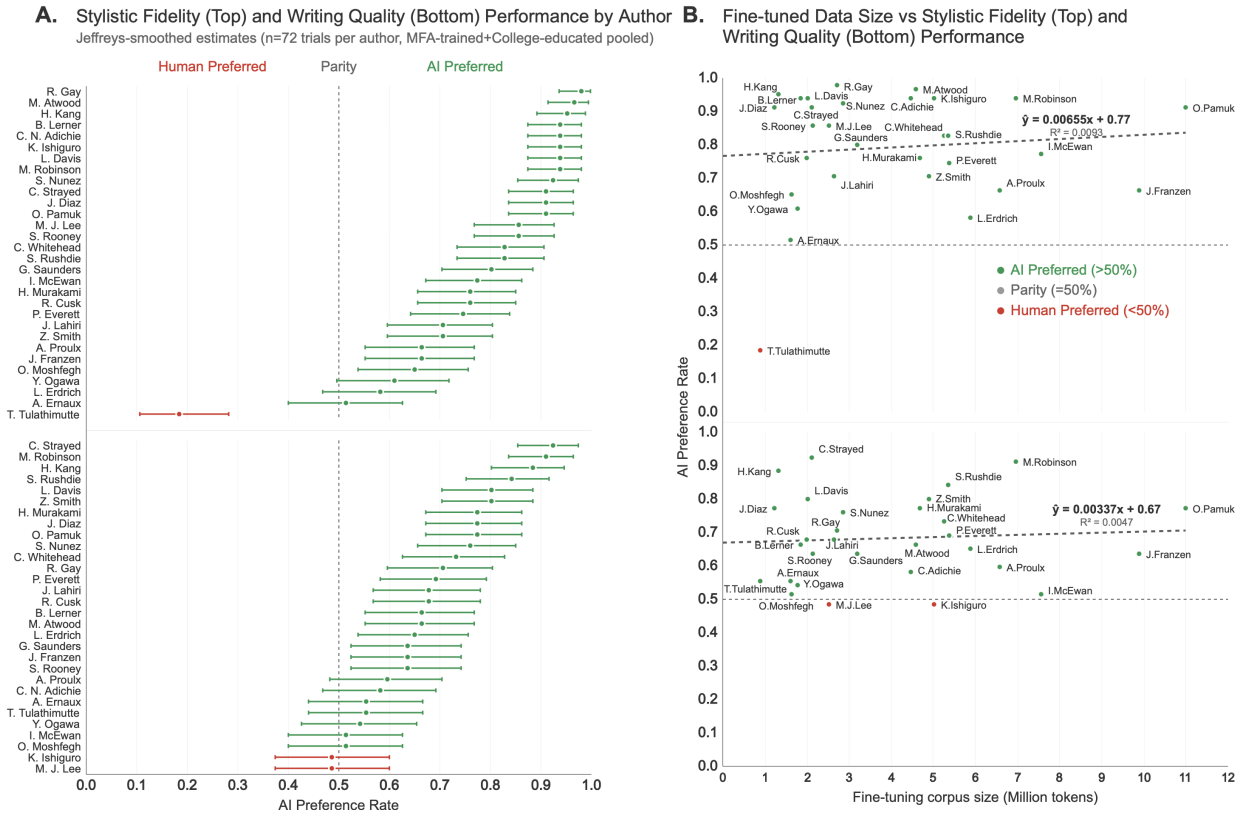


Figure 3: Author-level AI preference and its association with fine-tuning corpus size (fidelity and quality) (A) For each fine-tuned author, the share of blinded pairwise trials in which the AI excerpt was preferred over the human (MFA expert) on stylistic fidelity (top) and overall quality (bottom). Points show Jeffreys-prior estimates $(k + 0.5)/(n + 1)$; vertical bars are 95% Jeffreys intervals (Beta$(\frac{1}{2}, \frac{1}{2})$); the dotted line at 0.5 marks human–AI parity. Readers are pooled (MFA-trained and college-educated general readers). (B) AI preference rate versus the fine-tuning corpus size for that author (million tokens), shown for stylistic fidelity (top) and overall quality (bottom). Each point is a fine-tuned author; the line is an OLS fit with heteroskedasticity-robust standard errors (no CI displayed). Slopes are near zero in both panels, indicating little association between corpus size (in this range) and AI preference.

attenuated the negative relationship between detectability and preference (Pangram × Setting for style: $\beta = 2.56 \pm 0.81$, $p = 0.002$; for quality: $\beta = 2.90 \pm 0.88$, $p < 0.001$), though the degree of attenuation differed across reader groups, as indicated by significant three-way interactions (Pangram × Setting × Reader Type for style: $\beta = -1.77 \pm 0.84$, $p = 0.036$; for quality: $\beta = -3.26 \pm 0.91$, $p < 0.001$). Two-stage mediation analysis (Fig. 4A) demonstrated that among MFA-trained readers, where the detectability penalty is concentrated, stylometric features, particularly cliché density (See Section S3.2 in SI), mediated 16.4% of the detection effect on preference before fine-tuning but a statistically insignificant 1.3% (95% CI includes zero) afterward, indicating that fine-tuning reduces rather than merely masks artificial stylistic signatures. Because college-educated general readers did not penalize AI-detectable text in either condition, mediation analysis does not apply to this group.

Author-Level Performance Heterogeneity

Next, we disaggregated our data to unpack heterogeneity at the level of individual authors. Of 30 fine-tuned author models, 29 exceeded parity for stylistic fidelity (median win rate = 0.83, IQR: 0.70–0.94) and 28 for writing quality (median = 0.67, IQR: 0.58–0.77). Using 95% Jeffreys intervals, fine-tuned models significantly outperformed human writers for 26 authors on stylistic fidelity and 21 authors on writing quality (Fig. 3A). These performance differences showed no systematic relationship with fine-tuning corpus size (Fig. 3B). The fine-tuning premium, i.e., the increase in AI preference rates relative to in-context prompting, ranged from −12.2 to +67.4 percentage points for stylistic fidelity (29 of 30 positive) and from −14.1 to +46.1 percentage points for writing quality (24 of 30 positive). This "premium" likewise showed no correlation with fine-tuning corpus size (Pearson $r < 0.1$ for both outcomes; Fig. 4B).

Cost Analysis

The weak dependence of performance on fine-tuning corpus size has direct economic implications. Total AI generations costs (fine-tuning plus inference) ranged from $25 to $276 per author (median = $81), assuming API-based fine-tuning at $25 per million tokens plus $3 for generating 100,000 words of raw text (Fig. 4C). These costs represent approximately 0.3% of what MFA-trained

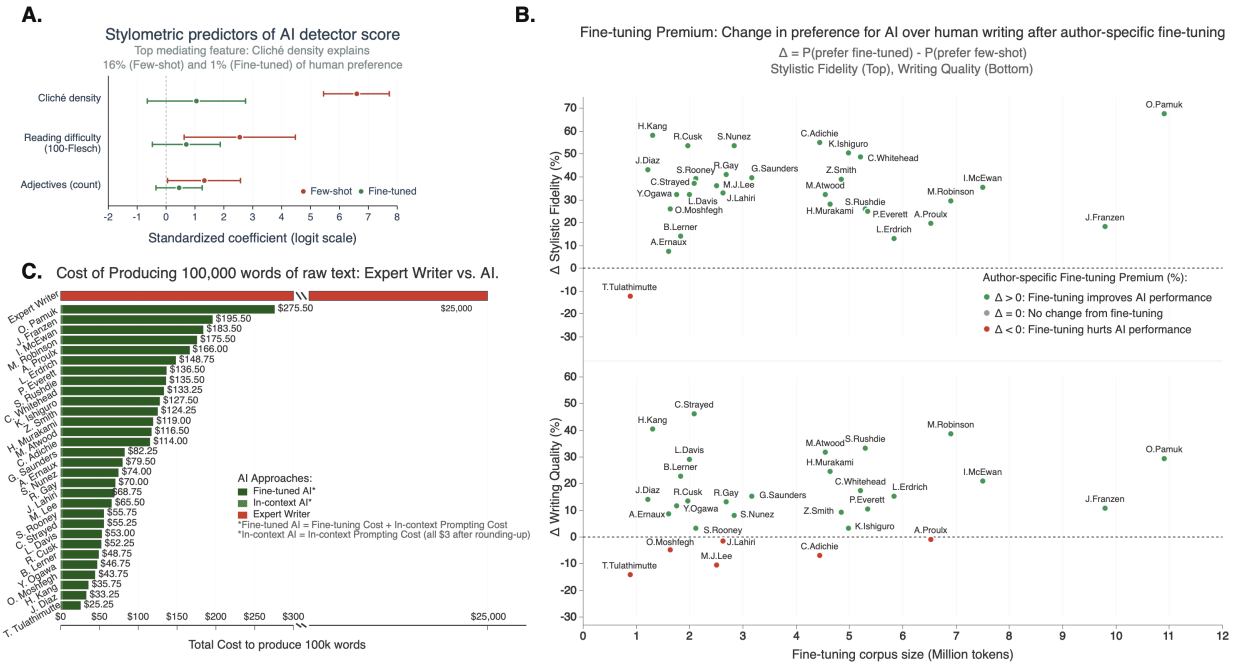


Figure 4: Fine-tuning substantially reduces stylometric signatures of AI text, improves stylistic fidelity and perceived writing quality over in-context prompting, and substantially cuts costs of producing first draft versus professional writers. (A) Mediation analysis linking stylometric features → AI-detector score → human preference. Standardized logistic coefficients with 95% CIs are shown for three features for in-context prompting (red) and fine-tuned models (green). Cliché density mediates 16.4% of the detector effect on choice for in-context prompting but only 1.3% for author fine-tuned models; all three features together mediate 25.4% vs −3.2% for in-context prompted and fine-tuned models respectively. (B) "Fine-tuning premium," defined as Δ = P(prefer fine-tuned over human) − P(prefer in-context over human), as a function of fine-tuning corpus size. Top: stylistic fidelity; bottom: writing quality. Points are authors; colors denote improvement (green, $\Delta > 0$), no change (gray), or degradation (red, $\Delta < 0$). Median Δ: +34.2 (fidelity) and +13.4 percentage points (quality). (C) Cost to produce 100,000 words of raw text vs. publishable prose. MFA-trained writers in our study would earn $25,000 for a 100k-word novel-length manuscript (red). By contrast, AI pipelines can generate 100k words of raw text for $25–$276 depending on fine-tuning corpus size (green bars = fine-tuning; hatched = in-context prompting, $3). This figure reflects direct compute/API costs only, not the additional human steering, chunking, and editing required to turn raw AI text into a cohesive publishable work. Authors ordered by total AI cost.

writers in our study would charge for a novel-length (100,000 words) manuscript. It should be noted that this comparison reflects raw generation costs before the human steering and editing required to transform AI outputs into publishable works. Despite this caveat, the minimal investment yielded outputs that achieved expert-level performance for most authors. Performance gains were uncorrelated with both corpus size and fine-tuning cost, indicating that computational scale did not drive improvements. The 99.7% reduction in raw generation costs, coupled with superior quality ratings for the majority of authors, underscores the potential for substantial producer-surplus shifts and market displacement.

Discussion

What are the implications of this research for the copyright infringement claims that authors have brought against AI companies alleging unauthorized use of their books in training datasets (*38, 39, 40*)? These cases raise the question whether copying millions of copyrighted books for AI training constitutes fair use when the resulting outputs do not themselves reproduce the copied works. The most significant consideration in evaluating this question is the fourth fair use factor: *"the impact of the use upon the actual or potential market for the copyrighted work."*[3]

Courts understand this factor to concern the extent to which works produced through copying serve as market substitutes for the original author's work (*41, 42*).

Our study has shown that AI-generated excerpts from in-context prompted models pre-trained on vast internet corpora (including millions of copyrighted works) were strongly disfavored by MFA-trained readers for both quality and stylistic fidelity, while college-educated general readers showed no preference on stylistic fidelity but already favored AI-generated text for writing quality. When models are fine-tuned on curated datasets consisting solely of individual authors' complete

[3]The fair use provision (section 107) of the US Copyright Act directs courts to consider four factors:

1. the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational
2. the nature of the copyrighted work;
3. the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
4. the effect of the use upon the potential market for or value of the copyrighted work. Courts understand this factor to concern the extent to which the works produced by the copying substitute for the author's work.

works, both groups decisively preferred the AI-generated excerpts over human-written examples. Notably, college-educated general readers, who are generally representative of the book-purchasing public, preferred fine-tuned AI by a substantially wider margin than MFA-trained readers (e.g., writing quality OR = 5.42 vs. 1.87). Moreover, the fine-tuned excerpts proved almost undetectable as AI-generated text (particularly when compared to outputs from in-context prompted models), while both MFA-trained and college-educated general readers preferred them over human writing in blind evaluations.

At first glance, a legal analyst might conclude that our findings are irrelevant to fair use because the outputs from the blind pairwise evaluation do not reproduce the copied works. While they may exhibit comparable literary quality and high stylistic fidelity to the originals, copyright law does not protect authors' style—only their expression (*41, 39*). These outputs may offer credible substitutes for an author's works, but so do human-authored works inspired by prior works. However, there is an important difference between human and AI-generated emulations: humans read; AI systems copy. Unlike human memory, which is not a verbatim storage device, all AI-generation requires predicate copying despite rhetoric equating human learning with machine "learning." (*43, 44, 45*)

The U.S. Copyright Office has recognized that such predicate copying may cause cognizable market harm through competing works that the inputs enable, potentially flooding the market and causing "market dilution" (*46*). While acknowledging this "market dilution" approach to the fourth fair use factor as "uncharted territory," the Office determined that both the statutory language and underlying concerns of the Copyright Act warrant this inquiry. *"... The speed and scale at which AI systems generate content pose a serious risk of diluting markets for works of the same kind as in their training data. That means more competition for sales of an author's works and more difficulty for audiences in finding them. If thousands of AI-generated romance novels are put on the market, fewer of the human-authored romance novels that the AI was trained on are likely to be sold... Market harm can also stem from AI models' generation of material stylistically similar to works in their training data... "* The Office emphasized that the effect of the copying impacts extant works by putting them in competition with AI-generated outputs. Copyrighted works become fodder for new productions targeting the same markets (*47*). Crucially, the Office did not claim that competing AI outputs copy the inputted works; rather, it examined the economic consequences of the predicate copying that enables these competing outputs. This focus on inputs remains essential

because, absent the initial copying, no infringement action exists against flooding markets with independently generated works—if a thousand humans write romance novels after reading Barbara Cartland's novels, they compete but do not infringe. The Copyright Office's expansive interpretation of "potential market for or value of the copied work" suggests that fair use might not excuse predicate copying even when it doesn't show up in the end product, if the copying's effect substitutes for source works.

In Kadrey v. Meta (*39*), an infringement action brought by 13 book authors against the copying of their books into the database underpinning Meta's Llama LLM, Judge Chhabria granted summary judgment to Meta but accepted the theory of market dilution: "*[I]ndirect substitution is still substitution: If someone bought a romance novel written by an LLM instead of a romance novel written by a human author, the LLM-generated novel is substituting for the human-written one. . . This case involves a technology that can generate literally millions of secondary works, with a miniscule fraction of the time and creativity used to create the original works it was trained on. No other use. . . has anything near the potential to flood the market with competing works the way that LLM training does."* Judge Chhabria effectively provided a road map for what authors would have to show to persuade a court that the AI inputs diluted their markets: First, is the AI system capable of generating such substitutional books now or in the near future? Second, what are the markets for the plaintiffs' books, and do the AI-generated books compete in those markets? Third, what impact does any such competition have on sales—does it displace those works entirely or merely chip away at demand, and are the effects likely to grow as AI-generated books proliferate and models improve? Fourth, how does the threat to the market for the plaintiffs' books differ between a world in which LLM developers may copy those books and a world in which they may not?

Our study's findings concerning reader preferences between human-authored and AI-generated works bear on all four considerations. They also demonstrate how LLMs have already gotten *"better and better at writing human-like text."* While Judge Chhabria speculated that the distinctiveness of an author's style renders works by well-known authors less susceptible to substitution, he observed that market dilution would vary by author prominence. Established authors with dedicated readerships (like Agatha Christie) would likely face minimal substitution, while AI-generated books could crowd out lesser-known or emerging authors, potentially preventing *"the next Agatha Christie from getting noticed or selling enough books to keep writing"*. Our work suggests otherwise. If

readers in fact prefer AI-generated emulations of authors whose market value lies in their distinctive voices, then the prospect of competition, especially from outputs of fine-tuned datasets, appears to be considerable. The comparatively low production costs of AI-generated texts relative to paying human authors (as shown in Figure 4C) further enhances the likelihood that AI platforms will in fact dilute the market for human-authored work.

These findings suggest that the creation of fine-tuned LLMs consisting of the collected copyrighted works (or a substantial number) of individual authors should not be fair use if the LLM is used to create outputs that emulate the author's works (*48*). As the Copyright Office observed, "[f]ine-tuning. . . usually narrows down the model's capabilities and might be more aligned with the original purpose of the copyrighted material," (*46*) and thus both less "transformative," and more likely to substitute for it. By contrast, the LLMs employed for in-context prompting do not target particular authors, and therefore can be put to a great variety of uses that do not risk diluting those authors' markets. Their claim to fair use seems accordingly stronger. But those models can generate author-emulations, and our study has shown that among college-educated general readers, who more closely represent the book-purchasing public, both in-context quality outputs and fine-tuned outputs show clear AI preference, indicating meaningful substitution potential even without author-specific training. Our findings have also been independently corroborated by a large-scale blind quiz conducted by The New York Times in which more than 86,000 readers compared AI and human-written passages and 54% preferred AI-generated text (*49*), as well as independent reporting in The New Yorker where even accomplished novelists struggled to distinguish fine-tuned AI emulations from human writing (*50*).

A reasonable solution might allow the inclusion of copyrighted works in the general-purpose dataset, but would require the model to implement guardrails that would disable it from generating non-parodic imitations of individual authors' oeuvres (*51, 52, 53*). As examples of guardrails, AI developers have implemented "refusal protocols" blocking outputs when prompts request content "in the style of" specific authors. Further, current reinforcement learning techniques can be easily modified to steer models away from stylistic imitation. However, such guardrails remain imperfect: refusal protocols can be circumvented through prompt rephrasing, and reinforcement learning–based steering may not generalize across all forms of stylistic emulation. Another solution, particularly where the lower quality of in-context prompting reduces the prospect of market

dilution, might be to condition a ruling of fair use on the prominent disclosure of the output's AI origin. This solution assumes that the public, informed that the output was not human-authored, will be less inclined to select the AI substitute; transparency should diminish the competition between human-authored and machine-generated offerings.

Whether disclosure would in fact diminish the competitive advantage of AI-generated works is a question our study cannot directly address, because all evaluations were conducted under blinded conditions. Whether the strong preferences observed here would persist when readers know the text is AI-generated remains an open empirical question requiring a separate, unblinded experimental design. Moreover, even if disclosure were to reduce the quality preference to some degree, relative cost may independently sustain consumer substitution. If AI-generated works enter the market at substantially lower price points than human-authored works—as our cost analysis suggests is feasible—consumers may select the AI version even knowing its provenance, because the quality-to-price ratio remains attractive. The interaction of disclosure, price, and reader preference in actual market conditions therefore warrants further investigation.

Methods

Experimental Setup

We recruited 28 expert writers affiliated with top MFA programs in the United States, including from programs such as Iowa Writers Workshop, Helen Zell Writers' Program at University of Michigan, MFA Program in Creative Writing at New York University, Columbia University School of Arts, and MFA in Creative Writing at BU, paying them $75 for writing each excerpt. These MFA candidates, along with the LLMs described earlier, emulated the style and voice of 50 award-winning authors representing diverse cultural backgrounds and distinct literary voices (full list in Table 2 in SI). The writing prompt provided to MFA candidates contained three components: (i) 20 sample excerpts spanning an author's complete body of work, (ii) textual descriptions of the author's distinctive style and voice, and (iii) detailed content specifications about the original author-written excerpt to be emulated. The selection of the 50 authors and the development of all writing prompts were completed in collaboration with five English Literature PhD students who

analyzed each author's literary voice and created the verbalized style descriptions in addition to carefully choosing representative excerpts written by the original author (prompt shown in Figure 5 in SI). For the purpose of the study, MFA-trained expert writers had no time limitation to submit these excerpts and were encouraged to take as long as required to produce their best work.

In-context Prompting

Larger context windows in LLMs allow for processing significantly more information at once, leading to improved accuracy, understanding, and complex reasoning capabilities. For our task this is crucial as it lets the model process the entire writing prompt including the 20 sample excerpts, style descriptions and content specifications during inference. Taking advantage of this capability in GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro, we generated LLM emulations using the identical writing prompt provided to MFA candidates (as described above). In this condition, models received no additional training on the target authors' works; they relied solely on the information contained within the prompt. Thus for AI Condition 1 (in-context prompting), we obtained 150 ⟨Human-AI⟩ pairs: 150 human-written excerpts (3 MFA writers $\times$ 50 authors) paired with 150 AI-generated excerpts. Each human excerpt was paired with one AI excerpt for the same author, with the AI excerpts distributed equally across the three models (50 excerpts each from GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro).

Fine-tuning

For AI Condition 2 (fine-tuning), we selected 30 living authors from our pool. This decision was made specifically to examine the potential economic impact of generative AI on the livelihoods of living authors while also considering the substantial computational costs associated with fine-tuning models on individual authors. Fine-tuning on books presents several technical challenges. Each book typically contains 50,000 to 80,000 words, and using such long sequences as training examples wastes computational capacity because models still struggle with very long inputs and long-context adaptation remains non-trivial. To preserve general capabilities in supervised fine-tuning, it is generally advised to use diverse, shorter samples. Breaking books into context-independent excerpts increases batch diversity and the number of distinct gradient signals per token budget, while also reducing overfitting to a single narrative flow and still capturing the salient stylistic and

local patterns characteristic of each author's voice.

We purchased ePub files of these 30 authors' complete works, converted them to plain text, and segmented them into 250–650 word context-independent excerpts. For each excerpt, we extracted content descriptions using GPT-4o with the following prompt: *"Describe in detail what is happening in this excerpt. Mention the characters and whether the voice is in first or third person for majority of the excerpt. Maintain the order of sentences while describing."* Since only GPT-4o (among our three models) supported API-based fine-tuning at the time of our study, we fine-tuned 30 author-specific GPT-4o models using input-output pairs structured as: "Write a [[n]] word excerpt about the content below emulating the style and voice of [[authorname]]\n\n[[content]]: [[excerpt]]" (see Figure 6 in SI for details). This technique is commonly referred to as **instruction back-translation** (*54*), where the model learns to generate text matching a target style given content specifications. Importantly, the original author excerpts on which the human-written emulations were based were excluded from the fine-tuning data to ensure fairness in comparison. This approach yielded 90 ⟨Human-AI⟩ pairs: each of the 30 fine-tuned GPT-4o models generated one excerpt per author, and each AI excerpt was paired with all three MFA-written excerpts for that author (30 authors × 3 human excerpts = 90 pairs). During inference, we verified that no generated excerpt regurgitated verbatim expressions from the original training data. ROUGE-L scores (*55*) ranged from 0.16 to 0.23, indicating minimal lexical overlap between AI-generated and original author-written excerpts. Refer to Section S1.3 for more details on the fine-tuning procedure and validation.

Human Participants for Writing and Evaluation Task

These ⟨Human-AI⟩ pairs from both AI conditions (in-context prompting and fine-tuning) were evaluated by 28 experts (the same MFA candidates mentioned above) and 516 college-educated general readers recruited from Prolific,[4] one of the leading crowdsourcing platforms for research participants. As mentioned earlier, MFA candidates served dual roles in our study: they created human-written excerpts for comparison with AI outputs, and they also evaluated all excerpts in blind pairwise comparisons (never evaluating their own work). We compensated MFA-trained writers $75 for writing each excerpt and an additional $75 for evaluating a batch of 10 writing quality comparisons or $100 for evaluating a batch of 10 stylistic fidelity comparisons (the latter requiring

[4] https://www.prolific.com/

comparison against original author excerpts).

Our MFA-trained writer sample represents the emerging literary elite whose professional judgment shapes contemporary publishing. Among our recruited MFA candidates, several have since been awarded Stanford Stegner Fellowships (one of the most competitive fiction writing fellowships in the United States), received Rhodes Scholarships, assumed editorial positions at prestigious literary magazines such as *Joyland*, been awarded the Publishers Weekly Star Watch 2025 Honor and secured publishing contracts with major houses including W.W. Norton and Harper Collins. These accomplishments underscore that our experts represent not merely MFA candidates, but established and emerging voices whose aesthetic judgments carry professional authority in the literary marketplace.

All recruited MFAs were residing in the United States at the time of the study. This geographic restriction was necessary for administrative compliance, as U.S. tax regulations require declaration of compensation above certain thresholds, and the university's payment infrastructure could not accommodate international participants without Individual Taxpayer Identification Numbers (ITINs) or Social Security Numbers (SSNs). Despite this constraint, our expert sample demonstrated substantial demographic diversity: 65% did not identify as cisgender men, and participants self-identified across multiple ethnic backgrounds, including South Asian, White, East Asian, and Black. This diversity helps ensure that our findings are not artifacts of a narrow demographic perspective.

For the 516 college-educated general readers recruited from Prolific, we restricted participation to native-born residents of English-speaking countries (USA and UK only) ensuring that all evaluators had native-level fluency in English. Additional eligibility criteria included maintaining a 100% task acceptance rate on the platform (indicating high performance on previous tasks) and having a college degree (ensuring a baseline level of literacy). We compensated the college-educated general readers according to Prolific's wage standards with approximately $15 per hour for their evaluation time. Unlike MFA-trained readers, college-educated general readers represented general educated audiences without specialized training in creative writing or literary analysis; hence their compensation was lower than MFAs.

Blind Evaluation

Experts never evaluated their own excerpts. Each ⟨Human-AI⟩ pair was assessed by three MFA-trained readers and by 19 college-educated general readers in the in-context condition or 21 college-educated general readers in the fine-tuned condition. All reported analyses are conducted at the reader-judgment level, with inter-rater agreement summarized separately using Fleiss' kappa within each reader group. In total, we obtained 6,600 pairwise evaluations (3,300 for quality, 3,300 for style) for AI Condition 1 (in-context prompting) and 4,320 evaluations (2,160 for quality, 2,160 for style) for AI Condition 2 (fine-tuning). The evaluation tasks differed by outcome: for quality evaluation, we showed only the ⟨Human-AI⟩ pair and asked readers to judge which excerpt was better written overall; for stylistic fidelity evaluation, we included the original author-written excerpt alongside the pair and asked readers to judge which excerpt better captured that author's distinctive style and voice (see Figures 12 and 13 in SI for screenshots of both interfaces).

Beyond selecting a preferred excerpt, both groups of readers provided 2–3 sentence explanations grounded in textual evidence to justify their choices (*56*) (see Figures 20–23 in SI for examples). We required these written rationales for two reasons. First, without observing why a choice was made, the latent factors driving that preference remain unclear—hidden user context, demographic or value variation, task ambiguity, or near-tie similarity can inflate effective label noise. Providing rationales or auxiliary side information has been empirically shown to improve label reliability, data efficiency, and debiasing (*57*). Second, for college-educated general readers specifically, written explanations serve as a quality control mechanism, helping us identify participants who may be selecting preferences randomly rather than engaging seriously with the task. Readers who provided generic or off-topic rationales, or whose explanations contradicted their stated preferences, were flagged for potential exclusion.

To ensure annotation quality, we implemented several attention checks. We recorded timestamps for each evaluation to identify and exclude participants who rushed through tasks (spending less than 30 seconds per comparison). Additionally, we screened all written rationales using Pangram,[5] a state-of-the-art AI detection tool, and excluded any participants who used generative AI to compose their responses (*58*). This screening was necessary because AI-generated rationales would undermine the validity of human preference judgments. Our study was approved by the University

[5] https://www.pangram.com/

of Michigan IRB (HUM00264127) and was preregistered at OSF.[6] Informed consent was obtained from all participants. As noted earlier, we compensated MFAs $75 per 10-excerpt batch for quality judgments and $100 per 10-excerpt batch for stylistic fidelity evaluations (the higher rate reflecting the additional cognitive load of comparing against original author excerpts).

Empirical Analyses

We tested three preregistered hypotheses using logistic regression models with CR2 cluster-robust standard errors clustered by reader. Clustering accounts for the fact that each reader evaluated multiple excerpt pairs, which induces within-reader correlation in preference judgments. The CR2 adjustment provides more accurate inference in settings with a modest number of clusters (28 MFA-trained expert readers, 516 college-educated general readers) (*34*).

For H1 (can baseline LLMs match human expert performance?), we compared human writing to the average performance across GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro in the in-context prompting condition. Because each human excerpt was paired with only one AI excerpt from a single model, we pooled AI excerpts across the three models and tested whether readers showed systematic preference for human versus AI writing on average. For H2 (can fine-tuned models match human expert performance?), we directly compared fine-tuned GPT-4o excerpts to human writing. Contrasts were computed separately for MFA-trained readers and college-educated general readers within each outcome (stylistic fidelity, writing quality), yielding four tests per hypothesis (2 reader groups $\times$ 2 outcomes). To control the family-wise error rate across these multiple comparisons, we applied Holm correction separately within each hypothesis-outcome combination. This procedure adjusts p-values to maintain valid inference when testing multiple contrasts that address the same research question.

For H3 (does AI detectability correlate with human preference, and does fine-tuning remove this correlation?), we modeled the relationship between Pangram AI-detection scores and reader preference. Specifically, we tested whether higher detection scores (indicating more AI-like text) predicted lower preference for AI excerpts in the in-context prompting condition, and whether fine-tuning attenuated this penalty via AI condition $\times$ detection score interactions. A significant negative coefficient on detection score in the in-context condition, coupled with a significant positive

[6] `https://osf.io/zt4ad`

interaction term, would indicate that fine-tuning removes the detectability-based preference penalty that AI excerpts suffer under in-context prompting. All model specifications, including the full regression equations, covariate adjustments, and robustness checks, are described in detail in SI Sections S4–S8.

Data availability

Anonymized trial-level data supporting the findings of this study are available at GitHub at the link (https://github.com/tuhinjubcse/Author-Style-Personalization/tree/main). Additional details are provided in SI, Section S10.

Code availability

All analysis and figure-generation code is available at GitHub (https://github.com/tuhinjubcse/Author-Style-Personalization/tree/main). Core analyses can be reproduced by running the numbered R scripts in sequence. Analyses were conducted in R 4.3.1 with key packages including `clubSandwich` and `emmeans`. Exact package versions and run instructions are provided in SI, Section S10.

Ethics statement

All procedures involving human participants were approved by the University of Michigan Institutional Review Board (HUM00264127). Informed consent was obtained from all participants, who were compensated for their time. The study was preregistered at OSF (https://osf.io/zt4ad).

Acknowledgements

We thank Jared Brent Harbor (Columbia Law School, J.D. Class of 2027; M.F.A. in Theatre Management and Producing, Columbia University School of the Arts) for research and editorial assistance.

References and Notes

1. Association of American Publishers. Industry statistics. https://publishers.org/data-and-statistics/industry-statistics/ (2025). Accessed: July 2, 2025.

2. Publishers Weekly. Book publishing sales rose 6.5% in 2024, per preliminary data. *Publishers Weekly* (2025). URL https://www.publishersweekly.com/pw/by-topic/industry-news/financial-reporting/article/97224-book-publishing-sales-rose-6-5-in-2024-per-preliminary-data.html. Accessed: September 7, 2025.

3. Reisner, A. What i found in a database meta uses to train generative ai. *The Atlantic* URL https://www.theatlantic.com/technology/archive/2023/09/books3-ai-training-meta-copyright-infringement-lawsuit/675411/. Accessed: July 2, 2025.

4. Samuelson, P. Generative ai meets copyright. *Science* **381**, 158–161 (2023).

5. United States District Court for the Northern District of California. Order on fair use. Court Opinion No. C 24-05417 WHA, Doc. 231, United States District Court, Northern District of California. URL https://storage.courtlistener.com/recap/gov.uscourts.cand.434709/gov.uscourts.cand.434709.231.0.pdf. Accessed: July 2, 2025.

6. Gero, K. I. *et al.* Creative writers' attitudes on writing as training data for large language models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–16 (2025).

7. Noy, S. & Zhang, W. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* **381**, 187–192 (2023).

8. Trinh, T. H., Wu, Y., Le, Q. V., He, H. & Luong, T. Solving olympiad geometry without human demonstrations. *Nature* **625**, 476–482 (2024).

9. Codeforces: Programming competitions and contests, programming community. https://codeforces.com/ (2025). Accessed: September 7, 2025.

10. Kolata, G. A.i. chatbots defeated doctors at diagnosing illness. *The New York Times* URL https://www.nytimes.com/2024/11/17/health/chatgpt-ai-doctors-diagnosis.html. Accessed: July 2, 2025.

11. OpenAI. Introducing healthbench. https://openai.com/index/healthbench/ (2025). Accessed: July 2, 2025.

12. Tomlinson, K., Jaffe, S., Wang, W., Counts, S. & Suri, S. Working with ai: Measuring the occupational implications of generative ai. *arXiv preprint arXiv:2507.07935* (2025).

13. Handa, K. *et al.* Which economic tasks are performed with ai? evidence from millions of claude conversations. *arXiv preprint arXiv:2503.04761* (2025).

14. Chatterji, A. *et al.* How people use chatgpt. Working Paper 34255, National Bureau of Economic Research (2025). URL http://www.nber.org/papers/w34255.

15. U.S. Bureau of Labor Statistics. Occupational Employment and Wage Statistics: Writers and Authors. https://www.bls.gov/oes/2023/may/oes273041.html (2023). Accessed: September 7, 2025.

16. Against AI: An Open Letter from Writers to Publishers. *Literary Hub*, https://lithub.com/against-ai-an-open-letter-from-writers-to-publishers/ (2024). Accessed: September 7, 2025.

17. Chakrabarty, T., Laban, P., Agarwal, D., Muresan, S. & Wu, C.-S. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–34 (2024).

18. Chakrabarty, T., Laban, P. & Wu, C.-S. Can ai writing be salvaged? mitigating idiosyncrasies and improving human-ai alignment in the writing process through edits. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–33 (2025).

19. Doshi, A. R. & Hauser, O. P. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances* **10**, eadn5290 (2024).

20. Laquintano, T. & Vee, A. Ai and the everyday writer. *PMLA* **139**, 527–532 (2024).

21. Walsh, M., Preus, A. & Gronski, E. Does chatgpt have a poetic style? *arXiv preprint arXiv:2410.15299* (2024).

22. Vara, V. Confessions of a viral ai writer. *WIRED* (2023). URL https://www.wired.com/story/confessions-viral-ai-writer-chatgpt/. Accessed: 2025-06-21.

23. Chiang, T. Why a.I. isn't going to make art. *The New Yorker* (2024). URL https://www.newyorker.com/culture/the-weekend-essay/why-ai-isnt-going-to-make-art. The Weekend Essay.

24. Tangermann, V. Readers annoyed when fantasy novel accidentally leaves ai prompt in published version, showing request to copy another writer's style. *Futurism* (2025). URL https://futurism.com/fantasy-novel-ai-prompt-copy-style. Accessed: 2025-06-21.

25. Literary Prizes Under Scrutiny. *Poets & Writers*, https://www.pw.org/content/literary_prizes_under_scrutiny (2025). Accessed: September 7, 2025.

26. Delaney, E. J. Where great writers are made: Assessing america's top graduate writing programs. *The Atlantic* **300** (2007). URL https://www.theatlantic.com/magazine/archive/2007/08/where-great-writers-are-made/306032/. Accessed: 07 July 2025.

27. Vara, V. The a.i. romance factory. URL https://www.bloomberg.com/features/2025-ai-romance-factory/. Feature.

28. Alter, A. The new fabio is claude: How A.I. is reshaping romance novels. *The New York Times* (2026). URL https://www.nytimes.com/2026/02/08/business/ai-claude-romance-books.html. Accessed: 2026-03-09.

29. Jones, C. Amazon is the world's biggest online book marketplace. it's filled with ai knockoffs. URL https://www.rollingstone.com/culture/culture-features/amazon-ai-book-knockoffs-1235450690/. Culture Feature.

30. Knibbs, K. Scammy ai-generated book rewrites are flooding amazon. URL https://www.wired.com/story/scammy-ai-generated-books-flooding-amazon/. Business.

31. Li, Z., Liang, C., Peng, J. & Yin, M. How does the disclosure of AI assistance affect the perceptions of writing? In Al-Onaizan, Y., Bansal, M. & Chen, Y.-N. (eds.) *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 4849–4868 (Association for Computational Linguistics, Miami, Florida, USA, 2024). URL https://aclanthology.org/2024.emnlp-main.279/.

32. Sarkar, A. Ai could have written this: Birth of a classist slur in knowledge work. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–12 (2025).

33. Horton Jr, C. B., White, M. W. & Iyengar, S. S. Bias against ai art can enhance perceptions of human creativity. *Scientific reports* **13**, 19001 (2023).

34. Pustejovsky, J. E. & Tipton, E. Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics* **36**, 672–683 (2018).

35. Russell, J., Karpinska, M. & Iyyer, M. People who frequently use chatgpt for writing tasks are accurate and robust detectors of ai-generated text. *arXiv preprint arXiv:2501.15654* (2025).

36. Jabarian, B. & Imas, A. Artificial writing and automated detection (2025). URL https://ssrn.com/abstract=5407424. SSRN working paper, Abstract ID 5407424.

37. Naddaf, M. Ai tool detects llm-generated text in research papers and peer reviews. *Nature* (2025). URL https://www.nature.com/articles/d41586-025-02936-6. News.

38. Bartz et al. v. anthropic pbc (2025). URL https://www.courtlistener.com/docket/69058235/bartz-v-anthropic-pbc/. Settlement reached after court granted partial summary judgment on fair use for training but denied on piracy claims.

39. Kadrey et al. v. meta platforms, inc. (2025). URL https://law.justia.com/cases/federal/district-courts/california/candce/3:2023cv03417/415175/598/. Order denying plaintiffs' motion for partial summary judgment and granting Meta's cross-motion on fair use grounds.

40. In re mosaic llm litigation (2025). URL https://www.courtlistener.com/docket/68325564/onan-v-databricks-inc/. Consolidated cases against Databricks and MosaicML for alleged use of pirated books in training LLMs.

41. Andy warhol foundation for the visual arts, inc. v. goldsmith (2023). URL https://www.supremecourt.gov/opinions/22pdf/21-869_87ad.pdf. Holding that the first fair use factor focuses on whether the use shares the same purpose or supersedes the original work.

42. Campbell v. acuff-rose music, inc. (1994). URL https://www.supremecourt.gov/opinions/boundvolumes/510bv.pdf. Establishing that market substitution is central to fair use analysis under the fourth factor.

43. Guile, D. & Popov, J. Machine learning and human learning: a socio-cultural and-material perspective on their relationship and the implications for researching working and learning. *AI & SOCIETY* **40**, 325–338 (2025).

44. Mitchell, M. & Krakauer, D. C. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences* **120**, e2215907120 (2023). URL https://www.pnas.org/doi/10.1073/pnas.2215907120.

45. Song, Y. *et al.* Inferring neural activity before plasticity as a foundation for learning beyond backpropagation. *Nature neuroscience* **27**, 348–358 (2024).

46. U.S. Copyright Office. Copyright and artificial intelligence part 3: Generative ai training report. Tech. Rep., U.S. Copyright Office (2024). URL https://www.copyright.gov/

ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf. Pre-publication version analyzing copyright implications of AI training.

47. Walsh, M. & Antoniak, M. The goodreads "classics": a computational study of readers, amazon, and crowdsourced amateur criticism. *Journal of Cultural Analytics* **6**, 243–287 (2021).

48. Ohm, P. Focusing on fine-tuning: Understanding the four pathways for shaping generative ai. *Science and Technology Law Review* **25** (2024).

49. Roose, K. & Thompson, S. A. Who's a better writer: A.I. or humans? Take our quiz. *The New York Times* (2026). URL https://www.nytimes.com/interactive/2026/03/09/business/ai-writing-quiz.html. Interactive feature. Accessed: March 12, 2026.

50. Vara, V. What if readers like A.I.-generated fiction? *The New Yorker* (2025). URL https://www.newyorker.com/culture/the-weekend-essay/what-if-readers-like-ai-generated-fiction. The Weekend Essay. Accessed: March 12, 2026.

51. Liu, X. *et al.* Shield: Evaluation and defense strategies for copyright compliance in llm text generation. *arXiv preprint arXiv:2406.12975* (2024).

52. Chen, T. *et al.* Parapo: Aligning language models to reduce verbatim reproduction of pre-training data. *arXiv preprint arXiv:2504.14452* (2025).

53. Jaech, A. *et al.* Openai o1 system card. *arXiv preprint arXiv:2412.16720* (2024).

54. Li, X. *et al.* Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259* (2023).

55. Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81 (Association for Computational Linguistics, Barcelona, Spain, 2004). URL https://aclanthology.org/W04-1013/.

56. McDonnell, T., Lease, M., Kutlu, M. & Elsayed, T. Why is that relevant? collecting annotator rationales for relevance judgments. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 4, 139–148 (2016).

57. Just, H. A., Jin, M., Sahu, A., Phan, H. & Jia, R. Data-centric human preference optimization with rationales. *arXiv preprint arXiv:2407.14477* (2024).

58. Veselovsky, V., Ribeiro, M. H. & West, R. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899* (2023).

Supplementary Information for Readers Prefer Outputs of AI Trained on Copyrighted Books over Expert Human Writers

Tuhin Chakrabarty[1], Jane C. Ginsburg[2], Paramveer Dhillon[3,4]

[1]Department of Computer Science, Stony Brook University.

[2]Columbia Law School.

[3]School of Information Science, University of Michigan.

[4]MIT Initiative on the Digital Economy.

Corresponding authors: tchakrabarty@cs.stonybrook.edu, ginsburg@law.columbia.edu, dhillonp@umich.edu

Materials and Methods

S1: Details about Writing Task

S1.1 Author List

Table 2 shows the list of 50 authors that were chosen by English Literature Ph.D. students. These chosen author list consists of canon-plus-global giants (Ernest Hemingway, Virginia Woolf, William Faulkner, Gabriel Garcia Márquez, Stephen King, Haruki Murakami, Kazuo Ishiguro), a strong set of contemporary prominent literary voices (Margaret Atwood, Ian McEwan, Jonathan Franzen, Colson Whitehead, George Saunders, Louise Erdrich, Octavia Butler, Salman Rushdie, Maya Angelou, Percival Everett), and a group of critically acclaimed / emerging authors (Ottessa Mosfegh, Tony Tulathimutte, Roxane Gay). Additionally our author list is culturally diverse where several of the authors write primarily in a non-English language (Han Kang, Yoko Ogawa, Annie Ernaux). Roughly two-thirds (34) have secured at least one major international or national prize (e.g., Nobel, Booker, Pulitzer, National Book Award, MacArthur Fellowship, Women's Prize, International Booker, Hugo/Nebula). 8 of the authors are Nobel Prize winners in Literature and 8 are Pulitzer Prize winners.

Table 1: List of Authors

#	Author	#	Author	#	Author
1	Alice Munro	19	J.D. Salinger	37	Philip Roth
2	Annie Ernaux (✓)	20	Jhumpa Lahiri (✓)	38	Rachel Cusk
3	Annie Proulx (✓)	21	Joan Didion	39	Roxane Gay (✓)
4	Ben Lerner (✓)	22	Jonathan Franzen (✓)	40	Sally Rooney (✓)
5	Charles Bukowski	23	Junot Díaz (✓)	41	Salman Rushdie (✓)
6	Cheryl Strayed (✓)	24	Kazuo Ishiguro (✓)	42	Shirley Jackson
7	Chimamanda Ngozi Adichie (✓)	25	Louise Erdrich (✓)	43	Sigrid Nunez (✓)
8	Colson Whitehead (✓)	26	Lydia Davis (✓)	44	Stephen King
9	Cormac McCarthy	27	Margaret Atwood (✓)	45	Tony Tulathimutte (✓)
10	David Foster Wallace	28	Marilynne Robinson (✓)	46	V. S. Naipaul
11	Ernest Hemingway	29	Maya Angelou	47	Virginia Woolf
12	Flannery O'Connor	30	Milan Kundera	48	William Faulkner
13	Gabriel García Márquez	31	Min Jin Lee (✓)	49	Yoko Ogawa (✓)
14	George Saunders (✓)	32	Nora Ephron	50	Zadie Smith (✓)
15	Han Kang (✓)	33	Octavia Butler		
16	Haruki Murakami (✓)	34	Orhan Pamuk (✓)		
17	Hunter S. Thompson	35	Ottessa Moshfegh (✓)		
18	Ian McEwan (✓)	36	Percival Everett (✓)		

Table 2: Author list for our pool of 50 authors. (✓) denotes authors who were used in fine-tuning experiment

S1.2 Writing Prompt

Our writing prompt can be seen in Figure 5). The same prompt was provided to both experts(MFA candidates) and LLMs.

S1.3 Fine-tuning details

For fine-tuning we bought digital ePub versions of these authors' books and transformed them into plain text files. If the epub file was basically a wrapper around scanned page images, then we ignored that epub. The number of books written by each author vary a lot. For instance *Tony Tulathimutte* has written only two books *Private Citizens* and *Rejection* so we could only fine-tune GPT-4o on two of them. While for *Haruki Murakami* we could fine-tune on 22 books *A Wild Sheep Chase*, *After Dark*, *After the Quake*, *Blind Willow, Sleeping Woman*, *Colorless Tsukuru Tazaki and his Years of Pilgrimage*, *Dance Dance Dance*, *First Person Singular*, *Hard-boiled Wonderland and the End of the World*, *Hear the Wind Sing*, *Kafka on the Shore*, *Killing Commendatore*, *Men Without Women*, *Norwegian Wood*, *Novelist as a Vocation*, *One and Two*, *Pinball, 1973*, *South of the Border,*

Writing Prompt

You will be given 20 excerpts written by Chimamanda Ngozi Adichie. Very carefully observe their style and voice. Then you will be given some content and you need to write an excerpt about that content using the style and voice of this author

Excerpt_1:
Princeton, in the summer, smelled of nothing [...] But she did not like that she had to go to Trenton to braid her hair.

Excerpt_2:
"How are you?" Aunty Ifeka asked [...] "Your life belongs to you and you alone, soso gi."

Excerpt_3:
Amaka used to write the office of the Head of State [...] It is like telling a crawling baby who tries to walk, and then falls back on his buttocks, to stay there.

Excerpt_4:
We teach girls to shrink themselves, to make themselves smaller [...] I know an unmarried woman in Nigeria who, when she goes to conferences, wears a wedding ring because she wants her colleagues to—according to her—"give her respect."

Excerpt_5:
The weeks passed and the water started running again [...] Since she had lost her savings, then she could not possibly lose her sister too; the custodians of fate were not that wicked.

Excerpt_6:
Grief is a cruel kind of education [...] One morning, Okey calls me a little earlier than usual and I think, Just tell me, tell me immediately, who has died now.

Excerpt_7:
"I sifted through my memories, as though through debris, trying to find a reason [...] "I just don't understand it. It's as if an artery burst inside him and suddenly his whole body is wired differently and he is no longer the person he was," I told Mmiliaku.

Excerpt_8:
So this guy said to Professor Hunk, "White privilege is nonsense. How can I be privileged?" [...] And Professor Hunk replied—that is exactly what white privilege is, that you can say that. Race doesn't really exist for you because it has never been a barrier.

Excerpt_9:
It was the last summer you spent in Nigeria, the summer before your parents' divorce [...] Grandmama told you the snake was called the echi eteka, "Tomorrow Is Too Far."

Excerpt_10:
Before, our driver, Kevin, would pick me up first at Daughters of the Immaculate Heart [...] "Mama is pregnant," I said. "We will take care of the baby; we will protect him."

Excerpt_11:
She stood in line outside the American embassy in Lagos [...] the splash on his chest so red she wanted to scold him about playing with the palm oil in the kitchen. He was only four years old.

Excerpt_12:
"What bigger picture?" Master asked [...] "I am Nigerian because a white man created Nigeria and gave me that identity. But I was Igbo before the white man came."

Excerpt_13:
I have always longed to be known, truly known, by another human being [...] On Zoom calls, everyone echoed, reaching but not touching, the distance between us all further hollowed out.

Excerpt_14:
Why does the image of two red butterflies on a T-shirt make me cry? [...] Grief has, as one of its many egregious components, the onset of doubt. Yes, my father truly was lovely.

Excerpt_15:
Inside the house, Obinna went straight to the children's room [...] He wanted to stop thinking of what could have been, of another life he could have had. And he wanted Amara to come home.

Excerpt_16:
That evening she received an e-mail: YOUR TALK WAS BALONEY. YOU ARE A RACIST [...] Her post about trying to date online, "What's Love Got to Do with It?" continued to draw comments, like something sticky, after many months.

Excerpt_17:
Okoloma was one of my greatest childhood friends [...] He told me that people were saying my novel was feminist, and his advice to me was that I should never call myself a feminist since feminists are women who are unhappy because they cannot find husbands. So I decided to call myself a Happy Feminist.

Excerpt_18:
"You did not tie your scarf well," Jaja says to Mama [...] "We'll plant new orange trees in Abba when we come back, and Jaja will plant purple hibiscus, too, and I'll plant ixora so we can suck the juices of the flowers."

Excerpt_19:
Olanna had wanted to give the scent of his mother's visit some time to diffuse [...] he whispered, "We will have a brilliant child, nkem, a brilliant child," and she said, Yes, yes. But they did not conceive a child, she knew.

Excerpt_20:
He was attractive, the kind of attractiveness that grew on you [...] "Or we could act. I want to act." He paused, shifted. "Ifem, I'm chasing you."

Stylistic Features:

The author uses both Igbo and English in their works, with Igbo phrases shown in italics and followed by an English translation [...] In developing characters, the author often exaggerates attitudes to contrast the differences between traditional and western cultures [...] The author's style often focuses on strong women and adds a gendered perspective to topics previously explored by other authors, such as colonialism, religion, and power relationships [...] They often incorporate humour, and use anecdotes, irony and satire to underscore a particular point of view. Their works about African diaspora consistently examine themes of belonging, adaptation and discrimination.

Content:

This excerpt is written primarily in the third person, but it contains a lengthy first-person monologue delivered by the character Ifemelu. Ifemelu, emboldened by white wine, speaks candidly about race, specifically the experience of being Black in America and dating a white person [...] Her words are bold, unfiltered, and challenge the comfort of her "nice liberal" audience. The scene unfolds at a dinner party attended by liberal guests [...] The excerpt ends as Ifemelu, unsure of her standing but spurred on by attention and drink, prepares to share more, this time about her ex-boyfriend Curt. This moment signals an impending emotional and narrative unraveling, rooted in race, intimacy, and truth-telling.

Now write a 400 word excerpt about the content above emulating the style and voice of the author as seen in the 20 excerpts and the mentioned stylistic features.

Figure 5: In-Context Writing Prompt used in AI Condition 1.

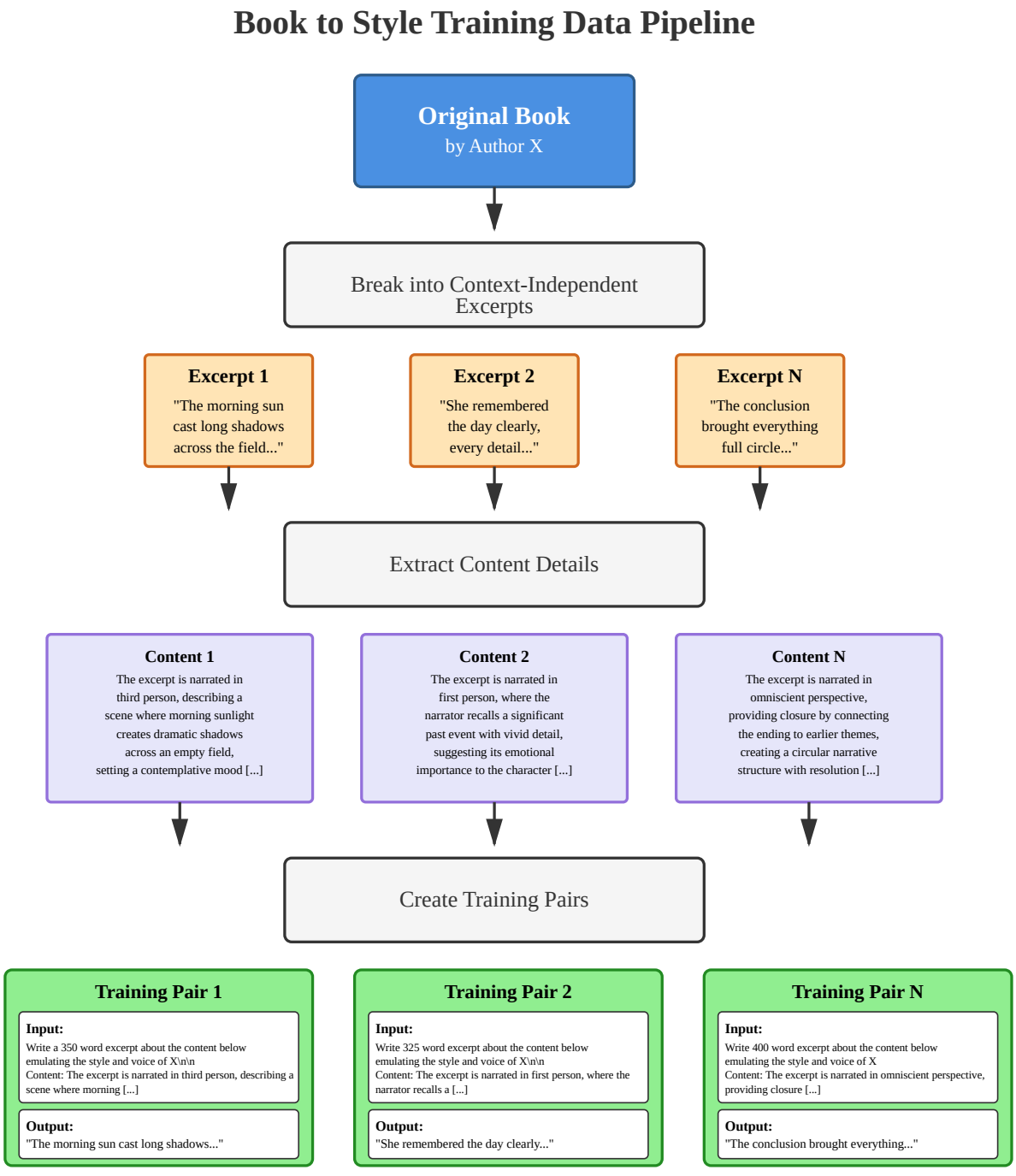


Figure 6: The pipeline used to fine-tune ChatGPT on an author's entire oeuvre

West of the Sun, *Sputnik Sweetheart*, *The City and Its Uncertain Walls*, *The Elephant Vanishes*, *The Wind-Up Bird Chronicle*, *What I Talk About When I Talk About Running*, *Wind/Pinball: Two Novels*

Our entire fine-tuning pipeline can be seen in Figure 6. We first convert the epub files to txt using `https://github.com/kevinboone/epub2txt2`. We then segment the entire book into context independent excerpts. At a first pass we naively split the entire book text by existing double-newlines and rejoin them to enforce excerpt size bounds(250-650 words). For the rare cases where the naive splitting would lead to excerpts longer than 650 words we used GPT-4o to segment them again using the prompt *Segment it into excerpts of minimum length 300-350 words such that each excerpt is grammatical from the start and doesn't feel abruptly cut off. There should be zero deletion and break into excerpts at grammatically natural places. Maintain the original word count. Avoid breaking into too many small excerpts. Start directly. Don't say Here's or Here is ....*

After fine-tuning is completed at inference time we can then simply generate an excerpt conditioned on a similar instruction that contains a novel content. We specifically excluded the original author written excerpt used for our experiments during fine-tuning so that the model does not get an unfair advantage. We also ensured that the generated outputs do not contain any memorized snippets from the original author written excerpt. In the rare occasion that a model regurgitated verbatim snippets/ngrams from original author written text we resampled it and manually verified that there is no verbatim overlap before using it for evaluation. While supervised fine-tuning in most cases ensures that the generated output contains all the information in the extracted content details, in the rare occasion that it does not we resample/regenerate it again. Last but not least sometimes supervised fine-tuning can lead to ungrammatical output or output with minor inconsistencies. To make sure these mistakes don't impact human evaluation, we performed a post processing step using GPT-4o using the following prompt *Fix grammar, tense, typo, spelling or punctuation error or any other awkward construction/ logical inconsistency*

For *Margaret Atwood, Kazuo Ishiguro, Salman Rushdie and Haruki Murakami* we fine-tuned GPT-4o for 1 epoch as they had multiple books/longer books. For the rest of the authors we fine-tuned for 3 epochs. We use default fine-tuning parameters set by OpenAI in addition to setting batch size as 1 and LR multiplier as 2. Figure 19 shows that fine-tuned models don't regurgitate a lot of verbatim text from the training corpus (author's entire oeuvre). We should also note that some of

Excerpt

My father tried to kill my mother one Sunday in June, in the early afternoon. I had been to Mass at a quarter to twelve as usual. I must have brought back some cakes from the baker in the new shopping precinct—a cluster of temporary buildings erected after the war while reconstruction was under way. When I got home, I took off my Sunday clothes and slipped on a dress that washed easily. After the customers had left and the shutters had been pinned down over the store window, we had lunch, probably with the radio on, because at that hour there was a funny program called Courtroom, in which Yves Deniaud played some wretched subordinate continually charged with the most preposterous offenses and condemned to ridiculous sentences by a judge with a quavering voice. My mother was in a bad temper. The argument she started with my father as soon as she sat down lasted throughout the meal. After the table was cleared and the oilcloth wiped clean, she continued to fire criticism at my father, turning round and round in the tiny kitchen—squeezed in between the café, the store and the steps leading upstairs—as she always did when she was upset. My father was still seated at the table, saying nothing, his head turned toward the window. Suddenly he began to wheeze and was seized with convulsive shaking. He stood up and I saw him grab hold of my mother and drag her through the café, shouting in a hoarse, unfamiliar voice. I rushed upstairs and threw myself on to the bed, my face buried in a cushion. Then I heard my mother scream: "My daughter!" Her voice came from the cellar adjoining the café. I rushed downstairs, shouting "Help!" as loud as I could. In the poorly-lit cellar, my father had grabbed my mother by the shoulders, or maybe the neck. In his other hand, he was holding the scythe for cutting firewood which he had wrenched away from the block where it belonged. At this point all I can remember are sobs and screams. Then the three of us are back in the kitchen again. My father is sitting by the window, my mother is standing near the cooker and I am crouching at the foot of the stairs. I can't stop crying. My father wasn't his normal self; his hands were still trembling and he had that unfamiliar voice. He kept on repeating, "Why are you crying? I didn't do anything to you." I can recall saying this sentence, "You'll breathe disaster on me. My mother was saying, "Come on, it's over."

Extract Content

The paragraph is written in the first person, primarily from the perspective of a daughter recounting a traumatic event involving her parents. The main characters are the narrator, her father, and her mother. The father attempted to harm the mother one Sunday in June. The narrator had returned home from attending Mass and having lunch with her parents when an argument between her parents escalated. The mother, visibly upset, continued to criticize the father throughout the meal and afterward. The situation intensified when the father, after being silent, suddenly began to shake and shouted in a hoarse voice, dragging the mother through the café. The narrator fled upstairs but soon heard her mother scream from the cellar, prompting her to call for help. In the dimly lit cellar, the father had the mother by the shoulders or neck and held a scythe, intending harm. The scene ends with all three back in the kitchen, the narrator sobbing, the father acting unusually, and the mother trying to calm the situation by insisting it's over.

Figure 7: Original author written excerpt and extracted content using an LLM



Figure 8: Total training token distribution for each author

these words are also part of the Content in the instruction which would anyway appear in the text and should not be penalized. For all generation(prompting or fine-tuning default temperature=1.0 was used)

S1.4 Why recruit MFAs as expert writers

As mentioned earlier MFA is typically recognized as a terminal degree for practitioners of visual art, design, dance, photography, theatre, film/video, new media, and creative writing—meaning that it is considered the highest degree in its field, qualifying an individual to become a professor at the university level in these disciplines. As can be seen in Figures 10, 9 and 11 most top writers have an MFA degree. While there are exceptional writers without any formal training the space for such individuals is enormous and difficult to systematically sample, whereas MFA graduates can be more readily identified and tracked through institutional records.

Author	MFA Program	Primary Category	Major Award(s)
Ada Limón	NYU	Poetry	NBCC; U.S. Poet Laureate
Adam Haslett	Iowa	Fiction	Pulitzer & NBA Finalist
Adam Johnson	McNeese State	Fiction	Pulitzer Prize; National Book Award
Aimee Bender	UC Irvine	Fiction	O. Henry; World Fantasy Finalist
Alexander Chee	Iowa	Fiction	Whiting Award
Andrew Porter	Iowa	Fiction	Flannery O'Connor Award
Andrew Sean Greer	Montana	Fiction	Pulitzer Prize
Angela Flournoy	Iowa	Fiction	NBA Finalist
Ann Patchett	Iowa	Fiction	PEN/Faulkner; Women's Prize
Anne Boyer	Wichita State	Nonfiction	Pulitzer Prize (General Nonfiction)
Anthony Doerr	Bowling Green	Fiction	Pulitzer Prize; Carnegie Medal
Anthony Marra	Iowa	Fiction	NBCC John Leonard; Whiting
Ben Lerner	Brown	Fiction/Poetry	MacArthur; NBCC (Poetry)
Brandon Taylor	Iowa	Fiction	Booker Shortlist
Carmen Maria Machado	Iowa	Fiction/Nonfiction	Lambda; NBA Finalist
Celeste Ng	Michigan	Fiction	ALA Medal; Mass. Book Award
Chang-rae Lee	Michigan	Fiction	PEN/Hemingway; Dayton
Cheryl Strayed	Oregon	Nonfiction	Alex Award; Film adaptation (Wild)
Chimamanda Ngozi Adichie	Johns Hopkins	Fiction	Women's Prize (Orange); NBCC
Chris Adrian	Iowa	Fiction	Guggenheim; O. Henry
Curtis Sittenfeld	Iowa	Fiction	James Tait Black shortlist
Daniel Alarcón	Iowa	Fiction	PEN/Faulkner Finalist; MacArthur
Denis Johnson	Iowa	Fiction	National Book Award
Edwidge Danticat	Brown	Fiction/Nonfiction	NBCC; American Book Award; MacArthur
Elissa Washuta	Iowa	Nonfiction	PNW Book Award; Lambda Finalist
Elizabeth Acevedo	University of Maryland	Fiction/YA	National Book Award
Elizabeth Strout	Syracuse	Fiction	Pulitzer; Booker Shortlist
Emily Ruskovich	Iowa	Fiction	International Dublin Literary Award
Eula Biss	Iowa	Nonfiction	NBCC Award
Flannery O'Connor	Iowa	Fiction	National Book Award
Garth Greenwell	Iowa	Fiction/Nonfiction	NBCC Finalist; British Book Award Shortlist
George Saunders	Syracuse	Fiction	Booker Prize; Story Prize
Heidi Julavits	Columbia	Fiction/Nonfiction	PEN Syndicated Fiction
Jamel Brinkley	Iowa	Fiction	Story Prize Finalist; NBCC John Leonard Finalist
Jericho Brown	University of New Orleans	Poetry	Pulitzer Prize (Poetry)
Jesmyn Ward	Michigan	Fiction	National Book Award (×2)
Jhumpa Lahiri	Boston University	Fiction	Pulitzer Prize
Jo Ann Beard	Iowa	Nonfiction	Whiting; Guggenheim
John D'Agata	Iowa	Nonfiction	Pushcart; NEA
John Irving	Iowa	Fiction	National Book Award
Jorie Graham	Iowa	Poetry	Pulitzer Prize (Poetry)
Joshua Ferris	UC Irvine	Fiction	PEN/Hemingway; NBA Finalist
Joy Harjo	Iowa	Poetry	U.S. Poet Laureate; Ruth Lilly
Julie Buntin	NYU	Fiction	Notable Book lists
Julie Otsuka	Columbia	Fiction	PEN/Faulkner Award
Julie Schumacher	Cornell	Fiction	Thurber Prize
Junot Díaz	Cornell	Fiction	Pulitzer Prize
Karan Mahajan	Michigan	Fiction	NBA Finalist
Karen Russell	Columbia	Fiction	MacArthur; Pulitzer Finalist
Kelly Link	UNC Greensboro	Fiction	MacArthur; World Fantasy
Kevin Brockmeier	Iowa	Fiction	O. Henry (×multiple)
Kevin Powers	UT Austin (Michener)	Fiction	PEN/Hemingway; Guardian First Book Award
Kevin Wilson	UT Austin (Michener)	Fiction	Shirley Jackson Award
Kevin Young	Brown	Poetry/Nonfiction	Lenore Marshall Prize
Kiese Laymon	Indiana University	Nonfiction	Andrew Carnegie Medal
Lauren Groff	Wisconsin–Madison	Fiction	NBA Finalist; Prix de l'Académie de Bretagne
Leslie Jamison	Iowa	Nonfiction	NBCC Finalist; Guggenheim
Ling Ma	Cornell	Fiction	Kirkus Prize (Fiction)
Lorrie Moore	Cornell	Fiction	Irish Times Int'l Fiction Prize
Marilyn Chin	Iowa	Poetry	Anisfield-Wolf Award
Mary Karr	Goddard	Nonfiction	NBCC (Memoir)
Meghan Daum	Columbia	Nonfiction	National Magazine Award; Guggenheim
Melissa Febos	Sarah Lawrence	Nonfiction	NBCC (Criticism); Lambda
Michael Chabon	UC Irvine	Fiction	Pulitzer Prize
Michael Cunningham	Iowa	Fiction	Pulitzer; PEN/Faulkner
Nam Le	Iowa	Fiction	Dylan Thomas Prize
Natasha Trethewey	UMass Amherst	Poetry	Pulitzer Prize (Poetry)
Nathan Englander	Iowa	Fiction	Frank O'Connor Award; PEN/Malamud
Nathan Hill	UMass Amherst	Fiction	L.A. Times First Fiction
Nick Flynn	NYU	Nonfiction/Poetry	PEN/Joyce Osterweil; Guggenheim
Ocean Vuong	NYU	Fiction/Poetry	T.S. Eliot Prize; MacArthur
Ottessa Moshfegh	Brown	Fiction	Booker Shortlist; PEN/Hemingway
Paul Beatty	Brooklyn College	Fiction	Booker Prize
Paul Harding	Iowa	Fiction	Pulitzer Prize
Phil Klay	Hunter College	Fiction	National Book Award
Rachel Kushner	Columbia	Fiction	NBA Finalist; Prix Médicis Étranger
Richard Ford	UC Irvine	Fiction	Pulitzer; PEN/Faulkner
Rita Dove	Iowa	Poetry	Pulitzer Prize (Poetry)
Sigrid Nunez	Columbia	Fiction	National Book Award
Stuart Dybek	Iowa	Fiction	MacArthur; PEN/Malamud
T.C. Boyle	Iowa	Fiction	PEN/Faulkner Award
Tea Obreht	Cornell	Fiction	Women's Prize (Orange)
Tommy Orange	IAIA	Fiction	PEN/Hemingway; Pulitzer Finalist
Tony Tulathimutte	Iowa	Fiction	Whiting Award; O. Henry; NBA Longlist
Tracy K. Smith	Columbia	Poetry	Pulitzer Prize (Poetry)
Yaa Gyasi	Iowa	Fiction	NBCC John Leonard; PEN/Hemingway Finalist
Yiyun Li	Iowa	Fiction	MacArthur; PEN/Hemingway
ZZ Packer	Iowa	Fiction	PEN/Malamud

Figure 9: Table showing how several award winning authors from US all had MFA degree

School	Writer	Distinctions
Boston University	Jhumpa Lahiri	Pulitzer Prize for *Interpreter of Maladies*
	Weike Wang	PEN/Hemingway; Whiting
	Lisa Taddeo	2× Pushcart; #1 NYT bestseller
Stony Brook	Helen Simonson	NYT bestseller *Major Pettigrew's Last Stand*
	Genevieve Sly Crane	Whiting Award
	Vanessa Cuti	Best American Short Stories 2021
University of Michigan	Jesmyn Ward	2× National Book Award; MacArthur
	Celeste Ng	ALA Alex Award; NYT bestseller
	Chigozie Obioma	2× Booker finalist
	Elizabeth Kostova	Hopwood Awards; *The Historian*
NYU	Ocean Vuong	T. S. Eliot Prize; MacArthur
	Garth Risk Hallberg	*City on Fire*; NBCC finalist
Columbia University	Sigrid Nunez	National Book Award
	Julie Otsuka	PEN/Faulkner
	Victor LaValle	World Fantasy; Shirley Jackson
	Emma Cline	Plimpton Prize; NBCC finalist
	Rivka Galchen	Rona Jaffe; New Yorker staff writer
Iowa Writers' Workshop	Flannery O'Connor	National Book Award (*Collected Stories*)
	John Irving	Oscar (screenplay); global bestseller
	T. C. Boyle	PEN/Faulkner
	Paul Harding	Pulitzer Prize
	Yiyun Li	PEN/Hemingway; MacArthur
	Carmen Maria Machado	NBCC John Leonard; Folio Prize
	Garth Greenwell	PEN/Faulkner (2025)
	Anthony Marra	NBCC John Leonard; Anisfield-Wolf

Figure 10: Table showing a few accomplished alumni with MFA degree from programs where we recruited our participants

NewYorker's 20 Under 40 List with Age and MFA Breakdown

Chimamanda Ngozi Adichie, 32 — Johns Hopkins
Chris Adrian, 39 — Iowa
Daniel Alarcón, 33 — Iowa
David Bezmozgis, 37 — No Writing MFA
Sarah Shun-lien Bynum, 38 — Iowa
Joshua Ferris, 35 — UCI
Jonathan Safran Foer, 33 — No MFA
Nell Freudenberger, 35 — NYU
Rivka Galchen, 34 — Columbia
Nicole Krauss, 35 — No MFA
Yiyun Li, 37 — Iowa
Dinaw Mengestu, 31 — Columbia
Philipp Meyer, 36 — Michener Center
C. E. Morgan, 33 — No MFA
Téa Obreht, 24 — Cornell
Z Z Packer, 37 — Iowa
Karen Russell, 28 — Columbia
Salvatore Scibona, 35 — Iowa
Gary Shteyngart, 37 — Hunter
Wells Tower, 37 — Columbia

Figure 11: Table showing breakdown of New Yorkers 20 under 40 Fiction List

S2: Details about Evaluation Task

S2.1 Evaluation Interface

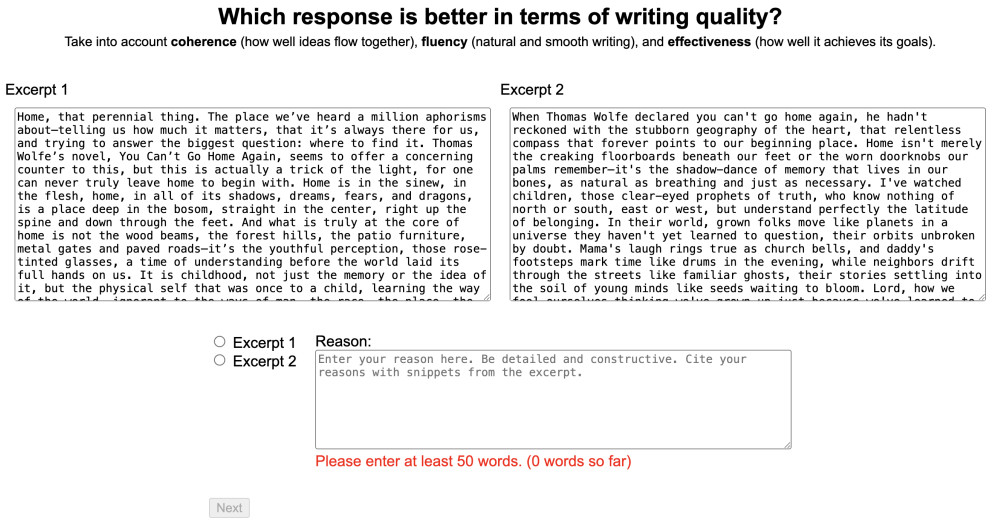


Figure 12: Evaluation Interface for writing quality evaluation where readers just see two excerpts without any disclaimer of its source and choose their preference supported by a reason

Figures 12 and 13 shows the evaluation interface shown to both MFA-trained and college-educated general readers. Readers read two excerpts for quality evaluation and three excerpts for stylistic fidelity evaluation given that the stylistic emulation is evaluated with respect to the original author written excerpt.

S2.2 How much is faithful style/voice emulation important for good writing?

As can be seen in Figure 14 writing quality has a strong correlation with a faithful style/voice fidelity for In-context Prompting condition compared to Fine-tuning. The high expert correlation for In-context Prompting suggests that models cannot emulate an author's style properly just by prompting and that experts use stylistic cues as a major factor in quality judgments - likely detecting "AI-ness" (clichés, purple prose, too much exposition, lack of subtext, mixed metaphors) in the style that makes them rate quality lower too. With fine-tuning, experts can appreciate quality independent of stylistic fidelity, suggesting that as AI loses its tell-tale signs experts can evaluate each dimension on its own merits. For college-educated general readers the correlation only drops slightly from 0.52 (In-context prompting) to 0.45 (fine-tuned), compared to experts' dramatic drop from 0.64

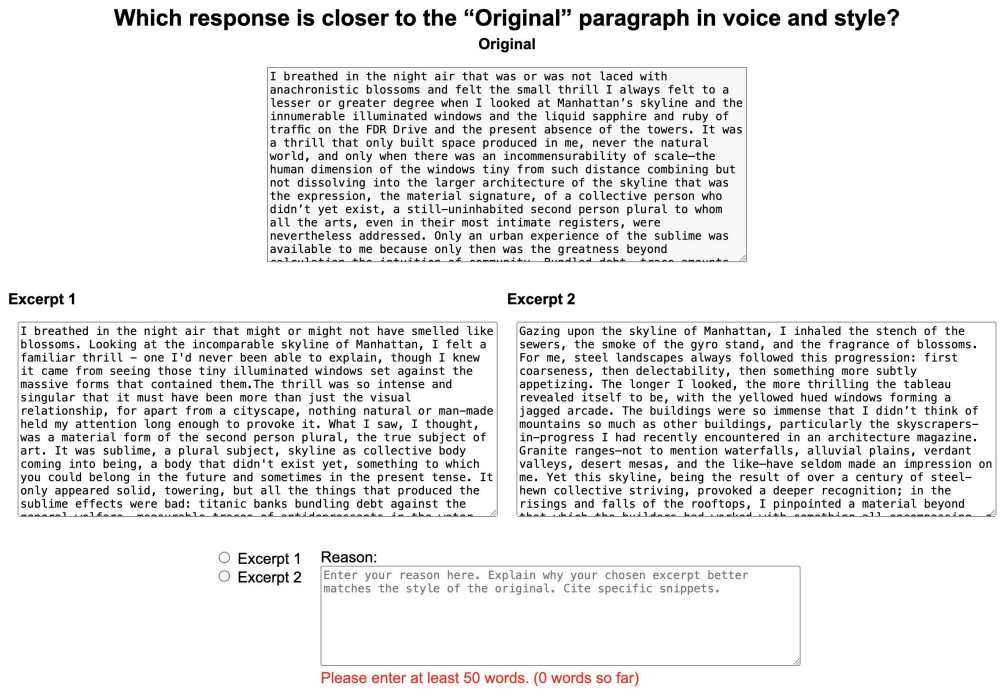
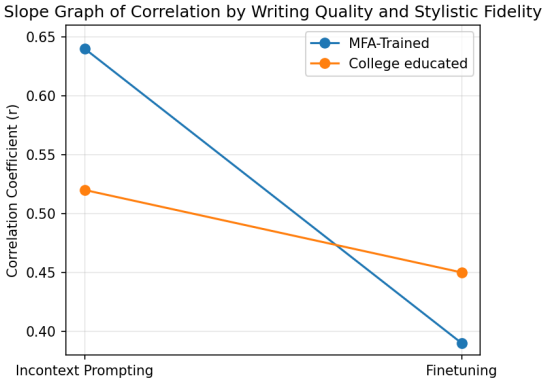


Figure 13: Evaluation Interface for stylistic fidelity evaluation where readers just see three excerpts (original and two emulations) without any disclaimer of its source and choose their preference supported by a reason



	Finetuning	Incontext Prompting
MFA-Trained	8.7e-08	9.6e-36
College educated	2.7e-10	2.1e-22

p-values

Figure 14: Pearson Correlation Coefficient between writing quality and style judgments

to 0.39. This also means that college-educated general readers might focus on more surface-level features such as grammar, clarity or flow rather than subtle stylistic markers that distinguish AI from human writing.

S2.4 Example of emulations

Figure 15,16, 17 and 18 shows emulations of the original excerpt written by Colson Whitehead, Ottessa Moshfegh, Jonathan Franzen and Han Kang. There are 3 distinct MFA emulations per author (MFA1, MFA2, MFA3). For the In-context Prompting setup we prompt Claude 3.5 Sonnet (AI1), Gemini 1.5 Pro (AI2) and GPT-4o(AI3) to produce excerpts that can be pitted against MFA written excerpts. In fine-tuning setup we fine-tune GPT-4o and compare the fine-tuned output AI4 against MFA1, MFA2, MFA3. The output from fine-tuned GPT-4o (AI4) was generated using default hyperparameters.

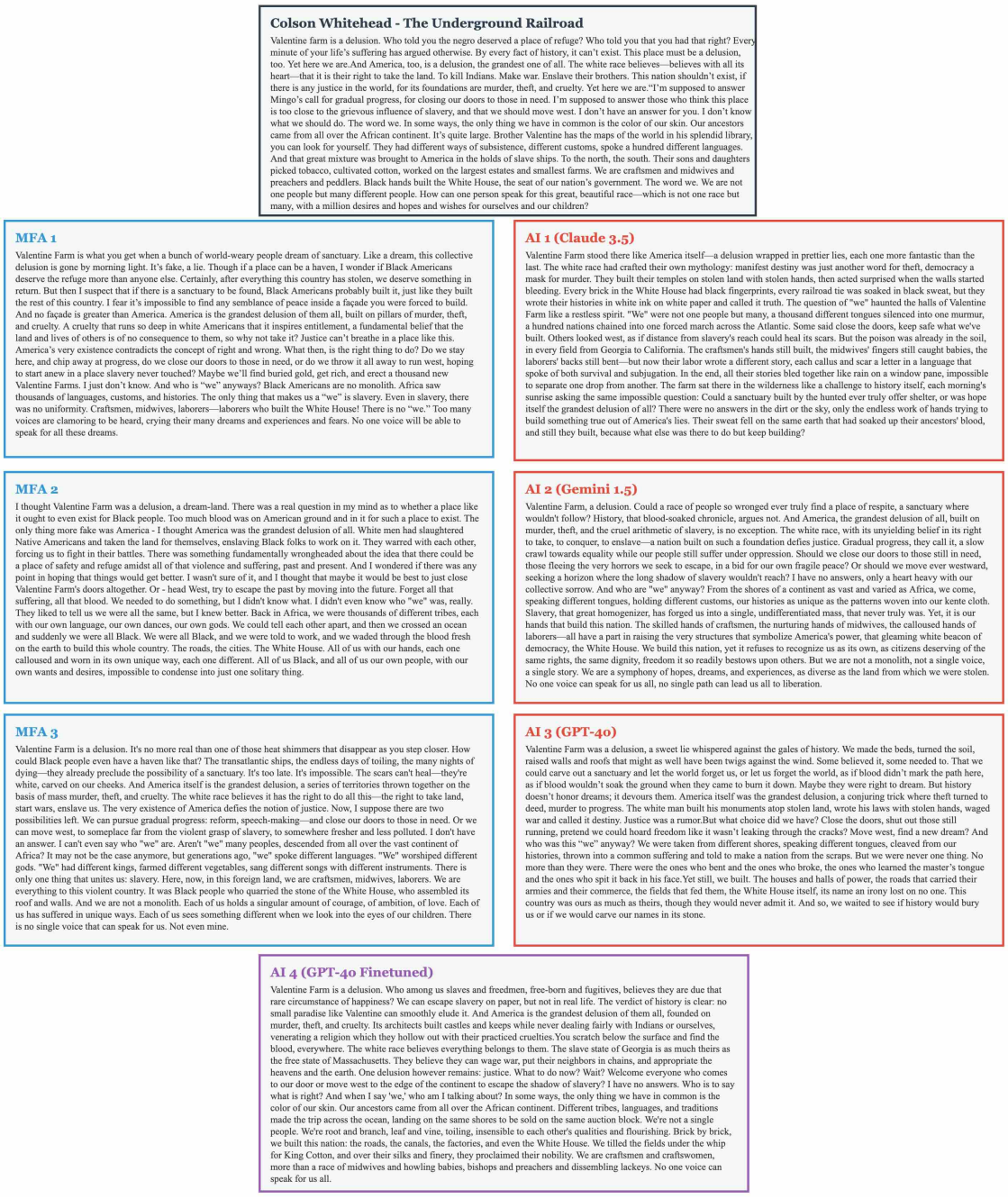


Figure 15: In AI Condition 1(In-context prompt) we contrast MFA1/2/3 vs AI1/2/3. In AI Condition 2(Fine-tuning) we contrast MFA1/2/3 vs AI4



Figure 16: In AI Condition 1(In-context prompt) we contrast MFA1/2/3 vs AI1/2/3. In AI Condition 2(Fine-tuning) we contrast MFA1/2/3 vs AI4



Figure 17: In AI Condition 1(In-context prompt) we contrast MFA1/2/3 vs AI1/2/3. In AI Condition 2(Fine-tuning) we contrast MFA1/2/3 vs AI4



Figure 18: In AI Condition 1(In-context prompt) we contrast MFA1/2/3 vs AI1/2/3. In AI Condition 2(Fine-tuning) we contrast MFA1/2/3 vs AI4

Writer	Overlap Words	Phrases	Overlap %
Kazuo Ishiguro	20	4	10.2%
Ian McEwan	22	4	9.5%
Colson Whitehead	30	2	9.2%
Haruki Murakami	20	2	8.2%
Chimamanda Ngozi Adichie	33	5	7.8%
Annie Ernaux	15	3	6.2%
Roxanne Gay	23	4	5.9%
Sigrid Nunez	15	3	5.2%
Margaret Atwood	15	3	4.9%
Salley Rooney	15	3	4.8%
Lydia Davis	10	2	4.8%
Percival Everett	18	3	4.7%
Zadie Smith	11	2	4.7%
Jhumpa Lahiri	10	2	4.4%
Orhan Pamuk	13	2	4.0%
Min Jin Lee	11	2	3.2%
Junot Diaz	10	2	3.2%
Jonathan Franzen	11	2	3.0%
Han Kang	5	1	2.5%
Cheryl Strayed	5	1	2.4%
Salman Rushdie	5	1	2.3%
Annie Proulx	5	1	1.8%
Yoko Ogawa	5	1	1.6%
Rachel Cusk	5	1	1.4%
Marilynne Robinson	5	1	1.3%
George Saunders	0	0	0.0%
Ottessa Mosfegh	0	0	0.0%
Tony Tulathimutte	0	0	0.0%
Ben Lerner	0	0	0.0%
Louise Erdrich	0	0	0.0%

Figure 19: Total overlap percentage calculated by number of words in generated output that come from authors overall corpus

S2.5 Dissecting Preference Evaluations from MFA-trained readers

Figure 20 and 21 shows the writing quality preference evaluation results from expert readers. Expert preferences are often grounded in solid reasoning and they often agree on similar snippets to support their reasoning. For instance look at **uncooked spaghetti** as an overwrought/awkward metaphor highlighted by both MFA-trained Reader 1 and 3 while **tarot card predicting his fate** highlighted by MFA-trained Reader 2 and 3. Grounding preferences in reasoning helps us uncover their mental models. It is also worth noticing that the preference often shifts towards AI once it is fine-tuned on authors entire oeuvre. Fine-tuning helps get rid of the officious disembodied robovoice that is characteristic of ChatGPT. Here experts praise the model for idiosyncratic narrative voice. By default due to post-training guardrails GPT-4o generates a rather safe and somewhat polite text in the in-context prompting setup. But what is particularly impressive here is when fine-tuned the vocabulary of the model generated text veers towards ribald and somewhat profane text that is characteristic of Tulathimutte's voice (particularly Rejection from where the text is drawn). In Figure 22 we see that experts prefer the MFA emulation as more faithful to Junot Diaz's voice. In fact the excerpt written by Gemini-1.5-Pro is oddly poetic and rife with purple prose and cliches which is nothing like Junot Diaz's voice that is often marked by a dazzling hash of Spanish, English, slang, literary flourishes, and pure virginal dorkiness as noted by all three MFA-trained readers. However when fine-tuned on authors' entire oeuvre (Figure 23) we see that the model learns these references and uses them in a rather wry and humorous manner (*papi chulo*, *fifty pendejas*, *little pito off*, *las plagas*) that is appreciated by MFA-trained readers leading them to prefer AI over MFAs.

S3: Details about AI detection and Stylometry

S3.1 AI detection thresholds

Based on AI likelihood scores from Pangram and GPTZero across 330 (150 MFA, 150 In-context AI, 30 Fine-tuned AI) evaluations we see that Pangram is bimodal. This means that AI likelihood scores are 0 most of the times for human written text. While there are few spikes, a conservative threshold of 0.9 drastically reduces any chance of a False Positive. Similarly, AI likelihood scores are 1.0 most of the times for AI written text with few exceptions. GPTZero is less accurate compared



Figure 20: MFA vs AI emulation of Tony Tulathimutte, read by 3 expert(MFA-trained) readers all of whom agree that MFA written excerpt is superior in terms of writing quality.

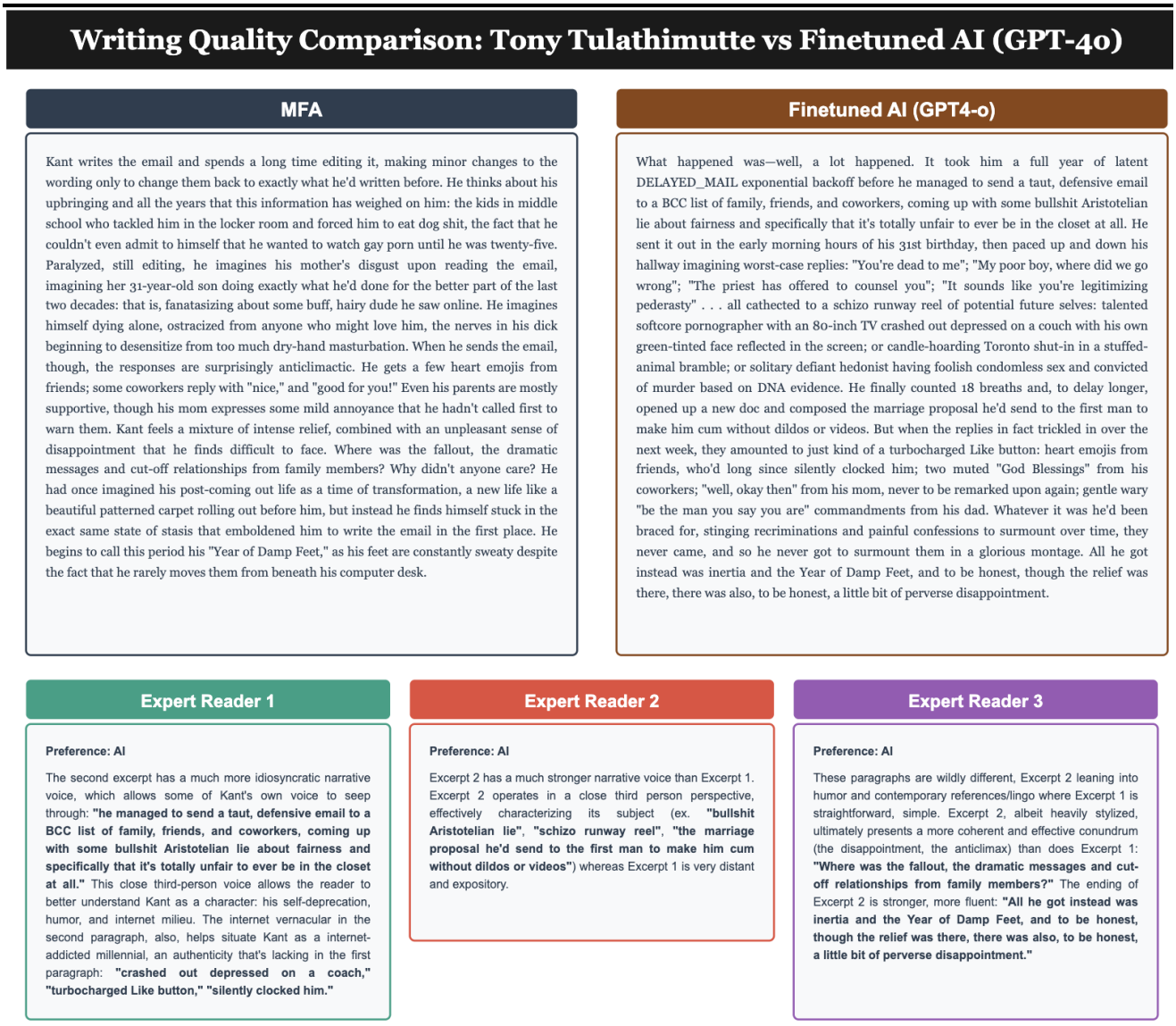


Figure 21: MFA vs AI emulation of Tony Tulathimutte, judged by 3 expert(MFA-trained) readers all of whom agree that Fine-tuned AI written excerpt is superior in terms of writing quality



Figure 22: MFA vs AI emulation of Junot Diaz, judged by 3 expert readers(MFA-trained) all of whom agree that MFA written excerpt is superior in terms of stylistic fidelity.

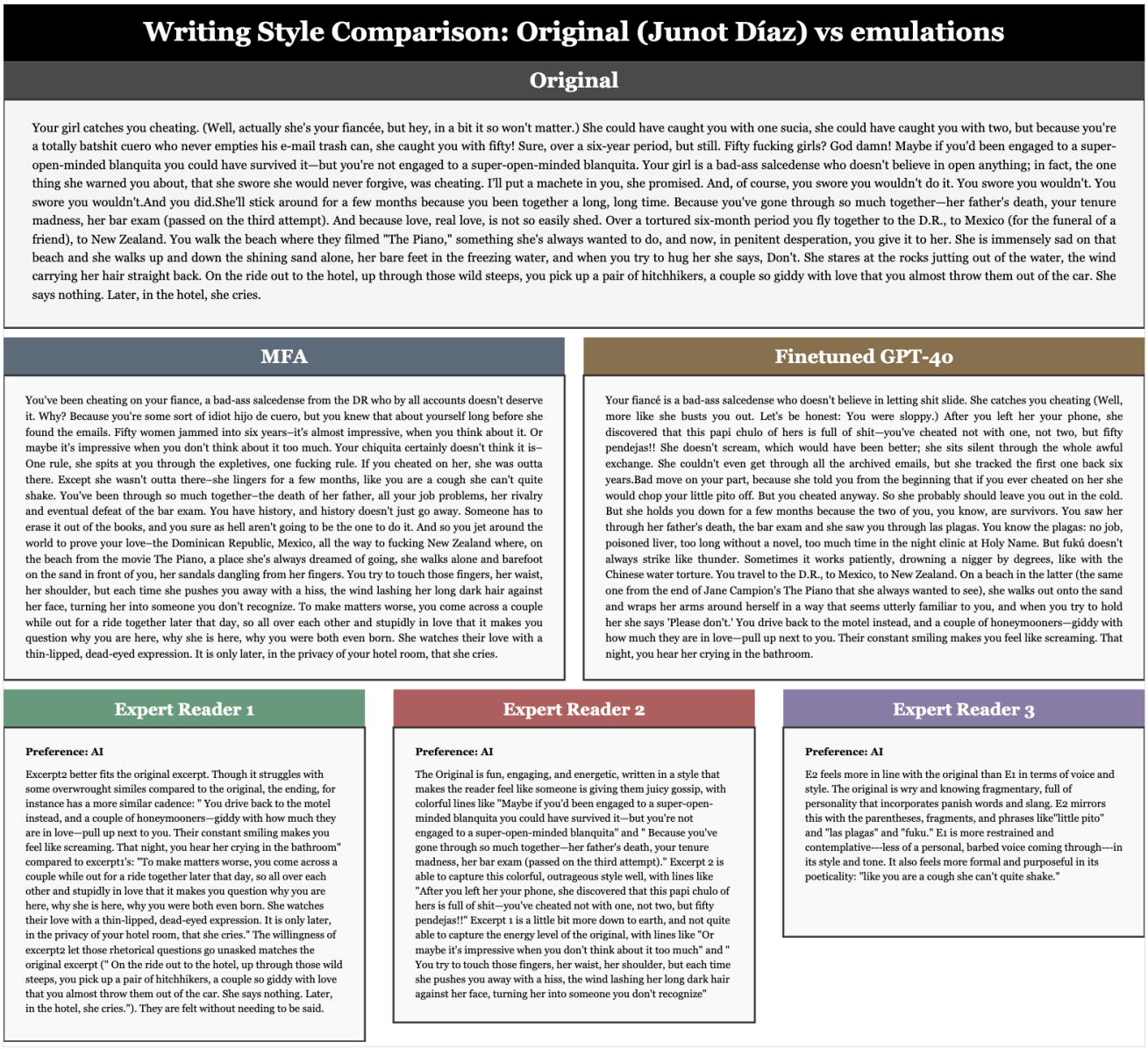


Figure 23: MFA vs AI emulation of Junot Diaz, judged by 3 expert readers(MFA-trained) all of whom agree that Fine-tuned AI written excerpt is superior in terms of stylistic fidelity

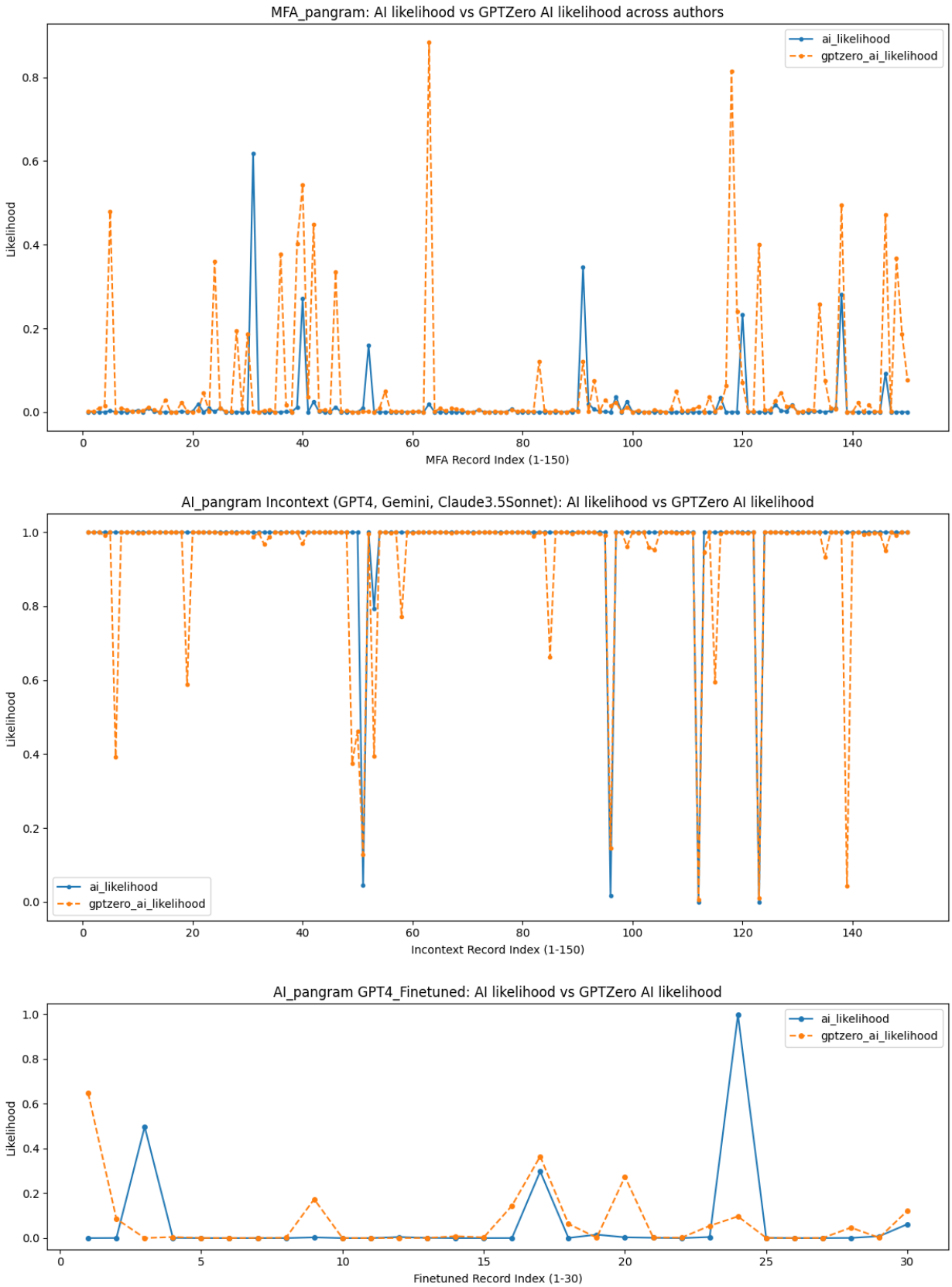


Figure 24: Pangram AI likelihood scores vs GPTZero AI Likelihood scores for MFA, In-context and Fine-tuned

to Pangram, however a threshold of 0.9 holds as a good threshold for it too. GPTZero considers fine-tuned AI generation to be more human like compared to Pangram.

S3.2 Calculating Cliché Density

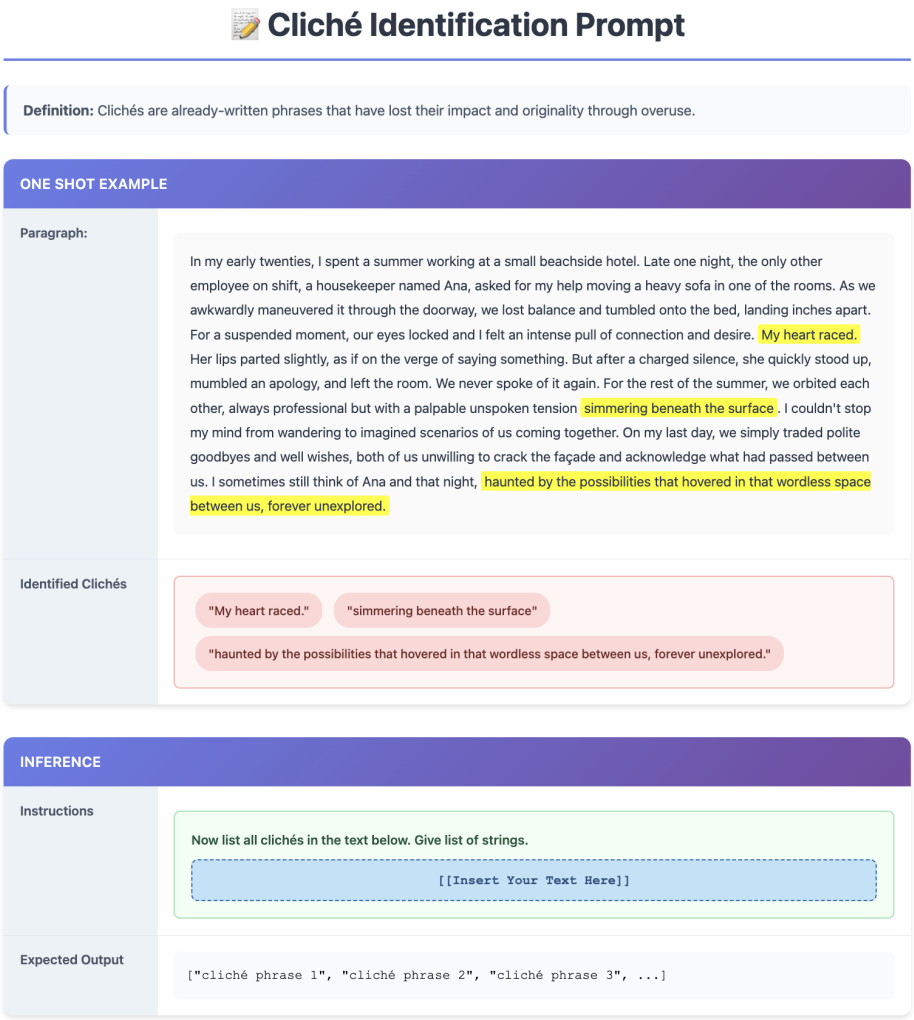


Figure 25: Prompt to identify Clichés

Language models are good at identifying specific patterns. Taking advantage of this we prompt Claude 4.1 Opus to generate list of clichés given an excerpt. Figure 25 shows the prompt used to identify Clichés. However, sometimes language models can overgenerate and penalize expressions that are not clichés. To address this, two authors of the papers separately annotated which of the expressions are actually clichés from the list. We take intersection of their individual list as the final list of clichés per paragraph. To calculate Cliché Density we then used following formula

$$\text{Cliché Density} = \left(\frac{\text{Total Words in Clichés}}{\text{Total Word Count of Excerpt}} \right) \times 100 \tag{1}$$

S4: Statistical Modeling and Analysis

S4.1: Data Structure and Notation

The analysis employs trial-level data in long format where each row represents a reader's preference decision for a single excerpt within a pairwise comparison. Each pairwise judgment contributes two rows (one per excerpt with $i \in \{1, 2\}$ sharing the same j, k). Let $Y_{ijk} \in \{0, 1\}$ denote the preference outcome where $Y_{ijk} = 1$ if excerpt i is preferred by reader j in comparison k, and 0 otherwise. The dataset comprises:

- In-context prompting: 6,600 *judgments*; 13,200 long-format *rows*.
- Fine-tuning: 4,320 *judgments*; 8,640 long-format *rows*.

Key variables include writer type $W_i \in \{\text{Human, GPT-4o, Claude, Gemini, GPT-4o-FT}\}$, reader type $J_j \in \{\text{MFA-trained, College-educated}\}$, and for H3, the Pangram AI-detection score $s_i \in [0, 1]$. Throughout, Human and MFA-trained readers serve as reference categories unless otherwise specified. Standard errors use CR2 with readers as the clustering unit; rows within a judgment pair are not independent, but in this design clustering by reader (the source of repeated measures) is the dominant dependence.

S4.2: Primary Models for H1 and H2

We test our preregistered hypotheses using fixed-effects logistic regression with CR2 cluster-robust standard errors. The CR2 estimator provides improved finite-sample performance compared to standard cluster-robust estimators, particularly important given our relatively small number of readers.

S4.2.1: H1: Baseline LLMs vs Human (In-context Setting)

For H1, we analyze only the in-context prompting trials containing Human and baseline LLMs (GPT-4o, Claude, Gemini), excluding fine-tuned excerpts. We fit the logistic regression:

$$\begin{aligned}\text{logit}\, P(Y_{ijk} = 1) = \alpha + \sum_{m \in \mathcal{M}} \beta_m \mathbf{1}[W_i = m] + \gamma \mathbf{1}[J_j = \text{College-educated}] \\ + \sum_{m \in \mathcal{M}} \phi_m \mathbf{1}[W_i = m] \cdot \mathbf{1}[J_j = \text{College-educated}]\end{aligned} \tag{2}$$

where $\mathcal{M} = \{\text{GPT-4o}, \text{Claude}, \text{Gemini}\}$ and Human serves as the reference category. The interaction terms ϕ_m capture differential preferences between MFA-trained and college-educated general readers. The preregistered H1 contrast tests whether humans outperform the average of baseline LLMs:

$$\Delta_g^{\text{H1}} = \frac{1}{3} \sum_{m \in \mathcal{M}} \eta(m, g) - \eta(\text{Human}, g) \tag{3}$$

where $\eta(w, g)$ denotes the linear predictor (fitted log-odds) for writer w and reader group $g \in \{\text{MFA-trained}, \text{College-educated}\}$. The odds ratio $\text{OR}_g^{\text{H1}} = \exp(\Delta_g^{\text{H1}})$ quantifies the relative preference, with values < 1 indicating human preference and values > 1 indicating AI preference.

S4.2.2: H2: Fine-tuned GPT-4o vs Human

For H2, we analyze only the fine-tuning trials containing Human and GPT-4o-FT excerpts. We fit:

$$\begin{aligned}\text{logit}\, P(Y_{ijk} = 1) = \alpha + \beta_{\text{FT}} \mathbf{1}[W_i = \text{GPT-4o-FT}] + \gamma \mathbf{1}[J_j = \text{College-educated}] \\ + \phi_{\text{FT}} \mathbf{1}[W_i = \text{GPT-4o-FT}] \cdot \mathbf{1}[J_j = \text{College-educated}]\end{aligned} \tag{4}$$

The H2 contrast directly compares fine-tuned GPT-4o to human writers:

$$\Delta_g^{\text{H2}} = \eta(\text{GPT-4o-FT}, g) - \eta(\text{Human}, g) \tag{5}$$

yielding $\text{OR}_g^{\text{H2}} = \exp(\Delta_g^{\text{H2}})$. This tests whether fine-tuning achieves parity (OR ≈ 1) or superiority (OR > 1) compared to human experts.

All contrasts and confidence intervals use Wald (normal) inference with the CR2 covariance

matrix. Specifically, 95% CIs are computed as estimate ±1.96· SE on the log-odds scale, then exponentiated. Holm correction is applied across the two reader-group contrasts within each outcome (style, quality) and hypothesis (H1, H2).

S4.3: H3: AI Detection and Preference

To test whether AI detectability influences preferences and whether fine-tuning removes this relationship, we model:

$$\begin{aligned}\text{logit}\, P(Y_{ijk} = 1) =& \alpha + \beta_1 s_i + \beta_2 \mathbf{1}[\text{Setting}_i = \text{FT}] + \beta_3 \mathbf{1}[J_j = \text{College-educated}] \\ &+ \beta_{12} s_i \cdot \mathbf{1}[\text{Setting}_i = \text{FT}] + \beta_{13} s_i \cdot \mathbf{1}[J_j = \text{College-educated}] \\ &+ \beta_{23} \mathbf{1}[\text{Setting}_i = \text{FT}] \cdot \mathbf{1}[J_j = \text{College-educated}] \\ &+ \beta_{123} s_i \cdot \mathbf{1}[\text{Setting}_i = \text{FT}] \cdot \mathbf{1}[J_j = \text{College-educated}] \end{aligned} \tag{6}$$

where s_i is the Pangram score (continuous, 0-1) and Setting indicates in-context vs fine-tuned. The coefficient β_1 captures the detection penalty in the baseline condition—how much AI detectability reduces preference. The interaction term β_{12} tests whether fine-tuning attenuates this relationship, with a positive value indicating that fine-tuning removes the penalty associated with AI detection.

S4.4: Exploratory Analyses

S4.4.1: Stylometric Mediation

The mediation analysis examines which textual features explain the link between AI detection and human preferences. We follow a two-stage approach. First, we regress Pangram scores on stylometric features using elastic net regularization to handle multicollinearity:

$$s_i = \alpha + \sum_k \lambda_k X_{ik} + \epsilon_i \tag{7}$$

where X_{ik} includes cliché density, sentence-length variance, and parts-of-speech proportions. Features surviving regularization are then included alongside Pangram scores in the preference

model to decompose the total effect into direct and mediated components. The proportion mediated is calculated using the product-of-coefficients method.

S4.4.2: Author-Level Heterogeneity

To assess variability across the 30 fine-tuned authors, we compute empirical preference rates using Jeffreys prior estimates to stabilize small-sample authors:

$$\hat{p}_a = \frac{k_a + 0.5}{n_a + 1} \tag{8}$$

with 95% intervals from Beta($\frac{1}{2}$, $\frac{1}{2}$) priors. The relationship between preference rates and fine-tuning corpus size (in millions of tokens) is assessed via OLS regression with heteroskedasticity-robust standard errors.

S4.5: Mapping to Figures

The model outputs map directly to figures in the main text:

- **Figures 2A-B**: Exponentiated contrasts OR_g^{H1} and OR_g^{H2} with 95% CIs computed on log-odds scale then transformed
- **Figures 2C-D**: Predicted probabilities $p(w, g) = \text{logit}^{-1}(\eta(w, g))$ with delta-method confidence intervals
- **Figure 2F**: Model-implied probabilities from H3 evaluated at $s \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$
- **Figure 3**: Author-level preference rates $\hat{p}_a$ plotted against corpus size with OLS regression lines
- **Figure 4A**: Stylometric mediation analysis showing proportion of Pangram effect explained by textual features
- **Figure 4B**: Fine-tuning premium (difference in AI preference between fine-tuned and in-context) versus corpus size
- **Figure 4C**: Cost comparison for producing 100,000 words of text

Inter-rater agreement quantifies consistency beyond chance using Fleiss' kappa, appropriate here as each pair was evaluated by multiple readers (3 MFA-trained and 19 in-context/21 fine-tuning college-educated general readers):

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{9}$$

where $\bar{P}$ represents mean observed agreement and $\bar{P}_e$ expected agreement by chance. MFA-trained readers achieved moderate agreement ($\kappa = 0.41 - 0.67$) while college-educated general readers showed minimal agreement ($\kappa = 0.08$–0.26), reflecting the subjective nature of literary evaluation.

Full code and exact package versions used to generate Figures 2-4 are provided in the Code & Data Availability section (S10) and the OSF repository.

S5.1: Cell Counts by Experimental Conditions

Unless noted otherwise, counts below refer to *reader-level long-format observations* entering the logistic models (two rows per judgment—one row per alternative in the pair). In the in-context prompting setting, **150** human–AI pairs were evaluated (evenly split across three baseline models: 50 Human vs GPT-4o, 50 Human vs Claude 3.5 Sonnet, and 50 Human vs Gemini 1.5 Pro). Because each pair was evaluated by 3 MFA-trained readers and 19 college-educated general readers, this yields **3,300** judgments per outcome and **6,600** long-format rows. In the fine-tuned setting, **90** human–AI pairs were evaluated; with 3 MFA-trained readers and 21 college-educated general readers per pair, this yields **2,160** judgments per outcome and **4,320** long-format rows. Human rows are three times more frequent than any single AI baseline in the in-context prompting setting because every pair contains a Human excerpt, while the 150 pairs are split evenly across the three baseline models. In the fine-tuned setting, Human and GPT-4o fine-tuned rows occur equally often.

S5.2: Writer Category Mapping

Table 7 provides the mapping from raw data labels to the analysis categories used in Section S4. This mapping ensures reproducibility and clarifies the distinction between baseline (in-context) and fine-tuned AI conditions.

Table 3: Observation counts for stylistic fidelity in the in-context prompting setting. Each row is a long-format observation entering the H1 analysis.

Outcome	Setting	Writer type	Reader type	*n*
style	In_Context	Human	MFA-trained	450
style	In_Context	Human	College-educated	2850
style	In_Context	GPT-4o_baseline	MFA-trained	150
style	In_Context	GPT-4o_baseline	College-educated	950
style	In_Context	Claude_baseline	MFA-trained	150
style	In_Context	Claude_baseline	College-educated	950
style	In_Context	Gemini_baseline	MFA-trained	150
style	In_Context	Gemini_baseline	College-educated	950

Table 4: Observation counts for stylistic fidelity in the fine-tuned setting. Balanced counts (Human = GPT-4o_finetuned) reflect the one-to-one comparison design for H2.

Outcome	Setting	Writer type	Reader type	*n*
style	Fine_tuned	Human	MFA-trained	270
style	Fine_tuned	Human	College-educated	1890
style	Fine_tuned	GPT-4o_finetuned	MFA-trained	270
style	Fine_tuned	GPT-4o_finetuned	College-educated	1890

Table 5: Observation counts for writing quality in the in-context prompting setting (same Human:AI baseline ratio as stylistic fidelity).

Outcome	Setting	Writer type	Reader type	*n*
quality	In_Context	Human	MFA-trained	450
quality	In_Context	Human	College-educated	2850
quality	In_Context	GPT-4o_baseline	MFA-trained	150
quality	In_Context	GPT-4o_baseline	College-educated	950
quality	In_Context	Claude_baseline	MFA-trained	150
quality	In_Context	Claude_baseline	College-educated	950
quality	In_Context	Gemini_baseline	MFA-trained	150
quality	In_Context	Gemini_baseline	College-educated	950

Table 6: Observation counts for writing quality in the fine-tuned setting.

Outcome	Setting	Writer type	Reader type	*n*
quality	Fine_tuned	Human	MFA-trained	270
quality	Fine_tuned	Human	College-educated	1890
quality	Fine_tuned	GPT-4o_finetuned	MFA-trained	270
quality	Fine_tuned	GPT-4o_finetuned	College-educated	1890

Table 7: Mapping of writer categories from raw data fields to analysis labels.

Writer category	Description	Mapping rule (from raw fields)
Human	Human-written (MFA authors)	excerpt_type = `Human`
GPT-4o_baseline	GPT-4o in-context prompting	AI excerpt; excerpt_model $\in$ {`GPT-4o`}; setting = `In_Context`
Claude_baseline	Claude 3.5 Sonnet (in-context)	AI excerpt; excerpt_model = `Claude3.5Sonnet`; setting = `In_Context`
Gemini_baseline	Gemini 1.5 Pro (in-context)	AI excerpt; excerpt_model = `Gemini_1.5_Pro`; setting = `In_Context`
GPT-4o_finetuned	GPT-4o fine-tuned	AI excerpt; excerpt_model = `GPT-4o_Finetuned`

S6: Main GLM Results and Contrasts (Figure 2)

This section presents the numerical results from the logistic regression models specified in Section S4. All models use CR2 cluster-robust standard errors with readers as the clustering unit. The reference category throughout is Human $\times$ MFA-trained. These tables provide the complete statistical foundation for Figure 2 in the main text.

S6.1: Model Coefficients

Table 8 reports the full coefficient tables for all four primary models. The dramatic sign reversal between in-context prompting and fine-tuned settings is immediately apparent: negative coefficients for AI writers in in-context prompting (indicating human preference) become positive in fine-tuning (indicating AI preference).

The writer type $\times$ reader type interactions in the in-context prompting models reveal that college-educated general readers are substantially more favorable to AI-generated text than MFA-trained readers, and the updated fine-tuned models show that this difference persists after fine-tuning rather than disappearing (style p=0.021; quality p=6.41 $\times$ $10^{-}5$).

Table 8: GLM coefficients with CR2 robust standard errors for all primary models. Negative coefficients indicate preference for the reference category (Human), while positive coefficients indicate preference for AI. The sign reversal between in-context prompting and fine-tuned models represents the core finding. In fine-tuned models, the writer-type × reader-type interaction is significant, indicating that college-educated general readers favor AI by a wider margin than MFA-trained readers.

Term	Est. (log-odds)	SE	z	p
Style — In-context Prompting (n = 6,600)				
(Intercept)	0.901	0.148	6.098	1.08e-09
writer_typeGPT-4o_baseline	−1.655	0.402	−4.121	3.78e-05
writer_typeClaude_baseline	−1.390	0.276	−5.039	4.67e-07
writer_typeGemini_baseline	−2.510	0.426	−5.894	3.77e-09
reader_typeCollege-educated	−0.929	0.153	−6.053	1.43e-09
writer_typeGPT-4o_baseline:reader_typeCollege-educated	1.784	0.414	4.307	1.66e-05
writer_typeClaude_baseline:reader_typeCollege-educated	1.583	0.292	5.412	6.22e-08
writer_typeGemini_baseline:reader_typeCollege-educated	2.357	0.438	5.384	7.27e-08
Style — Fine-tuned (n = 4,320)				
(Intercept)	−1.050	0.141	−7.435	1.04e-13
writer_typeGPT-4o_finetuned	2.100	0.282	7.435	1.04e-13
reader_typeCollege-educated	−0.356	0.154	−2.308	0.021
writer_typeGPT-4o_finetuned:reader_typeCollege-educated	0.713	0.309	2.308	0.021
Quality — In-context Prompting (n = 6,600)				
(Intercept)	0.978	0.174	5.634	1.76e-08
writer_typeGPT-4o_baseline	−2.406	0.473	−5.083	3.71e-07
writer_typeClaude_baseline	−1.246	0.368	−3.390	7.00e-04
writer_typeGemini_baseline	−2.406	0.421	−5.711	1.13e-08
reader_typeCollege-educated	−1.275	0.179	−7.127	1.03e-12
writer_typeGPT-4o_baseline:reader_typeCollege-educated	3.197	0.484	6.601	4.08e-11
writer_typeClaude_baseline:reader_typeCollege-educated	1.935	0.381	5.073	3.91e-07
writer_typeGemini_baseline:reader_typeCollege-educated	2.716	0.433	6.268	3.65e-10
Quality — Fine-tuned (n = 4,320)				
(Intercept)	−0.314	0.122	−2.571	0.010
writer_typeGPT-4o_finetuned	0.627	0.244	2.571	0.010
reader_typeCollege-educated	−0.531	0.133	−3.997	6.41e-05
writer_typeGPT-4o_finetuned:reader_typeCollege-educated	1.062	0.266	3.997	6.41e-05

S6.2: Primary Hypothesis Tests

Table 9 presents the preregistered contrasts testing H1 and H2, corresponding to panels A–B in Figure 2. The odds ratios reveal a striking reversal: MFA-trained readers' 6–8 fold preference for human writing in in-context prompting conditions (OR = 0.16 for style, 0.13 for quality) transforms into an 8-fold preference for AI in fine-tuning (OR = 8.16 for style). College-educated general readers show directionally consistent but substantially larger shifts, particularly after fine-tuning (OR = 16.65 for style, 5.42 for quality), indicating that this reader group favors AI-generated text by a wider margin than MFA-trained readers.

Table 9: Primary contrasts testing H1 (baseline LLMs vs Human) and H2 (fine-tuned vs Human). Holm correction applied within each outcome and hypothesis across reader types. The OR column shows odds ratios with values < 1 favoring humans and > 1 favoring AI.

Outcome	Hyp.	Reader	Est. (log-odds)	SE (log-odds)	p	p_{Holm}	OR exp(Est.)	95% CI (OR scale)
style	H1	MFA-trained	−1.852	0.314	3.65e-09	7.31e-09	0.157	[0.08, 0.29]
style	H1	College-educated	0.056	0.083	0.501	0.501	1.058	[0.90, 1.25]
style	H2	MFA-trained	2.100	0.282	1.04e-13	1.04e-13	8.163	[4.69, 14.20]
style	H2	College-educated	2.813	0.125	4.91e-112	9.81e-112	16.652	[13.03, 21.28]
quality	H1	MFA-trained	−2.020	0.381	1.19e-07	1.19e-07	0.133	[0.06, 0.28]
quality	H1	College-educated	0.597	0.087	6.00e-12	1.20e-11	1.816	[1.53, 2.15]
quality	H2	MFA-trained	0.627	0.244	1.02e-02	1.02e-02	1.873	[1.16, 3.02]
quality	H2	College-educated	1.690	0.105	4.63e-58	9.27e-58	5.417	[4.41, 6.66]

S6.3: Model-Predicted Probabilities

Table 10 reports the predicted probabilities displayed in Figure 2 panels C–D. These probabilities, derived from the inverse-logit transformation of linear predictors, show that while MFA-trained readers strongly prefer humans over all baseline models in in-context prompting settings (probabilities 0.17–0.43), fine-tuning elevates AI preference to 0.74 for MFA-trained readers and 0.80 for college-educated general readers on stylistic fidelity, and to 0.58 versus 0.70 on writing quality. Unlike the in-context condition, both reader groups now favor fine-tuned AI, but college-educated general readers do so by a significantly wider margin.

Table 10: Predicted probabilities of selecting AI excerpts by model and reader type, corresponding to Figure 2C–D. Values above 0.5 indicate AI preference. After fine-tuning, both reader groups prefer AI, but college-educated general readers show higher predicted probabilities than MFA-trained readers.

Outcome	Setting	Model	Reader	$\hat{p}$	95% CI
style	In_Context	GPT-4o (In-Context)	MFA-trained	0.320	[0.21, 0.45]
style	In_Context	GPT-4o (In-Context)	College-educated	0.525	[0.49, 0.56]
style	In_Context	Claude (In-Context)	MFA-trained	0.380	[0.30, 0.46]
style	In_Context	Claude (In-Context)	College-educated	0.541	[0.51, 0.57]
style	In_Context	Gemini (In-Context)	MFA-trained	0.167	[0.10, 0.27]
style	In_Context	Gemini (In-Context)	College-educated	0.455	[0.42, 0.49]
style	Fine_tuned	GPT-4o (Fine-tuned)	MFA-trained	0.741	[0.68, 0.79]
style	Fine_tuned	GPT-4o (Fine-tuned)	College-educated	0.803	[0.78, 0.82]
quality	In_Context	GPT-4o (In-Context)	MFA-trained	0.193	[0.11, 0.31]
quality	In_Context	GPT-4o (In-Context)	College-educated	0.621	[0.59, 0.65]
quality	In_Context	Claude (In-Context)	MFA-trained	0.433	[0.33, 0.54]
quality	In_Context	Claude (In-Context)	College-educated	0.597	[0.56, 0.63]
quality	In_Context	Gemini (In-Context)	MFA-trained	0.193	[0.12, 0.29]
quality	In_Context	Gemini (In-Context)	College-educated	0.503	[0.47, 0.54]
quality	Fine_tuned	GPT-4o (Fine-tuned)	MFA-trained	0.578	[0.52, 0.64]
quality	Fine_tuned	GPT-4o (Fine-tuned)	College-educated	0.699	[0.68, 0.72]

S6.4: Interaction Diagnostics

Tables 11 and 12 examine the writer-type × reader-type interactions in detail. The significant interactions in in-context prompting models (all $p < 0.001$) quantify college-educated general readers' greater tolerance for AI-generated text. Unlike the previous analysis with a smaller college-educated sample, the interactions in fine-tuned models are now also significant (style $p = 0.021$; quality $p = 6.41 \times 10^{-5}$), indicating that college-educated general readers favor fine-tuned AI by a wider margin than MFA-trained readers rather than converging to similar preference levels.

The results in this section provide compelling statistical evidence for the paper's central finding: fine-tuning on author-specific corpora fundamentally transforms AI-generated text from clearly inferior (as judged by experts) to preferred over human writing. Rather than converging after fine-tuning, both reader groups prefer fine-tuned AI, but college-educated general readers do so by a significantly wider margin than MFA-trained readers. This pattern suggests that fine-tuning produces improvements that are meaningful to both audiences while leaving a persistent difference in how strongly each audience favors the AI output.

Table 11: Individual interaction terms showing differential preferences between MFA-trained and college-educated general readers. Large positive coefficients indicate college-educated general readers are more favorable to AI. In fine-tuned models, interactions are significant, indicating that reader-group differences persist after fine-tuning.

Outcome	Setting	Term	Est.	SE	z	p
style	In_Context	GPT-4o_baseline:reader_typeCollege-educated	1.784	0.414	4.307	1.66e-05
style	In_Context	Claude_baseline:reader_typeCollege-educated	1.583	0.292	5.412	6.22e-08
style	In_Context	Gemini_baseline:reader_typeCollege-educated	2.357	0.438	5.384	7.27e-08
style	Fine_tuned	GPT-4o_finetuned:reader_typeCollege-educated	0.713	0.309	2.308	0.021
quality	In_Context	GPT-4o_baseline:reader_typeCollege-educated	3.197	0.484	6.601	4.08e-11
quality	In_Context	Claude_baseline:reader_typeCollege-educated	1.935	0.381	5.073	3.91e-07
quality	In_Context	Gemini_baseline:reader_typeCollege-educated	2.716	0.433	6.268	3.65e-10
quality	Fine_tuned	GPT-4o_finetuned:reader_typeCollege-educated	1.062	0.266	3.997	6.41e-05

Table 12: Joint Wald tests for writer-type × reader-type interactions. Interactions are significant in both in-context prompting and fine-tuned models, indicating that college-educated general readers are more favorable to AI than MFA-trained readers in both conditions.

Outcome	Setting	Term	df	χ^2	p
style	In_Context	writer_type:reader_type	3	36.606	5.57e-08
style	Fine_tuned	writer_type:reader_type	1	5.328	0.021
quality	In_Context	writer_type:reader_type	3	46.884	3.68e-10
quality	Fine_tuned	writer_type:reader_type	1	15.978	6.41e-05

S7: Author-Level Heterogeneity (Figure 3)

This section examines variation in AI preference rates across the 30 fine-tuned authors. Despite substantial differences in training corpus sizes (0.89M to 10.9M tokens) and fine-tuning costs ($22 to $273), we find no detectable relationship between data quantity and model performance within this token range, suggesting that even authors with limited published works can be effectively emulated.

S7.1: Per-Author Success Rates

Table 13 presents author-level AI preference rates using Jeffreys prior estimates to stabilize small-sample authors. Each author now has 72 pooled trials (9 from MFA-trained readers and 63 from college-educated general readers), so the pooled rates are weighted approximately 87.5% toward the college-educated group. The results reveal striking heterogeneity. For **style**, AI preference rates range from 18% (Tony Tulathimutte) to 98% (Roxane Gay), with 29 of 30 authors showing AI superiority (rate > 0.5). For **quality**, the range spans 49% (Kazuo Ishiguro and Min Jin Lee) to 92% (Cheryl Strayed), with 28 of 30 authors favoring AI.

Notably, some authors show divergent patterns across outcomes. Tony Tulathimutte represents an extreme case: readers found his *style* uniquely difficult to emulate (18% AI preference) yet rated AI *quality* as acceptable (55%). This suggests certain idiosyncratic voices resist algorithmic mimicry even when technical writing competence is achieved.

Table 13: Per-author AI preference rates (Jeffreys-smoothed and raw) ranked by performance. Values above 0.5 indicate AI preference. Each author has 72 pooled trials (9 MFA-trained, 63 college-educated general readers). Note the wide variation across authors and the divergence between **style** and quality rankings for some authors.

Outcome	Author	Rank	AI Win Rate (Jeffreys)	Human Win Rate (Jeffreys)	AI Win Rate (Raw)
quality	Cheryl Strayed	1	0.92	0.08	0.931
quality	Marilynne Robinson	2	0.91	0.09	0.917
quality	Han Kang	3	0.88	0.12	0.889
quality	Salman Rushdie	4	0.84	0.16	0.847
quality	Lydia Davis	5	0.80	0.20	0.806
quality	Zadie Smith	6	0.80	0.20	0.806

Continued on next page

Outcome	Author	Rank	AI Win Rate (Jeffreys)	Human Win Rate (Jeffreys)	AI Win Rate (Raw)
quality	Haruki Murakami	7	0.77	0.23	0.778
quality	Junot Diaz	8	0.77	0.23	0.778
quality	Orhan Pamuk	9	0.77	0.23	0.778
quality	Sigrid Nunez	10	0.76	0.24	0.764
quality	Colson Whitehead	11	0.73	0.27	0.736
quality	Roxane Gay	12	0.71	0.29	0.708
quality	Percival Everett	13	0.69	0.31	0.694
quality	Jhumpa Lahiri	14	0.68	0.32	0.681
quality	Rachel Cusk	15	0.68	0.32	0.681
quality	Ben Lerner	16	0.66	0.34	0.667
quality	Margaret Atwood	17	0.66	0.34	0.667
quality	Louise Erdrich	18	0.65	0.35	0.653
quality	George Saunders	19	0.64	0.36	0.639
quality	Jonathan Franzen	20	0.64	0.36	0.639
quality	Sally Rooney	21	0.64	0.36	0.639
quality	Annie Proulx	22	0.60	0.40	0.597
quality	Chimamanda Ngozi Adichie	23	0.58	0.42	0.583
quality	Annie Ernaux	24	0.55	0.45	0.556
quality	Tony Tulathimutte	25	0.55	0.45	0.556
quality	Yoko Ogawa	26	0.54	0.46	0.542
quality	Ian McEwan	27	0.51	0.49	0.514
quality	Ottessa Moshfegh	28	0.51	0.49	0.514
quality	Kazuo Ishiguro	29	0.49	0.51	0.486
quality	Min Jin Lee	30	0.49	0.51	0.486
style	Roxane Gay	1	0.98	0.02	0.986
style	Margaret Atwood	2	0.97	0.03	0.972
style	Han Kang	3	0.95	0.05	0.958
style	Ben Lerner	4	0.94	0.06	0.944
style	Chimamanda Ngozi Adichie	5	0.94	0.06	0.944
style	Kazuo Ishiguro	6	0.94	0.06	0.944
style	Lydia Davis	7	0.94	0.06	0.944
style	Marilynne Robinson	8	0.94	0.06	0.944
style	Sigrid Nunez	9	0.92	0.08	0.931
style	Cheryl Strayed	10	0.91	0.09	0.917
style	Junot Diaz	11	0.91	0.09	0.917
style	Orhan Pamuk	12	0.91	0.09	0.917
style	Min Jin Lee	13	0.86	0.14	0.861
style	Sally Rooney	14	0.86	0.14	0.861
style	Colson Whitehead	15	0.83	0.17	0.833

Continued on next page

Outcome	Author	Rank	AI Win Rate (Jeffreys)	Human Win Rate (Jeffreys)	AI Win Rate (Raw)
style	Salman Rushdie	16	0.83	0.17	0.833
style	George Saunders	17	0.80	0.20	0.806
style	Ian McEwan	18	0.77	0.23	0.778
style	Haruki Murakami	19	0.76	0.24	0.764
style	Rachel Cusk	20	0.76	0.24	0.764
style	Percival Everett	21	0.75	0.25	0.750
style	Jhumpa Lahiri	22	0.71	0.29	0.708
style	Zadie Smith	23	0.71	0.29	0.708
style	Annie Proulx	24	0.66	0.34	0.667
style	Jonathan Franzen	25	0.66	0.34	0.667
style	Ottessa Moshfegh	26	0.65	0.35	0.653
style	Yoko Ogawa	27	0.61	0.39	0.611
style	Louise Erdrich	28	0.58	0.42	0.583
style	Annie Ernaux	29	0.51	0.49	0.514
style	Tony Tulathimutte	30	0.18	0.82	0.181

Note. Jeffreys prior adds 0.5 to successes and 0.5 to failures; differences from raw are small but avoid 0/1 edge cases.

S7.2: Relationship with Corpus Size

Table 14 presents OLS regression results examining the relationship between fine-tuning corpus size (ranging from 0.89M tokens for Tulathimutte to 10.9M for Pamuk) and AI preference rates. The near-zero slopes with confidence intervals spanning zero and minimal R^2 values indicate no detectable relationship within this token range between training data quantity and model performance.

Table 14: OLS regression of author-level AI win rate on fine-tuning corpus size (millions of tokens). Slopes are near zero with CIs spanning zero and very low R^2, indicating no detectable relationship between corpus size and AI preference rate.

Outcome	Slope	Slope CI (Low)	Slope CI (High)	Intercept	Intercept CI (Low)	Intercept CI (High)	R^2	N Authors
quality	0.0029	−0.0173	0.0231	0.676	0.580	0.771	0.0031	30
style	0.0089	−0.0194	0.0373	0.760	0.626	0.894	0.0153	30

Non-parametric check. Spearman $\rho = 0.052$ ($p = 0.791$) for quality and $\rho = -0.040$ ($p = 0.839$) for style.

This finding has economic implications: authors with limited published works (costing $22–$50 to fine-tune) can be emulated as effectively as prolific authors (costing $200+). The absence of any detectable corpus-size effect ($R^2 \approx 0$) suggests that capturing an author's voice depends more on stylistic consistency than sheer data volume.

S8: Detection Mechanisms and Economic Analysis (Figure 4)

This section examines the mechanisms linking AI detectability to human preferences (H3) and quantifies the economic implications of fine-tuning. We show that fine-tuning reverses and substantially attenuates the negative association between AI detection and preference observed in in-context prompting, while achieving performance that readers prefer at ≈99.7% lower cost than human writers.

S8.1: AI Detection and Preference Relationship

Table 15 presents the full model examining how AI detectability (Pangram score) influences preferences. The key finding is the significant interaction between *pangram_score* and *setting* (quality: $\hat{\beta} \approx 2.90$, $p < 0.001$), indicating that fine-tuning attenuates the negative association between detectability and preference observed in in-context prompting. The significant three-way interactions (pangram_score × setting × reader_type) further indicate that the detectability penalty and its attenuation differ across reader groups, with college-educated general readers showing less sensitivity to detectability than MFA-trained readers even before fine-tuning.

Table 15: Pangram GLM coefficients with CR2 robust standard errors (clustered by reader). Negative coefficients for *pangram_score* in in-context prompting indicate detection reduces preference; positive interactions with *setting* show that fine-tuning attenuates this penalty. Significant three-way interactions indicate the detectability penalty differs across reader groups.

Outcome	*N*	*N* Readers	Term	Estimate	SE	*z*	*p*	OR	OR (Low)	OR (High)
quality	10920	542	(Intercept)	0.990	0.169	5.87	4.43e-09	2.69	1.93	3.74
quality	10920	542	*pangram_score*	−2.010	0.333	−6.03	1.65e-09	0.13	0.07	0.26
quality	10920	542	*setting* (Fine-tuned)	−1.025	0.175	−5.86	4.64e-09	0.36	0.26	0.51
quality	10920	542	*reader_type* (College-educated)	−1.288	0.174	−7.40	1.36e-13	0.28	0.20	0.39
quality	10920	542	pangram_score × setting	2.904	0.876	3.32	9.13e-04	18.20	3.28	102.00
quality	10920	542	pangram_score × reader_type	2.613	0.344	7.59	3.24e-14	13.60	6.94	26.80
quality	10920	542	setting × reader_type	1.313	0.180	7.28	3.34e-13	3.72	2.61	5.29
quality	10920	542	pangram_score × setting × reader_type	−3.258	0.908	−3.59	3.31e-04	0.04	0.01	0.23
style	10920	535	(Intercept)	0.909	0.146	6.23	4.79e-10	2.48	1.86	3.31
style	10920	535	*pangram_score*	−1.845	0.292	−6.32	2.57e-10	0.16	0.09	0.28
style	10920	535	*setting* (Fine-tuned)	−0.938	0.158	−5.95	2.65e-09	0.39	0.29	0.53
style	10920	535	*reader_type* (College-educated)	−0.925	0.152	−6.10	1.09e-09	0.40	0.29	0.53

Continued on next page

Outcome	N	N Readers	Term	Estimate	SE	z	p	OR	OR (Low)	OR (High)
style	10920	535	pangram_score × setting	2.556	0.810	3.15	1.61e-03	12.90	2.63	63.00
style	10920	535	pangram_score × reader_type	1.878	0.304	6.18	6.26e-10	6.54	3.61	11.90
style	10920	535	setting × reader_type	0.921	0.163	5.65	1.63e-08	2.51	1.82	3.46
style	10920	535	pangram_score × setting × reader_type	−1.766	0.844	−2.09	3.64e-02	0.17	0.03	0.89

Notes. CR2 robust standard errors clustered by reader. Readers: 28 MFA-trained and 516 college-educated general readers.

S8.2: Predicted Probabilities Across Detection Levels

Tables 16 and 17 show model-predicted probabilities at different AI detection levels, corresponding to Figure 2F. For MFA-trained readers evaluating *quality*, high detection scores (0.9) reduce AI preference to ∼31% in in-context prompting but have a point estimate of ∼68% after fine-tuning (95% CI 0.37–0.89), demonstrating that fine-tuning makes outputs robust to detection-based skepticism. College-educated general readers show a weaker detectability penalty in the in-context condition and a modest positive slope after fine-tuning, with predicted probabilities reaching 0.55–0.67 at high detection scores.

Table 16: Predicted AI preference probabilities for **style** at different Pangram detection scores. Fine-tuning attenuates the relationship with detection score for MFA-trained readers (with wider uncertainty at high detection). College-educated general readers show a weak relationship with detectability in the in-context condition and a positive slope after fine-tuning.

Bin	Setting	Reader Type	Prob	Prob (Low)	Prob (High)
0.1	In-context Prompting	MFA-trained	0.674	0.621	0.722
0.5	In-context Prompting	MFA-trained	0.497	0.494	0.500
0.9	In-context Prompting	MFA-trained	0.320	0.273	0.372
0.1	Fine-tuned	MFA-trained	0.511	0.491	0.530
0.5	Fine-tuned	MFA-trained	0.581	0.435	0.714
0.9	Fine-tuned	MFA-trained	0.648	0.380	0.847
0.1	In-context Prompting	College-educated	0.497	0.481	0.513
0.5	In-context Prompting	College-educated	0.500	0.500	0.500
0.9	In-context Prompting	College-educated	0.503	0.487	0.520
0.1	Fine-tuned	College-educated	0.512	0.505	0.519
0.5	Fine-tuned	College-educated	0.593	0.544	0.641
0.9	Fine-tuned	College-educated	0.670	0.581	0.748

Table 17: Predicted AI preference probabilities for **quality**. MFA-trained readers show strong detection sensitivity in in-context prompting (~0.69 → ~0.31) that attenuates after fine-tuning (point estimate at 0.9 ~0.68; 95% CI 0.37–0.89). College-educated general readers show a weaker in-context penalty and modest positive slopes after fine-tuning.

Bin	Setting	Reader Type	Prob	Prob (Low)	Prob (High)
0.1	In-context Prompting	MFA-trained	0.688	0.628	0.742
0.5	In-context Prompting	MFA-trained	0.496	0.493	0.499
0.9	In-context Prompting	MFA-trained	0.306	0.254	0.363
0.1	Fine-tuned	MFA-trained	0.514	0.490	0.537
0.5	Fine-tuned	MFA-trained	0.602	0.431	0.751
0.9	Fine-tuned	MFA-trained	0.683	0.373	0.887
0.1	In-context Prompting	College-educated	0.441	0.424	0.457
0.5	In-context Prompting	College-educated	0.501	0.500	0.501
0.9	In-context Prompting	College-educated	0.561	0.544	0.578
0.1	Fine-tuned	College-educated	0.504	0.497	0.510
0.5	Fine-tuned	College-educated	0.529	0.479	0.578
0.9	Fine-tuned	College-educated	0.553	0.461	0.643

S8.3: Stylometric Correlates and Mediation

Table 18 shows that cliché density has the strongest correlation with AI detection (Pearson $r = 0.60$), while readability ease is negatively correlated ($r = -0.23$), suggesting AI text is marked by formulaic phrases but simpler syntax. Among MFA-trained readers, where the detectability penalty is concentrated, approximately 16% of the detection effect on preference is mediated through cliché density in the in-context prompting condition; after fine-tuning, this mediation drops to a statistically insignificant 1.3%, indicating that fine-tuning substantially reduces rather than merely masks these stylistic signatures. Because college-educated general readers did not penalize AI-detectable text in either condition, mediation analysis does not apply to this group.

Table 18: Correlations between Pangram AI detection scores and stylometric features. These correlations are computed at the excerpt level ($N = 330$ unique excerpts) and are unchanged by the expanded reader sample.

Metric	Pearson r	Spearman ρ	N
readability_ease	−0.232	−0.312	330
cliche_density	0.600	0.602	330
total_adjective_count	0.104	0.129	330
num_cliches	0.614	0.633	330

S8.4: Economic Analysis

Table 19 documents the cost structure for AI-based text generation. With fine-tuning costs ranging from $22.25 to $272.50 (median $77.88) plus $3 for inference to generate 100,000 words, the

total AI cost represents approximately 0.3% of the $25,000 a professional writer would charge. This ~99.7% cost reduction, combined with reader preference for fine-tuned outputs, quantifies the potential economic disruption to creative writing markets.

Table 19: Cost comparison for generating 100,000 words. Fine-tuning plus inference costs ($25–$276) represent about 0.1% to 1.1% of professional writer compensation ($25,000), with a median of approximately 0.3%.

N Authors	Fine-tune Min	Fine-tune Med	Fine-tune Max	Total Min	Total Med	Total Max
30	$22.25	$77.88	$272.50	$25.25	$80.88	$275.50

Note. Inference cost of $3 per 100k words follows Figure 4c; per-token assumptions are documented in the repository.

S9. Deviations from Preregistration

This section documents deviations from the preregistered analysis plan (https://osf.io/zt4ad). All other aspects were implemented exactly as specified.

- **Model framework.** The preregistration specified mixed-effects logistic regression with random intercepts for readers and prompts. We implemented logistic GLMs with CR2 cluster-robust standard errors (clustered by reader) due to convergence issues and singularity warnings in the mixed-effects models. Point estimates from both approaches were similar when convergence was achieved, and inference on the preregistered contrasts (H1, H2) remained unchanged.

- **Sample sizes (exceeded targets).** We successfully recruited more participants than planned:

  - MFA-trained readers: 28 (planned: 25)
  - College-educated general readers: 516 (planned: 120)
  - Fine-tuned authors: 30 (planned minimum: 10)

  All analyses used the full realized sample. No stopping rules were applied or violated; the additional data strengthens the reliability of our findings.

- **Additional analyses (not preregistered).** We added:

- Writer type × Reader type interaction tests (Type-III Wald with robust covariance) for transparency
- Predicted probability visualizations (Figures 2C-D) to complement the odds ratio panels

These additions provide fuller context but do not alter the conclusions drawn from the preregistered H1 and H2 contrasts.

S10. Code and Data Availability

All code and data necessary to reproduce Figures 2-4 and the associated statistical analyses are publicly available:

- **GitHub Repository:** `https://github.com/tuhinjubcse/Author-Style-Personalization/tree/main`
- **Preregistration:** `https://osf.io/zt4ad` (Version 2, July 2025)
- **Key Analysis Scripts:** The repository contains R scripts for data processing, model fitting, and figure generation. Core analyses can be reproduced by running the numbered scripts in sequence.
- **Environment:** Analyses were conducted in R 4.3.1 with key packages including `clubSandwich` (CR2 robust SEs) and `emmeans` (contrasts). Full package versions are documented in the repository.
- **Data:** Anonymized trial-level data in long format, with documentation of all variable definitions and transformations applied.