GENERATIVE AI FLOODS AND DILUTES THE MARKET FOR BOOKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative AI can produce book-length works of fiction at near-zero cost. These books are typically tagged as low-quality "slop" that buyers will ignore, and won't carry much commercial weight in the market. We test that assumption by running AI detection across 14,419 self-published genre-fiction books sold on Amazon from 2023 to 2026, matched to daily sales records through June 2026. None of these books disclose whether or not they contain AI-produced content. We find that books for which we detected substantial AI text ($> 25\%$) make up a large share of the catalog but a smaller share of sales. Even so, they reach commercial scale, winning a growing share of sales over time and taking more of the scarce top-rank positions once held by books with no detected AI text. Over this period, the number of books with observed sales in a quarter grew 19.2-fold, while quarterly revenue grew only 8.9-fold. The market added selling books faster than it added revenue, and revenue per selling book fell across most genres. Books with no detected AI text lost the most ground in genres with high AI diffusion, especially in genres with high Kindle Unlimited availability. Among top-selling books, those with substantial AI text contain more rare expressions found in existing books than do books with no AI text. Generative AI can thus reshape a creative market through scale rather than quality. Our results bear directly on the market-effect question at the center of the fair-use defense to copyright infringement.

1 INTRODUCTION

Generative AI has significantly reduced the cost of producing long form, book-length fiction (Pham et al., 2026). Online communities have formed around fiction writing (Gupta et al., 2026), where members trade detailed recipes for turning a one-line premise into a finished novel.[1] Readers and critics, however, often dismiss these outputs as "slop" (Shaib et al., 2025)—cheap and formulaic work that buyers will ignore. But a book being low average quality does not necessarily imply commercial irrelevance. Due to its inexpensive nature, hundreds of producers can draft and list books that are partially or fully generated by AI. At sufficient volume, even an average product can attract the attention of buyers and siphon away sales that could have otherwise gone to competing titles. Many of these books reach readers without disclosure (Russell et al., 2026a), and the brief previews shown on platforms make it hard for readers to discern which books are AI generated. Whether such books can reshape a market is therefore an empirical question. One author, identified as "Coral Hart," reportedly released more than 200 romance novels across 21 pen names in a single year and sold roughly 50,000 copies (Alter, 2026a). In 2026, Hachette canceled the horror novel *Shy Girl* after an independent analysis found it to be largely AI-generated (Alter, 2026b;c). More recently, Jerry Falade's debut crime novel "Call Me I'll Hide The Body" was sold to Macmillan's Minotaur imprint for a reported $2 million after a fourteen-way auction only to be withdrawn by the author's own agents due to AI accusations.[2] These episodes provide strong evidence that such incidents can happen in real life, but individual cases cannot account for the market-wide effects. This paper is primarily aimed at measuring how generative AI has impacted a real book market once its output reached the catalog and competed for sales.

[1] See, for example, the communities at `https://www.reddit.com/r/WritingWithAI/` and `https://plotprose.com/`.

[2] `https://www.theguardian.com/books/2026/jul/31/crime-novel-deal-collapses-questions-ai-jerry-falade-call-me-ill-hide-the-body`

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We study self-published genre-fiction that is primarily sold on Amazon. Self-publishing now accounts for the bulk of books produced in the U.S. In 2025, the number of self-published titles registered with Bowker exceeded 3.5 million—a 38.7% increase from 2024—and represented roughly 84% of the 4.17 million ISBN-registered titles released that year (Milliot, 2026). Most of these books are released on the market through Amazon's Kindle Direct Publishing (KDP) platform. Unlike traditional publishing, KDP allows producers to add titles at a low marginal cost because there is no editorial gatekeeping involved and the listing is nearly free (WordsRated, 2023). Within this broader market, genre fiction faces a high risk of being disrupted by generative AI. Readers of genre fiction come back for the same tropes and an influx of new books, which can easily be churned out by generative AI (Russell et al., 2026b; Chakrabarty et al., 2024). Self-published genre fiction is therefore a leading-edge market in which AI-generated text can appear at scale. The mechanism is, however, not unique to fiction. In fact, some of the earliest documented evidence of the flooding effect was in the non-fiction category with a New York Times investigation finding travel guides inundated by AI-generated titles under fabricated author identities (Kugel & Hiltner, 2023). We collected a stratified random sample of 14,419 ebooks that were published between January 2023 and March 2026, together with their daily sales through June 2026 from a proprietary panel maintained by a major "Big Five" publisher. The panel tracks approximately 500,000 Amazon ASINs accounting for at least 95% of daily ebook unit volume.

In a recent study, Reimers & Waldfogel (2026) showed that the number of new book releases has nearly tripled after 2022 and that books that are detected as AI (based on 1,000 word previews) earn lower ratings and weaker sales compared to books that are fully human written. In contrast, we ask a different question: what happens to the market that receives these books? While they detect AI from previews and study how flagged titles themselves perform, we detect AI across the full text of each book and measure what those titles do to the books around them. When a massive number of books that are cheap to produce enter the market at a rate faster than the sales, revenue and top ranks available to absorb them, the revenue and attention available to the average book decline. We call this dilution. This has gained significant importance in copyright litigation over AI training, where market effects have become a key, perhaps decisive, issue. In Kadrey v. Meta (Chhabria, 2025), Judge Chhabria singled out market dilution (the erosion of a market through rapid production of substitutable works) as a theory that could determine the outcome of a fair-use defense in such cases, even as he noted that the supporting evidence was absent in the case before him.

We therefore ask four questions. First, do books with substantial AI text only occupy the bottom of the market, or are they competitive enough that they sell and rank alongside books with no AI text in them? Second, does their entry expand the market or reduce the sales and visibility available to books with no detected AI text? Third, do producers who use AI release books at inordinately high rates beyond what human-only writing would enable? Finally, among the high-earning books, do those with substantial AI text draw on more distinctive language of existing books than those with no AI text? Our findings show a market rebuilt by scale. Books containing substantial AI text enter the catalog faster than the revenue available to absorb them and in turn decrease the revenue earned by individual books while competing for the top ranks once held by books with no AI text. Last but not least, books with AI text that succeed draw most heavily on the distinctive language of existing books. This restructuring of a creative labor market bears directly on the market-effect question now central to copyright law. [3]

2 Data and Measurement

2.1 Observed market panel

We study sales of newly self-published genre-fiction ebooks released between January 1, 2023, and March 31, 2026. We rely on a proprietary sales panel provided by one of the "Big Five" publishers that includes daily sales on Amazon until June 29, 2026. We chose to start with the entire population of such titles and then reduced the search space by excluding titles whose Amazon sales rank never fell below 1.5 million at any point during the observation period.[4] From the remaining titles, we drew a random sample of 15,000 books.

[3] We will release code, experimental set up and anonymized data publicly available for reproducibility.

[4] A book whose sales rank never fell below 1.5M essentially sold 0 copies and hence is not a good candidate for quantifying dilution.

We clustered the raw Amazon genre labels into eight broad groups: General Romance, Sports Romance, Speculative/Adventure Romance, Crime/Suspense Romance, Fantasy/Supernatural/Horror, Mystery/Thriller/Crime, Science Fiction/Adventure/Dystopian, and General/Contemporary Fiction. These genre groupings were fixed prior to performing any empirical work. We ensure that the proportion of each genre in our sample of books reflects how these genres are represented in a full population of new releases. We also sample these books uniformly at random across release dates. It should be noted that all titles are chosen without knowing their content before hand such that potential AI detection results would not bias our sample.

After dropping titles with no recorded sales data in our panel, we ultimately used 14,419 titles in our analyses (details in Appendix A). To calculate daily revenue for title i on day t we use $R_{it} = Q_{it} \times P_{it}$, where Q_{it} equals the quantity of units sold and P_{it} equals the consumer price for those units. We replace any P_{it} that are missing or non-positive with the median valid price within the time period of the title's launch window.

For each book title in our sample we defined the Kindle Unlimited (KU) flag based on whether the title ever participated in the program. Finally, we constructed author identity using information contained in the byline field of each book title.

2.2 Full-text AI classification

We detect AI generated text using Pangram v3.3 (Masrour et al., 2026). Pangram is one of the most accurate AI detectors audited by several independent research groups (Jabarian & Imas, 2025; Russell et al., 2025; Shah & Levy, 2026). We pass each chapter of a book to Pangram and define the AI-detection score S_i as the percentage of text windows that are not labeled as "Human Written". We group books into three bands by S_i. A book with *no AI* text has no flagged windows. Books with *light AI* text have a flagged share above 0% and at most 25% while books with *substantial AI* text have a flagged share greater than 25%. Pangram's held-out evaluation reports an N-weighted false-positive rate of 0.04% and a false-negative rate of 0.14%. In our own audit of 100 self-published genre fiction books published before 2022 we found no books flagged as AI. Further details in Appendix B.

2.3 Market measures

How do the title shares of each group translate into sales and revenue? We use three measures for that—title share, sales share, and revenue share of each band. Much of what follows uses a fixed window around the time of the title's launch, defined as the observed preorder days before the title goes on sale along with days 0 through 90 after it, $\mathcal{L}_i = \{t : t < O_i\} \cup \{t : 0 \leq t - O_i \leq 90\}$ where O_i is the title's on-sale date. We measure dilution in two ways. The first measure tracks the market on a quarterly basis: for each quarter we count released titles, titles with positive sales, total sales, and total revenue, and index each series to its 2023 Q1 level. The second measure compares each title over a common time frame: launch-window revenue per selling title, defined as total launch-window revenue divided by the number of titles with positive observed unit sales in that window, which we report by release cohort, AI text band, and genre. Finally, we use the Amazon sales rank as an indicator of market positioning, where a lower number indicates a better position. Within each genre-week cell, we take each book's median observed sales rank, order titles by that median, and define Top-25 rank slots. To measure turnover, we compare the sets of Top-25 book titles from one quarter to the next. Let T^{25}_{gq} denote the titles holding Top-25 rank slots in genre g during quarter q. Retention of books with no AI text is the fraction of the prior quarter's no-AI Top-25 titles that remain in the current quarter's Top-25, and the entrant share is the fraction of new Top-25 entrants with substantial AI text. Let N and A denote books with no AI text and substantial AI text, respectively.

$$\text{Retention}_{gq} = \frac{|T^{25}_{g,q-1} \cap T^{25}_{gq} \cap N|}{|T^{25}_{g,q-1} \cap N|}, \quad \text{EntrantShare}_{gq} = \frac{|(T^{25}_{gq} \setminus T^{25}_{g,q-1}) \cap A|}{|T^{25}_{gq} \setminus T^{25}_{g,q-1}|}. \tag{1}$$

Appendix C gives the full definitions, together with the churn, AI-exposure, and Kindle Unlimited analyses.

3 RESULTS AND FINDINGS

3.1 DIFFUSION: BOOKS WITH SUBSTANTIAL AI TEXT SPREAD BUT SELL LESS

Between January 2023 and March 2026, books with substantial AI text made up a rising share of new releases across all eight genre clusters, and the all-genre average rose with them (Fig. 1A). Their numbers did not translate into sales at the same rate. Books with substantial AI text make up 20% of all books in our sample but earn only 12.1% of sales and 11.3% of revenue, while books with no AI text account for 62.9% of books and a larger 71.7% of sales and 72.5% of revenue (Fig. 1C). Books with light AI text fall between the two and account for 17.1% of books and a near-equal 16.2% of sales and 16.2% of revenue. This implies that the shortfall in sales falls on the books with substantial AI text and does not depend on detected AI text in general. The gap becomes wider toward the top of the market. Books with no AI text hold 63% of all launch-window positions and 73% of the best-selling top 5%, while the share held by books with substantial AI text falls from 20% of all books to 10% of that top 5% (Fig. 1B). But it is worth noting that even a 12% share of sales is crucial, especially when it comes from 20% of the catalog that competes for a revenue pool which grew far more slowly than the catalog itself.

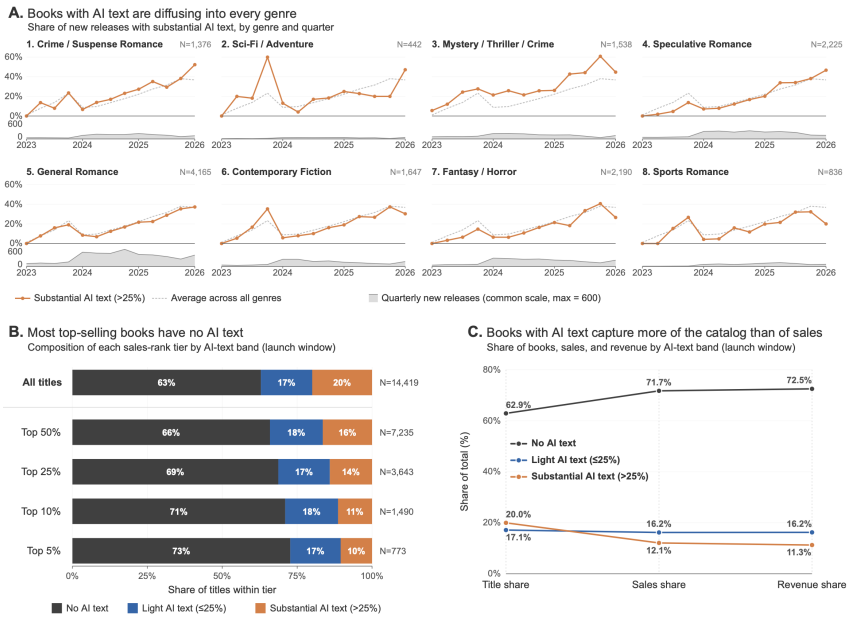


Figure 1: **Diffusion and sales share by AI text band.** Here, N is the number of books in each genre cluster. For Figure B, the Top K% tiers are formed by ranking books by launch-window sales within genre-quarter cells and pooling, so tier sizes differ slightly from exact percentages of the total 14,419 books.

3.2 DISPLACEMENT: BOOKS WITH SUBSTANTIAL AI TEXT REACH THE TOP RANKS

Although such books sell weakly on average on their own, books with substantial AI text did not stay at the margins. Their share of observed sales rose over the period, from near zero in early 2023 to roughly 20% by 2026 Q2, and the share held by books with light AI text rose alongside it to about 19% (Fig. 2A). These books also reached the scarce positions at the top of the market. Their share of constructed Top-25 rank slots grew throughout the period, with books with light and substantial AI text together holding roughly 40% by 2026 Q2 (Fig. 2B). The same rise appears in every genre cluster, most steeply in Crime/Suspense Romance, Science Fiction/Adventure/Dystopian, and Mystery/Thriller/Crime (Fig. 2C).

After the first comparable quarter, retention of books with no AI text in the Top-25 fell to a trough near 28% and recovered only unevenly thereafter, ending near 62%, while the share of new Top-25 entrants with substantial AI text rose from near zero to 31% (Fig. 2D; Appendix C). Thus, turnover in the Top-25 sets increasingly brought in books with substantial AI text in them, even though the majority of the top-ranked sampled books remained the ones with no AI text in them. The market

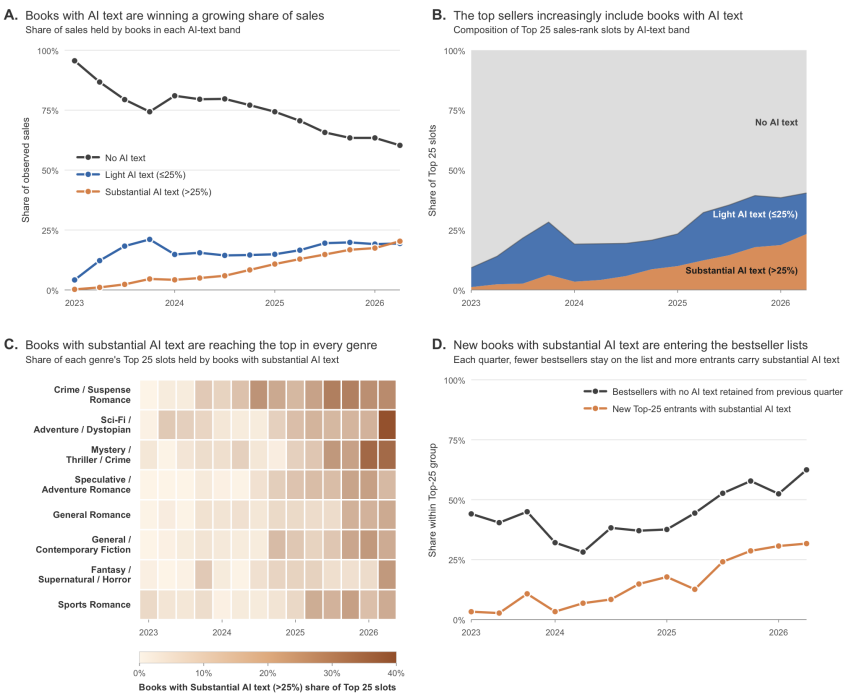


Figure 2: **Sales share, Top-25 rank slots, and turnover by AI text band.** Top-25 rank slots are constructed among sampled titles within each genre-week and aggregated to quarters (Appendix C). In (D), the two lines measure different quantities and need not sum to 100%.

now holds many books with substantial AI text that sell little, along with a smaller set that reaches the very top.

3.3 Dilution: Revenue per book falls, even for books with no AI text

The number of books released expanded faster than the sales and revenue available to absorb them. Between January 2023 and March 2026, the cumulative catalog of released books grew 38.3-fold and the number of books selling in a quarter grew 19.2-fold. Meanwhile, the quarterly sales and revenue grew only 7.3-fold and 8.9-fold (Fig. 3A). To ensure a fixed observation window despite the catalog expansion, we measure revenue over the launch window and group books into release cohorts. We compare the same window across cohorts, such that later cohorts are not penalized for being on sale for a shorter amount of time. Across these fixed launch windows, revenue per book that sold at least one copy declined on net from the 2023 cohort to the 2025 cohort, after a rise in 2024 that reached every band. Across cohorts, the revenue lead that books with no AI text held over books with substantial AI text narrowed, as returns to books with no AI text fell toward those of the other bands (Fig. 3B). The decline reaches across the market. Revenue per book fell in six of eight genres for all books (Fig. 3C) and in seven of eight for books with no AI text alone (Fig. 3D), when comparing the 2023 and 2025 launch cohorts. The drop in revenue per book cannot be explained as a mix shift where the average drops only because the catalog is filled with many lower-revenue AI books, because the books with no AI text even earn less in later cohorts as well.

The losses become more prominent where exposure to AI is highest. Across genre-months, the share of Top-25 rank slots held by books with no AI text falls as the share of the released catalog with substantial AI text rises, from about 88% in the lowest-exposure bin to about 63% in the highest (Fig. 4A). The fall is steeper in genres where more of the titles are available on Kindle Unlimited. There, readers draw from a shared subscription pool, so each new book with substantial AI text competes for the same borrows. In these genres, after adjusting for AI exposure and calendar month, the lead that books with no AI text hold over books with substantial AI text is 8.5 percentage points smaller in sales share, and 8.4 points smaller in revenue share, than in genres where fewer titles are available on Kindle Unlimited (Fig. 4B; Appendix C). The same pattern holds at the low-exposure end. In Fantasy / Supernatural / Horror, the genre AI reached latest and least, revenue per

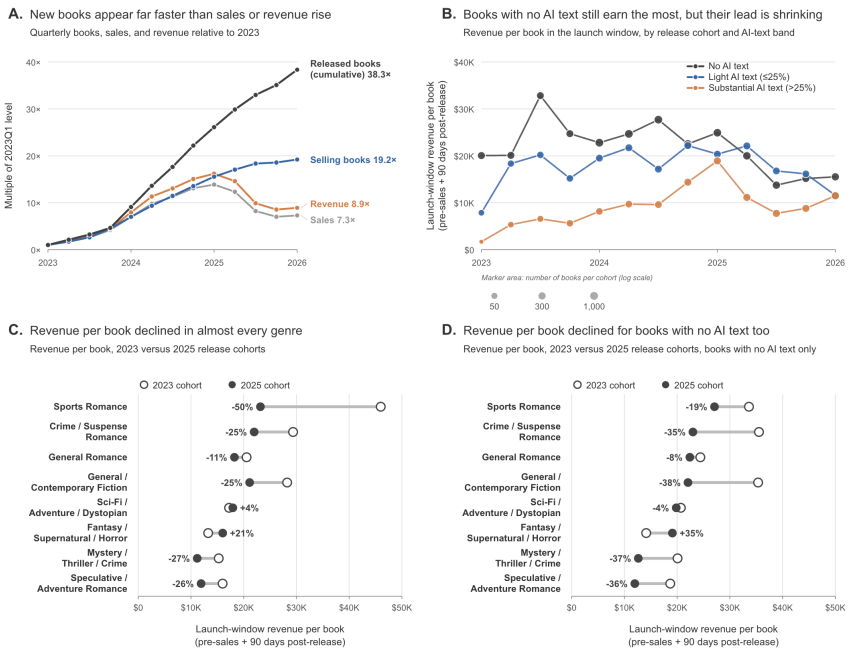


Figure 3: **Catalog growth and launch-window revenue per selling title.** In (A), the released-books line is cumulative, while selling titles, unit sales, and revenue are measured within each quarter; all series are indexed to 2023 Q1. Revenue per book in (B)-(D) is launch-window revenue per selling title.

book for books with no AI text rose 35% across cohorts even as it fell in the high-exposure genres (Fig. 3D). Returns to books with no AI text thus move with AI exposure across the market. While observational, these patterns align with what dilution would predict.

3.4 A SUPPLY SHOCK: AI ADOPTION INCREASES AUTHOR OUTPUT

For each author, their adoption month is computed as the first month in which they publish a book containing substantial AI text. We compare each author's monthly output before adoption with months 1–12 following it, and measure author and book revenue over a 12-month horizon (Appendix D). The rise in output can be attributed to adopters who stuck with AI, instead of every author who tried it. Of the 824 authors in the event-time panel, 311 raised their monthly output after adoption, but among authors who released additional books with substantial AI text after their first such release, 287 of 385 did (Fig. 4C). This expansion was uneven but not confined to a handful of authors. Among these 385 identities, the top quartile produced 60.5% of such post-release books (Gini 0.48). The most prolific authors, however, published at extreme volume, reaching dozens of books a year, and they produced almost entirely books with substantial AI text, with a few exceptions (Fig. 4D). These authors also reached real commercial scale. Among the 15 author identities generating the most revenue from books with substantial AI text, books associated with the top identity generated $1.7M in gross consumer revenue across 8 titles (Fig. 4E). The 15 highest-revenue individual such books each generated more than $220K, the top one reaching $643K on 80,431 sales (Fig. 4F). Because output and revenue are measured for observed byline identity, authors who publish under several pen names are split across these counts, and our approach very likely understates the true amount of production and earnings. [5]

3.5 SELECTION: RARE-EXPRESSION OVERLAP WITH EXISTING BOOKS

The books that are substantially AI and succeed, share a textual signature. To identify it, we measure how much of a book is covered by rare expressions that already occur in existing books, following work that attributes distinctive language in AI-generated text to large pre-training corpora (Lu et al.,

[5]Our revenue measures are also calculated before platform commissions, which means the actual author earnings are lower.

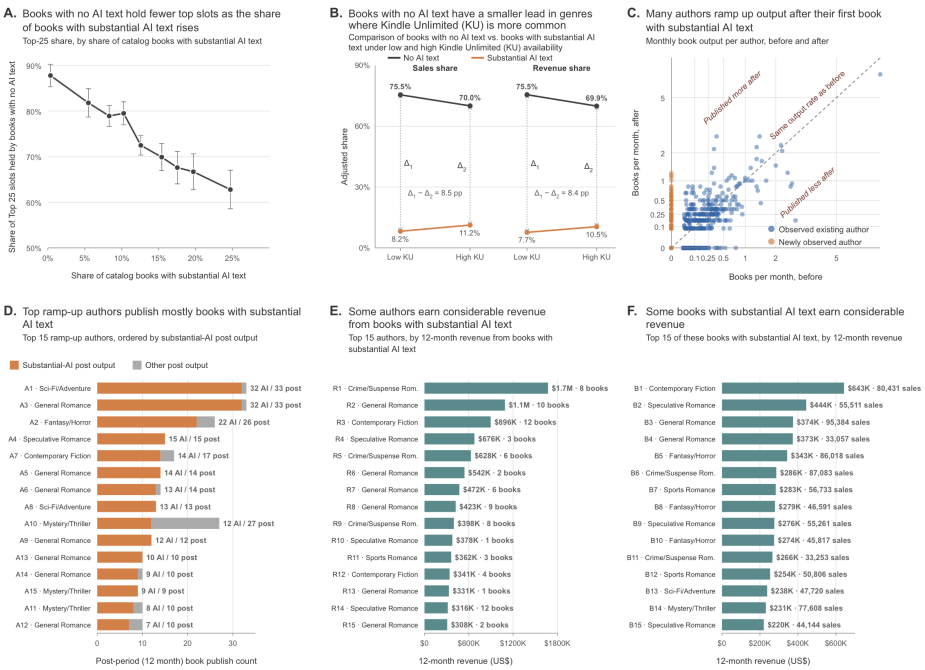


Figure 4: **AI exposure, Kindle Unlimited, and author output.** Error bars in (A) show 95% confidence intervals; shares in (B) are adjusted for AI exposure and calendar month. In (C)–(F), output and revenue are measured over the 12 months after each author's first release with substantial AI text (using available follow-up in C–D), or over each book's first 12 months on sale in (F)(Appendix D). Revenue is gross consumer spending.

2025; Liu et al., 2025). We call an expression rare when it appears in at most five Google Books volumes and is absent from a 4.7-trillion-token snapshot of the internet. The first condition focuses only on language that is present in a handful of books instead of looking into generic and common phrasing, and the second gets rid of expressions that circulate widely online. We calculate rare-expression coverage for each book as the share of its tokens that are part of at least one rare expression (Liu et al., 2026) (Appendix E). Therefore, coverage reflects overlap with the distinctive language of particular books [6].

Considering the universe of top-selling books, the ones that are detected as having substantial AI fill more of their text with these rare expressions than books with no AI text do (Fig. 5A). For the Top-50 tier, the mean incorporation of rare expressions is 45.0% for books that are substantially AI vs 37.7% for books with no AI text, that make up for a gap of 7.2 percentage points. The gaps are 5.5 points at Top-100 and 4.3 points at Top-200 (Welch's t-test, $p < 10^{-8}$). Comparatively, incorporation of rare expressions for books with no AI text stays flat across tiers, from 37.7% to 37.2%. This means that the gap is indicative of the substantially-AI detected books themselves instead of the revenue ranking. To see if such overlap corresponds to commercial success, we fit a linear model to the top 200 books from each group:

$$C_i = \alpha + \beta_1 AI_i + \beta_2 \log_{10}(Revenue_i) + \beta_3 AI_i \times \log_{10}(Revenue_i) + \epsilon_i. \quad (2)$$

Here, C_i is rare-expression coverage and AI_i is a binary variable marking books with substantial AI text (i.e., >25%). The parameter β_2 represents the revenue slope for books with no AI text and $\beta_2 + \beta_3$ the slope for books with substantial AI text, so β_3 measures how much more steeply coverage rises with revenue for AI-heavy books than for books with no AI text. This overlap also correlates with commercial success, among books that are substantially AI. For these books, overlap rises with revenue at 7.6 percentage points per tenfold increase ($\beta_2 + \beta_3$; $p = 10^{-3}$), while the slope for books with no AI text is not distinguishable from zero at 1.1 points (β_2; $p = 0.67$); the difference between the two slopes is marginal (β_3; $p = 0.063$) (Fig. 5B). The steeper slope for the substantially-AI books survives genre fixed effects, so it does not follow from genre composition.

[6] Distinctiveness here does not necessarily imply literary quality, however. Nor does it necessarily mean human-written.

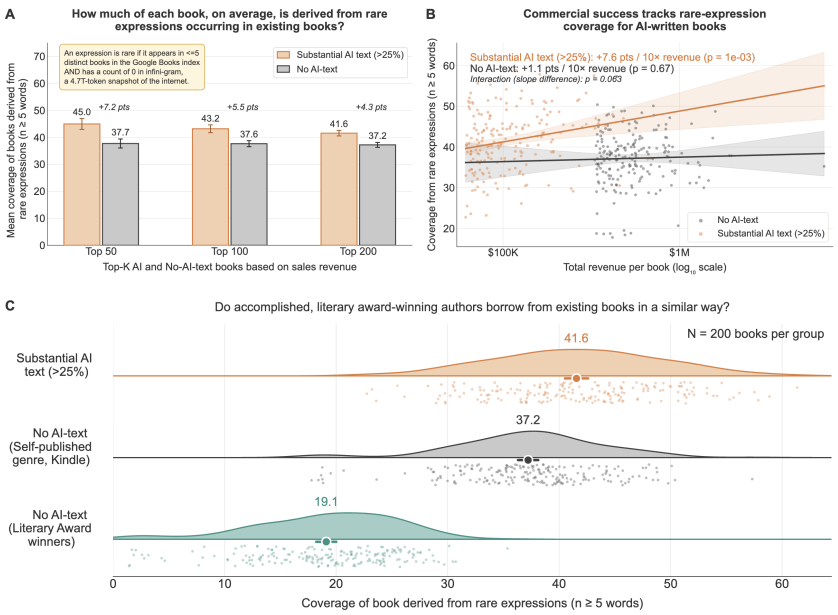


Figure 5: **Rare-expression coverage among top-selling books.** Books are ranked by total revenue within each group, and (B)–(C) use the top 200 per group. Error bars and shaded bands show 95% CIs; lines in (B) are OLS fits, and clouds in (C) are kernel density estimates.

To test whether drawing specific expressions from existing books is inherent to authorship, we add a third reference group of 200 literary award-winning books across literary, historical, and speculative fiction (Appendix E). Award-winning fiction, often held up as a hallmark of originality, has a mean coverage of just 19.1%, roughly half the rate for books that are substantially AI (41.6%) and well below self-published bestsellers with no AI text (37.2%); all pairwise differences are statistically significant (Welch's t-test; $p = 1.2 \times 10^{-13}$ between the two groups of bestsellers and $p < 10^{-70}$ for both comparisons against award winners; Fig. 5C).

The rare expressions derived from existing books in books with substantial AI text are particularly embodied in nature such as "*cheekbones more pronounced, the shadows beneath her eyes deeper*", "*The question circled his mind like a vulture*", "*cross my eyes and puff out my cheeks like a demented chipmunk*" and "*my stomach flip-flopping like a fish out of water*". These expressions route meaning through touch, smell, and visceral sensation. Our findings align with recent work from Russell et al. (2026b) showing how AI overwrites the body and senses, conveying feeling through physical sensation and environment while rarely labeling emotion explicitly.

4 Discussion

Research on generative AI in creative work has focused primarily on whether machine output is good enough to substitute for human work (Chakrabarty et al., 2025; Chakrabarty & Dhillon, 2026). Our results, however, point to a different problem for creative markets, one that is agnostic to the quality of AI work. Our analysis of more than 14,000 self-published genre-fiction books sold on Amazon between 2023-26 shows that the catalog grew far faster than the sales and revenue. That said, total revenue also rose over this period, but at a much slower pace, and as a result the revenue per book declined. Books containing AI text were largely responsible for this substantial increase in the growth, but most such books sold very few copies. We further observe that books with no AI text earned less per book in later cohorts as well, so our results cannot be explained by mix-shift alone (Foster et al., 2001; Baily et al., 1992), where the average falls only because the catalog is filled with many low-earning AI books. Per-book revenue fell most in the genres where AI text had diffused the most. These patterns are consistent with dilution as each book earns less after substitutable books crowd a market.

Two factors shaped this outcome, one on each side of the market. On the supply side, AI adoption increased output among authors who continued using it. On the demand side, a smaller subset of books containing substantial AI text attracted substantial consumer demand and reached the top of

the market. Rare-expression overlap with existing books was greater among these high-revenue books than among top-selling books with no detected AI text, and within the substantial-AI group, overlap rose with revenue. This pattern is consistent with a derivative-production channel (Landes & Posner, 1989; 2003), though the measure captures textual overlap and not copying of any particular book. These two forces compounded each other: cheap production lengthened the tail of supply, and the language overlap with existing books characterized the smaller set of AI books that reached the top. It should however be noted that on the demand side, consumers might benefit from the greater variety and lower prices of books, so the net welfare effect is unclear. But for the questions we address, i.e., the returns to human creative effort, and the market-dilution issue at the heart of copyright law, the harm to authors is what is at stake.

These findings bear on questions now before the courts. In recent litigation over training large language models on copyrighted books, market effect has emerged as the central unresolved empirical question. The "market dilution" theory endorsed by the Copyright Office (U.S. Copyright Office, 2024) and entertained by Judge Chhabria in Kadrey v. Meta (Chhabria, 2025), recognizes that human-authored works may be "crowded out" by AI-generated outputs which may substitute for sales of the human-authored works on which the AI platform trained (i.e., human-authored works copied into the training data). Liability (if any) for infringement results not from the platform's production of outputs that compete with human-authored works, but from the massive upstream copying that enabled the competing outputs.[7] In Kadrey v. Meta, concerning proof of market dilution, Judge Chhabria inquired whether AI-generated and human-written books compete in the same markets, and "what impact does this competition actually have on sales of the books it competes with? Whatever the effects have been thus far, are they likely to increase in the future, as more and more AI-generated books are written, and as LLMs get better and better at writing human-like text?" This study examines the market for self-published genre titles on Amazon, a market in which human-written books and books with substantial AI content compete. It documents patterns consistent with a depressive effect of the massive entry of AI content titles on the sales of human-authored books. Because the effect is one of scale, one may surmise that "as more and more AI-generated books are written," the crowding effect will increase.

Our comparisons are associational. We compare release cohorts and relate outcomes to genre-level AI exposure, and genres differ along many dimensions, so we read these comparisons as observational. We offer some possible explanations for competition through scale in Appendix F. By dilution we mean two declines we observe as the catalog fills with AI books, in per-book revenue and in the market share of books with no detected AI text. The decline within books with no detected AI text might seem consistent with a broad market trend, but the pattern reverses in Fantasy / Supernatural / Horror, the genre AI reached latest and least, which makes it unlikely to be caused by a market-wide trend. And the magnitudes we report are specific to the single platform we observe, though their direction is not. We should observe similar outcomes to those in our study in other domains where a near-limitless supply of cheap works meets a fixed reserve of reader attention. Music, stock media (Goldberg & Lam, 2025), food delivery platforms (Dai et al., 2026), and short-form writing markets share several of these features. Our work provides the first systematic evidence of these effects in self-published genre fiction, and our results have direct implications for choices now facing authors, platforms, and courts.

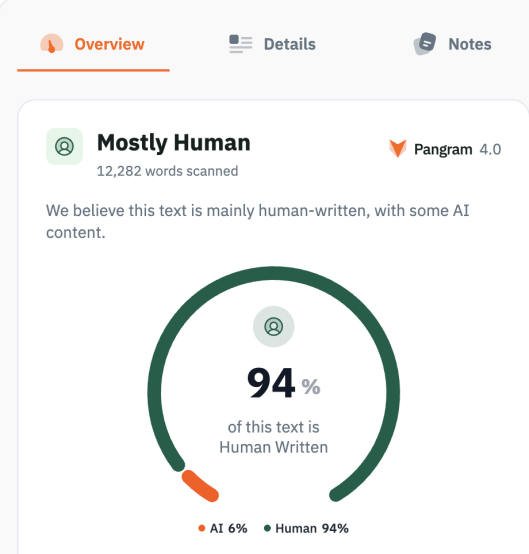
5 Conclusion

We present a study involving 14,419 self-published books and show that books with substantial AI text make up a large share of the catalog but a smaller share of sales. These books with substantial AI text while commonly dismissed as lower-quality slop reach commercial scale, winning a growing share of sales over time and taking more of the scarce top-rank positions once held by books with no detected AI text. We also show how revenue per selling book declined across most genres and more specifically in genres with higher amount of AI diffusion. Among top-selling books, those with substantial AI text contain more rare expressions found in existing books than do books with no AI text. These observational findings are consistent with market dilution and highlight how production at scale can reshape creative markets despite weaker average sales.

[7] The outputs may also compete with human-written works that the AI system did not copy, but that substitution effect is not within the purview of copyright law: competition without copying is not infringement

AI USE STATEMENT

In this work, we used generative AI tools for shortening some of the text to fit the page limit as well as minor line-level copy editing. We have not used generative AI tools for ideation or implementation or writing any significant portion of the paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI. Last but not least at the time of submission Pangram 4 flagged 94% of the content as Human Written.



REFERENCES

Alexandra Alter. The new fabio is claude: The romance industry, always at the vanguard of technological change, is rapidly adapting to A.I. *The New York Times*, February 2026a. URL https://www.nytimes.com/2026/02/08/business/ai-claude-romance-books.html.

Alexandra Alter. Horror novel 'shy girl' canceled over suspected a.i. use. *The New York Times*, March 2026b. URL https://www.nytimes.com/2026/03/19/books/shy-girl-book-ai.html.

Alexandra Alter. A.I. is writing fiction. Publishers are unprepared. *The New York Times*, March 2026c. URL https://www.nytimes.com/2026/03/19/books/ai-fiction-shy-girl.html.

Martin Neil Baily, Charles Hulten, and David Campbell. Productivity dynamics in manufacturing plants. *Brookings Papers on Economic Activity: Microeconomics*, 1992:187–267, 1992.

Tuhin Chakrabarty and Paramveer S Dhillon. Can good writing be generative? expert-level ai writing emerges through fine-tuning on high quality books. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1–27, 2026.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 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*, pp. 1–34, 2024.

Tuhin Chakrabarty, Jane C Ginsburg, and Paramveer Dhillon. Readers prefer outputs of ai trained on copyrighted books over expert human writers. *arXiv preprint arXiv:2510.13939*, 2025.

Brian X Chen. I tested a popular a.i. slop detector. it felt empowering. *The New York Times*, 2026. https://www.nytimes.com/2026/08/13/technology/personaltech/pangram-ai-detector-test.html.

Vince Chhabria. Kadrey et al. v. meta platforms, inc. (No. 3:23-cv-03417-VC), 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.

Jingwei Dai, Beibei Li, and Rahul Telang. Generative ai content governance and market outcomes: Evidence from food delivery platforms. *Available at SSRN 6711039*, 2026.

Lucia Foster, John Haltiwanger, and C. J. Krizan. Aggregate productivity growth: Lessons from microeconomic evidence. In Charles R. Hulten, Edwin R. Dean, and Michael J. Harper (eds.), *New Developments in Productivity Analysis*, pp. 303–372. University of Chicago Press, 2001.

Samuel Goldberg and H. Tai Lam. Generative AI and creative goods: Market expansion, crowd-out, and copyright. Stanford university graduate school of business research paper, Stanford University Graduate School of Business, February 2025. URL http://dx.doi.org/10.2139/ssrn.5152649. Available at SSRN: https://ssrn.com/abstract=5152649.

Neel Gupta, Maria Antoniak, and Melanie Walsh. Ai fiction in the wild. *arXiv preprint arXiv:2606.22748*, 2026.

Brian Jabarian and Alex Imas. Artificial writing and automated detection. Technical report, National Bureau of Economic Research, 2025.

Seth Kugel and Stephen Hiltner. A new frontier for travel scammers: A.I.-generated guidebooks. *The New York Times*, 2023. https://www.nytimes.com/2023/08/05/travel/amazon-guidebooks-artificial-intelligence.html.

William M. Landes and Richard A. Posner. An economic analysis of copyright law. *The Journal of Legal Studies*, 18(2):325–363, 1989.

William M. Landes and Richard A. Posner. *The Economic Structure of Intellectual Property Law*. Harvard University Press, Cambridge, MA, 2003.

Jiacheng Liu, Taylor Blanton, Yanai Elazar, Sewon Min, Yen-Sung Chen, Arnavi Chheda-Kothary, Huy Tran, Byron Bischoff, Eric Marsh, Michael Schmitz, et al. Olmotrace: Tracing language model outputs back to trillions of training tokens. pp. 178–188, 2025.

Xinyue Liu, Niloofar Mireshghallah, Jane C Ginsburg, and Tuhin Chakrabarty. Alignment whack-a-mole: Finetuning activates verbatim recall of copyrighted books in large language models. *arXiv preprint arXiv:2603.20957*, 2026.

Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, et al. Ai as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text. In *International Conference on Learning Representations*, volume 2025, pp. 91005–91056, 2025.

Elyas Masrour, Max Spero, Bradley Emi, and Katherine Thai. Pangram 3.3 model card. https://www.pangram.com/research/model-card/pangram-3-3, May 2026. Accessed: 2026-07-05.

Jim Milliot. Book output topped four million in 2025. *Publishers Weekly*, March 2026. URL https://www.publishersweekly.com/pw/by-topic/industry-news/publisher-news/article/99943-book-output-topped-4-million-in-2025.html.

Chau Minh Pham, Yapei Chang, and Mohit Iyyer. Autofiction, 2026. URL https://www.autofiction.ai/. Web platform.

Imke Reimers and Joel Waldfogel. Ai and the quantity and quality of creative products: Have llms boosted creation of valuable books? Technical report, National Bureau of Economic Research, 2026.

Jenna Russell, Marzena Karpinska, and Mohit Iyyer. People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5342–5373, Vienna, Austria, 2025. Association for Computational Linguistics.

Jenna Russell, Marzena Karpinska, Destiny Akinode, James Zhou, Katherine Thai, Bradley Emi, Max Spero, and Mohit Iyyer. Ai use in american newspapers is widespread, uneven, and rarely disclosed. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14554–14580, 2026a.

Jenna Russell, Rishanth Rajendhran, Chau Minh Pham, Mohit Iyyer, and John Wieting. Storyscope: Investigating idiosyncrasies in ai fiction. *arXiv preprint arXiv:2604.03136*, 2026b.

Anand Shah and Joshua Levy. Access to justice in the age of ai: Evidence from us federal courts. 2026.

Chantal Shaib, Tuhin Chakrabarty, Diego Garcia-Olano, and Byron C Wallace. Measuring ai" slop" in text. *arXiv preprint arXiv:2509.19163*, 2025.

U.S. Copyright Office. Copyright and artificial intelligence part 3: Generative AI training report. Technical report, U.S. Copyright Office, December 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.

Marijke Van Vlasselaer, Filip Van Droogenbroeck, and Bram Spruyt. Who wrote this? evaluating the reliability of ai detection tools in higher education. *International Journal for Educational Integrity*, 22(1):16, 2026.

WordsRated. Amazon publishing statistics. https://wordsrated.com/amazon-publishing-statistics/, 2023. WordsRated industry statistics.

A Data and sample

A.1 Sample construction

We study sales of newly self-published genre-fiction ebooks released between January 1, 2023, and March 31, 2026.[8] We have a proprietary sales panel that includes daily sales as well as Amazon sales rankings until June 29, 2026, which is 90 days after the last release date in our window, so that even titles released at the end of the period can be tracked through their full launch window. The sales panel was provided by one of the "Big Five" publishers. We chose to start with the entire population of such titles and then reduced the search space by excluding titles whose Amazon sales rank never fell below 1.5 million at any point during the observation period. This removed any book listings that had no sales. From the remaining titles, we drew a random sample of 15,000 books, which was the largest number we could process given the cost of running AI detection on full texts. We sampled books in proportion to the share each genre makes up among new releases (General Romance 28.8%, Speculative / Adventure Romance 15.5%, Fantasy / Supernatural / Horror 15.2%, General / Contemporary Fiction 11.4%, Mystery / Thriller / Crime 10.7%, Crime / Suspense Romance 9.6%, Sports Romance 5.8%, Science Fiction / Adventure / Dystopian 3.1%), so that our sample mirrors the mix of genres in the full population of new releases. We also sampled uniformly at random across release dates. Importantly, we chose titles without knowing their content, so the results of our AI detection could not influence which books ended up in the sample. Finally, some of our selected titles had no usable sales data recorded in the panel and were dropped. Thus, we ultimately used 14,419 titles in our analyses.

Our panel consists of only those titles that appear in Amazon's sales and ranking feeds, thereby limiting our observations to only those titles that actively participate in Amazon's marketplace. All the estimates reported in the paper, such as the catalog shares or the dilution estimates are based on this observed universe of books only.

We secured full texts for each of the selected books in our corpus lawfully by emailing authors for epubs, borrowing available ebooks online from local libraries such as Libby/Hoopla, as well as purchasing remaining ebooks from Amazon. A few of these titles have since been picked up by traditional publishing houses over the study period.

A.2 Panel construction and variables

In general, throughout this paper we refer to the "title" as the ASIN level catalog entry that we are analyzing and refer to the "book" as the actual work being read. Each title has its unique Amazon Standard Identification Number (ASIN), along with author and title metadata, an on-sale date, subject and genre categories, Kindle Unlimited fields, daily Amazon unit sales, consumer price, Amazon sales rank, and a score giving the share of its full text detected as AI-generated (Appendix B).

We parse the date, numeric, and score columns from the raw data into their respective types, trim the title ID strings to remove excess whitespace, and remove empty IDs. We then create a pair of connected tables, one at the title level and one at the title-day level. The first table is a list of all titles with each title having one row in this table. For any given title, if there are multiple rows that contain different information (i.e., more than one possible value) for any of the title metadata we retain the most frequent non-missing value for each field. The second table is a list of all days by title. In this case each row represents an observation of a particular title on a specific day. Any duplicate rows of this nature are collapsed down to just one observation for each day.

To calculate daily revenue for title i on day t we use:

$$R_{it} = Q_{it} \times P_{it},$$

where Q_{it} equals the quantity of units sold, P_{it} equals the consumer price for those units. We replace any P_{it} that are missing or non-positive with the median valid price within the time period of the title's launch window. All titles have complete unit sales and price data during the entire launch window (the observed preorder days before the on-sale date plus days 0 through 90 after release).

[8] During publisher verification we identified a small number of traditionally published editions that accounts for roughly 0.3% of the corpus. We retain them in the analysis.

Amazon provides a detailed list of genres and subjects for books published on its platform. However, its genre classification is too fine-grained for performing analyses at the level of genre-month since many cells will have too few books. So, we clustered the raw Amazon genre labels into the eight broad groups listed above, and assigned the remainder to a miscellaneous "Other" category that we exclude from the main figures. These genre groupings were developed to retain as much commercially relevant signal as possible and they ensured that each genre-month cell contains enough books to support sufficiently powered analyses. These genre groupings were fixed prior to performing any empirical work. Table 1 contains the conversion of raw Amazon genre labels to the eight groups defined above.

Table 1: Genre-cluster crosswalk used in the frozen analysis.

Cluster	Raw-label families
General Romance	Contemporary, small-town, billionaire, comedy, holiday, historical, gothic, military, western romance, erotica, and residual romance labels
Speculative / Adventure Romance	Fantasy romance, paranormal romance, science-fiction romance, and action/adventure romance
Crime / Suspense Romance	Romantic suspense and mafia romance
Sports Romance	Sports romance and hockey romance
Fantasy / Supernatural / Horror	Fantasy, dark fantasy, epic fantasy, LitRPG/progression fantasy, supernatural fiction, paranormal fantasy, urban fantasy, and horror
Mystery / Thriller / Crime	Crime fiction, mystery/suspense, thriller, psychological thriller, and cozy mystery
Science Fiction / Adventure / Dystopian	Science fiction, adventure fiction, dystopian, and post-apocalyptic fiction
General / Contemporary Fiction	Contemporary fiction, women's fiction, LGBTQ+ fiction, historical fiction, literary fiction, young adult fiction, humorous fiction, and miscellaneous fiction

Next, for each book title in our sample we defined the Kindle Unlimited (KU) flag based on whether the title ever participated in the program. So, when performing the genre-month analyses we computed the KU-share as the fraction of titles released in that genre in that month that had the KU flag checked. Unfortunately, our panel data does not break down the number of unit sales of a book into KU page reads and à la carte purchases.

Finally, we constructed author identity using information contained in the byline field of each book title. So, books with missing bylines were not part of our author-level analysis. We anonymized the authors in the figures in the paper as A1, A2, and so on. It is worth noting that one potential shortcoming of our setup is that an author may write under several pen names and several people may share one byline. Hence, the byline count may underestimate how concentrated a given author's production is.

B AI DETECTION AND OUR 25% CUTOFF

B.1 DETECTING AI TEXT WITH PANGRAM

We use Pangram, one of the most accurate AI detectors audited by several independent research groups (Jabarian & Imas, 2025; Russell et al., 2025; Shah & Levy, 2026; Van Vlasselaer et al., 2026; Chen, 2026) and now used by the popular online publishing platform Substack.[9] Each book carries a full-text AI score from the Pangram detector. We segment each book into individual chapters after excluding Contents and Acknowledgements, Dedications, Copyright pages, Author bios, Previews and pass each chapter to Pangram's API. Pangram's detection algorithm returns a list of analyzed text windows spanning the input chapter. Each window includes text, label, ai_assistance_score, confidence, start_index, end_index, word_count, and token_length. Across all chapters of a book, we count the windows whose label is not "Human Written" and define the AI-detection score as the percentage of windows so flagged,

$$S_i \;=\; 100 \times \frac{|\{\, w \in W_i : \ell(w) \neq \text{“Human Written”} \,\}|}{|W_i|},$$

where W_i is the set of analyzed windows for book i and $\ell(w)$ the label of window w. We assign each title to one of three bands based on S_i,

$$\text{No AI text}_i = \mathbf{1}[S_i = 0], \quad \text{Light AI text}_i = \mathbf{1}[0 < S_i \leq 25], \quad \text{Substantial AI text}_i = \mathbf{1}[S_i > 25].$$

[9] https://post.substack.com/p/against-claudefishing

A book with no AI text has no flagged windows; one with light AI text has a flagged share above 0% and at most 25%; one with substantial AI text has a flagged share greater than 25%. These bands are discretized versions of continuous scores S_i. For our study, we used Pangram version 3.3 (Masrour et al., 2026) which was released in May 2026. This latest version shows improved performance (recall) on long-form (2,000+ words) text, humanized text, and has a smaller false positive rate (FPR) on non-native English writing. Its performance has been validated on a sample containing 566,745 texts drawn from 15 domains. Held-out evaluation reports an N-weighted false-positive rate of 0.04% and a false-negative rate of 0.14%. We also audited Pangram's false positive rate on a held out set of 100 self-published genre fiction books pre-2022 and none of the books were flagged (exact binomial 95% CI for the book-level false-positive rate: 0–3.6%).

B.2 Does the 25% cutoff matter?

We call a book as containing substantial AI text when more than 25% of its analyzed text windows are flagged by the detector. This leaves us with an open question on how much of our result depends on this threshold. Table 2 provides detail on varying the detectability thresholds and reports the share of titles, sales, and revenue that fall in the substantial band at each setting. The picture holds throughout. This means that while a higher cutoff puts fewer books in the substantial AI band, at every setting the band takes a smaller share of sales and revenue than of titles.

Table 2: The substantial band at different cutoffs, over the launch window.

Cutoff	Titles	Title share (%)	Sales share (%)	Revenue share (%)
$S > 10\%$	3,586	24.9	16.1	15.2
$S > 20\%$	3,066	21.3	13.1	12.2
$S > 25\%$	2,880	20.0	12.1	11.3
$S > 30\%$	2,741	19.0	11.0	10.2
$S > 40\%$	2,440	16.9	9.4	8.6
$S > 50\%$	2,168	15.0	7.9	7.1

C Market, rank, and exposure measures

C.1 Market outcomes and dilution

We calculate all major market variables from observed total values. How do the title shares of each group translate into sales and revenue? We use three measures for that—title share, sales share, and revenue share of each band. For band h,

$$\text{title-share}_h = N_h/N$$

$$\text{sales-share}_h = \left(\sum_{i\in h}\sum_t Q_{it}\right) \Big/ \left(\sum_i \sum_t Q_{it}\right)$$

$$\text{revenue-share}_h = \left(\sum_{i\in h}\sum_t R_{it}\right) \Big/ \left(\sum_i \sum_t R_{it}\right)$$

Much of what follows uses a fixed window around the time of the title's launch, defined as the observed preorder days before the title goes on sale along with days 0 through 90 after it,

$$\mathcal{L}_i = \{t : t < O_i\} \cup \{t : 0 \le t - O_i \le 90\},$$

where O_i is the title's on-sale date, so that every title gets the same amount of follow-up time. We count preorder days as part of this window because Amazon records sales and rankings for some books before they officially go on sale. Within each book's 90-day launch window, we measure its total sales, total revenue, the number of days on which it made at least one sale, and its best ranking. For our analysis pertaining to the books in the upper tail, i.e., top 50%, 25%, 10%, 5%, and 1% we sort the titles by launch sales.

We measure dilution in two ways. The first measure tracks the market on a quarterly basis. For each quarter q we count released titles, titles with positive sales, total sales, and total revenue, and index each series to its 2023 Q1 level. We also compute revenue per selling title, i.e., titles that sold at least one copy,

$$\text{RevPerSellingTitle}_q = \frac{\sum_i \sum_{t \in q} R_{it}}{\#\{i : \sum_{t \in q} Q_{it} > 0\}}.$$

Because every title on sale in a given quarter enters this measure, it represents a snapshot of the market in which older and newer titles compete simultaneously. The second measure compares each title over a common time frame. Launch-window revenue per selling title is defined as total launch-window revenue divided by the number of titles with positive observed unit sales in that window. We compute the launch window revenue for each book that sold at least one copy over that time period, then, group titles by release quarter, and retain only the release cohorts with complete launch-window data. So, our measure only accounts for a release quarter when its last-released title can be observed through day 90 after its on-sale date. Finally, we report this metric for the entire marketplace, broken down by AI text band and genre, and for books without any AI content.

C.2 Rank-slot and exposure analyses

We use the Amazon sales rank as an indicator of market positioning, where a lower number indicates a better position. We cannot see the internal signals that influence impressions or placements of books on the platform, such as recommendation algorithms, search result rankings, and paid advertising, so our rank-slot analyses relate solely to observed sales-rank positions.

Within each genre-week cell, we take each book's median observed sales rank, order titles by that median, and define Top-10 and Top-25 rank slots. For the genre heatmap, we compute the share of Top-25 slots held by books with substantial AI text in each genre-quarter cell.

Next, for the top-rank churn analysis (Figure 2D), we compare the sets of Top-25 book titles from one quarter to the next. Let T^{25}_{gq} denote the titles holding Top-25 rank slots in genre g during quarter q. Retention of books with no AI text is the fraction of the prior quarter's no-AI Top-25 titles that remain in the current quarter's Top-25, and the entrant share is the fraction of new Top-25 entrants with substantial AI text,

$$\text{Retention}_{gq} = \frac{\left|\{i \in T^{25}_{g,q-1} : i \text{ has no AI text}\} \cap T^{25}_{gq}\right|}{\left|\{i \in T^{25}_{g,q-1} : i \text{ has no AI text}\}\right|},$$

$$\text{EntrantShare}_{gq} = \frac{\left|\{i \in T^{25}_{gq} \setminus T^{25}_{g,q-1} : i \text{ has substantial AI text}\}\right|}{\left|T^{25}_{gq} \setminus T^{25}_{g,q-1}\right|}.$$

Because the first quarter does not have any preceding quarter to compare against, we exclude it from our analyses of churn.

To perform AI-exposure analyses described in Figures 4A and B, we first construct a genre-month level panel. Then, for every cell (genre-month) in that panel we calculate, (a) the share of the catalog containing books with substantial AI text; (b) the share of sales of books with no AI text; (c) the share of Top-25 slots occupied by books with no AI text; (d) the Kindle Unlimited (KU) share of books. Next, in Figure 4A, we group all the genre-month cells into bins based on the share of books with substantial AI in each of them and plot the average Top-25 share held by books with no AI text in each bin. The 95% confidence intervals are calculated as $\bar{y} \pm 1.96\,\text{SE}$.

Then, in Figure 4B, we test how the no-AI lead changes with Kindle Unlimited availability. We rank the genre clusters by their KU share and divide them into a KU-light group and a KU-heavy group. We drop the middle. Within each KU group, we define the lead as the difference in sales or revenue share between books with no AI text and books with substantial AI text. We then estimate how it differs between the two groups with a linear regression on genre-month observations with month and exposure-bin fixed effects. The main figure reports the adjusted shares of each band in each KU group, and the difference in the no-AI lead between the groups, for both sales and revenue. We compute confidence intervals by resampling genre clusters. We read this difference as heterogeneity by KU availability across genres, and not as a causal effect of Kindle Unlimited.

D PRODUCER-SCALING ANALYSIS

We use the byline identifier developed above to track author output around the time-period that they adopted AI. For each author, their adoption month is computed as the first month in which they publish a book containing substantial AI text. Then, for every author who produced at least one such book during the study period, we build an author-month panel containing the titles they released in each month, separately accounting for the titles with substantial AI text and other titles.

We report the number of titles that each author produced in event time, relative to each author's AI adoption month. The pre-adoption period covers the time prior to when the author adopted AI and the post-adoption period covers months 1-12 following their adoption. We ignore the adoption month, because it can conflate the first title with substantial AI text with other titles, potentially with no AI text, from the same production burst. We also exclude authors whose panels do not have at least 3 months in the post-adoption period. For each remaining author, we calculate the output per month before and after adoption, the titles with substantial AI text and other titles published in the post-adoption period, and the change in monthly output across the two periods.

Further, we define observed existing authors as those authors who had at least one observed title prior to their adoption month, and newly observed authors as those who did not. This second category describes only our study panel and does not indicate that the author published nothing elsewhere.

We rank all authors by the increase in their average monthly volume of titles published after adoption and display the 15 greatest gains, breaking down their total post-adoption output into titles with substantial AI text and other titles. We assign each of these authors a single dominant post-adoption genre, the genre in which they publish the most titles after adoption.

We rank author identities by the revenue their books with substantial AI text generate during the 12 months following adoption, and we rank individual books with substantial AI text by the revenue each earns in its own first 12 months on sale. In both these analyses we keep only those authors and books that had a complete 12-month follow-up, and we display the top 15 of each. This 12-month horizon differs from the 90-day launch window used in the market-level analyses. Concentration metrics for these two rankings are reported in Table 3; specifically, we report both the Gini coefficient and the inverse-Herfindahl effective count.

Table 3: Concentration of output and revenue after adoption. Revenue is gross consumer spending, not author earnings.

Quantity	Population	Total	Gini	Effective count	Top-15 share (%)
Substantial-AI-text books published after adoption	385 bylines	1,310 books	0.476	165.4	20.2
Same, among authors who raised their output	287 bylines	1,022 books	0.483	125.4	24.0
Revenue from those books	466 bylines	$25.5M	0.737	71.2	34.7
Revenue per book over its first 12 months	1,753 books	$32.4M	0.730	297.3	14.7

E RARE-EXPRESSION OVERLAP

We measure rare-expression overlap using the full text of the top-revenue books (across entire observation period) with substantial AI text ($> 25\%$) and those with no AI text. Each book was broken down into word sequences to enumerate all expressions containing at least five words. A particular expression is considered to be an example of rare existing-book language if both of the below criteria were met: it appeared in no more than five different volumes within the Google Books index; it did not appear anywhere within infini-gram, a 4.7-trillion-token web snapshot. The first criterion identifies language used in a very limited set of publications; the second eliminates expressions which occur frequently in general web-based language. The individual books being tested are themselves excluded from the Google Books index.

Following this, we calculate rare-expression coverage for each book as the ratio of the number of tokens flagged as being part of at least one rare expression to the total number of tokens within that volume (Liu et al., 2026). Coverage is measured at the token level: every word within a given book is either covered by at least one rare expression, or not. Because we flag words rather than counting

instances of expressions, overlapping matches do not inflate the total, and coverage remains bounded between 0 and 100%.

To see if such overlap corresponds to commercial success, we fit a linear model to the top 200 books from each group:

$$C_i = \alpha + \beta_1 AI_i + \beta_2 \log_{10}(Revenue_i) + \beta_3 AI_i \times \log_{10}(Revenue_i) + \epsilon_i.$$

Here, C_i is rare-expression coverage and AI_i is a binary variable marking books with substantial AI text (i.e., $>25\%$). The parameter β_2 represents the revenue slope for books with no AI text and $\beta_2 + \beta_3$ the slope for books with substantial AI text, so β_3 measures how much more steeply coverage rises with revenue for AI-heavy books than for books with no AI text. We compare Top-K means with Welch's t-tests and estimate the model above using ordinary least squares (OLS), adding genre fixed effects to our controlled model; we also provide effect sizes in percentage points of book coverage.

As a reference group, we add 200 literary award-winning or nominated books across literary, historical, and speculative fiction that won or were nominated for the Booker Prize, the Pulitzer Prize, the National Book Award, the National Book Critics Circle Award, the PEN/Faulkner Award, and the Nobel Prize.

F Possible explanations for competition through scale

This study does not attribute a reason for competition through scale; it merely documents the phenomenon. One might speculate that the absence of disclosure of significant AI content may give the new entrants a competitive advantage, or allow them to elude a competitive disadvantage, assuming that readers informed of the presence of AI content will be less inclined to buy the book. We do not, however, have the data that would permit us to draw that conclusion. It is also possible that the pen names of some producers of AI books acquire a kind of brand name recognition: if AI-aided "Alice Author" has a following, the surfeit of books generated under that pseudonym, relative to human authors, may better satisfy audience demand for more titles with less waiting for new titles to appear, and may thus increase sales attributable both to volume and to brand loyalty. This outcome may resemble results obtained by brand-name authors who employ a stable of human ghostwriters to keep up with reader demand, but, as with AI book production generally, the scale and speed of production far outpace the endeavors of even multiple humans. Here again, however, we lack data addressing the brand-name effect of particularly prolific AI pseudonyms.