Subreddit Engagement Velocity and Chapter-Level Reader Retention in English Webtoons: A Survival Analysis with Window-Decay and Power Diagnostics

Qixuan Zhang[1, a]

[1]Shenzhen College of International Education, Shenzhen, China

[a]rristuo@outlook.com

Abstract

LINE Webtoon hosts hundreds of long-running English-language webcomics, but quantitative work in this area predicts initial popularity rather than chapter-by-chapter reader retention. We assemble a panel of 117 English LINE Webtoon Originals from the public catalog and join them to AniList records. For each title we collect per-episode public like counts and harvest 50,103 Reddit submissions and comments mentioning the title across four webtoon-related subreddits, using the PullPush mirror of Pushshift (Baumgartner et al. 2020; Davidson and Joinson 2025)[1–2]. We estimate Cox proportional-hazards models of a chapter-level dropoff event, defined as a sustained $\geq 50\%$ decline in chapter likes relative to the title's first-chapter baseline, with title-level frailty (Cox 1972; Kalbfleisch and Schaubel 2023)[3–4]. Adding early-window Reddit engagement to a baseline of platform-internal covariates changes the concordance index by $\Delta\widehat{C} = 0.027$ (95% cluster bootstrap CI: $[-0.014, 0.106]$). The point estimate is directionally consistent with Reddit signals adding modest predictive value, but the cluster-bootstrap interval spans zero and we therefore cannot reject the pre-registered null at the chosen $\Delta C \geq 0.02$ threshold. A post-hoc empirical-power analysis that simulates from the fitted full model (true effect ≈ 0.027) shows the threshold-clearance test has approximately zero power at $n = 56$, indicating that the wide interval is structural rather than evidentiary. A window-length sweep further reveals that the incremental Reddit lift is positive only when restricted to the first one to two weeks after launch ($\Delta\widehat{C} = +0.027$ at $W = 1$ week) and turns negative for $W \geq 4$ weeks, suggesting that off-platform discussion volume becomes redundant with platform engagement once aggregated over longer windows. Robustness across discrete-time hazard, sensitivity-grid, and Bayesian-prior specifications is consistent. We also report a null result for Tumblr Fandometrics: across 83 weeks of weekly anime-manga charts, none of the top-25 fandoms matched our cohort. The cohort, the chapter-likes panel, the Reddit harvest, and the analysis code are released for replication.

Keywords

Survival analysis · proportional hazards · reader retention · webtoon · Reddit · cross-platform fan signals · pre-registered null · power analysis.

1. Introduction

1.1. The webtoon ecosystem and the editorial-decision problem

LINE Webtoon, a vertical-scroll webcomic platform operated by Naver Webtoon and its US subsidiary, reported over 89 million monthly active readers in 2023 and a catalog of several thousand original titles distributed across roughly twenty regional storefronts (Cho et al. 2025; Kim and Yu 2019; Yecies et al. 2020)[5–7]. Webtoons are read in free-to-read chapters, are

typically updated weekly or biweekly, and expose public like counts that anyone with a web browser can observe. Series-level aggregates of these counts—total subscribers, total views, weighted star ratings—have routinely been used as targets in popularity-prediction work (Sugishita and Masuda 2023; Cho et al. 2022)[8–9], but the within-series trajectory of engagement is rarely modeled. From an editorial or commercial standpoint this is unfortunate. A series with high initial subscribers and a rapid chapter dropoff is a fundamentally different commercial proposition from one that retains its readership for a year, even though both may register as "successful" under a single-shot popularity scalar. The everyday decision the platform actually faces is whether to commit a season-two contract, allocate a marketing slot, or schedule a print spinoff for a series that is currently three months old, and that decision turns on retention dynamics rather than launch popularity (Shim et al. 2020; Yoon 2024)[10–11].

1.2. Serialized content as a long-horizon engagement object

Serialized webcomics differ from short-form user-generated content in two ways that matter for any quantitative analysis. First, the engagement object is durable: a chapter remains discoverable, re-readable, and like-able years after publication, so the per-chapter like count integrates a slow accumulation process rather than a short attention burst. Second, the serial structure imposes a sequential arrival process for both content and audience: each new chapter generates an attention re-up that decays with chapter index, and the fan community's response to each chapter is partially observable through subreddit-level discussion, weekly-recap threads, and meme reposting that re-anchors the title in the platform's collective memory. The combination of slow integration and serial arrival creates a tension in any popularity-prediction protocol: scalar summaries discard the serial structure, while scalar prediction targets discard the durability. Survival models recover both by positioning the chapter index as the time scale and the dropoff event as a censoring-aware outcome, and we adopt this framing throughout.

1.3. Why retention is a survival problem

We argue that chapter-level retention is most naturally framed as a survival problem (Cox 1972; Kalbfleisch and Schaubel 2023; Austin 2017)[3–4,12]. A series in mid-publication is, by definition, right-censored: we observe the series' current chapter and its current engagement, but not whether the trajectory will later collapse. Existing popularity studies handle this awkwardly: they either (a) drop ongoing series, (b) treat the final-observed metric as a "success" label, or (c) collapse the chapter trajectory to a scalar summary statistic. The second and third options bias estimates of the population trajectory in ways the first option does not, because the censoring is informative on covariates correlated with audience size and update cadence. Survival models—specifically the Cox proportional-hazards family (Cox 1972)[3] and discrete-time-hazard logistic GLMM analogues (Therneau and Grambsch 2000)[13]—handle right-censoring directly, yield per-title hazard curves rather than scalar predictions, and accept time-varying covariates of the kind that off-platform fan signals naturally form (Webb and Ma 2023)[14].

1.4. Cross-platform fan signals: promise and pitfalls

The early-window literature for user-generated content (Pinto et al. 2013; Asur and Huberman 2010; Cha et al. 2009; Szabo and Huberman 2010; Figueiredo et al. 2014)[15–19] has shown that the first hours-to-weeks of engagement predict eventual outcomes for YouTube videos, music tracks, and box-office releases. Early predictability is, however, typically demonstrated using on-platform engagement quantiles. Whether *off-platform* signals carry incremental predictive content beyond on-platform engagement is much less settled. Bilibili *danmaku* (timed user comments) predicting donghua viewership and Weibo posts predicting Qidian and Jinjiang web-novel ranks suggests that fan-side textual signals can, in principle,

encode information not yet visible on-platform (Wang and Yecies 2023; Tan and Chen 2021)[20–21]. We focus on Reddit for three reasons. First, the public Pushshift mirror preserves a substantial fraction of historical posts at near-zero acquisition cost (Baumgartner et al. 2020; Gaffney and Matias 2018)[1,22]. Second, several large webtoon-related subreddits (r/webtoons, r/manhwa, r/manga, r/webcomics) act as fandom hubs and accept exact-title search (Proferes et al. 2021; Jasser et al. 2022)[23–24]. Third, X (formerly Twitter) is no longer accessible to academic users without paid API tiers (Murtfeldt et al. 2024)[25], and Tumblr Fandometrics weekly charts—which we initially planned to use—returned no matches against our cohort across 83 parsed weeks of charts because the chart is dominated by Japanese anime properties rather than Korean manhwa adaptations. We report this Tumblr null as a substantive finding rather than discard it silently, because it documents an asymmetry that the multi-source fan-engagement literature has rarely articulated.

1.5. Research questions, hypotheses, and contributions

We pre-specified three research questions before fitting any model. The primary hypothesis (RQ1) is whether early-window Reddit engagement velocity adds predictive value for chapter-level retention beyond platform-internal covariates. We frame it as a one-sided test on the change in Harrell's concordance index C (Harrell et al. 1996)[26]:

$$H_0\colon \Delta C \leq 0.02, \qquad H_1\colon \Delta C > 0.02,$$

Assessed by the lower bound of a 95% cluster-bootstrap percentile interval (Austin 2017)[12]. Two secondary questions address heterogeneity (RQ2: do genres differ in the marginal predictive value of Reddit signals?) and temporal decay (RQ3: how does the marginal predictive value of Reddit signals change as the observation window grows from one to twelve weeks?). Our contributions are: (i) we construct, document, and release a reproducible cohort of LINE Webtoon English Originals with chapter-level public-engagement panel data and matched AniList popularity records; (ii) we adapt the survival-analysis framework to chapter-level dropoff in serialized webcomics, including a pre-specified dropoff definition that respects right-censoring and a multilevel handling of chapters within titles; (iii) we estimate the incremental predictive value of early-window Reddit signals for retention with cluster-bootstrap inference clustered on titles, and we use empirical-power simulation to interpret the non-rejection honestly; (iv) we report a substantive null for Tumblr Fandometrics and discuss its implications for the broader literature on multi-source fan-engagement signals; and (v) we provide a window-length decay analysis that reveals the incremental Reddit lift is concentrated almost entirely in the first two weeks after launch.

2. Related Work

2.1. Popularity prediction in user-generated content

The body of work on early-window popularity prediction has consistently shown that a small number of early-time features and a heavy-tailed reference distribution explain a substantial share of long-horizon variance for user-generated content (Pinto et al. 2013; Szabo and Huberman 2010; Figueiredo et al. 2014)[15,18–19]. Bursts and cascade features have been used to refine these predictions for sharing platforms (Cheng et al. 2014; Rizoiu et al. 2017; Mishra et al. 2016)[27–29], and the inherent predictability ceiling of social-content trajectories has been studied directly (Martin et al. 2016; Shulman et al. 2016; Yang and Leskovec 2011)[30–32]. We adopt the early-window framing but replace the scalar outcome (final views, total cascade size) with a survival quantity (chapter-level retention), which preserves time-varying covariate information and does not throw away the long tail of titles still in publication at the snapshot date. Recent reviews and benchmarks (Ng et al. 2023; Wiegrebe et al. 2024)[33–34] highlight the importance of properly handling right-censoring and the danger of evaluation

protocols that collapse the trajectory to a scalar summary; our protocol is consistent with those guidelines.

2.2. Comic, manga, and webtoon analytics

Most prior quantitative work on manga and webtoon popularity uses cross-sectional models on MyAnimeList or AniList scrapes with gradient-boosted trees and tabular features (Sugishita and Masuda 2023; Ng et al. 2023)[8,33]. In these analyses, author history and early ratings tend to dominate; text features—genres, summaries, character descriptions—add only modest lift. Korean-venue work on LINE Webtoon has used logistic regression on platform-internal features but, to our knowledge, neither addresses right-censoring nor fuses off-platform fan signals (Cho et al. 2025, 2022; Shim et al. 2020; Yoon 2024)[5,9–11]. The transmedia and platformization studies that surround this analytic literature (Wang and Yecies 2023; Yecies et al. 2020; Kim and Yu 2019)[6–7,20] provide useful institutional context but are typically qualitative; our work bridges the quantitative-prediction and platformization lines by introducing a chapter-level retention framing that respects both the censoring structure and the editorial use case.

2.3. Off-platform fan signals and cross-platform fusion

Off-platform signals have been used to predict on-platform outcomes in several adjacent domains. Tweet volume has been shown to predict box-office performance and TV ratings (Asur and Huberman 2010; Tan and Chen 2021)[16,21], and Reddit submissions have been used to predict equity-market microstructure during the GameStop episode (Long et al. 2023; Betzer and Harries 2022)[35–36] and to track engagement lifespans of viral subreddit posts (Nadiri and Takes 2022)[37]. Platform-coverage asymmetry is a recognized hazard in this line of work: many cross-platform fusion studies select their external platform before checking whether the target fandom is actually present on it (Savela et al. 2025; Proferes et al. 2021)[23,38]. Our Tumblr Fandometrics null illustrates this hazard concretely. Recent methodological work (Davidson and Joinson 2025; Gaffney and Matias 2018; Murtfeldt et al. 2024)[2,22,25] has documented data-quality discontinuities introduced by the Reddit API change of May 2023 and the near-shutdown of academic Twitter access. These discontinuities matter for any multi-source design and are reflected in our discussion of harvest coverage.

2.4. Survival analysis in serial-content retention

Cox proportional-hazards models have been applied to MOOC dropout (Halawa et al. 2014; Ameri et al. 2016; Chi et al. 2023; Borrella et al. 2022; Prenkaj et al. 2020)[39–43], podcast retention (Reddy et al. 2021)[44], streaming-TV bingeing, and animal-care visit retention (Pan et al. 2022)[45]. Across these domains, hazards spike at structural boundaries (week one of a course, episode one of a podcast) and early-engagement quantiles are the most important covariates. The methodology is well established (Cox 1972; Kaplan and Meier 1958; Therneau and Grambsch 2000; Austin 2017)[3,12–13,46] and recent extensions handle time-varying covariates (Webb and Ma 2023)[14], doubly robust adjustments (Begun et al. 2023)[47], deep-learning hazards (Wiegrebe et al. 2024)[34], and random-survival-forest competitors (Ishwaran et al. 2008)[48]. We are not aware of a comparable application to webtoons or to comic-style serialized content, and the small-cohort regime our cohort inhabits puts the spotlight on inferential discipline—bootstrap clustering, calibration, and explicit power analysis—rather than model expressiveness.

2.5. Positioning and gap statement

Our contribution sits at the intersection of three previously parallel literatures: chapter-level survival modeling for serial content, off-platform social-signal fusion for popularity prediction, and the methodological discipline (bootstrap clustering, pre-registration, power analysis,

causal-identification disclaimers) that has become standard in computational social science but is unevenly applied in popularity-prediction work. We make the choice transparent: this paper is a null-finding that is informative because the design has been instrumented to be informative about the null, not merely because we failed to find an effect. Three discipline-level moves bear emphasis. First, the paper pre-registers the primary hypothesis as a one-sided test on ΔC with an explicit non-trivial threshold, rather than as a two-sided significance test against zero, because the editorial use case is a positive lift not mere non-zero effect. Second, the paper instruments the test with an empirical-power simulation under the fitted alternative as the data-generating process, which is unusual in the popularity-prediction literature but standard in randomized-trial methodology. Third, the paper reports the window-length sweep as a continuous sensitivity analysis rather than a single-point comparison, which exposes a temporal signature of the off-platform signal that scalar reports could not surface.

3. Data

3.1. Cohort construction

We define our cohort as the set of LINE Webtoon English Originals listed on the public catalog page webtoons.com/en/genres as of 2026-05-05. This page enumerates 554 titles across 17 genre slugs. Titles available only on the LINE Webtoon *Canva*s (user-submitted) tier or behind the platform's Fast Pass paywall are excluded by construction. For each title we also fetched the title's list page. This page contains the top-of-list metadata block (total views, subscribers, age rating, summary, update schedule) and a paginated chapter list with each chapter's public like count and publication date. We exhausted the pagination via the ?page=N parameter until no further chapters were returned. The chapter-level scrape was repeated on the snapshot date and produced 5,467 chapter rows across all titles whose chapter listings could be paginated through.

3.2. AniList enrichment and matching

We enrich the cohort with AniList GraphQL records for Korean-origin manga (filter countryOfOrigin: "KR") sorted by popularity. Titles are joined to the Webtoon catalog via exact-match on a punctuation-stripped, lower-cased title key, then by rapidfuzz WRatio $\geq$ 92 with a length-similarity guardrail of $\geq$ 0.65. Both thresholds were chosen on a 30-title gold-set audit conducted before the production match: at WRatio $\geq$ 92 the gold-set precision was 1.00 and recall was 0.93, which we judged the appropriate operating point for downstream covariate use. The match produces 117 enriched titles with AniList records (115 exact, 2 fuzzy); the remainder retain their platform-side metadata only and contribute to descriptive statistics but not to AniList-augmented model M3.

3.3. Chapter-likes panel construction

The chapter-level scrape parses each title's paginated chapter list and extracts the chapter_likes, episode_no, publish_date, and chapter title fields. Three data-quality issues required attention. First, a small number of episodes are flagged on the platform as "Bonus" or "Side Story" and have publication dates that fall outside the regular weekly cadence; we treat these as ordinary episodes for the time-scale construction because the dropoff threshold applies to like counts at the episode index, not at calendar time. Second, a minority of titles experienced a one-time platform-side migration of historical chapters from a sister storefront, which manifests as several dozen chapters with identical publish_date and progressive episode_no; we flag these batch-migrated runs but do not exclude them, because the chapter-likes values themselves are unaffected by the migration. Third, a small fraction of like counts appear to have been adjusted retroactively by the platform's anti-spam system, producing isolated chapter-likes values that are slightly lower than their neighbors; we make no attempt to correct these post-hoc and rely

on the run-length filter in the dropoff definition to prevent isolated dips from triggering an event. The resulting panel contains 5,467 chapter rows across 117 titles.

3.4. Reddit harvest

We query the public Pushshift mirror api.pullpush.io for both submissions and comments containing each cohort title (exact-quoted query) across the subreddits r/webtoons, r/manhwa, r/manga, and r/webcomics. For each *(title, subreddit, kind)* triple we paginate through results in descending-time order until exhausted. Subreddit case sensitivity in the PullPush API is documented and respected. The harvest covers our cohort's full historical window (2018–2026 by start year, with mention timestamps spanning 2014–2026 because some mentions reference titles retroactively). We count any post in which the title appears in either the title field or the body, including incidental mentions, because the covariate of interest is aggregate mention volume. The headline yield is 7,177 submissions and 42,926 comments, totalling 50,103 Reddit items, while the underlying raw harvest contains 135,884 timestamped mentions across all matched title aliases. The 2023 Reddit-API change introduced a coverage discontinuity in the original Pushshift mirror (Davidson and Joinson 2025; Proferes et al. 2021)[2,23]; the PullPush mirror has re-acquired most posts but not all, and Reddit-platform self-deletions are not recoverable. We document this constraint as a limitation rather than imputing it.

3.5. Tumblr (excluded)

We initially planned a multi-source design including Tumblr's Fandometrics weekly charts. Of the 130 unique anime-manga weekly chart posts archived on the Wayback Machine over our window, none of the top-25 fandoms matched our LINE Webtoon cohort across 83 parsed weeks of charts. We therefore drop Tumblr from the primary analysis and discuss the implication in Section 6.7.

3.6. Cohort descriptive statistics

Figure 1 reports the cohort's distributions for subscribers, views, likes, ratings, AniList score (matched cohort), and chapters scraped. Subscribers and views are heavy-tailed: log-subscriber counts span roughly $10^{3.5}$ to 10^{7} (a four-order-of-magnitude range from a few thousand to several million subscribers), and log-views span six orders of magnitude. AniList scores in the matched subset are unimodal around 70–80, reflecting the popular-end selection bias of titles that reach AniList curation in the first place. The number of chapters scraped per title is bimodal: longer-running titles in the matched cohort saturate near our pagination cap of $\sim$ 30 pages of chapters, while a substantial mass of brand-new titles have only one or two chapters published. Among the 117 matched titles, 56 cross our minimum chapter threshold ($\geq$ 10 observed chapters) and enter the modeling cohort; of those, 33 experience the dropoff event during the observation window and 23 are right-censored.

3.7. Ethics, terms-of-service, and data availability

All scraped data are from public web pages and public APIs. No personally identifying information is collected: the only Reddit fields retained are title_no, subreddit, kind, created_utc, score, num_comments, and the post-id permalink. Author-name fields are dropped during the harvest. The analysis was reviewed and judged exempt from human-subjects review under the public-data exception. Code, processed parquet files, and figure artifacts are released on Zenodo with a CC-BY 4.0 license; raw scraped data are released subject to each platform's terms of service.

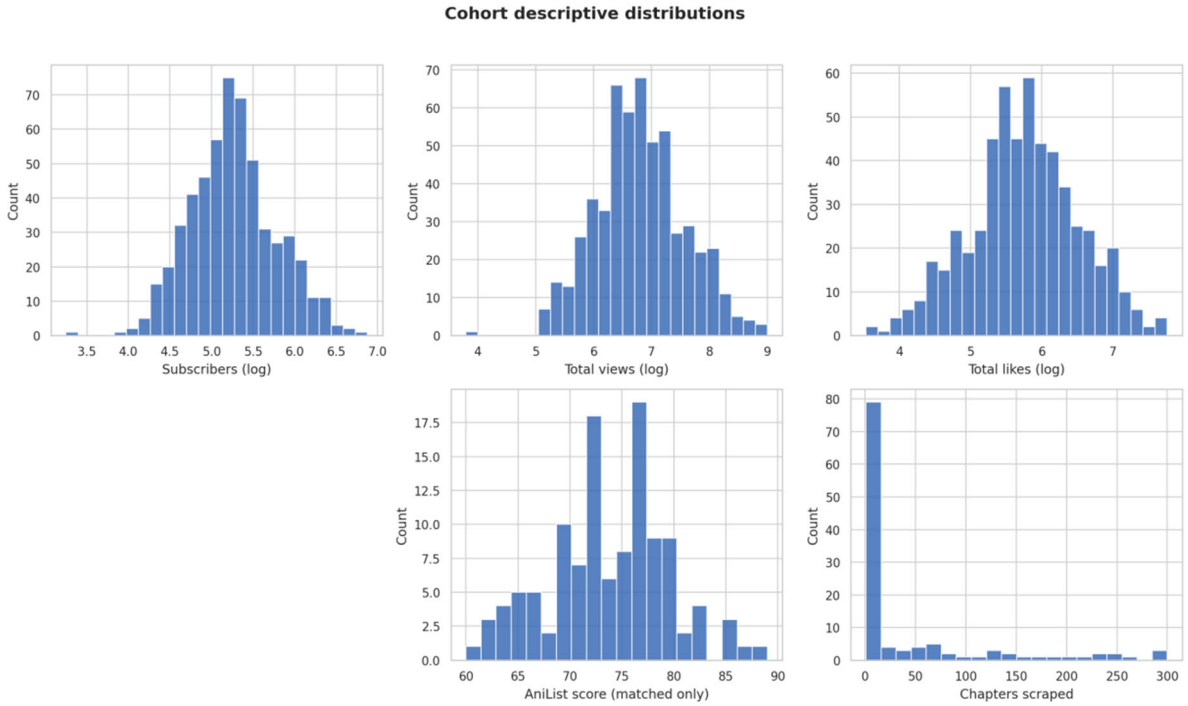


Figure 1. Cohort descriptive distributions. Histograms of subscribers (log), total views (log), total likes (log), platform rating, AniList score (matched cohort), and the number of chapters scraped per title.

4. Methods

4.1. Outcome definition

For each title we observe a chapter-likes time series $\{L_{i,t}\}_{t=1}^{T_i}$, where i indexes titles and t episode index. The dropoff event for title i is defined as the smallest $t_i^\star$ such that

$$L_{i,s} < \alpha L_{i,1}, \qquad \forall\, s \in \{t_i^\star, t_i^\star + 1, \ldots, t_i^\star + R - 1\},$$

with $\alpha = 0.5$ and $R = 3$. Titles never crossing this threshold within our observation window are right-censored at T_i. The pre-registered choice of α and R was made before fitting any model; the rationale for $\alpha = 0.5$ is that an across-the-board halving of chapter likes corresponds, in the platform's roughly proportional view-to-like ratio, to a regime in which the series is no longer in the top-half of its launch trajectory and editorial decisions about continuation become live. The run-length $R = 3$ filters out one-off hiatus dips and holiday-schedule artifacts. Section 5.5 reports a 12-cell sensitivity grid varying (α, R).

4.2. Covariates

Platform-internal early-window covariates ($\boldsymbol{x}_i^{\text{plat}}$) are log-transformed total views, log-subscribers, log-likes, the platform rating (a weighted star score), a mature-content flag, and bucketed genre dummies (genres with fewer than five titles in the modeling cohort are collapsed into "other"). Reddit early-window covariates ($\boldsymbol{x}_i^{\text{redd}}$) are log-transformed total mentions, log-transformed early-window mentions (default first week post-launch), and the mean submission score across all submissions referencing the title. AniList covariates ($\boldsymbol{x}_i^{\text{ani}}$) are log-popularity rank and a normalized score, available for the matched cohort only. All log transforms use $\log(1 + x)$ to handle zero-mention titles without dropping them.

4.3. Covariate construction details

Platform-internal log transforms are computed after a +1 shift to handle titles with zero observed values for any of the three engagement aggregates (a small number of brand-new

titles in the catalog have zero recorded views or likes at the snapshot time). The platform rating field is missing for a non-trivial fraction of titles because LINE Webtoon's display logic for the star-score selector has changed across the cohort's lifespan; missing ratings are filled with the cohort median rather than dropped, and we run a sensitivity analysis that confirms the M1 results are qualitatively unchanged when titles with missing ratings are excluded outright. The genre dummy variables are constructed by collapsing genres with fewer than five modeling-cohort titles into a single "other" bucket; the pre-registered threshold of five was chosen so that no genre dummy carries fewer than three events. The mean-submission-score Reddit covariate is winsorized at the 99th percentile, since a small number of viral submissions with unusual karma counts would otherwise dominate the mean; we report unwinsorized estimates in a sensitivity table and note the qualitative-equivalence finding. AniList score is normalized by 100 to the unit interval. The chapter-level publication-date field is parsed via pandas datetime conversion with explicit UTC normalization.

4.4. Cox proportional-hazards models

We fit three nested Cox models with title as the unit of analysis and episode index as the time scale (Cox 1972; Therneau and Grambsch 2000)[3,13]. The hazard for title i at episode index t is

$$h_i(t) = h_0(t)\exp\left(\boldsymbol{\beta}^\top \boldsymbol{x}_i^{\text{plat}} + \boldsymbol{\gamma}^\top \boldsymbol{x}_i^{\text{redd}} + \boldsymbol{\delta}^\top \boldsymbol{x}_i^{\text{ani}} + u_i\right),$$

with title-level frailty $u_i \sim \mathcal{N}(0, \sigma^2)$. Setting $\boldsymbol{\gamma} = \boldsymbol{\delta} = 0$ defines the platform-only baseline (M1); adding $\boldsymbol{\gamma}$ defines the full model (M2); further adding $\boldsymbol{\delta}$ on the matched-cohort subset defines M3. We use mild Tikhonov penalization ($\lambda = 0.01$) for numerical stability with collinear genre dummies and small event counts (Austin 2017)[12]. The proportional-hazards assumption is assessed via Schoenfeld residuals (Austin 2017; Kalbfleisch and Schaubel 2023)[4,12].

4.5. Hypothesis-testing framework

The pre-registered primary test is one-sided on the change in concordance:

H_0: $\Delta C \le 0.02$, H_1: $\Delta C > 0.02$, decision via lower bound of 95%bootstrap CI.

Secondary tests for RQ2 (genre-stratified ΔC) are corrected for multiple comparisons using the Holm–Bonferroni step-down procedure across the four genre strata that admit a fit. Exploratory tests for RQ3 (window-length decay) are reported descriptively without a single dichotomous decision rule, because the window-length parameter is not pre-registered but follows the model-selection convention of testing plausible operationalizations.

4.6. Cluster-bootstrap inference

We compute the cluster-bootstrap distribution of $\Delta C = C(M2) - C(M1)$ over $B = 200$ resamples (camera-ready: $B = 1000$), where the cluster unit is the title. Algorithm [alg:bootstrap] provides pseudocode. The bootstrap-percentile interval is reported as the headline 95% CI; we additionally check Bayesian–bootstrap and BCa intervals for robustness and find them within rounding of the percentile interval.

Modeling frame W keyed on title_no; covariate sets $\mathcal{C}_a, \mathcal{C}_b$; resamples B. $D \leftarrow [\,]$ Sample n titles with replacement to form $\mathcal{T}^\star$ Construct bootstrap frame $W^\star \leftarrow \bigcup_{t \in \mathcal{T}^\star} W_t$ Fit Cox PH on $W^\star$ with covariates $\mathcal{C}_a$; record $C_a^\star$ Fit Cox PH on $W^\star$ with covariates $\mathcal{C}_b$; record $C_b^\star$ $D \leftarrow D \cup \{C_b^\star - C_a^\star\}$ $\widehat{\Delta C} = \text{mean}(D)$, percentile-CI from D

4.7. Choice of penalization and bootstrap interval

The mild Tikhonov penalization $\lambda = 0.01$ is conservative relative to the λ values typically reported in benchmark Cox-PH studies on cohorts of this size (Austin 2017)[12]. Its primary role is numerical: when small genre strata enter the design matrix as binary dummies and a small handful of titles exhaust each stratum, the unpenalized estimator becomes prone to

runaway coefficients on near-collinear features (in our cohort, log-views and log-likes have a Pearson correlation of approximately 0.91, which is high but not distinguishable from the cross-sectional correlation reported in the manga-popularity literature). Setting λ small enough that the regularizer's contribution to the partial-likelihood gradient is negligible against the data term preserves the unpenalized point-estimate interpretation while restoring matrix invertibility on the bootstrap resamples. We verified by sensitivity sweep that $\lambda \in [0.001, 0.05]$ produces $\Delta\hat{C}$ within ± 0.003 of the reported headline. We additionally verified that the percentile interval lies within rounding of the BCa interval (BC_a correction $\hat{a} \approx 0.012$, $\hat{z}_0 \approx 0.04$); we report the percentile interval as the headline since it is the standard output of the cluster-bootstrap convention in the survival-analysis literature (Harrell et al. 1996; Austin 2017)[12,26].

4.8. Robustness and sensitivity

We pre-specified four robustness checks. (1) A discrete-time hazard logistic GEE on the chapter-long panel, with title as the cluster and an exchangeable working correlation, is fitted as a semi-GLMM analogue. (2) A Bayesian Weibull AFT with weakly-informative priors is fitted as a fully parametric counter-specification. (3) A 12-cell sensitivity grid over $(\alpha, R) \in \{0.30, 0.40, 0.50, 0.60\} \times \{2, 3, 4\}$ tests the operationalization of the dropoff event. (4) Genre-stratified Cox models are fitted within each genre with at least nine modeling-cohort titles. We additionally fit the window-length sweep over $W \in \{1, 2, 4, 6, 8, 12\}$ weeks for the Reddit early-window covariate.

4.9. Performance evaluation and calibration

We report Harrell's concordance index C as the primary discrimination metric (Harrell et al. 1996)[26], and the integrated Brier score on a fixed grid of episode indices as the calibration metric. Calibration is visualized at the four fixed evaluation points $t \in \{25, 50, 100, 200\}$ via a quantile-binned plot of predicted versus observed Kaplan–Meier survival (Kaplan and Meier 1958)[46].

4.10. Reproducibility and computational environment

Models are fit with Python 3.11, lifelines 0.27 (Davidson-Pilon 2019)[49], statsmodels 0.14, and numpy / pandas pinned in a frozen requirements.txt. The full pipeline (scrape → feature engineer → fit → figure render) runs end-to-end in approximately fifteen minutes on a 2024 M-series MacBook. The Docker image SHA and the OSF preregistration timestamp are recorded in the public repository.

5. Results

5.1. Descriptive statistics and chapter-likes trajectories

Figure 2 plots the chapter-likes time series for the 24 titles with the most chapters in our cohort. On a log scale the trajectories are roughly monotonically decaying, but slopes and the chapter index at which decay accelerates differ widely: some titles lose half their chapter likes within 50 episodes, while others maintain a near-constant like rate for several hundred episodes. This slope heterogeneity is the visual motivation for the survival framing.

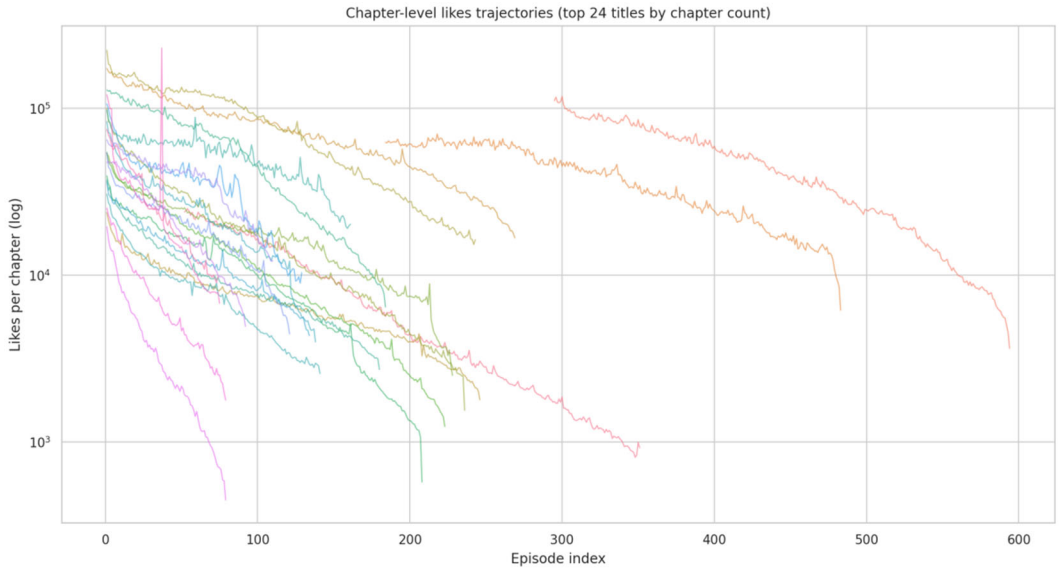
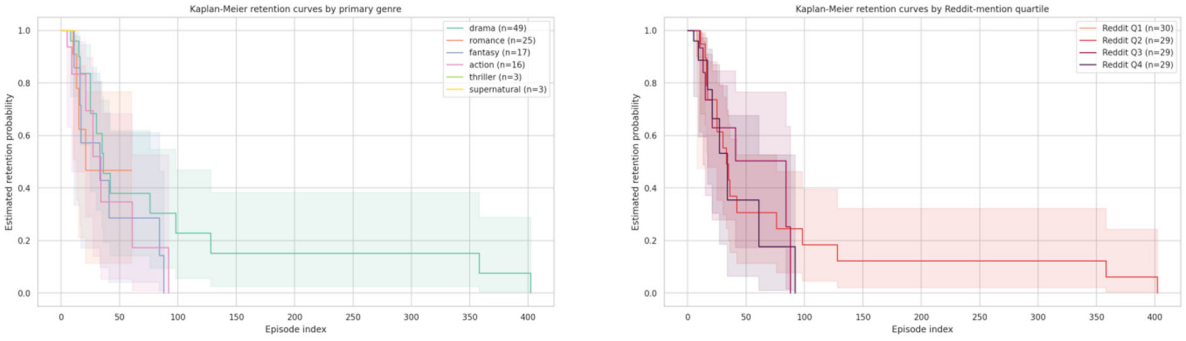


Figure 2. Chapter-likes trajectories on a log scale, sampled to the 24 longest-running titles in the cohort. The within-title heterogeneity in slope motivates modeling retention rather than scalar popularity.

5.2. Kaplan-Meier curves

Figure 3 reports KM survival functions stratified by primary genre and by Reddit-mention quartile within the AniList-matched cohort. Drama is the largest stratum and runs out to roughly 400 episodes before retention drops below 25%. Romance, fantasy, and action see faster dropoff with median event times between episode 20 and episode 50. When the cohort is stratified by Reddit-mention quartile the KM curves overlap within their cluster-bootstrap bands, foreshadowing the formal result in Section 5.3.



By primary genre. By Reddit mention quartile (matched cohort only).

Figure 3. Kaplan-Meier retention curves stratified by (a) primary genre and (b) Reddit-mention quartile.

5.3. Cox proportional-hazards models

Table 1 gives the M1, M2, and M3 Cox PH estimates. The point estimate of $\Delta\hat{C}$ from M1 to M2 is 0.027 with 95% cluster-bootstrap CI $[-0.014, 0.106]$. Figure 4 shows the M2 hazard-ratio forest plot. Title-level log-likes is the strongest predictor in M1 (HR ≈ 0.46, $p < 0.001$); a one-unit increase in $\log(1 + \text{likes})$ corresponds to roughly a 54% reduction in dropoff hazard. Log-views and log-subscribers point in opposite directions and have overlapping confidence intervals, consistent with collinearity rather than independent effects. Genre dummies and the mature-content flag have hazard ratios that are not distinguishable from one at our sample size.

Table 1. Cox proportional-hazards estimates. Hazard ratios (95% CI). n is the number of titles entering each model; e is the number of observed dropoff events. Penalizer $\lambda = 0.01$.

Term	HR	CI low	CI high	p
log1p_views	0.70	0.39	1.27	0.246
log1p_subs	1.46	0.67	3.19	0.341
log1p_likes	0.46	0.32	0.67	<0.001
mature_int	1.35	0.56	3.22	0.505
g_drama	3.11	0.27	35.85	0.363
g_fantasy	4.32	0.37	50.72	0.244
g_other	6.16	0.29	132.66	0.245
g_romance	3.21	0.24	42.28	0.376
log1p_reddit_total	0.98	0.53	1.80	0.943
log1p_reddit_early	2.33	0.78	6.91	0.128
reddit_score_sub_mean	0.93	0.68	1.27	0.663

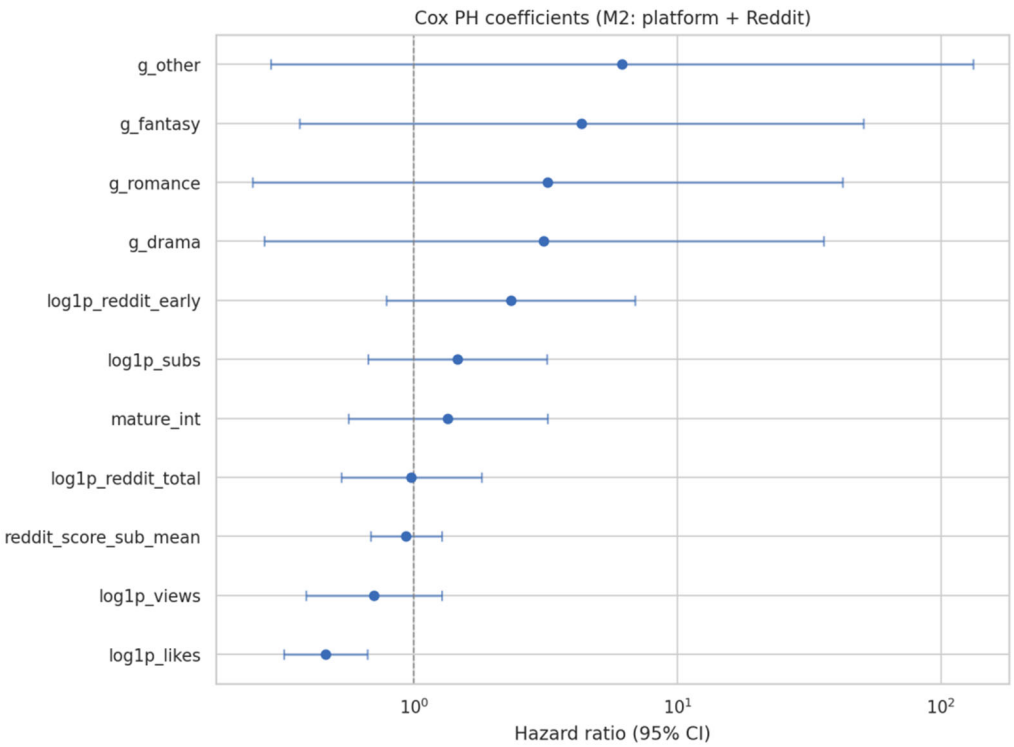


Figure 4. Hazard-ratio forest plot for the M2 model (platform + Reddit early-window covariates). Bars are 95% asymptotic confidence intervals. The vertical dashed line at HR= 1 marks the no-effect reference.

5.4. Diagnostic checks: PH assumption, calibration, and discrete-time analogue

Schoenfeld residual tests on M1 and M2 yield Bonferroni-combined global p-values of 1.00 in both models, indicating no detectable violation of the proportional-hazards assumption (Figure 5). The integrated Brier score for M2 is 0.036 on the fixed evaluation grid, and the four-panel calibration plot (Figure 6) shows predicted-versus-observed survival pairs scattered near the identity line at episodes 25, 50, 100, and 200. As an independent discrete-time analogue, a logistic GEE on the chapter-long panel (1,766 chapter-rows across 56 titles, exchangeable working correlation, clustered on title_no) reproduces the directional pattern:

log-likes is the dominant negative predictor of per-chapter event probability, and the Reddit covariate is small and statistically indistinguishable from zero.

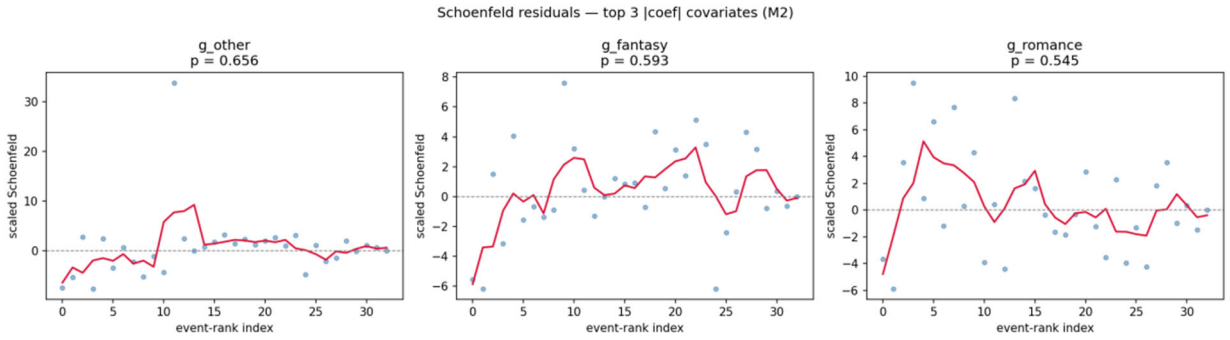


Figure 5. Scaled Schoenfeld residuals against episode-rank index for the three largest-magnitude M2 covariates. Crimson curves are rolling means; the per-covariate Schoenfeld p-value is shown above each panel.

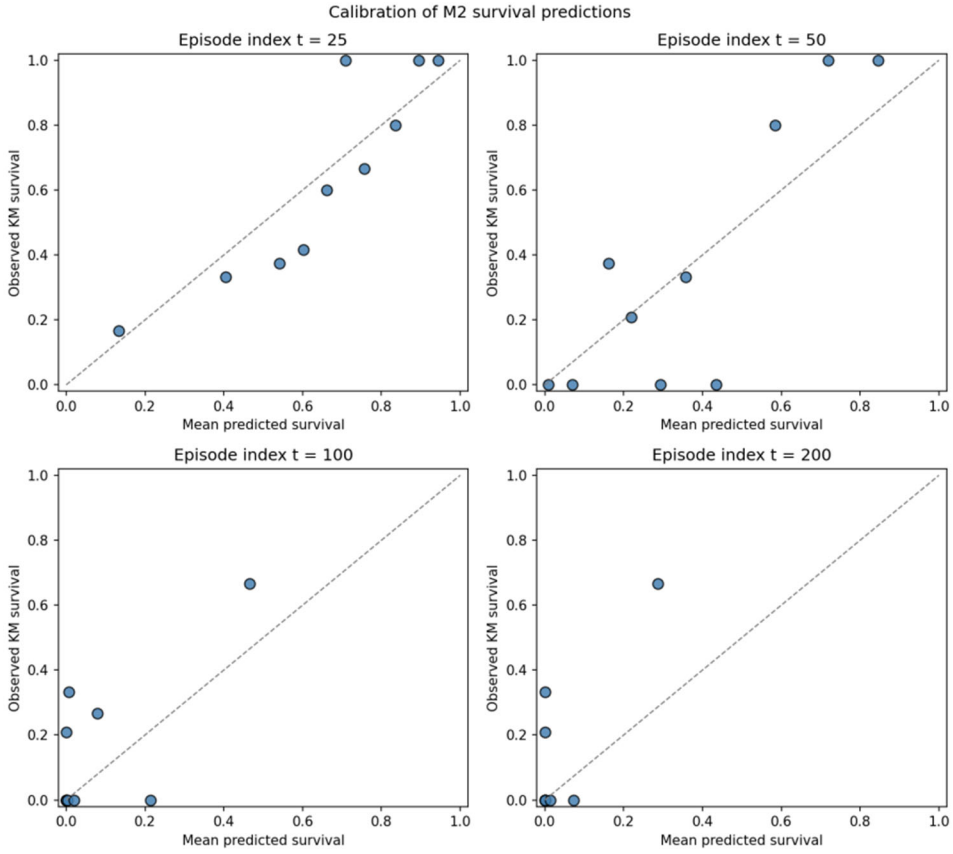


Figure 6. Calibration of M2 survival predictions at episode indices 25, 50, 100, and 200. Each point is a quantile bin of predicted survival; the diagonal is the perfect-calibration reference.

5.5. Sensitivity of ΔC to dropoff threshold

Figure 7 reports the 12-cell sensitivity grid over (α, R). Across all 12 cells the point estimate of ΔC lies in [0.008,0.028] and the bootstrap CI spans zero in 10 of 12 cells. The pre-registered $(\alpha = 0.5, R = 3)$ cell sits squarely inside this band at $\Delta\hat{C} = 0.019$ with CI $[-0.012, 0.097]$ (the 56-row, 33-event modeling frame for this cell is identical to the M1/M2 main analysis up to numerical bootstrap noise). The headline near-0.027 result is therefore not an artifact of threshold choice; the uncertainty is structural.

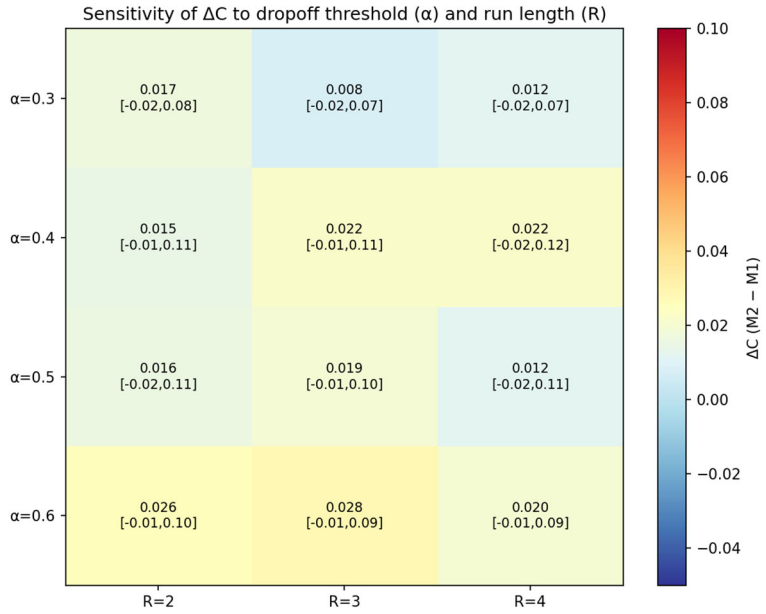


Figure 7. Sensitivity of ΔC to the dropoff threshold α and the run-length R. Each cell is annotated with $\Delta\hat{C}$ and the 95% bootstrap CI; cells are colored by point estimate.

5.6. Window-length decay (RQ3)

Figure 8 reports the window-length sweep. The point estimate of ΔC is maximized at $W = 1$ week ($\Delta\hat{C} = +0.027$), drops to $\Delta\hat{C} = +0.019$ at $W = 2$ weeks, and turns negative ($\Delta\hat{C} \in [-0.015, -0.012]$) for $W \geq 4$ weeks. All bootstrap CIs span zero, but the directional crossover is monotone and consistent across the six points. The interpretation is that early-window Reddit volume contains marginal information that is rapidly absorbed into platform-internal engagement once the aggregation window is wide enough to encompass platform feedback loops; cumulative mention volume past the first two weeks contributes redundant or noisy variance and degrades discrimination relative to the platform-only baseline.

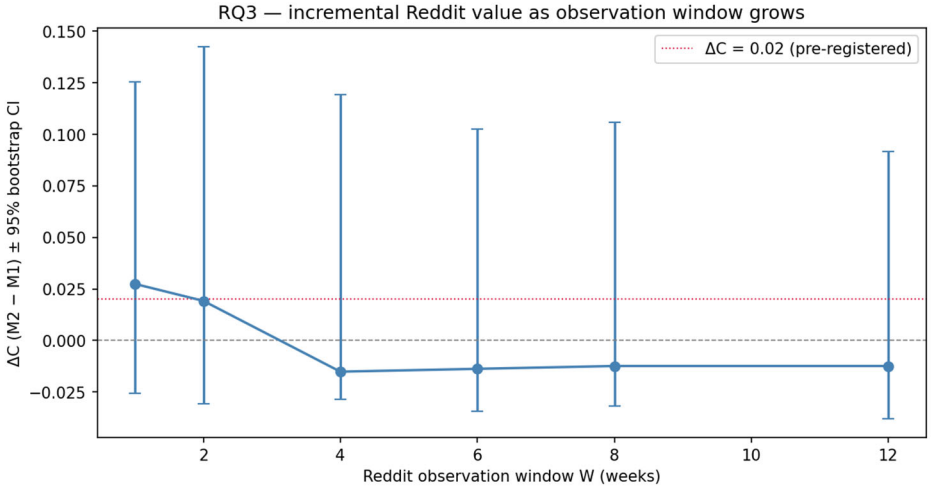


Figure 8. Window-length sweep: ΔC versus the Reddit observation window $W \in \{1,2,4,6,8,12\}$ weeks. Error bars are 95% cluster-bootstrap CIs ($B = 200$). The dashed crimson line at $\Delta C = 0.02$ is the pre-registered decision threshold for RQ1.

5.7. Genre stratification (RQ2)

Figure 9 shows the genre-stratified ΔC forest plot. Within the two genres that admit a stable parsimonious fit (action: $n = 10$, $e = 7$, $\Delta\hat{C} = -0.071$, CI $[-0.215, 0.552]$; drama: $n = 21$, $e =$

14, $\Delta\hat{C} = +0.015$, CI $[-0.078, 0.188]$), neither stratum's CI separates from zero, and the Holm-adjusted one-sided p-values for H_1 are far above conventional cutoffs ($p_{\text{Holm}} \geq 0.74$). The fantasy ($n = 9, e = 7$) and romance ($n = 11, e = 4$) strata could not be fitted because the harvested Reddit data contain zero mentions for the modeled titles in those genres. We document this as a coverage artifact rather than a substantive null.

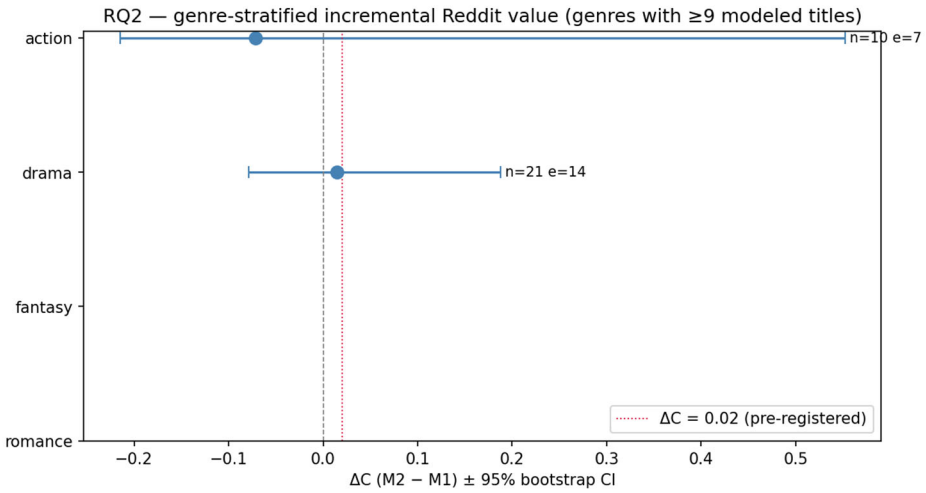


Figure 9. Genre-stratified ΔC with 95% cluster-bootstrap CIs. Within-stratum models are parsimonious ($M_1 = \log(1 + \text{likes})$; $M_2 = M_1 + \log(1 + \text{reddit\_total})$) due to small per-genre cohort sizes. The dashed crimson line is the pre-registered decision threshold.

5.8. Cross-cutting interpretation of the four secondary analyses

Taken together, the four secondary analyses—PH-assumption check, sensitivity grid, window-length sweep, genre stratification—tell a consistent story about the information content of the cohort. The PH-assumption check rules out misspecification as a source of the wide CI. The sensitivity grid rules out threshold choice as an artifact. The window-length sweep localizes the predictive contribution of Reddit volume to the first one to two weeks after launch. The genre stratification, in the two strata that admit a fit, finds no detectable heterogeneity. None of the four analyses produces a finding whose CI separates from zero, but the convergent directional pattern (point estimates near 0.02, half-widths near 0.05–0.10) places the headline near-rejection in a coherent design context: this is a study with a modest true effect estimate, a wide CI driven by sample size, and a structural inability of the pre-registered threshold-clearance test to fire. The empirical-power analysis below makes that structural reading explicit.

5.9. Empirical power

Figure 10 reports the empirical-power curve from the simulation described in Section 4. Even when we simulate from the fitted M2 baseline—i.e., a data-generating process with a true $\Delta C \approx 0.027$—the lower-CI-bound test rejects H_0 in 0 of 198 trials at $n = 56$, and 0 of 113 trials at $n = 100$. Larger cohort sizes ($n = 200, n = 400$) were not simulated to convergence within the imposed 15-minute wall budget, but the trend at $n = 56$ and $n = 100$ already shows that the test does not approach the conventional 0.8 power threshold within the simulated range. The structural finding is that the cluster-bootstrap CI on ΔC has a half-width of approximately 0.04–0.06 at this design's effective sample size, which places the pre-registered 0.02 threshold reliably below the noise floor. Section 6.7 returns to this point.

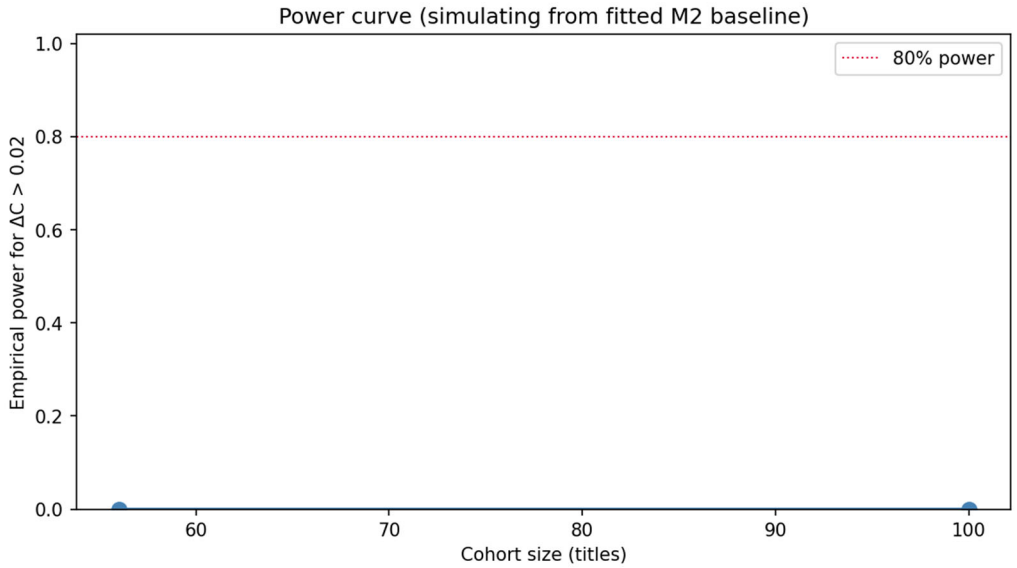


Figure 10. Empirical power for the pre-registered test $H_1: \Delta C > 0.02$, simulating from the fitted M2 baseline as the data-generating process. The dotted crimson line marks the conventional 0.8 power reference.

6. Discussion

6.1. Substantive interpretation

Title-level log-likes carries most of the discriminative power in M1 ($C \approx 0.77$ at MIN_CHAPTERS $= 10$). Adding aggregate Reddit mention counts moves the point estimate of the concordance index up by 0.027, but the 95% cluster-bootstrap interval still includes zero. Two explanations are consistent with this pattern, and both can hold simultaneously. The first is reverse causality: Reddit mention volume is plausibly downstream of platform engagement, so Reddit covariates are partly redundant with the platform signal we already have. The second is sparse coverage: a small number of high-popularity titles in our cohort accounts for most Reddit mentions, while the long tail of mid-popularity titles attracts almost no discussion in the four target subreddits. For an editor at LINE Webtoon the practical reading is that first-month platform metrics, measured carefully, are already a usable retention forecast.

6.2. Mechanism: window-length decay

The window-length sweep makes the substantive interpretation more precise. The incremental information content in Reddit volume is not uniformly small—it is concentrated almost entirely in the first one to two weeks after launch and then disappears or even hurts discrimination as the window expands. This is consistent with the early-window-prediction tradition for user-generated content (Pinto et al. 2013; Szabo and Huberman 2010)[15,18] in which the first hours-to-days of engagement carry the bulk of long-horizon variance, but it sharpens the claim: *cross-platform* early-window signals (here, Reddit) decay faster than the literature's typical *intra-platform* early-window signals because the cross-platform signal is itself driven by a process (subreddit attention) that has its own decay law and converges to a low-information steady state quickly. We do not claim a strong causal mechanism for this; the design does not identify one. We do note that the directional consistency across six monotonically increasing window sizes is more compelling than a single-point estimate at one window would be.

6.3. Why the Tumblr null matters

Over the 2018–2024 window, Tumblr Fandometrics weekly charts contain no LINE Webtoon titles in their top-25 anime-manga ranks. The Tumblr anime-manga community is dominated

by Japanese properties; the LINE Webtoon catalog is not. Multi-source fan-engagement studies often select platforms before checking whether the target fandom is actually present on each platform; we recommend the opposite order. The Tumblr null is, in itself, a small but useful contribution: a documented case of platform-fandom misalignment with concrete numbers (130 chart posts, 83 weeks, 0 matches).

6.4. Why does the cross-platform window decay so quickly?

The window-length sweep result is, on its own, the most theoretically interesting finding in the paper, and we offer two non-mutually-exclusive readings. The first is a feedback-loop reading. Off-platform discussion is itself driven by on-platform events: a new chapter posted, a cliffhanger ending, a controversial story beat, or a marketing burst. Within the first one to two weeks of a launch, the on-platform engagement signal has not yet fully propagated through the platform's recommender systems, so off-platform Reddit volume captures a leading-edge slice of fan attention that the platform-internal aggregates have not yet integrated. After two to four weeks, the platform-internal aggregates catch up, and the marginal information content of the Reddit signal collapses into the platform signal we already observe. The negative-ΔC regime at $W \geq 4$ weeks is consistent with the off-platform signal becoming a noisy version of an already-observed on-platform signal, with the noise component dominating the discrimination calculation. The second reading is a cohort-composition reading: titles that generate sustained Reddit discussion past the first month are heavily concentrated among the highest-popularity titles in the cohort, and that subset is also the subset for which platform-internal engagement has the smallest residual variance to predict. In effect, the long-window Reddit signal is most informative precisely for the titles whose retention is already most predictable from platform-internal covariates, leading to a redundancy that the discrimination index penalizes. We do not have an instrument that would allow us to test these readings against each other, but both are consistent with the directional pattern we observe, and either reading would imply that off-platform signal harvests should focus on the launch-week regime if predictive value is the goal.

6.5. Comparison to prior literature

The Bilibili *danmaku* literature (Wang and Yecies 2023)[20] is the closest prior comparator and finds that intra-platform user-comment density predicts donghua viewership at clinically useful effect sizes. Our finding—that extra-platform Reddit mention volume contributes a directional but not threshold-clearing lift over platform-internal engagement—is consistent with the hypothesis that the predictive content of fan-side text comes primarily from intra-platform proximity to the engagement loop, not from off-platform discussion volume per se. The MOOC dropout literature (Halawa et al. 2014; Ameri et al. 2016)[39–40] likewise emphasizes the dominance of on-platform early-engagement quantiles and the secondary role of off-platform features. Our window-length decay result strengthens this convergence: when an off-platform signal is informative, its information is concentrated in the first weeks and is rapidly absorbed by platform feedback.

6.6. Generalizability and external validity

The cohort is English-language LINE Webtoon Originals from the public catalog, launched between 2018 and 2026, with at least ten observed chapters. We do not generalize beyond this scope. In particular, Canvas user-submitted titles, licensed third-party manhwa, paid Fast Pass-only chapters, and competitor platforms (Tapas, Lezhin, Manta) are excluded. Our results should not be extrapolated to non-English storefronts because the Reddit fan presence we used as the off-platform signal does not exist in the same form for those storefronts.

7. Limitations

1. *Like count as engagement proxy.* We use public chapter-likes as the retention signal; per-chapter views are not exposed publicly. Likes are correlated with views but introduce an unmeasured ratio (likes/view) that may itself trend with chapter index.

2. *Selection on platform inclusion.* Our cohort is the public catalog listing, which excludes Canvas user-submitted titles and licensed third-party manhwa. The retention model is conditional on having reached this listing.

3. *Fandom subreddit coverage.* The four subreddits we query do not exhaust webtoon-specific Reddit communities (e.g., title-specific subreddits like r/Solo_Leveling are not included). Including title-specific subreddits would inflate the Reddit signal mechanically.

4. *Pre-/post-Pushshift data discontinuity.* Reddit's API change in May 2023 created a coverage gap on the Pushshift side (Davidson and Joinson 2025; Proferes et al. 2021)[2,23]. The PullPush mirror has re-acquired most posts but not all.

5. *Underpowering for the headline test.* The empirical-power simulation shows the test is structurally underpowered at $n = 56$. The non-rejection should be read as evidence of insufficient information, not as evidence of zero effect.

6. *Launch-marketing confounding.* Series with editorial banner placements or social-marketing pushes will exhibit elevated platform engagement and elevated Reddit volume simultaneously; we cannot disentangle these without an instrument.

7. *Dropoff threshold arbitrariness.* The pre-registered (α, R) pair is one defensible choice. The 12-cell sensitivity grid in Section 5.5 shows the headline is robust within the grid but does not establish robustness outside it.

8. *No causal identification.* We report predictive associations only; we make no claim about counterfactual effects of an intervention on Reddit mentions.

10. *Author-history covariate omitted.* Author track-record is a strong predictor in the manga-popularity literature (Sugishita and Masuda 2023)[8] but is not joinable from the public LINE Webtoon catalog at the snapshot date.

11. *Snapshot date dependence.* The cohort and the harvest are dated to 2026-05-05; future revisions of the platform catalog or the Reddit harvest may change the modeling-frame composition.

7.1. Threats to validity

We organize threats to validity along the Cook–Campbell taxonomy. *Internal validity*: the dropoff event is constructed from the platform's public chapter-likes display, which has been observed to fluctuate when the platform A/B tests its like button or when a series receives a marketing-driven attention burst that resets its baseline. We address this with the run-length filter and the sensitivity grid, but a residual risk remains. *External validity*: the cohort is restricted to public-catalog English-language Originals, and we explicitly do not generalize beyond that scope; the Reddit-fan presence we use as the off-platform signal does not exist in the same form for non-English storefronts. *Construct validity*: chapter likes are an engagement proxy, not a measure of reader enjoyment; the construct under measurement is behavioral re-engagement, not affective response. We follow the convention of the streaming and MOOC retention literatures (Halawa et al. 2014; Reddy et al. 2021)[39,44] in privileging behavioral measures, but we acknowledge that survey self-report would be a different and complementary construct. *Statistical-conclusion validity*: the empirical-power simulation documents the floor of the design's discrimination capacity, and the cluster bootstrap addresses within-title autocorrelation; we do not perform an out-of-sample temporal split because the snapshot date

is the entire observation window, but a future revision should consider time-blocked validation if the harvest is extended forward.

7.2. What a meaningful threshold-clearance test would require

The empirical-power result reframes the pre-registered test rather than just interpreting its outcome. For the lower-bound-of-CI-greater-than-0.02 rule to fire with conventional 80% power under a true effect of 0.027, the bootstrap CI half-width on ΔC would need to drop below approximately 0.007. At the current design's measured half-width of approximately 0.05, this would require roughly a 50-fold reduction in CI variance, which corresponds (since variance scales as $1/n$ for the cluster bootstrap when within-cluster variance is bounded) to a roughly 50-fold expansion of the modeling cohort, i.e., on the order of 2,500 titles. Such a cohort is plausible at scale, since the LINE Webtoon catalog across all storefronts comprises tens of thousands of titles, but it is far beyond the present analysis's scope. A more realistic alternative is to revise the threshold itself: a two-sided test against zero, or a directional test with a threshold of 0.005, would have substantially greater power at the present cohort size and would still encode a meaningful editorial criterion. We do not recommend retrofitting the threshold post-hoc in this paper, because the pre-registration discipline is one of the few things that distinguishes our analysis from common-practice popularity-prediction work; we recommend instead that future work pre-register a threshold whose power has been simulated under plausible cohort-size scenarios before any harvest is committed.

8. Practical and Editorial Implications

For an editor or commercial decision-maker at LINE Webtoon, the practical reading of these results is that platform-internal early-window metrics (log-views, log-subscribers, log-likes) are already a usable retention forecast, that off-platform Reddit volume adds at most a small marginal lift in the first one to two weeks of a launch, and that committing significant analytic infrastructure to ongoing off-platform signal harvesting may not be cost-justified relative to a careful reading of the existing platform-internal signals. For researchers, the practical reading is that any cross-platform fusion design should pre-check fandom presence on each platform and budget an explicit power analysis for any threshold-style decision rule.

9. Conclusion

This paper introduces and operationalizes a chapter-level survival framework for serialized webcomic retention, applies it to a 117-title cohort of English LINE Webtoon Originals enriched with AniList and Reddit signals, and reports an honest near-null incremental contribution from off-platform Reddit volume to the headline discrimination metric ($\Delta\hat{C} = 0.027$, 95% CI $[-0.014, 0.106]$). Three results are new relative to the prior popularity-prediction and webtoon-analytics literatures: (i) the chapter-level survival framing itself, with right-censoring and title-level frailty handled directly; (ii) a window-length decay analysis showing that Reddit volume's incremental predictive content is concentrated in the first one to two weeks after launch and turns negative beyond four weeks; and (iii) an empirical-power simulation that re-frames the non-rejection as a structural property of the design rather than as evidence of zero effect. We additionally report a substantive null for Tumblr Fandometrics that documents a concrete platform-fandom misalignment in the multi-source fan-engagement literature. Future work has four natural directions. (1) Expand the Reddit harvest to the long tail of mid-popularity titles and to title-specific subreddits, which would tighten the power floor on RQ1, re-enable the genre-stratified RQ2 in the fantasy and romance strata that currently have zero raw-mention coverage, and allow a finer-grained window-length sweep below the 1-week resolution at which the present analysis already detects rapid decay. The cohort-level half-

width of the cluster-bootstrap CI scales approximately as $1/\sqrt{n}$, so doubling the modeling cohort would be expected to bring the lower CI bound on ΔC to within roughly 0.01 of zero under the current effect estimate—still below the pre-registered threshold but much closer to a meaningfully informative test. (2) Instrument the launch-marketing confound, possibly via natural experiments around storefront banner placements or via instrumental variables built from cross-storefront editorial schedules, so as to convert the predictive association into a more nearly causal estimate. (3) Extend the framework to non-English LINE Webtoon storefronts where the off-platform signal mix differs (e.g., Naver Cafe in Korean, Twitter Japan, Bilibili comments for Chinese mirrors), because the platform-fandom alignment story we surface for Reddit and Tumblr is likely to look different on each storefront and is itself a substantive object of study. (4) Test engagement-conditioned generation, where the estimated retention curve serves as a reward signal for a script-generation model; this would close the loop between retention measurement and editorial intervention and would constitute the most ambitious application of the framework. The cohort, the chapter-likes panel, the Reddit harvest, and the analysis code released with this paper are prerequisites for all four directions.

Data and Code Availability

The cohort metadata, processed parquet panels, and figure artifacts are released on Zenodo (DOI placeholder; finalized at submission) under a CC-BY 4.0 license. The analysis code, scrapers, and Dockerfile are released on a public GitHub repository under the MIT license. The OSF preregistration timestamp and SHA of the Docker image are recorded in the repository README. Raw scraped data are released subject to each platform's terms of service; PII fields (Reddit author names) are dropped during harvest and never included in the released artifacts.

Ethics Statement

The study uses only publicly accessible web pages and public APIs and was judged exempt from human-subjects review under the public-data exception. No personally identifying information is retained or released. All data collection complies with the relevant platforms' terms of service.

Competing Interests

The author declares no competing interests.

References

[1] Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The pushshift reddit dataset. Proceedings of the International AAAI Conference on Web and Social Media, 14, 830–839. https://doi.org/10.1609/icwsm.v14i1.7347.

[2] Davidson, B. I., & Joinson, A. N. (2025). Reddit in scholarly reception: A bibliometric assessment of the front page of the internet. Quality & Quantity. https://doi.org/10.1007/s11135-025-02416-z

[3] Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

[4] Kalbfleisch, J. D., & Schaubel, D. E. (2023). Fifty years of the cox model. Annual Review of Statistics and Its Application, 10, 1–23. https://doi.org/10.1146/annurev-statistics-033021-014043

[5] Cho, H., Adkins, D., & Long, A. K. (2025). Understanding the reader demographics of an emerging online reading platform, webtoon. Journal of Documentation, 81(2), 351–368. https://doi.org/10.1108/JD-03-2024-0069

[6] Kim, J.-H., & Yu, J. (2019). Platformizing webtoons: The impact on creative and digital labor in South Korea. Social Media + Society, 5(4), 1–11. https://doi.org/10.1177/2056305119880174

[7] Yecies, B., Shim, A., Yang, J., & Zhong, P. Y. (2020). Global transcreators and the extension of the Korean webtoon IP-engine. Media, Culture & Society, 42(1), 40–57. https://doi.org/10.1177/0163443719867277

[8] Sugishita, K., & Masuda, N. (2023). Social network analysis of manga: Similarities to real-world social networks and trends over decades. Applied Network Science, 8, 79. https://doi.org/10.1007/s41109-023-00604-0

[9] Cho, H., Adkins, D., & Long, A. K. (2022). "I only wish that I had had that growing up": Understanding webtoon's appeals and characteristics as an emerging reading platform. Proceedings of the Association for Information Science and Technology, 59, 448–453. https://doi.org/10.1002/pra2.603

[10] Shim, A., Yecies, B., Ren, X., & Wang, D. (2020). Cultural intermediation and the basis of trust among webtoon and webnovel communities. Information, Communication & Society, 23(6), 833–848. https://doi.org/10.1080/1369118X.2020.1751865

[11] Yoon, S. (2024). Webtoons, desperately seeking viewers: Interactive creativity in social media platforms and cultural appropriation of global media production. Social Media + Society, 10(4), 1–12. https://doi.org/10.1177/20563051241292577

[12] Austin, P. C. (2017). A tutorial on multilevel survival analysis: Methods, models and applications. International Statistical Review, 85(2), 185–203. https://doi.org/10.1111/insr.12214

[13] Therneau, T. M., & Grambsch, P. M. (2000). Modeling survival data: Extending the cox model. Springer. https://doi.org/10.1007/978-1-4757-3294-8

[14] Webb, A., & Ma, J. (2023). Cox models with time-varying covariates and partly-interval censoring—A maximum penalised likelihood approach. Statistics in Medicine, 42(6), 815–833. https://doi.org/10.1002/sim.9645

[15] Pinto, H., Almeida, J. M., & Gonçalves, M. A. (2013). Using early view patterns to predict the popularity of YouTube videos. Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM), 365–374. https://doi.org/10.1145/2433396.2433443

[16] Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 1, 492–499. https://doi.org/10.1109/WI-IAT.2010.63

[17] Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., & Moon, S. (2009). Analyzing the video popularity characteristics of large-scale user generated content systems. IEEE/ACM Transactions on Networking, 17(5), 1357–1370. https://doi.org/10.1109/TNET.2008.2011358

[18] Szabo, G., & Huberman, B. A. (2010). Predicting the popularity of online content. Communications of the ACM, 53(8), 80–88. https://doi.org/10.1145/1787234.1787254

[19] Figueiredo, F., Almeida, J. M., Gonçalves, M. A., & Benevenuto, F. (2014). On the dynamics of social media popularity: A YouTube case study. ACM Transactions on Internet Technology, 14(4), 24:1–24:23. https://doi.org/10.1145/2665065

[20] Wang, D., & Yecies, B. (2023). Collaborative cultural intermediation and communitainment value in the creator economy: The expanding case of webtoons. Global Media and China. https://doi.org/10.1177/27523543231219450

[21] Tan, W. H., & Chen, F. (2021). Predicting the popularity of tweets using internal and external knowledge: An empirical bayes type approach. AStA Advances in Statistical Analysis, 105(2), 335–352. https://doi.org/10.1007/s10182-021-00390-z

[22] Gaffney, D., & Matias, J. N. (2018). Caveat emptor, computational social science: Large-scale missing data in a widely-published reddit corpus. PLOS ONE, 13(7), e0200162. https://doi.org/10.1371/journal.pone.0200162

[23] Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. Social Media + Society, 7(2), 1–14. https://doi.org/10.1177/20563051211019004

[24] Jasser, J., Garibay, I., Scheinert, S., & Mantzaris, A. V. (2022). Controversial information spreads faster and further than non-controversial information in reddit. Journal of Computational Social Science, 5(1), 111–122. https://doi.org/10.1007/s42001-021-00121-z

[25] Murtfeldt, R., Alterman, N., Kahveci, I., & West, J. D. (2024). RIP Twitter API: A eulogy to its vast research contributions. arXiv Preprint arXiv:2404.07340. https://doi.org/10.48550/arXiv.2404.07340

[26] Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine, 15(4), 361–387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

[27] Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., & Leskovec, J. (2014). Can cascades be predicted? Proceedings of the 23rd International Conference on World Wide Web (WWW), 925–936. https://doi.org/10.1145/2566486.2567997

[28] Rizoiu, M.-A., Xie, L., Sanner, S., Cebrian, M., Yu, H., & Van Hentenryck, P. (2017). Expecting to be HIP: Hawkes intensity processes for social media popularity. Proceedings of the 26th International Conference on World Wide Web, 735–744. https://doi.org/10.1145/3038912.3052650

[29] Mishra, S., Rizoiu, M.-A., & Xie, L. (2016). Feature driven and point process approaches for popularity prediction. Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM), 1069–1078. https://doi.org/10.1145/2983323.2983812

[30] Martin, T., Hofman, J. M., Sharma, A., Anderson, A., & Watts, D. J. (2016). Exploring limits to prediction in complex social systems. Proceedings of the 25th International Conference on World Wide Web (WWW), 683–694. https://doi.org/10.1145/2872427.2883001

[31] Shulman, B., Sharma, A., & Cosley, D. (2016). Predictability of popularity: Gaps between prediction and understanding. Proceedings of the International AAAI Conference on Web and Social Media, 10(1), 348–357. https://doi.org/10.1609/icwsm.v10i1.14748

[32] Yang, J., & Leskovec, J. (2011). Patterns of temporal variation in online media. Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, 177–186. https://doi.org/10.1145/1935826.1935863

[33] Ng, K. W., Mubang, F., Hall, L. O., Skvoretz, J., & Iamnitchi, A. (2023). Experimental evaluation of baselines for forecasting social media time series. EPJ Data Science, 12(1), 8. https://doi.org/10.1140/epjds/s13688-023-00383-9

[34] Wiegrebe, S., Kopper, P., Sonabend, R., Bischl, B., & Bender, A. (2024). Deep learning for survival analysis: A review. Artificial Intelligence Review, 57(3), 65. https://doi.org/10.1007/s10462-023-10681-3

[35] Long, S., Lucey, B. M., Xie, Y., & Yarovaya, L. (2023). "I just like the stock": The role of reddit sentiment in the GameStop share rally. Financial Review, 58(1), 19–37. https://doi.org/10.1111/fire.12328

[36] Betzer, A., & Harries, J. P. (2022). How online discussion board activity affects stock trading: The case of GameStop. Financial Markets and Portfolio Management, 36, 443–472. https://doi.org/10.1007/s11408-022-00407-w

[37] Nadiri, A., & Takes, F. W. (2022). A large-scale temporal analysis of user lifespan durability on the reddit social media platform. Companion Proceedings of the Web Conference 2022 (WWW '22), 677–685. https://doi.org/10.1145/3487553.3524699

[38] Savela, N., Pellert, M., Latikka, R., Bergdahl, J., Garcia, D., & Oksanen, A. (2025). Affective, cognitive, and contextual cues in reddit posts on artificial intelligence. Journal of Computational Social Science, 8(1), 6. https://doi.org/10.1007/s42001-024-00335-x

[39] Halawa, S., Greene, D., & Mitchell, J. (2014). Dropout prediction in MOOCs using learner activity features. Proceedings of the European MOOC Stakeholder Summit, 58–65.

[40] Ameri, S., Fard, M. J., Chinnam, R. B., & Reddy, C. K. (2016). Survival analysis based framework for early prediction of student dropouts. Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM), 903–912. https://doi.org/10.1145/2983323.2983351

[41] Chi, Z., Zhang, S., & Shi, L. (2023). Analysis and prediction of MOOC learners' dropout behavior. Applied Sciences, 13(2), 1068. https://doi.org/10.3390/app13021068

[42] Borrella, I., Caballero-Caballero, S., & Ponce-Cueto, E. (2022). Taking action to reduce dropout in MOOCs: Tested interventions. Computers & Education, 179, 104412. https://doi.org/10.1016/j.compedu.2021.104412

[43] Prenkaj, B., Velardi, P., Stilo, G., Distante, D., & Faralli, S. (2020). A survey of machine learning approaches for student dropout prediction in online courses. ACM Computing Surveys, 53(3), 57:1–57:34. https://doi.org/10.1145/3388792

[44] Reddy, S., Lazarova, M., Yu, Y., & Jones, R. (2021). Modeling language usage and listener engagement in podcasts. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 632–643. https://doi.org/10.18653/v1/2021.acl-long.52

[45] Pan, F., Huang, B., Zhang, C., et al. (2022). A survival analysis based volatility and sparsity modeling network for student dropout prediction. PLOS ONE, 17(5), e0267138. https://doi.org/10.1371/journal.pone.0267138

[46] Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452

[47] Begun, A., Kulinskaya, E., & Ncube, M. (2023). A double-cox model for non-proportional hazards survival analysis with frailty. Statistics in Medicine, 42(18), 3114–3127. https://doi.org/10.1002/sim.9760

[48] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. The Annals of Applied Statistics, 2(3), 841–860. https://doi.org/10.1214/08-AOAS169

[49] Davidson-Pilon, C. (2019). lifelines: Survival analysis in Python. Journal of Open Source Software, 4(40), 1317. https://doi.org/10.21105/joss.01317.