What Can Reviews Reliably Tell Us About Books? A Per-Attribute Audit and Tiered Annotated Dataset Across 1.7M Titles

Anonymous ACL submission

Abstract

Aggregated reader reviews are increasingly used to measure book attributes across computational literary studies, recommendation, and narratological-benchmark design, yet the field lacks a per-attribute audit of what review aggregates can reliably recover. Using the HUMANREADER survey corpus (Anonymous, 2026) as ground truth, we audit nine review-derived feature methods against 391 attributes. Most apparent predictive performance is popularity-confounded. Among the methods we evaluate, only LLM-curated keyword densities retain held-out content signal after partialling out $\log(1 + n_{\text{reviews}})$, and only on 7.9% of attributes. The recoverable subset is explained by an attribute's "lexical anchorability", i.e. how often reviews mention keywords in the attribute's semantic neighborhood, with high-anchorability attributes about three times more likely to reach our predictive-performance threshold (47% of the top anchorability quartile, 17% of the bottom). We release **HumanReader-Reviews-1.7M**, a per-book annotation set of review-based attribute predictions, separating attributes by whether the predictions reflect review content or book popularity. It covers 74 attributes reliably predicted by review aggregates (7 from review content alone and 67 where book popularity also contributes to the review-based predicted annotation), with per-method test statistics released for the remainder. Overall, we find that raw held-out R^2 overstates per-book content recovery by an order of magnitude; researchers building review-derived book annotations should report popularity-residualized ΔR^2 over a $\log(1 + n_{\text{reviews}})$ baseline.

1 Introduction

Can aggregated reader reviews recover, per book, whether the prose is plain or ornate, whether the narrator is reliable, or whether the plot sags in the middle? Across computational literary studies (Holur et al., 2021; Luțan and Bădică, 2022; Boot and Koolen, 2020; Bizzoni et al., 2023), large-scale recommendation (Wan and McAuley, 2018), and narratological-benchmark design (Hamilton et al., 2025b), the answer is generally treated as yes, and researchers extract features from aggregated reviews to use as measurements of book attributes that would otherwise require an expert reading. The motivating scale is hard to overstate, since the Goodreads Book Graph alone (Wan and McAuley, 2018; Wan et al., 2019) contains roughly 1.7 million English-reviewed books and over 10 million English reviews, well beyond what expert panels can annotate at scale and beyond what fits in even a million-token LLM context window (Hamilton et al., 2025a; Karpinska et al., 2024). What review aggregates can actually recover, however, has not been audited at the per-attribute level.

This paper provides such an audit. We take 391 attributes from the HUMANREADER dataset of structured, human-labeled narrative features for hundreds of fiction novels (Anonymous, 2026),[1] spanning character, plot, narrator, setting, style, and reader response. We map per-attribute recoverability against nine review-derived feature methods, each paired with Ridge or Lasso regression: LLM-curated keyword densities, TF-IDF features, sentence-transformer embeddings, review metadata variants, emotion features from GoEmotions, aspect-based sentiment via VADER, NMF topics, stylometric and readability features, and a regression distilled from LLM-on-reviews labels (§4.1).

Of the 205 attributes that pass our predictability filters, namely whether an attribute has enough variance and reader agreement to be predictable in

[1] This work is submitted concurrently with Anonymous (2026), which introduces the HUMANREADER corpus itself: instrument design, inter-reader agreement, ICC-based reliability. The contributions are non-overlapping; the companion paper documents the survey, while this paper takes the corpus as ground truth and maps which per-attribute targets aggregated reviews can recover.

principle, 74 pass our validation threshold[2] and 131 fail it under every method we evaluate. After partialling out $\log(1 + n_{\text{reviews}})$, only 7 retain content-driven signal. The recoverable subset concentrates in mood descriptors, genre items, narrator-type labels, and writing-style descriptors; plot-structural attributes (where turning points fall, where the plot drags) and gestalt judgments (overall enjoyment, character-versus-plot orientation) fail under every method we evaluate.

This pattern tracks an attribute's *lexical anchorability*, namely the per-book frequency with which reviews mention keywords in the attribute's LLM-curated semantic neighborhood. "Lighthearted", whose neighborhood is *cheerful, charming, cozy, sweet, warm*, reaches held-out test $R^2 \approx 0.5$ and survives popularity residualization; "overall enjoyment" lacks a distinctive review vocabulary and fails under every method we evaluate. Across attributes, anchorability moves pass rate from 17% in the lowest quartile to 47% in the highest, at Spearman $\rho = +0.27$ on test R^2 ($p < 10^{-3}$).

We release **HumanReader-Reviews-1.7M**, a per-book annotation set of review-based attribute predictions, separating attributes by whether the predictions reflect review content or book popularity. It contains predictions for 7 content-driven attributes and 67 where popularity contributes substantially to the review-based predicted annotation, with per-method test statistics released for the remaining 131 in-scope attributes that no evaluated method predicts; the release tiering and per-attribute metadata are described in Section 6. Taken together, we find that raw held-out R^2 overstates per-book content recovery by an order of magnitude in this regime, suggesting that researchers building review-derived book annotations should report popularity-residualized ΔR^2 over a $\log(1 + n_{\text{reviews}})$ baseline.

2 Related Work

Our work draws on three lines of prior research, namely book reviews as a measurement instrument, long-form narrative comprehension benchmarks, and annotator disagreement as signal.

A first line of work mines aggregated book-review corpora to recover book-level properties (Holur et al., 2021; Luțan and Bădică, 2022; Boot and Koolen, 2020; Bizzoni et al., 2023).[3] Closest to ours, Rosenbusch and Korthals (2025) conclude from 624,320 Goodreads reviews that reviews carry up to ten times more reviewer than book information. We extend this from an average claim about overall rating to a per-attribute recoverability map.

A second line evaluates long-context LLMs on narrative comprehension at varying lengths,[4] with Hamilton et al. (2025b) cataloging narrative events, style, perspective, and revelation as underserved across 78 benchmarks. We contribute a per-attribute predictability map for the style and perspective dimensions they identify, sidestepping book-text inference at scale.

A third line treats annotator disagreement on subjective tasks as irreducible signal (Pavlick and Kwiatkowski, 2019; Plank, 2022; Cabitza et al., 2023); within narrative tasks, Subbiah et al. (2024) and Marco et al. (2025) demonstrate irreducible reader variance on faithfulness and creative-writing-evaluation tasks. We operationalize this signal at book length, since per-book population means with bootstrap intervals determine which attributes pass our validation threshold and reach the release.

3 Data

Our analysis combines (1) the HUMANREADER survey corpus, which provides per-reader responses to 101 closed-form post-reading questions across 387 fictional novels, resolving to 391 per-book modeling attributes that supply our ground-truth targets (§3.1), and (2) the Goodreads Book Graph (Wan and McAuley, 2018; Wan et al., 2019), which provides 13.0M English reviews across 1,707,435 books and supplies both the review text from which we derive features and the books we release predictions for.

3.1 HUMANREADER Ground Truth and Modeling Units

HUMANREADER (Anonymous, 2026) comprises 387 fictional novels with 809 finished post-reading

[2] Held-out test $R^2 > 0.10$ or $|\rho| > 0.15$, with multiple-comparison correction; both filters defined in §4.3.

[3] Allied work treats Goodreads as a measurement instrument and documents its biases (Wan and McAuley, 2018; Koolen et al., 2020; Hu et al., 2022; Walsh and Antoniak, 2021).

[4] NarrativeQA (Kočiský et al., 2018), NovelQA (Wang et al., 2025), NoCha (Karpinska et al., 2024), BookSum (Kryściński et al., 2022), BooookScore (Chang et al., 2024), LongStoryEval (Yang and Jin, 2025), FairyTaleQA (Xu et al., 2022), and PsychDepth (Harel-Canada et al., 2024).

sessions from 714 distinct habitual readers recruited from book-related subreddits, each randomly assigned a book (with reassignment for any reader who had previously read it) and surveyed before, during, and after reading; under this design, per-book target means reflect book content rather than reader self-selection, which is the identifying assumption we maintain throughout. We use the post-reading instrument as ground truth; it comprises 101 closed-form questions across seven categories, namely overall enjoyment, genre tags, and mood descriptors (Overview; response options include *Lighthearted*, *Bittersweet*); turning-point location and event structure (Plot); protagonist properties and character-vs-plot orientation (Character); narrator type (Narrator; response options include *first-person*, *unreliable*); era, place, and world-building (Setting); writing-style descriptors and readability (Style; response options include *Humorous*, *Dark*); and post-reading reactions (Reflection). For full instrument design and inter-reader agreement statistics, we refer the reader to Anonymous (2026).

We construct per-book modeling targets from the 101 questions at one of two grains. Numeric and ordinal questions have one attribute per question, with target equal to the unweighted mean of reader responses on the integer encoding (36 numeric + 38 ordinal = 74 such attributes). Categorical and checkbox questions yield one attribute per (question, response choice) pair, with target equal to the share of readers who selected that choice (83 from categorical and 234 from checkbox questions, totaling 317 such attributes). The full expansion yields 391 per-book modeling attributes across the 101 questions. We restrict training to books with $n_{\text{readers}} \geq 2$ for the relevant unit and weight each (book, unit) row by n_{readers}, which is the inverse-variance weight for a per-book sample mean under the constant within-book reader-noise assumption (see Appendix A).

3.2 Goodreads Join and Variance Ceiling

We match HUMANREADER titles to Goodreads records via title normalization in a streaming pass, preferring English-language candidates with the highest ratings count among ties. The procedure matches 291 of 387 books (75%);[5] the unmatched 96 skew toward less-popular and non-English titles (Section 8). After applying per-unit $n_{\text{readers}} \geq 2$ filters, the working subset for training is $\sim$ 286 books. After dropping units with fewer than 10 mapped books and books with $n_{\text{readers}} = 1$, the typical training population is $\sim$229 books per unit (median across the 391 attributes), with $\sim$183 in each per-seed train fold after the 80/10/10 split.

Per-book targets are sample means of reader responses, with sampling noise placing an upper bound on the test R^2 any model can achieve (the per-attribute *variance ceiling*); the formula and per-attribute values are in Appendix A. The median ceiling across units is 0.73 (with 1 indicating full per-book reader agreement and 0 indicating noise-dominated targets). This is the same per-question reliability decomposition reported as intraclass correlation (ICC) in Anonymous (2026), with ceiling $\to 1$ corresponding to ICC $\to 1$; per-category mean inter-reader agreement (IAA) reported there ranges from 0.54 (Overview) to 0.75 (Narrator), informing the achievable per-category release-pass rate (Section 5.3).

4 Feature Methods and Validation Protocol

We frame the prediction task per attribute, namely one modeling unit per question for numeric and ordinal items, and one per (question, response choice) pair for categorical and checkbox items (§3.1). For each unit we predict the per-book target y_b from features derived from book b's aggregated reviews; the per-attribute results are reported in Section 5. The remainder of this section specifies the nine feature methods we evaluate (§4.1), the training and bootstrap-evaluation protocol (§4.2), and the release threshold that gates which per-attribute predictions are eligible for release (§4.3).

4.1 Feature Methods

We evaluate three tiers of review-derived feature methods organized by per-prediction inference cost, since the planned release covers 1,707,435 books and per-book inference cost is the binding constraint at this scale. A popularity-only baseline of $\log(1 + n_{\text{reviews}})$ per book is partialled out in the headline comparisons.

Cheap tier. To establish lightweight baselines that scale to full-corpus inference, we evaluate two cheap-tier methods. LLM-curated keyword densities prompt GPT-5 once per attribute to generate keywords reflecting each polarity (60 keywords in

[5] Manual spot-check of 20 random matched pairs found all 20 correct (17 exact title matches, 3 with minor variation in articles or punctuation; identical author in every case).

one call for monotonic items, 30 per choice for non-monotonic; see Appendix B); per book and per bundle we compute a keyword density and a review-coverage density, giving four features per monotonic unit and two per non-monotonic. Review meta-features use the review count, the mean review length, the fraction of reviews under 50 characters and over 500 characters, and $\log(1+n_{\text{reviews}})$; the meta-plus-ratings variant adds the per-book mean and standard deviation of star ratings.

Mid tier. To test whether richer text representations recover more than cheap-tier features, we evaluate two mid-tier methods. TF-IDF + Lasso fits a `TfidfVectorizer` on a Goodreads subsample stratified by review count (`min_df=20`, `max_df=0.95`, top-10000 features) and mean-pools TF vectors per book; the downstream regression is Lasso with a per-seed alpha sweep. Sentence-transformer embeddings encode up to 200 reviews per book with `all-MiniLM-L6-v2` (384 dim) and mean-pool per book.

Expensive tier. To upper-bound what a per-book LLM call could recover, we include LLM-on-reviews as a silver-label[6] validation reference rather than a release candidate, since per-book LLM inference across the full 1.7M-book corpus would be prohibitively expensive at our release scale. For a subsample of 163 training books stratified by review count and 23 representative concepts, we sample 20 reviews per book and prompt GPT-5 with the survey question and reviews. A distilled regression (Ridge on the LLM-on-reviews silver labels, using keyword-density and review-meta features as inputs) covers the 35 units with sufficient label coverage and serves as an upper-bound check for the cheap-tier methods; we will release the distilled regression as a separate exploratory artifact upon acceptance.

Additional methods. To probe whether more specialized feature families recover attributes that broad-purpose features miss, we evaluate four further methods, each targeting a specific dimension of the review text: emotion (RoBERTa-GoEmotions (Demszky et al., 2020) logits per review, mean-pooled per book), aspect sentiment (ABSA via keyword-anchored VADER (Hutto and Gilbert, 2014) over eight book aspects), topic structure (NMF (Lee and Seung, 1999) distributions over the TF-IDF vocabulary), and writing style (stylometric and readability features). Implementation details will be in the released code upon acceptance.

[6]Following weak-supervision terminology, *silver labels* are automatically-generated labels used as a noisy proxy for human-annotated gold labels; we treat the LLM-on-reviews outputs as silver labels for the corresponding HumanReader survey responses.

4.2 Training and Evaluation Protocol

For each (attribute, feature method) pair we run the following protocol. We restrict to books with at least two reader responses and a Goodreads review-corpus match (Section 3.2), yielding a median of 229 books per unit; we caution that this is small for stable per-unit R^2 inference and report bootstrap intervals throughout. Each (book, unit) row is weighted by n_{readers} during fitting and evaluation, the inverse-variance weight for a per-book sample mean assuming constant within-book reader noise. We use an 80/10/10 random book-level split per seed and average across 5–10 seeds depending on method (5 for the full-corpus TF-IDF and NMF runs, 10 for all others). We select the winning model per (unit, method) by the seed-mean validation R^2, with ties broken in favor of Ridge over Lasso, and report test-set R^2 and Spearman ρ for the winner.

We compute bootstrap intervals by resampling the per-seed test R^2 values with replacement $B = 500$ times and reporting the 5th and 95th percentiles of the resampled mean (a 90% interval). A one-sided bootstrap p-value is the fraction of bootstrap means at or below zero. We apply Benjamini-Hochberg false discovery rate (BH-FDR) control at $q = 0.05$ across all (method, attribute, model-class) cells (Ridge and Lasso both contribute cells per method-attribute). For the winner we refit on the full training set and apply to the Goodreads corpus, clipping predictions to the registry-derived natural range; full-corpus methods cover all 1.7M books, GoEmotions and ABSA cover the matched training subset (§6).

4.3 Release Threshold

A unit is *releasable* for a given method if and only if all of the following hold:

- *Effect size:* mean test R^2 across seeds > 0.10 and bootstrap 5th-percentile $R^2 > 0$, or mean test $|\rho| > 0.15$ and bootstrap 5th-percentile Spearman exceeds zero in absolute value.
- *Multiplicity:* the unit's bootstrap one-sided p-value passes the BH-FDR critical value at

$q = 0.05$ across all (method, attribute) cells in the panel.

- *Predictability filters:* target variance across training books > 0.005, target range > 0.05, and variance ceiling (Section 3.2) > 0.30.

A releasable unit is additionally tagged *content-driven* if its popularity-residualized incremental ΔR^2 over the $\log(1 + n_{\text{reviews}})$ baseline exceeds 0.05; Section 6 structures the release around this criterion.

For the final released dataset we report the winning method per attribute, with selection by seed-mean validation R^2 and the reported metrics taken from the winner's test set.

5 Results

Our goal is not a leaderboard across feature methods but to map per-attribute recoverability and isolate its content-driven component. We report the popularity-confound audit first (§5.1), then the per-attribute recoverability map filtered through that audit (§5.2, §5.3), the lexical-anchorability mechanism that explains the recoverable subset (§5.4), and threshold-robustness checks (§5.5).

5.1 The Popularity-Content Confound

Before reporting per-attribute pass rates, we audit a natural confound. Any review-derived feature method may simply exploit the popularity-content correlation in the Goodreads corpus, since popular books receive more reviews and also cluster in specific genres and styles. To isolate the content-extraction contribution of each method, we partial out $\log(1 + n_{\text{reviews}})$ from both the target and the features within each training fold, fit the feature method on the residualized data, and report the difference between its held-out test R^2 and the popularity-only model's test R^2 as ΔR^2.[7]

In raw held-out R^2 (Figure 1a), several methods (review meta, sentence embeddings, review meta plus ratings) appear competitive with LLM-curated keyword densities. After partialling out popularity (Figure 1b), keyword densities dominate the positive right tail. Their incremental ΔR^2 exceeds 0.05 on 7.9% (29) of the 366 attributes for which the residualization pipeline produced a valid ΔR^2 (median incremental $\Delta = -0.011$); the next-best methods clear the same threshold on at most 0.6% of units each. We caution that the keyword-density advantage is partly by construction, since keywords are LLM-curated per response option. The direction of the finding (most methods fail popularity residualization) holds across methods, although the ordering reflects this construction. The eight other methods we evaluate show negative median incremental R^2 over the popularity baseline, with NMF topics and sentence embeddings exhibiting the largest overfitting penalties (median $\Delta = -29.1$ and -2.4 respectively). A per-attribute bias audit stratifying Goodreads books into review-count quintiles flags attributes as *popularity-leaning* when their across-quintile prediction-mean range exceeds 0.5σ while their popularity-residualized incremental ΔR^2 stays below 0.05; 20 of 73 auditable releasable attributes (27%) are flagged. The release (Section 6) tiers attributes by this content-versus-popularity distinction.

5.2 Per-Attribute Pass Rates

We next ask how many of the 391 HumanReader attributes pass our validation threshold under the best method per attribute, and how the wins distribute across methods. Of the 391 attributes, 186 fail one or more predictability filter (target variance, range, ceiling, or non-degenerate best fit), leaving 205 in-scope. Taking the best across methods per attribute (Figures 2 and 3), 74 units pass our validation threshold. All 74 pass via the Spearman $\rho > 0.15$ route; 4 of those additionally pass via the $R^2 > 0.10$ route. The asymmetry between the two routes reflects the variance ceiling cap, since the median ceiling of 0.73 leaves limited headroom for raw R^2 once feature-method noise is added. Across the 74 winners, the median bootstrap interval width on test R^2 is 0.24; 27 have width below 0.20 and 50 below 0.30, with 13 exhibiting wider intervals (above 0.40) primarily on small-n attributes.[8] Sentence embeddings win 35 of the 74

[7] We do not treat popularity-informed predictions as uninformative; both the content-driven and popularity-informed attributes are released and labeled accordingly (§6). The separation matters in three contexts: methodological claims of the form "review features recover X" implicitly assume content extraction rather than popularity covariation; long-tail or newly-released books do not yet carry the popularity signal popularity-informed attributes lean on; and generalization to other review platforms requires content extraction rather than corpus-specific popularity. Where these distinctions are not load-bearing, popularity-informed predictions may remain useful.

[8] The 4 strict $R^2 > 0.10$ winners are "Lighthearted", "Humorous", "Dark" (keyword densities, $R^2 \in [0.16, 0.18]$) and "how science-fictional was the world" (embeddings,

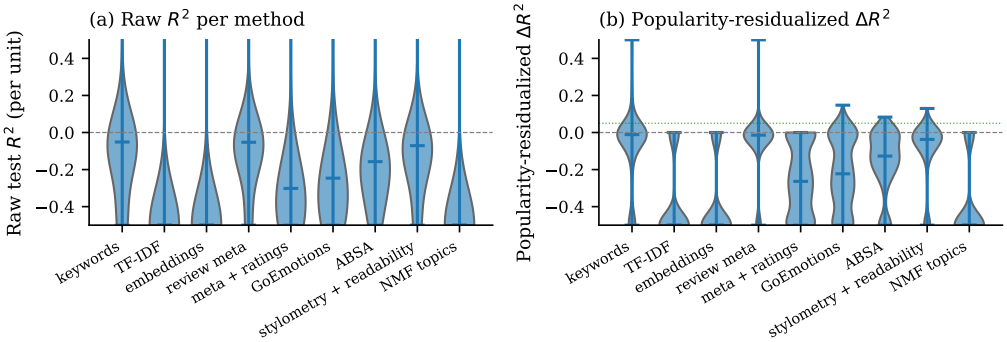

Figure 1: The popularity-content confound dominates the cost-tier comparison. **(a)** Distribution of raw held-out test R^2 per attribute, by method. Several methods (review meta, sentence embeddings, review meta plus ratings) appear competitive in raw R^2. **(b)** After partialling out $\log(1 + n_{\text{reviews}})$ from both features and target. Only LLM-curated keyword densities retain a positive right tail; other methods collapse to or below zero, with embedding-based methods showing the largest overfitting penalty.

units and GoEmotions wins 16; the remaining 23 split across 11 keyword-density winners, 5 each from review meta plus ratings and ABSA, and 2 from TF-IDF. Of these 74 winners only 7 meet the popularity-residualized content-driven criterion (§6), all won by keyword densities; the dominant embedding and GoEmotions winners fail popularity residualization. The top-tail by held-out R^2 is dominated by mood, genre, the "how fantastical was the world" item, and narrator-type checkboxes (Figure 5).

5.3 Per-Category Pass Rates

We aggregate per-attribute pass rates to the seven HumanReader categories to ask which kinds of attributes review aggregates can recover. After applying our predictability filters, pass rates vary substantially across the seven categories (Figure 4): Overview passes at 57% (26 of 46), Style at 48% (13 of 27), Reflection at 40% (2 of 5, too small to characterize), Character at 34% (11 of 32), Setting at 29% (9 of 31), Narrator at 27% (4 of 15), and Plot at 18% (9 of 49).

The cleaner empirical pattern is at the level of the individual response option's lexical anchorability (§5.4). Wherever readers describe a choice through its semantic neighborhood in review text (mood adjectives, genre tags, style descriptors, narrator-type labels), review-derived features recover the per-book share of readers selecting that choice; where the response option requires text inference or integrative judgment, no method we evaluate passes our validation threshold.

$R^2 = 0.20$); all four explain 21–27% of their variance ceiling.

5.4 Lexical Anchorability and Recoverability

The qualitative pattern in §5.3 is that recoverable attributes are those whose response options readers describe directly in review text. We operationalize this as a *lexical anchorability* score per attribute, defined as the mean across training-mapped books of the fraction of per-book reviews containing any keyword in the LLM-curated semantic neighborhood for that (concept, choice) pair. This neighborhood expands the literal response-option text into a semantic field (e.g., "Lighthearted" → *cheerful, charming, cozy, sweet, warm, ...*) and captures the natural register readers use, which the keyword-density method is designed to exploit.

Lexical anchorability is significantly associated with per-unit recoverability across the 205 in-scope attributes, at Spearman $\rho = +0.27$ on held-out test R^2 ($p < 10^{-3}$). Pass rate rises from 17% in the lowest-anchorability quartile to 47% in the highest; full scatter in Appendix H. The mechanism is the LLM-curated semantic neighborhood rather than the literal surface form alone, since the literal form is uncorrelated with recoverability at $\rho \approx 0$.[9]

5.5 Robustness Across Validation Gates and Keyword Regeneration

Because the headline counts depend on validation-threshold choices, we sweep alternative settings. Total winners range 45–81 across R^2, Spearman,

[9] Anchorability shares its lexical input with the keyword-density method that wins all 7 content-driven attributes, raising a circularity concern. We address this with a held-out lexicon test, re-computing anchorability under the v2 keyword bundle (Appendix I; mean Jaccard 0.14 with v1) and recovering Spearman $\rho = +0.31$ with the v1-feature test R^2 ($p < 10^{-5}$). Anchorability therefore tracks an attribute property rather than the specific v1 lexicon.

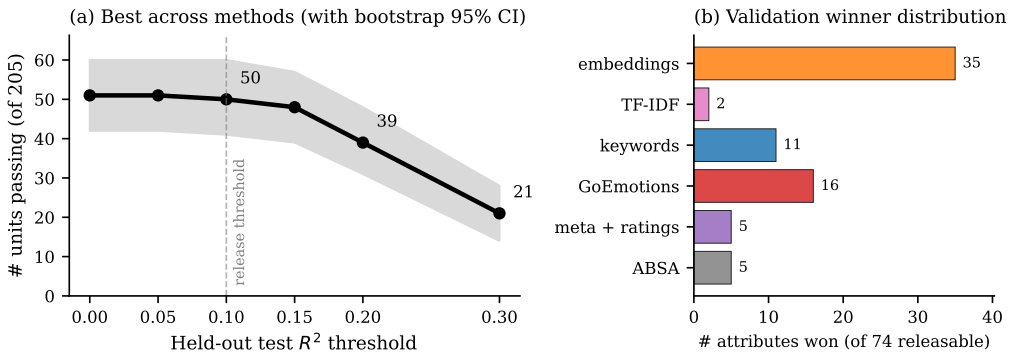

Figure 2: Panel (a): per-unit counts at each held-out R^2 threshold (best across methods by validation R^2). This is raw threshold exceedance pre-multiplicity; after BH-FDR adjustment, 4 units pass via the R^2 route. The dashed line marks $R^2 > 0.10$. Panel (b): per-method winner counts after BH-FDR adjustment (the 74-winner total).

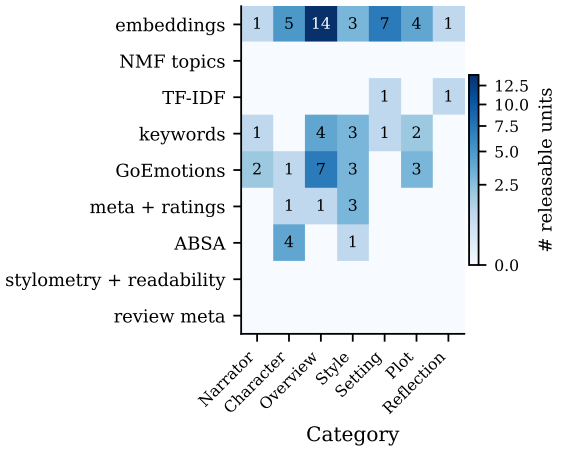

Figure 3: Number of releasable (concept, choice) attributes won by each method (rows) within each category (columns). Wins distribute across six feature families with no single method dominating a category; content-driven wins (survivors of popularity residualization, §5.1) concentrate in Overview and Style.

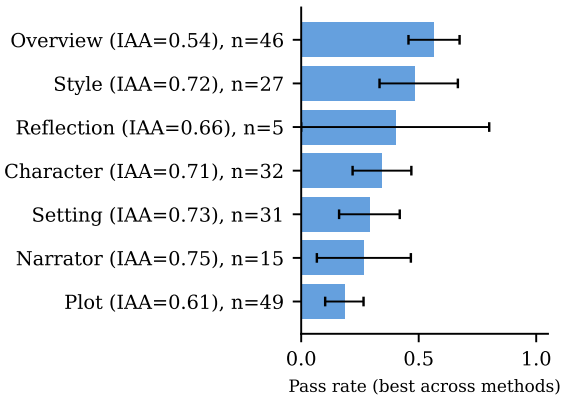

Figure 4: Per-category fraction of (concept, choice) attributes passing the predetermined validation threshold (best across methods), sorted descending. Bootstrap 90% intervals over units (5th/95th percentile). Parenthesized values are the per-category inter-reader agreement (IAA) from the dataset paper (Anonymous, 2026); reliable inference about an IAA–pass-rate relationship is precluded by $n = 7$ categories (Spearman $\rho = -0.36$, $p = 0.43$).

and BH-FDR variants, with the default 74 stable across all R^2 choices; the content-driven count varies 4–10 across alternative popularity-residualization thresholds (full table in Appendix G).

The content-driven finding rests on keyword bundles from a single GPT-5 prompt. A re-prompted v2 bundle (Appendix I; mean Jaccard 0.14 with v1) preserves the direction of signal, with 9 of 11 v1 winners remaining above the relaxed pre-multiplicity threshold.

As a per-winner null check, we re-fit each winner's recipe on 100 permutations of book-level target labels (Appendix J). Of the 74 winners, 59 exceed their permutation 95th percentile, with sentence-embedding and TF-IDF winners uniformly above their nulls.

6 The HumanReader-Reviews-1.7M Release

Given the popularity-confound audit (§5.1) and the lexical-anchorability mechanism (§5.4), we structure the release into three tiers reflecting how much of each attribute's apparent recoverability is content-driven. Tiers A and B include per-book predictions; Tier C is documented through per-method test statistics only. We apply the winning method per (question, choice) to the Goodreads Book Graph (Wan and McAuley, 2018; Wan et al., 2019) and plan to release per-book predictions upon acceptance, restricted to the 74 attributes that pass our validation threshold (Section 4.3).

Tier A (Content-driven, $n = 7$). Passes the validation threshold AND has positive held-out test R^2 AND has popularity-residualized incremental

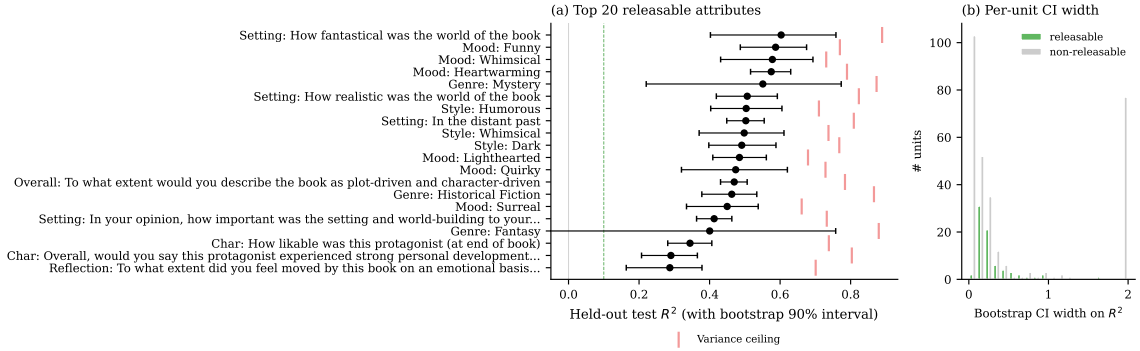

Figure 5: Top 20 releasable attributes sorted by held-out test R^2, with bootstrap 90% intervals (black bars) and variance ceilings (red ticks). Label prefixes *Mood*/*Genre* separate HumanReader Overview sub-categories; *Char*/*Narr* abbreviate Character/Narrator. Inset (b) shows bootstrap interval widths, releasable vs. non-releasable (clipped at 2.0).

$\Delta R^2 > 0.05$. All 7 Tier A attributes are won by LLM-curated keyword densities, concentrated in mood descriptors (*Lighthearted*, *Heartwarming*, *Bittersweet*, *Quirky*), writing-style descriptors (*Humorous*, *Dark*), and the numeric "how fantastical was the world" item. Tier A statistics are sensitive to the specific keyword bundle, since re-prompting GPT-5 yields different magnitudes (though the signal direction is preserved; Appendix I).

Tier B (Popularity-informed, $n = 67$). Passes the validation threshold but the predictions partly reflect popularity-correlated review-aggregate features rather than direct content extraction. Predictions on these attributes covary with $\log(1 + n_{\text{reviews}})$ in addition to any content signal.

Tier C (Documented, $n = 131$). In-scope after our predictability filters but fails the validation threshold under every method we evaluate (see Section 8 on method-palette completeness). Tier C concentrates in Plot, Setting, and Character, namely plot-structural items, world-building details, and integrative character judgments (§5.3).

Format. Upon acceptance we will release three artifacts, namely per-book predictions organized by winning method (one file per method, indexed by book identifier); a per-attribute metadata file covering all 205 in-scope units with winning method, bootstrap CIs, popularity-residualized ΔR^2, lexical-anchorability score, and tier; and per-method test statistics for every (method, unit) cell. Tier C units appear in the metadata with their full statistics but without per-book predictions.

7 Discussion and Conclusion

Review-derived features recover the per-book population mean wherever reviewers describe a response option in their own words: mood adjectives, genre labels, narrator-type descriptors. They fail under every method we evaluate when the response option requires text inference, such as turning-point location or ending type, or integrative judgment such as overall enjoyment or character-versus-plot orientation. The 131 Tier C attributes mark the empirical boundary of recovery *under the methods we evaluate*. We emphasize they are not provably unrecoverable; richer feature families such as fine-tuned cross-encoders, retrieval-augmented prompting, or stacked ensembles may yet extract signal our nine fail to capture, and our framework supports their evaluation on the same per-attribute panel. The released **HumanReader-Reviews-1.7M** annotation set covers the 7 content-driven and 67 popularity-informed attributes that survive validation.

For researchers building review-derived book annotations, popularity-residualized ΔR^2 over a $\log(1 + n_{\text{reviews}})$ baseline should be the standard content-extraction metric. Raw held-out R^2 overstates per-book content recovery by an order of magnitude. Among the methods we evaluate, only LLM-curated keyword densities survive popularity residualization, and only on 7.9% of attributes.

The popularity-residualization protocol should generalize to any review platform where review-count covaries with content type, and lexical anchorability is computable ex ante for any attribute. Together they provide a portable framework for evaluating which book attributes review aggregates can recover, on Goodreads and beyond.

8 Limitations

We catalog the principal limitations.

Sample and matching. The 291 of 387 HUMANREADER books that we mapped to Goodreads records are systematically more popular and more clearly English-language than the 96 unmatched titles, since the matching procedure relies on title normalization and prefers English-language candidates with high ratings counts. Released predictions on Goodreads books that fall outside the matched training distribution (very low review counts, non-English-leaning titles) should be interpreted with caution; we recommend filtering downstream applications to books with at least 10 English reviews.

Reader population. HUMANREADER respondents are habitual Anglophone readers recruited from book-related subreddits; per-attribute targets reflect this population's judgments and we do not test generalization to differently-recruited populations.

Corpus recency. The Goodreads corpus is a 2017 snapshot (Wan and McAuley, 2018). Since 2017 the platform has undergone substantial changes including platform redesign, increased automated reviews, and shifts in the active reviewer population post-Amazon-acquisition; released predictions should be interpreted as describing the 2017 review ecosystem, and downstream applications to current reviews should validate against a fresh sample.

Per-book reader-count noise. Most HUMANREADER books have between two and four reader responses per conceptual question. Per-book sample means are correspondingly noisy. We mitigate via inverse-variance weighting (under the constant within-book noise assumption) and report variance ceilings (Appendix A) that bound achievable R^2. Target stability on the $n_{\text{readers}} \geq 3$ subset is reported in Appendix F. Full R^2 re-validation on that subset is future work.

Inferential basis of release decisions. The seed-bootstrap does not formally adjust for the max-over-method selection step or for partial test-fold overlap across seeds. The permutation null in Appendix J provides a partial check at the per-winner level (re-fitting the winning recipe on permuted labels) but does not address selection. Selection-corrected inference (e.g., max-over-method permutation tests, out-of-fold stacking) is future work.

Popularity confound documented but not eliminated. We report popularity-residualized R^2 as the content-extraction metric and explicitly partial out $\log(1 + n_{\text{reviews}})$ in our headline comparisons. Higher-order popularity-content confounds (e.g., review-content style correlating with publisher marketing intensity) are not separately controlled.

Method-palette completeness. We test nine feature methods plus a distilled LLM-on-reviews variant. Alternative LLM-curated keyword routes (e.g., Claude or Gemini), fine-tuned cross-encoders, retrieval-augmented prompting, and out-of-fold stacked ensembles were not evaluated; our recoverability map is therefore an under-estimate of what is achievable under the method-comparison framing.

Training-data contamination. The LLM used for keyword-bundle generation and for the LLM-on-reviews validation reference (GPT-5) was likely trained on overlapping subsets of the Goodreads review corpus we feature against. Estimates from these two methods should therefore not be read as contamination-free; they may overstate what an LLM trained on a strictly disjoint corpus could achieve. The popularity-residualized comparisons against the non-LLM methods (TF-IDF, embeddings, review meta variants, GoEmotions, ABSA, NMF, stylometry) remain interpretable, since those methods are trained from scratch.

Validation threshold determined post-exploration. The effect-size thresholds ($R^2 > 0.10$ and $|\rho| > 0.15$), the BH-FDR $q = 0.05$, and the predictability-filter thresholds were set in advance of computing the held-out per-method R^2 values reported in the paper, but after preliminary inspection of per-attribute variance ceilings on the HUMANREADER corpus. We do not claim formal pre-registration.

9 Ethical Considerations

Released per-attribute predictions are population means estimated against the reader-survey ground truth of HUMANREADER; they should not be used for automated literary judgment in high-stakes contexts (educational assessment, content moderation, or acquisitions decisions), where the inter-reader variance documented in the dataset paper (Anonymous, 2026) reflects irreducible judgment differences. The Goodreads Book Graph (Wan and McAuley, 2018; Wan et al., 2019) and the HUMANREADER corpus carry their upstream licensing and consent terms; downstream use of **HumanReader-Reviews-1.7M** inherits both.

Use of AI assistants. We used Claude Code (Anthropic) for coding support during data analysis, figure generation, and manuscript formatting. All analyses were designed by the human authors and all numerical results were verified against the underlying data. The authors disclose no commercial conflicts of interest.

References

Anonymous. 2026. HumanReader: A dataset for evaluating descriptive and experiential comprehension in book-length narratives. Parallel submission to EMNLP 2026, under review; full paper included as supplementary material.

Yuri Bizzoni, Pascale Moreira, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2023. Sentimental matters: Predicting literary quality by sentiment analysis and stylometric features. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis (WASSA 2023)*, pages 11–18.

Peter Boot and Marijn Koolen. 2020. Captivating, splendid or instructive? assessing the impact of reading in online book reviews. *Scientific Study of Literature*, 10(1):66–93.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 6)*.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of llms. In *Proceedings of the International Conference on Learning Representations*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the ACL*, pages 4040–4054.

Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 2342–2351.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Sil Hamilton, Rebecca M. M. Hicke, Matthew Wilkens, and David Mimno. 2025a. Too long, didn't model: Decomposing llm long-context understanding with novels. *arXiv preprint arXiv:2505.14925*.

Sil Hamilton, Matthew Wilkens, and Andrew Piper. 2025b. Narrabench: A comprehensive framework for narrative benchmarking. *arXiv preprint arXiv:2510.09869*.

Fabrice Y Harel-Canada, Hanyu Zhou, Sreya Muppalla, Zeynep Senahan Yildiz, Miryung Kim, Amit Sahai, and Nanyun Peng. 2024. Measuring psychological depth in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17162–17196.

Pavan Holur, Shadi Shahsavari, Ehsan Ebrahimzadeh, Timothy R Tangherlini, and Vwani Roychowdhury. 2021. Modelling social readers: novel tools for addressing reception from online book reviews. *Royal Society Open Science*, 8(12):210797.

Yuerong Hu, Zoe LeBlanc, Jana Diesner, Ted Underwood, Glen Layne-Worthey, and J. Stephen Downie. 2022. Complexities associated with user-generated book reviews in digital libraries: Temporal, cultural, and political case studies. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*.

Clayton J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 216–225.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One thousand and one pairs: A "novel" challenge for long-context language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17048–17085.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Marijn Koolen, Peter Boot, and Joris van Zundert. 2020. Online book reviews and the computational modelling of reading impact. In *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*, volume 2723 of *CEUR Workshop Proceedings*.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. Booksum: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558.

Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Elena-Ruxandra Luțan and Costin Bădică. 2022. Emotion-based literature book classification using online reviews. *Electronics*, 11(20):3412.

Guillermo Marco, Julio Gonzalo, and Víctor Fresno. 2025. The reader is the metric: How textual features and reader profiles explain conflicting evaluations of AI creative writing. *arXiv preprint arXiv:2506.03310*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682.

Hannes Rosenbusch and Luke Korthals. 2025. Are some books better than others? *arXiv preprint arXiv:2503.02671*.

Melanie Subbiah, Faisal Ladhak, Akankshya Mishra, Griffin Thomas Adams, Lydia Chilton, and Kathleen McKeown. 2024. STORYSUMM: Evaluating faithfulness in story summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9988–10005.

Melanie Walsh and Maria Antoniak. 2021. The Goodreads "classics": A computational study of readers, Amazon, and crowdsourced amateur criticism. *Journal of Cultural Analytics*, 6(2).

Mengting Wan and Julian McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18)*, pages 86–94.

Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2605–2610.

Cunxiang Wang, Ruoxi Ning, Boqi Pan, Tonghui Wu, Qipeng Guo, Cheng Deng, Guangsheng Bao, Xiangkun Hu, Zheng Zhang, Qian Wang, and Yue Zhang. 2025. Novelqa: Benchmarking question answering on documents exceeding 200k tokens. In *Proceedings of the International Conference on Learning Representations*.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Bao Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: Fairytaleqa–an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 447–460.

Dingyi Yang and Qin Jin. 2025. What matters in evaluating book-length stories? a systematic study of long story evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16375–16398, Vienna, Austria. Association for Computational Linguistics.

A Per-Question Variance Ceilings

A.1 Definition

Each per-book target y_b is a sample mean over n_b reader responses. The sampling variance of this estimator is $\sigma^2_{\text{within}}/n_b$, where σ^2_{within} is the pooled within-book reader variance for the unit. For a model predicting the book population mean μ_b, the achievable R^2 on y_b is bounded by

$$R^2_{\max} = 1 - \frac{\sigma^2_{\text{within}}/\bar{n}_{\text{readers}}}{\text{Var}_b(\mu_b)},$$

which we estimate by substituting $\hat{\mu}_b = y_b$ in the denominator and computing σ^2_{within} via the pooled within-book sum of squares across all reader-book pairs in HUMANREADER. We refer to this as the *variance ceiling* for the unit. The pooled estimator assumes constant within-book reader noise; we use it because per-book within-noise estimates from $\bar{n}_{\text{readers}} \approx 2$–$4$ are too noisy to use as per-book denominators. The ceiling should be interpreted as a population-mean upper bound rather than a per-book one and may understate the true ceiling for polarizing books.

A.2 Distribution

Across the 391 attributes the variance ceiling has mean 0.71, median 0.73, with the 25th and 75th percentiles at 0.64 and 0.89. The minimum is 0.00 (units where the within-book variance equals the between-book variance, i.e., the per-book target is so noisy that no model can exceed an intercept). The maximum is 1.00 (units where every reader gives the same response per book, e.g., genre items in some books). Per-category statistics are reported in Table 1. Per-modeling-unit values will be released alongside the dataset upon acceptance.

Category	Mean	Median	n
Overview	0.72	0.73	54
Plot	0.68	0.68	52
Character	0.86	0.86	86
Narrator	0.90	1.00	52
Setting	0.70	0.75	46
Style	0.49	0.54	96
Reflection	0.69	0.70	5

Table 1: Per-category variance-ceiling statistics across the 391 attributes. Lower ceilings (Style) reflect per-book reader disagreement on style descriptors; higher ceilings (Character, Narrator) reflect strong per-book reader agreement.

A.3 Use in the validation threshold

The predetermined predictability filters (Section 4.3) require ceiling > 0.30 before a unit enters the analysis panel. Across the 391 attributes, 186 fail one or more predictability filter (predominantly checkbox items that no reader in the training set selects, units where reader disagreement is too large for a population-mean target to be predictable, and units where the best fit is degenerate on a near-constant target or numerically catastrophic from overfit). The remaining 205 units form the in-scope analysis panel.

B Keyword Bundle Generation

B.1 Prompt design

For each attribute we issue one prompt to a GPT-5-family model via the OpenAI responses API. For monotonic items we ask for 60 keywords in one call (30 low-end and 30 high-end); for non-monotonic items we ask for 30 keywords per (question, choice).

Monotonic prompt: `Generate keywords that would appear in a Goodreads review of a book where a reader would respond at the low and high ends of the following question: [QUESTION_TEXT]. Return a Python list with 60 entries: 30 low-end keywords followed by 30 high-end keywords. Use single-word or 2-token phrases, lowercased.`

Non-monotonic prompt: `Generate 30 single-word or 2-token phrases (lowercase) that would appear in a Goodreads review of a book where a reader would select "[CHOICE_LABEL]" on the question "[QUESTION_TEXT]". Return a Python list.`

DNF (did-not-finish) lexicon. A single global bundle of 11 keywords: `dnf`, `dnfed`, `dnf'd`, `did not finish`, `could not finish`, `couldn't finish`, `didn't finish`, `couldnt finish`, `didnt finish`, `put it down`, `stopped reading`. This bundle is used in two ways: (a) it produces its own per-book DNF-rate feature (the fraction of reviews containing any DNF keyword) added to every attribute's feature vector, and (b) it serves as an exclusion list when post-processing per-attribute bundles, so an attribute's keyword set never includes DNF terms.

B.2 Validation

Each bundle is post-processed by lower-casing, removing duplicates within bundle, dropping single-token stopwords (a, the, of, etc.), and deduplicating against the DNF lexicon. The final per-bundle keyword count varies from 18 to 30 with a median of 27. Prompt templates and full keyword bundles will be released alongside the dataset upon acceptance.

B.3 Feature computation

Per book b and per bundle B, the keyword density is

$$\rho_{kw}(b, B) = \frac{1}{|R_b|} \sum_{r \in R_b} \sum_{k \in B} \mathbb{1}[k \in r],$$

where R_b is the set of reviews of book b and matches are 1- to 3-gram lowercased with word-boundary regex. The review-coverage density is

$$\rho_{rev}(b, B) = \frac{|\{r \in R_b : \exists k \in B, k \in r\}|}{|R_b|}.$$

Monotonic units have four features per (book, unit) cell (ρ_{kw} and ρ_{rev} for both low-end and high-end bundles); non-monotonic units have two features per cell.

C English-Language Filter on Goodreads Reviews

We apply a two-stage filter to the 15M reviews in the 2017 Goodreads Book Graph.

Stage 1 (fast stopword check). We require at least one of {`the, a, an, and, or, but, in, on, at, to, for, of, with, by, is, are, was, were, this, that`} to appear in the first six space-separated tokens. This drops reviews that are pure tags, emoji-only, or non-English with no English stopword cooccurrence. Reject rate: $\approx 12\%$ of all reviews.

Stage 2 (probabilistic detection). For reviews that pass stage 1, we run `fast_langdetect` (a Python wrapper over fastText's lid.176 model) on the first 100 characters of the review. We retain reviews where the predicted-language confidence for "en" exceeds 0.5. Reject rate: $\approx 3\%$ of stage-1-passing reviews.

Post-filter the corpus contains 13.0M reviews across 1,707,435 books. We chunk the filtered corpus into 44 CSV files of approximately 200K–300K reviews each for streaming-friendly downstream processing.

D Bootstrap Intervals

Per (attribute, method) we observe S per-seed test R^2 values $\{r_1, \ldots, r_S\}$, where $S = 10$ for most methods and $S = 5$ for the full-corpus TF-IDF and NMF runs (Section 4.2). We resample these S values with replacement $B = 500$ times. For each resample b we compute the mean $\bar{r}^{(b)} = \frac{1}{S}\sum_{i=1}^{S} r_i^{(b)}$. The reported bootstrap interval is the 5th and 95th percentiles of $\{\bar{r}^{(b)}\}_{b=1}^{B}$ (a 90% interval). The one-sided bootstrap p-value is the fraction of bootstrap means at or below zero, $p = |\{b : \bar{r}^{(b)} \leq 0\}|/B$.

We bootstrap the per-seed test values (not the per-book residuals) to account for split-induced variance that is the dominant source of uncertainty at the median $n \approx 229$ training books per unit. Per-book residual bootstrapping would underestimate the seed-split variance.

E Benjamini-Hochberg FDR Procedure

Across the full panel of K (method, attribute, model-class) cells (Ridge and Lasso each contribute one cell per method-attribute pair) we collect the one-sided bootstrap p-values $\{p_1, \ldots, p_K\}$. We sort them ascending: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(K)}$. For target $q = 0.05$, we find the largest k such that $p_{(k)} \leq kq/K$ and reject the null for all hypotheses with $p_{(i)} \leq p_{(k)}$. A unit's R^2-route threshold requires both $\bar{r} > 0.10$ and BH significance on the R^2 p-value; the Spearman route is parallel.

F Robustness: $n_{\text{readers}} \geq 3$ Subset

To assess the impact of the $n_{\text{readers}} \geq 2$ training threshold, we recompute target statistics on the subset of (unit, book) pairs with $n_{\text{readers}} \geq 3$. Across the 205 in-scope units, the median number of training books with $n_{\text{readers}} \geq 3$ is 103; nearly all units retain at least 20 such books and at least 50. The within-book sampling noise drops by roughly 33% on this subset (from $1/2$ to $1/3$ readers averaged). Per-unit statistics for the subset will be released alongside the dataset upon acceptance.

G Robustness Panel Across Gates

Table 2 reports total winner counts and Tier A counts across alternative choices of the R^2 threshold, the Spearman threshold, the BH-FDR target q, and the popularity-residualized ΔR^2 threshold that defines Tier A. The qualitative findings (74 winners, 7 Tier A, keyword densities the sole popularity-survivor) are stable across these variations.

Variation	Setting	Winners (Tier A)
R^2 threshold	> 0.05	74 (7)
	> 0.10 (default)	74 (7)
	> 0.15	74 (7)
	> 0.20	74 (7)
Spearman threshold	> 0.10	78 (7)
	> 0.15 (default)	74 (7)
	> 0.20	60 (7)
	> 0.25	45 (7)
BH-FDR q	0.01	65 (7)
	0.05 (default)	74 (7)
	0.10	81 (7)
Tier A ΔR^2	> 0.03	74 (10)
	> 0.05 (default)	74 (7)
	> 0.10	74 (4)

Table 2: Robustness of the 74-winner and 7-Tier-A headline counts across alternative validation-threshold settings. The total winner count moves between 45 and 81 across all variations; Tier A is 4–10.

H Lexical Anchorability Scatter

I Keyword-Bundle Stability

To test how sensitive the keyword-density method is to LLM prompt design, we regenerated the keyword bundles under a re-prompted GPT-5 configuration (system prompt extended with "Be maximally lexically precise: favor distinctive content words over generic descriptors"). The regenerated v2 vocabulary covers the same 391 (concept, choice) attributes with 30 keywords each. Across units, the mean Jaccard similarity between v1 and v2 bundles is 0.142 (median 0.132, range 0.000–0.429). Tier A units exhibit slightly higher stability (mean Jaccard 0.215). We then rebuild keyword densities on the training-mapped subset under v2 and re-fit. Of the 11 v1 keyword-density winners, 9 remain among attributes passing the relaxed threshold ($R^2 > 0.10$ or $|\rho| > 0.15$ pre-multiplicity) under v2; only 1 retains test $R^2 > 0.05$, and the median v1 winner's R^2 drops by approximately 0.3. The signal direction is preserved across LLM regenerations but the absolute magnitudes are prompt-sensitive.

J Permutation Baseline

For each of the 74 validated winners we re-fit the winning recipe (winning feature method, winning model class, fixed regularization α) on 100 permutations of book-level target labels, averaging R^2 across 5 random 80/10/10 splits per permutation. The null distribution of test R^2 per winner spans the range produced by chance label alignments under the fixed pipeline. Per-winner permutation p-value is the fraction of permutations whose mean test R^2 meets or exceeds the observed value. With $N_{\text{perm}} = 100$ the minimum achievable p-value is 0.01; a reported $p = 0$ means no permutation beat the observed.

K Bias Audit on Released Predictions

For each releasable attribute we stratify Goodreads books into review-count quintiles and compute the mean predicted attribute value per stratum. The audit metric is the across-stratum range of stratum-means in units of the overall prediction standard deviation:

$$\text{stratum_range_z}(u) = \frac{\max_q \bar{y}_{u,q} - \min_q \bar{y}_{u,q}}{\sigma(\hat{y}_u)}.$$

We flag an attribute as *popularity-leaning* when stratum_range_z > 0.5 AND the popularity-residualized incremental ΔR^2 over the $\log(1 + n_{\text{reviews}})$ baseline is below 0.05. Of 73 auditable releasable attributes, 20 are flagged. One of the 74 releasable attributes (Narrator: Third person, won by GoEmotions) is excluded from the audit because its predictions have insufficient variation across review-count quintiles for stratum-range computation. Per-attribute audit values for the remaining 73 will be released alongside the dataset.

L Datasheet for the Released Dataset

Following the framework of Gebru et al. (2021) and Díaz et al. (2022), we record the seven datasheet sections below.

Motivation. The dataset provides per-book per-attribute predictions at a scale that expert human annotation cannot reach, as a complement to direct LLM-on-text comprehension benchmarks.

Composition. Each instance is a (Goodreads book, attribute) prediction with bootstrap CI. The dataset contains 74 validated attributes (7 Tier A content-driven plus 67 Tier B popularity-informed). 53 attributes won by full-corpus methods (keyword densities, TF-IDF, sentence embeddings, review meta plus ratings) cover all 1,707,435 books; 21 attributes won by GoEmotions or ABSA cover the matched training subset of ~ 286 books only. An additional 131 in-scope attributes (Tier C) will have per-method test statistics released alongside the dataset without per-book predictions.

Collection process. Ground-truth targets from the HUMANREADER corpus (2024–2025 reader

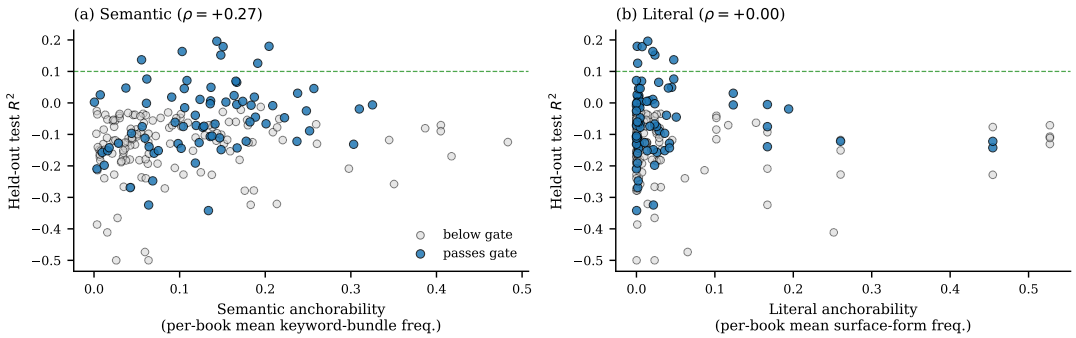

Figure 6: Lexical anchorability operationalizes the recoverability mechanism. **(a)** Per-unit held-out test R^2 versus semantic anchorability (per-book mean fraction of reviews containing any keyword in the LLM-curated semantic neighborhood). **(b)** Per-unit held-out test R^2 versus literal anchorability (per-book mean fraction of reviews containing morphological variants of the response-option text). The semantic neighborhood predicts recoverability ($\rho = +0.27$, $p < 10^{-3}$); the literal surface form alone does not ($\rho \approx 0$). Blue points pass the validation threshold; grey points do not. Outliers $|R^2| > 0.5$ clipped.

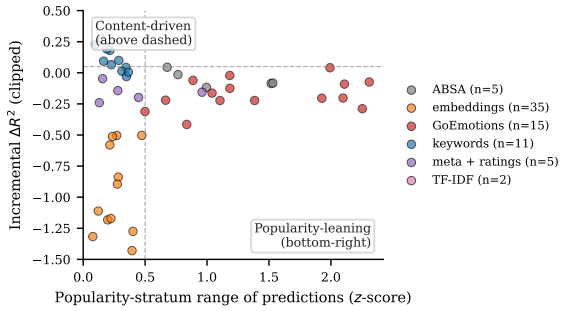

Figure 7: Per-attribute popularity-stratum bias scatter. Each point is a (method, attribute) pair; the x-axis is the across-quintile prediction-mean range (in z-units), the y-axis is the popularity-residualized incremental ΔR^2 over $\log(1 + n_{\text{reviews}})$. Points above the dashed $\Delta R^2 = 0.05$ line are content-driven; points in the bottom-right quadrant (high stratum range, low incremental ΔR^2) are popularity-leaning. Sentence embeddings dominate the low-incremental band by count; GoEmotions and ABSA winners predominantly fall in the popularity-leaning region.

sessions); review text from the 2017 Goodreads Book Graph (Wan and McAuley, 2018); predictions from per-attribute regression models trained 2026.

Preprocessing. Two-stage English-language filter (Appendix C); title-normalized HumanReader–Goodreads match; predictability filters exclude 186 of 391 attributes.

Uses. Recommended for attribute-conditioned book filtering and computational literary analysis at scale for the 53 attributes with full-corpus coverage (won by keyword densities, TF-IDF, sentence embeddings, or review meta plus ratings); the 21 attributes won by GoEmotions or ABSA cover only the matched training subset of ~ 286 books. Predictions are clipped to each attribute's natural range; since 70 of 74 winners pass via the Spearman validation route, natural-scale predictions are validated on rank order more than on absolute-scale calibration (the 4 R^2-route winners, §5.2, are better-calibrated absolutely). Also useful as a calibration target for review-aggregate prediction research. Not recommended for individual reader-judgment inference, since predictions are population means.

Bias and confounds. The popularity-content confound (Section 5.1) flags 20 of 73 auditable releasable attributes (27%) as popularity-leaning. The metadata CSV includes a per-attribute incremental ΔR^2 column.

Distribution and maintenance. On acceptance the dataset will be released under a license matching the upstream HumanReader and Goodreads Book Graph terms; the public URL will be inserted in the camera-ready version, and corrections and additions will be issued as versioned updates. The released models are static after publication; subsequent Goodreads snapshots would require retraining the per-attribute models, which is out of scope for this release.