

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

Samsun Knight
University of Toronto

Yakov Bart
Northeastern University

Ethan Tang
Northeastern University

Jory Gutman
Northeastern University

Abstract

Aggregated reader reviews are widely used as scalable proxies for book attributes in computational literary studies, recommendation, and narratological-benchmark design, but the field lacks a per-attribute audit of what review aggregates can reliably recover. We audit eight review-derived feature methods on 391 attributes, using the HUMANREADER survey corpus (Knight et al., 2026) as ground truth. Most apparent predictive performance is popularity-confounded. Only 34.9% of evaluable attributes have a method that adds signal beyond a popularity proxy (popularity-conditioned incremental $\Delta R^2 > 0.01$), and just 12 additionally pass held-out validation, multiplicity correction, and a stricter $\Delta R^2 > 0.05$ release gate. Recoverability is significantly associated with an attribute’s “lexical anchorability,” i.e. how often reviews mention keywords in the attribute’s semantic neighborhood, with high-anchorability attributes about 2.3 times more likely to clear validation (50% of the top anchorability quartile, 22% of the bottom). We release **HumanReader-Reviews-1.7M**¹, per-book predictions on 12 content-driven attributes (Tier A) and 77 validated but popularity-informed attributes (Tier B), with per-method test statistics on the 111 attributes no scalable method recovers (Tier C). Raw held-out R^2 is a liberal estimate of per-book content recovery, crediting any popularity-correlated content, while popularity-conditioned ΔR^2 is conservative in what it attributes to content; we report both. On a coverage-limited probe of 31 attributes, direct LLM-on-reviews annotation recovers signal the release-panel methods miss, but at a compute cost prohibiting per-book deployment across the 1.7M-book corpus; as a complement, we release an open-weight Qwen 2.5-32B companion scoring 9,990 books.

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 review features as measurements of book attributes that would otherwise require expert annotation. The appeal is scale: the Goodreads Book Graph alone (Wan and McAuley, 2018; Wan et al., 2019) contains roughly 1.7 million books and over 10 million reviews, well beyond what expert panels can annotate or 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 (Knight et al., 2026). These attributes cover Overview (genre and mood), Plot, Character, Narrator, Setting, Style, and Reflection. We map per-attribute recoverability with eight scalable review-derived feature methods (the release panel) plus a coverage-limited direct LLM-on-reviews validation reference. The eight feature methods span surface statistics (TF-IDF, NMF topics, review metadata with ratings, stylistic-readability), pretrained sentiment and semantic representations (GoEmotions, ABSA, sentence-transformer embeddings), and LLM-curated keyword densities, and each is paired per attribute with Ridge or Lasso regression (§4.1).

Of the 200 attributes that pass our predictability filters, which require enough mapped books,

¹<https://osf.io/3ncuz/>

target variance, and reader agreement for a per-book mean to be estimable, 89 pass our validation threshold² and 111 fail it under every one of the eight release-panel feature methods.³ At least one method clears popularity-conditioned incremental $\Delta R^2 > 0.01$ on 34.9% of evaluable attributes (105 of 301, a looser evaluable set than the 200 because the paired statistic needs only that both nested models be fittable); only 12 additionally clear the pre-specified $\Delta R^2 > 0.05$ release gate, held-out validation, and multiplicity correction. The recoverable subset concentrates in Overview descriptors, namely genre labels and mood adjectives, and in world-type Setting items; Character and writing-style descriptors are recovered at intermediate rates. Narrator attributes are the least recoverable: the narrator-tone adjectives readers volunteer in review text (e.g., *jaded*, *warm*) pass, while the narration-type labels themselves, first- versus third-person, do not pass. Fine-grained Plot structure is recovered only patchily, and the integrative reading-experience judgments, such as how enjoyable readers found the book, fail under every release-panel 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. The genre label “Mystery,” whose semantic neighborhood includes terms such as *detective*, *clue*, *investigation*, *suspect*, and *whodunit*, reaches held-out test $R^2 = 0.26$ under keyword densities (bootstrap 90% interval [0.09, 0.43]) and adds predictive value beyond the review-count baseline, entering the release on the Spearman route, since the R^2 route also requires multiplicity correction, while “overall enjoyment” lacks a distinctive review vocabulary and fails validation. Across attributes, anchorability is associated with release-eligibility at Spearman $\rho = 0.20$ ($p = 0.00442$).

We release **HumanReader-Reviews-1.7M**, a per-book annotation set that separates attributes by whether their feature families add predictive value beyond linear review count, namely predictions for 12 content-driven attributes (Tier A) and 77 validated but popularity-informed attributes (Tier B),

with per-method test statistics for the remaining 111 attributes no scalable method recovers (Tier C). The release tiering and per-attribute metadata are described in §6. Raw held-out R^2 is a liberal estimate of content recovery, and popularity-conditioned incremental ΔR^2 is conservative in what it attributes to content (§5.1). We report both, and neither bounds the true content share. We also provide an open-weight Qwen 2.5-32B companion annotation of 9,990 books, produced by prompting the model directly on sampled reviews. We audit the 228 registry attributes above the Qwen screening floor, a looser frame than the 200 in-scope for the main release, against HUMANREADER ground truth on 287 books, and deploy the 25 attributes clearing partial- $r \geq 0.30$. The companion covers a fraction of the 1.7M corpus but provides a reference point for direct-LLM annotation at scales beyond what expert panels can score.

2 Related Work

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).⁴ Closest to our audit, Rosenbusch and Korthals (2025) conclude from 624,320 Goodreads reviews that reviews contain up to ten times more information about the reviewer than about the book. We extend that finding from aggregate ratings and cross-review agreement to a per-attribute recoverability map.

A second line evaluates long-context LLMs on narrative comprehension at varying lengths,⁵ 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, Marco et al. (2025) traces conflicting creative-writing evaluations to stable differences in reader profiles rather than to

²Held-out test $R^2 > 0.10$ or $|\rho| > 0.15$, with multiple-comparison correction.

³A ninth comparison, direct GPT-5 annotation from sampled reviews, is evaluated on 31 units and is not a release-panel winner candidate, since per-book LLM inference is prohibitively expensive at the 1.7M-book release scale.

⁴Allied work treats Goodreads as a measurement instrument and documents its biases (Koolen et al., 2020; Hu et al., 2022; Walsh and Antoniuk, 2021).

⁵NarrativeQA (Kočíský et al., 2018), NovelQA (Wang et al., 2025), NoCha (Karpinska et al., 2024), BookSum (Kryściński et al., 2022), BoookScore (Chang et al., 2024), LongStoryEval (Yang and Jin, 2025), FairytaleQA (Xu et al., 2022), and PsychDepth (Harel-Canada et al., 2024).

noise, while Subbiah et al. (2024) finds that disagreement on story-summary faithfulness often reflects errors annotators miss until they are pointed out, so no single protocol establishes ground truth. Reader disagreement therefore mixes genuine perspective with detection failure, and our per-attribute reader-agreement ceilings bound the first without separating the two. 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

We combine (1) the HUMANREADER survey corpus, which provides per-reader responses to 100 closed-form post-reading questions across 384 fictional novels, resolving to 391 per-book modeling units that supply our ground-truth targets (§3.1),⁶ and (2) the Goodreads Book Graph (Wan and McAuley, 2018; Wan et al., 2019), from which our English-token heuristic retains 12,996,191 reviews across 1,707,435 books. Goodreads reviews are the source of our features, and the Goodreads books are the set for which we release predictions.

3.1 HUMANREADER Ground Truth and Attributes

HUMANREADER (Knight et al., 2026) comprises 384 fictional novels with 807 finished post-reading sessions from 711 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. We maintain throughout the identifying assumption that, under this design, per-book target means reflect book content rather than reader self-selection. We use the post-reading instrument as ground truth. The instrument comprises 100 closed-form questions across seven categories: Overview (overall enjoyment, genre tags, and mood descriptors, with response options such as *Lighthearted* and *Bittersweet*); Plot (turning-point location and event structure); Character (protagonist properties, conflict type, and character-arc trajectories); Narrator (narrator type, with response options such as *first-person* and *unreliable*); Setting (era, place, and world-building); Style (writing-style descriptors

⁶The companion paper (Knight et al., 2026) analyzes 100 closed-form post-reading items from a 101-item instrument; the 101st (a held-out fit-to-“ideal reader” item) is released with the corpus but excluded from analysis in both papers.

and readability, with response options such as *Humorous* and *Dark*); and Reflection (post-reading reactions). For full instrument design and inter-reader agreement statistics, we refer the reader to Knight et al. (2026).

We construct per-book modeling targets from the 100 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 units across the 100 questions. We restrict training to books with $n_{\text{readers}} \geq 2$ for the attribute and weight each (book, attribute) 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 E).

3.2 Goodreads Join and Variance Ceiling

Of the 384 survey books, 291 (75.8%) match a Goodreads record, giving 289 distinct physical books with review features, since two pairs of survey titles share a Goodreads record. We match on normalized titles in a streaming pass, preferring English-language candidates with the highest ratings count among ties. After applying the per-unit $n_{\text{readers}} \geq 2$ filter and pooling duplicate physical titles, the median number of training books per attribute is 194. Each seed then applies the 80/10/10 book-level split as described in Section 4.2.

Per-book targets are sample means of reader responses, so sampling noise places 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 E. The median ceiling across attributes is 0.44 (with 1 indicating full per-book reader agreement and 0 indicating noise-dominated targets).

4 Feature Methods and Validation Protocol

We predict each attribute’s per-book target y_b from features derived from book b ’s aggregated reviews, where the attributes are the modeling units defined in §3.1. In the remainder of this section, we de-

scribe the eight release-panel feature methods and the direct-LLM validation reference we evaluate (§4.1), our training and bootstrap-evaluation protocol (§4.2), the incremental-value specification (§4.3), and the validation threshold and release gate that determine which per-attribute predictions enter the release (§4.4).

4.1 Feature Methods

We order the review-derived feature methods by inference cost per prediction, since the planned release includes 1,707,435 books and per-book inference cost binds at that scale. We condition the main comparison on a popularity-only baseline ($\log(1 + n_{\text{reviews}})$ per book) through paired nested models.

Cheap tier. Both cheap-tier methods scale to full-corpus inference. The first, LLM-curated keyword densities, involves prompting GPT-5 once per attribute to generate keywords reflecting each polarity (60 keywords in one call for monotonic items, whose response options have a natural order, and 30 per choice for non-monotonic; see Appendix F). For each book and each keyword bundle, we then compute a keyword density and a review-coverage density, giving four features per monotonic attribute and two per non-monotonic attribute, plus the two global did-not-finish densities, for six and four in all. The second, review meta-features, uses review count, mean review length, fraction of reviews below 50 characters, fraction of reviews above 500 characters, and $\log(1 + n_{\text{reviews}})$. The meta-plus-ratings variant adds per-book rating-distribution and reviewer-cluster summaries, for 27 features in all, and the full list is in the replication package.

Mid tier. Two mid-tier methods test whether richer text representations recover more. TF-IDF fits a `TfidfVectorizer`, without IDF weighting, on a Goodreads subsample stratified by review count, for downstream Ridge/Lasso regression, so the family is a surface-frequency baseline rather than a tuned lexical representation (Appendix H). Sentence-transformer embeddings encode the first 200 retained reviews per book with `all-MiniLM-L6-v2`, normalize each 384-dimensional review vector, and mean-pool all 384 dimensions per book. Training and full-corpus inference use the same global review cap and pooling rule.

Expensive tier. We use direct LLM-on-reviews annotation to set an upper bound on what per-book

LLM calls could recover, as a validation reference rather than a release candidate. For a uniform random sample of 200 mapped training books, of which 163 have retained reviews, we sample up to 20 reviews per book and prompt GPT-5 with the survey question and the reviews. The canonical registry join yields 31 evaluable attributes with at least 20 books and non-degenerate ground-truth variation. Their direct-extraction recovery appears in Appendix Q. This coverage-limited probe serves as the ninth comparison but does not enter release-panel winner selection.⁷

Additional methods. Four further methods test whether specialized feature families recover attributes that broad-purpose features miss, each targeting one dimension of the review text: emotion, from RoBERTa-GoEmotions (Demszky et al., 2020) logits per review, mean-pooled at the book level (Appendix H); aspect sentiment, from keyword-anchored VADER (Hutto and Gilbert, 2014) on eight book aspects (ABSA, aspect-based sentiment analysis); topic structure, from NMF (Lee and Seung, 1999) with $K = 50$ components; and writing style, from stylistic-readability features. For full implementation details, see the replication package.

4.2 Training and Evaluation Protocol

We fit and evaluate each (attribute, feature method) pair as follows. We consider only books with at least two reader responses and a corresponding entry in the Goodreads review corpus (Section 3.2), and we require at least 16 eligible mapped books per attribute, which is the smallest sample that leaves at least two books in both validation and test folds under an 80/10/10 split. The retained units’ median of 194 books per attribute is small for stable per-attribute R^2 inference, so we report bootstrap intervals throughout. We apply n_{readers} weights to each (book, unit) row during fitting and evaluation, the inverse-variance weight motivated in §3.2. We use an 80/10/10 random split at the book level for train/validation/test drawn per seed, and average results over 5–10 seeds depending on method, 5 for TF-IDF and NMF topics and 10 for all others. For each seed, Ridge selects $\alpha \in \{0.1, 1, 10, 100\}$ on the validation fold.

⁷As a second-best that scales roughly sixty times further than the per-book GPT-5 probe, we use open-weight Qwen 2.5-32B at intermediate cost and deploy LLM-derived annotations to a Goodreads subsample rather than only probing them; see Appendix R.

Lasso likewise selects on validation: global-feature families use $\{0.0001, 0.001, 0.01, 0.05, 0.1\}$ and keyword densities use $\{0.001, 0.01, 0.05, 0.1, 0.5\}$. The model class for each (unit, method) is chosen by the highest seed-mean validation R^2 , with exact ties broken by preferring Ridge, and test-set metrics (test R^2 , test Spearman ρ) are reported for the winner. The full-data release refits use the median validation-selected α across seeds for the winning class.

We form 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 it to the Goodreads corpus, clipping predictions to the registry-derived natural range; full-corpus methods cover all 1.7M books, while GoEmotions and ABSA cover the matched training subset (§6).

4.3 Popularity-Conditioned Incremental Value

For the popularity-conditioned audit we compare paired nested Ridge models. Within each weighted 80/10/10 book split, the baseline model predicts the raw target using only $\log(1 + n_{\text{reviews}})$, and the combined model predicts the same target from the method features plus that popularity term. Ridge $\alpha \in \{0.1, 1, 10, 100\}$ is selected separately for each model on the validation fold. Feature standardization and, for feature families with more than 100 columns, the top-100 variance ranking are fitted on the training fold only. Both models are scored on the identical held-out test books. We define incremental $\Delta R^2 = R^2_{\text{method+popularity}} - R^2_{\text{popularity only}}$ per seed and average across ten seeds. Thus ΔR^2 measures held-out predictive value added by the review-feature family conditional on linear review-count information. We compare nested models on the raw target rather than projecting popularity out of the target and features in a first stage, since the nested form keeps ΔR^2 on the raw target's R^2 scale and adds one fewer fitted stage, at the cost of not being bounded by what the method explains of the popularity residual (Appendix B). An attribute's cell is excluded when the combined model's held-

out R^2 is numerically degenerate, since a paired difference between two ill-conditioned fits says nothing about content share; the survivor count is insensitive to where that cut is drawn (§5.1). We require at least 30 mapped books per attribute for this audit.

Raw held-out R^2 credits all predictive channels available to a feature method, including review-count covariation. The nested ΔR^2 diagnostic is narrower: it asks what the method adds beyond a linear log-review-count baseline on matched folds. The statistic can still contain nonlinear or proxy popularity channels, so we interpret ΔR^2 as popularity-conditioned incremental prediction rather than content purified of every attention-related confound.

4.4 Validation Threshold and Release Gate

An attribute 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 , or mean test $|\rho| > 0.15$. The magnitude threshold is two-sided, but the multiplicity test that accompanies it is one-sided for a positive association, so an attribute whose mean ρ is strongly negative clears the threshold and then fails the correction: in the realized panel the 6 in-scope attributes with mean $\rho < -0.15$ are all rejected, and the minimum among winners is 0.16.
- *Multiplicity*: the attribute's one-sided bootstrap p -value passes the BH-FDR critical value at $q = 0.05$ (Appendix J), applied across all (method, attribute, model-class) cells of the full 301-attribute panel. We run BH separately on the R^2 and Spearman p -values and admit an attribute by either route, so the joint error rate across the two routes is not itself controlled at q .
- *Predictability filters*: target variance across training books > 0.005 , target range > 0.05 , variance ceiling (Section 3.2) > 0.30 , and a non-degenerate best-across-methods fit, where we drop an attribute whose selected fit returns test $R^2 < -1$ (a numerical blow-up) or $R^2 > 0.99$ with a near-zero bootstrap interval (a near-constant target).⁸

⁸Appendix S lists the notable attributes these filters drop.

A releasable attribute is additionally tagged *content-driven* if its mean test R^2 is positive and its popularity-conditioned incremental ΔR^2 over the $\log(1 + n_{\text{reviews}})$ baseline exceeds the pre-specified 0.05 release gate; Section 6 structures the release around this criterion. The separate population diagnostic in Section 5.1 reports the more permissive $\Delta R^2 > 0.01$ threshold.

5 Results

We map per-attribute recoverability and isolate its content-driven component rather than ranking feature methods. 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 associated with 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. We therefore compute the held-out predictive value added by each feature family beyond a linear popularity-only baseline, following the paired nested-model specification in §4.3.⁹

Across the eight release-panel feature families, at least one family clears popularity-conditioned incremental $\Delta R^2 > 0.01$ on 34.9% of attributes (105 of 301 evaluable¹⁰), with the largest method-

specific counts from stylistic-readability features (38) and GoEmotions (36). Of these, 12 additionally pass the full held-out validation and multiplicity-correction criteria of §4.4 and clear the stricter pre-specified $\Delta R^2 > 0.05$ release gate, entering the release as content-driven Tier A, while the remaining 93 do not clear one or more release gates.

In raw held-out R^2 (Figure 3a, Appendix B), every family has a negative median: stylistic-readability features have the median closest to zero and keyword densities the only positive 95th percentile, while TF-IDF, sentence embeddings and the topic method have the heaviest negative tails. The 3 attributes clearing the strict $R^2 > 0.10$ route are exceptions to that pattern. In the nested-model comparison (Figure 3b), keyword densities and GoEmotions have the longest positive tails, though the keyword advantage is partly by construction, since the bundles are LLM-curated per response option. Most release-panel methods add little predictive value beyond review count. Because held-out R^2 differences are unbounded below, Table 1 (Appendix B) reports the complete artifact-derived distribution, and we do not interpret large negative values as a negative content share.

Raw R^2 and ΔR^2 answer different questions, and we report both so that neither is read as an attribute’s content share. Raw R^2 is a liberal estimate of content recovery: it credits the method for any held-out predictive signal, including signal that is really popularity. ΔR^2 is conservative in what it attributes to content, since we order the nested models so that popularity takes first claim on shared variance, withholding from the method any content that moves with popularity. Held-out R^2 is not a variance decomposition: the popularity-only model has negative held-out R^2 on 98% of evaluable attributes, so the paired difference adds that model’s own generalization failure back in, and ΔR^2 exceeds the method’s raw R^2 for 124 of 200 attributes. Neither statistic bounds the content share, which is not point-identified here. A two-stage residualized variant, which does bound the method’s credit, is much stricter: 8 of the 12 Tier A attributes clear its 0.05 gate (Appendix B).

a combined R^2 below -10^6 , and the count is insensitive to where the cut is drawn: the same 105 attributes result anywhere between discarding $\Delta R^2 > 1$ and tightening to a combined R^2 below -1 . The denominator stays all 301 attributes for which the statistic was computed, so the 34.9% is conservative: 60 do not contribute a non-degenerate cell and count as non-survivors.

⁹We 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 and newly-released books have not yet accumulated the popularity signal that popularity-informed attributes rely on; and generalization to other review platforms requires content extraction rather than corpus-specific popularity. Where these distinctions do not matter for the downstream use, popularity-informed predictions remain useful.

¹⁰The nested-model comparison evaluates on 301 attributes, more than the 200 in-scope for the main release, since the paired ΔR^2 statistic requires only that two Ridges be fit-table and their difference computable, and is not subject to the variance-ceiling, target-range or degenerate-fit gates that govern main-pipeline winner selection, though it imposes its own floor of 30 mapped books per attribute, stricter than the main pipeline’s 16. We screen out attributes whose combined method-plus-popularity Ridge is numerically degenerate on the held-out fold, since a paired difference between two ill-conditioned fits says nothing about content share; we cut at

The release tiers therefore separate attributes whose apparent performance tracks review volume from attributes that add signal conditional on it.

We also audit bias per attribute, stratifying Goodreads books into review-count quintiles and flagging a winner as *popularity-leaning* when its prediction means range more than 0.5σ across those quintiles while its popularity-conditioned incremental ΔR^2 stays below 0.05. Under this criterion 27 of 87 auditable (method, attribute) cells are flagged, and every flagged cell is a GoEmotions (22/27) or ABSA (5/11) winner. No flag falls on a sentence-embedding (0), keyword-density (0), review-meta (0), TF-IDF (0) or NMF-topic (0) winner; embeddings fall in the low-incremental band but their across-quintile means vary too little to meet the stratum-range criterion. Appendix P gives the per-stratum values and the coverage caveat for GoEmotions and ABSA, whose predictions cover matched books only. 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, 191 fail one or more predictability filter (the 16-mapped-book floor, target variance, range, ceiling, or non-degenerate best fit), leaving 200 in-scope. Taking the best across methods per attribute (Figure 2; per-threshold staircase in Appendix C), 89 units pass our validation threshold. All 89 pass via the Spearman $\rho > 0.15$ route, and only 3 also clear the $R^2 > 0.10$ route, since that route additionally requires multiplicity-corrected significance, which binds before the ceiling does.¹¹ Because the Spearman route does not require positive R^2 , 70 of the 89 winners have non-positive mean held-out test R^2 , all of them Tier B; those attributes are recovered in rank order rather than in squared error, and their released predictions should be interpreted in rank terms. Across the 89 winners, the median bootstrap interval width on test R^2 is 0.21; 44 have width below 0.20 and 58 below 0.30, with 22 exhibiting wider intervals (above 0.40) primarily on small- n attributes. The 89 wins distribute across

methods as GoEmotions 27, sentence embeddings 25, keyword densities 15, ABSA 11, NMF topics 5, TF-IDF 4, review meta plus ratings 2, and stylistic-readability 0. Of these 89 winners, 12 also clear the pre-specified popularity-conditioned incremental-value gate $\Delta R^2 > 0.05$ (§6): 6 keyword-density, 3 sentence-embedding, 2 ABSA, and 1 GoEmotions winners. The top-tail by held-out R^2 is dominated by genre and mood descriptors, plus the “how fantastical was the world” setting item (Figure 1; full release details in Appendix A).

5.3 Per-Category Pass Rates

Pass rates vary substantially across the seven HumanReader categories (Appendix D): Overview passes at 78% (31 of 40), Setting at 55% (12 of 22), Character at 46% (17 of 37), Style at 42% (10 of 24), Plot at 32% (12 of 38), Reflection at 20% (1 of 5, too small to characterize), and Narrator at 18% (6 of 34).

Wherever readers describe a choice through its semantic neighborhood in review text (mood adjectives, genre tags, style descriptors, world-type scales), 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.

5.4 Lexical Anchorability and Recoverability

Recovery concentrates in attributes whose response options readers describe directly in review text (§5.3). We operationalize that pattern 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., “Mystery” $\rightarrow$ *detective*, *clue*, *investigation*, *suspect*, *whodunit*, ...) and captures the natural register readers use, which the keyword-density method is designed to exploit.

Lexical anchorability is significantly associated with per-attribute release-eligibility across the 200 in-scope attributes with a computed anchorability score, at Spearman $\rho = 0.20$ on the binary pass indicator ($p = 0.00442$); the scatter appears in Appendix M. The association runs through the LLM-curated semantic neighborhood rather than the literal surface form: semantic anchorability correlates with $|\rho|$ on the release-eligibility Spearman

¹¹The 3 strict $R^2 > 0.10$ winners are the genre label “Historical Fiction” and the mood label “Mysterious” under sentence embeddings, and the Setting item “how fantastical was the world” under keyword densities.

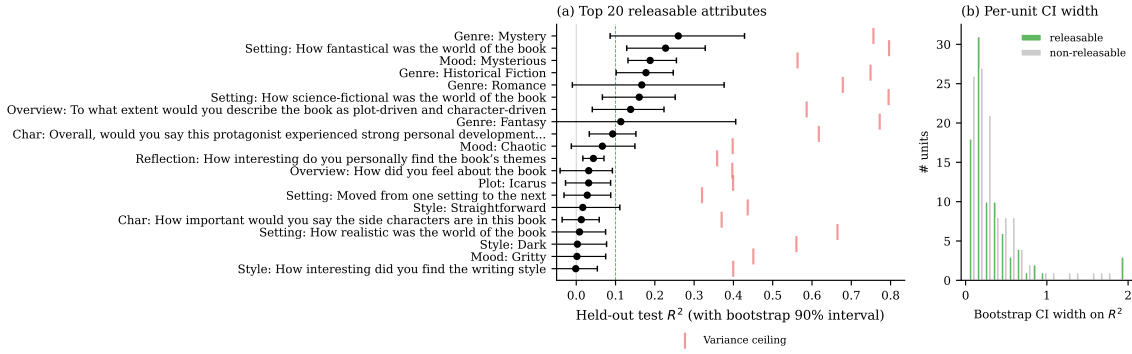


Figure 1: Top 20 releasable attributes sorted by held-out test R^2 , with bootstrap 90% intervals (black bars) and variance ceilings (red ticks). Label prefixes name HumanReader categories or Overview sub-categories (*Mood*, *Genre*, *Overview*:), and *Char* abbreviates Character; no Narrator item reaches the top-20. Of the top 20, 12 are Tier A (content-driven) and 8 are Tier B (popularity-informed); raw R^2 correlates with content-driven-ness but does not identify it. Panel (b) shows bootstrap CI widths, releasable vs. non-releasable (clipped at 2.0).

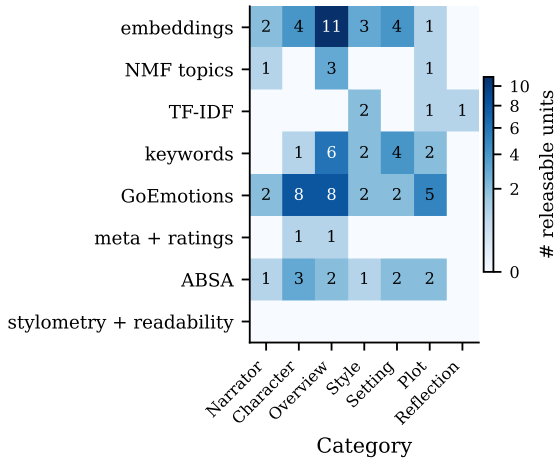


Figure 2: Number of releasable (concept, choice) attributes won by each method (rows) within each category (columns). Wins distribute across seven feature families, with GoEmotions and sentence embeddings the most frequent winners; content-driven wins (survivors of the popularity-conditioned incremental-value gate, §5.1) span keyword densities, sentence embeddings, ABSA, and GoEmotions.

route at $\rho = 0.15$ ($p = 0.0371$), while the literal surface form is not detectably correlated with $|\rho|$ ($\rho = 0.07$, $p = 0.34$).

To test whether the relationship survives a re-generated lexicon, we re-compute anchorability under a re-prompted keyword bundle (Appendix N), drawn from GPT-5 with a lexically distinct system message and sharing mean Jaccard 0.149 with v1. Correlating this v2 anchorability against the v1-feature test Spearman $|\rho|$ gives Spearman $\rho = 0.29$, $p = 0.0000275$; the attribute-level relationship transfers across lexicon regenerations. The specific per-attribute R^2 magnitudes transfer only partially. Rebuilding the keyword-density features on the v2 bundle and refitting yields a median signed v2-minus-v1 R^2 change of -0.026 but a median abso-

lute change of 0.148; only 1 of the 15 v1 keyword-density winners retains test $R^2 > 0.05$ under v2, while 11 remain above the pre-multiplicity threshold. A stability test with the same base model cannot rule out training-data contamination on the underlying books.

5.5 Robustness Across Validation Gates and Keyword Regeneration

Because the headline counts depend on validation-threshold choices, we sweep alternative settings. All winners pass via Spearman, so the R^2 gate is non-binding and total winners remain 89 across the reported R^2 choices: the Spearman sweep shifts the total from 48 to 96, the BH-FDR sweep from 78 to 96, and the content-driven count varies from 7 to 14 across alternative popularity-conditioned incremental-value thresholds (tabulated in Appendix L).

Six of the 12 Tier A winners rest on keyword bundles from a single GPT-5 prompt, while the rest use sentence-embedding, ABSA or GoEmotions features and are unaffected by the bundle. The re-prompted bundle of §5.4 leaves 11 of 15 keyword winners above the relaxed threshold.

As a per-winner null check, we re-fit each of 64 non-embedding winners' recipes on 100 permutations of book-level target labels (Appendix O). 23 exceed their permutation 95th percentile at a median per-winner p of 0.25, and all 9 Tier A winners covered by the permutation pass exceed it at $p \leq 0.01$; the overall rate is driven by Tier B. The 25 sentence-embedding winners are not covered, since re-fitting the 384-dimensional basis under permutation was outside our compute budget, so no permutation evidence is available for that family.

6 The HumanReader-Reviews-1.7M Release

Given the popularity-confound audit (§5.1) and the lexical-anchoring mechanism (§5.4), we release **HumanReader-Reviews-1.7M** as a three-tier annotation set spanning the 200 in-scope attributes. Tiers A and B give per-book predictions for 87 of the 89 attributes that pass validation, of which the 38 won by GoEmotions or ABSA cover 286 matched books rather than the full corpus; Tier C documents the remaining 111 through per-method test statistics only.

Tier A (Content-driven, $n = 12$). Passes the validation threshold with positive held-out test R^2 and popularity-conditioned incremental $\Delta R^2 > 0.05$. Comprises 6 keyword-density, 3 sentence-embedding, 2 ABSA, and 1 GoEmotions winners.

Tier B (Popularity-informed, $n = 77$). Passes the validation threshold but does not clear the incremental-value gate, so we do not establish that these predictions carry content signal beyond review volume; for 58 of these attributes the incremental point estimate is itself negative. The tier marks a failure to establish content value, not a demonstrated popularity gradient; our bias audit does not flag any popularity-leaning cell among the sentence-embedding, keyword-density or review-meta winners (§5.1). Two Tier B winners are withheld from the released predictions because their refits return a constant (Appendix A).

Tier C (Documented, $n = 111$). Passes our predictability filters but fails the validation threshold under every one of the eight release-panel feature methods. Concentrates in Narrator, Plot, and Character, namely the first- versus third-person narration-type labels, plot-structural items, and integrative character judgments (§5.3).

Alongside the tiered release, we ship a companion Qwen 2.5-32B annotation of 9,990 books, covering the 25 attributes clearing partial- $r \geq 0.30$ against HUMANREADER ground truth on 287 books. Appendix A details the release format, per-book coverage indicators, aggregation-gap caveats, and the companion probe’s validation and deployment protocol.

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, style descriptors, world-type Setting items,

and the narrator-tone adjectives that readers volunteer (e.g., *jaded*, *warm*), while the narration-type labels themselves, first- versus third-person, do not pass. Review-derived features fail under every one of the eight release-panel methods when the response option requires text inference (e.g., turning-point location or ending type) or integrative judgment (e.g., overall enjoyment). The 111 Tier C attributes mark the empirical boundary of recovery under this method palette. A ninth comparison (direct GPT-5 annotation from sampled reviews, §4.1) is a coverage-limited validation reference on 31 evaluable units rather than a release candidate.¹² The per-attribute panel supports evaluating the richer feature families discussed in Limitations. The released **HumanReader-Reviews-1.7M** annotation set covers the 12 content-driven Tier A and 77 Tier B popularity-informed attributes that survive validation.

For researchers building review-derived book annotations, popularity-conditioned incremental ΔR^2 over a $\log(1 + n_{\text{reviews}})$ baseline should be reported alongside raw held-out R^2 . Raw R^2 is a liberal estimate that credits popularity-correlated channels, while ΔR^2 is conservative in what it attributes to content. Neither bounds the content share, and the two are not numerically ordered (§5.1). At the population diagnostic ($\Delta R^2 > 0.01$), 105 of 301 attributes (34.9%) clear under at least one release-panel method, led by stylistic-readability features (38) and GoEmotions (36). At the fully validated release gate ($\Delta R^2 > 0.05$), 12 attributes enter Tier A as content-driven across four families (§6), with none won by the stylistic-readability features that lead the population diagnostic.¹³ Taken together, our results show that what a review aggregate can be trusted to predict varies attribute by attribute, and the panel and tiering let downstream users draw that distinction rather than assume it.

¹²As a second-best scaling roughly sixty times further than the per-book GPT-5 probe, we use open-weight Qwen 2.5-32B zero-shot on sampled reviews; it mostly re-recovers attributes the release panel already reaches rather than widening the frontier (Appendix R).

¹³Release-panel methods also fail on most narrative-structure attributes (opening hook, ending reversal, primary conflict type), though one story-shape option reaches Tier A under keyword densities.

Limitations

Sample and matching. The 291 of 384 HUMANREADER books that we mapped to Goodreads records need not be representative of the 93 unmatched titles, since the match depends on title normalization. Released predictions on books outside the matched training distribution (very low review counts, non-English-leaning titles) warrant caution; downstream applications should filter to books with at least 10 reviews in the released coverage table.

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. Released predictions therefore estimate what habitual Anglophone readers report, and are not validated for a broader reading population.

What Tier A does and does not establish. Tier A membership is a screen, not a significance claim. Its held-out R^2 ranges from 0.003 to 0.26, and only 3 of the 12 are BH-significant on R^2 ; all 12 qualify on the Spearman route. The incremental-value gate that separates Tier A from Tier B is a ten-seed point estimate reported without an interval. It is also computed on a Ridge model of the winning feature family, capped at its 100 highest-variance columns, rather than on the released recipe (§4.3), so it evidences incremental signal for the family rather than for the shipped predictor.

Corpus definition. Which attributes are recoverable is less stable than how many. Rebuilding the pipeline under a `fast_langdetect` English filter instead of our historical indicator heuristic (Appendix G) leaves the totals close (84 winners and 14 Tier A against 89 and 12), but the two panels share only 6 Tier A attributes and 58 release-pass attributes. No language gold standard is available to adjudicate between the two corpus definitions, so we ship the historical one and report the comparison rather than treating either panel’s membership as settled.

Corpus recency. The Goodreads corpus is a 2017 snapshot (Wan and McAuley, 2018). The platform has since been redesigned, hosts more automated reviews, and draws a different active reviewer population; 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. The per-book-unit reader count has median 2 and 95th percentile 4. 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 E) that bound achievable R^2 . Target stability on the $n_{\text{readers}} \geq 3$ subset is reported in Appendix K. 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 O provides a partial check at the per-winner level (re-fitting the winning recipe on permuted labels) but does not address selection. Of the 64 winners for which we ran the permutation pass, 23 exceed their permutation 95th percentile at a median per-winner p -value of 0.25; the release tiering therefore rests on the Spearman-route and BH-FDR criteria, and not on the stricter permutation null. Selection-corrected inference (e.g., max-over-method permutation tests, out-of-fold stacking) is future work.

Popularity confound documented but not eliminated. We compare review features plus $\log(1 + n_{\text{reviews}})$ against the log-review-count baseline alone. Higher-order and proxy popularity-content confounds (e.g., review-content style correlating with publisher marketing intensity) are not separately controlled, so the diagnostic is not a causal content share. NMF topics show the gap most sharply: their incremental ΔR^2 is negative at every percentile (Table 1, Appendix B) yet they win 5 Tier B attributes on the Spearman route.

Method-palette completeness. We test eight release-panel feature methods plus a coverage-limited direct LLM-on-reviews reference, for nine comparisons in total. 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 two LLMs used in this paper, GPT-5 (for keyword-bundle generation and the LLM-on-reviews validation reference) and Qwen 2.5-32B (for the open-weight companion release), were likely trained on overlapping subsets of the Goodreads review corpus we build features from. Estimates from these meth-

ods should therefore not be read as contamination-free; they may overstate what an LLM trained on a strictly disjoint corpus could achieve. Four of the non-LLM methods — TF-IDF, NMF topics, review meta plus ratings and stylistic-readability — are fit from scratch on our corpus, so their popularity-conditioned comparisons carry no contamination of this kind. The remaining three are not: sentence embeddings, GoEmotions and ABSA use pretrained encoders and lexicons whose training data we do not control, so they could in principle encode book knowledge acquired elsewhere. We do not separate that channel from review-derived signal for those three.

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.

Ethical Considerations

Released per-attribute predictions are population means; they should not be used for automated literary judgment in high-stakes contexts (educational assessment, content moderation, acquisitions), where the inter-reader variance documented in the dataset paper (Knight et al., 2026) reflects irreducible judgment differences. **HumanReader-Reviews-1.7M** inherits the upstream licensing and consent terms of the Goodreads Book Graph (Wan and McAuley, 2018; Wan et al., 2019) and HUMANREADER. We used Claude Code (Anthropic) for coding support; all analyses were designed by the human authors and numerical results verified against underlying data. The authors have no commercial conflicts of interest to disclose.

References

- 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. BoookScore: 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.
- Samsun Knight, Yakov Bart, Jory Gutman, and Jimin Lee. 2026. HumanReader and HumanViewer: Two datasets of perceived narrative attributes in book-length fiction and feature-length film. Concurrent cross-medium dataset paper; introduces the HumanReader book-side corpus (used here as ground truth) alongside the HumanViewer film-side corpus.
- 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, Xiangukun 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 Release Details

We apply the winning method per (question, choice) to the Goodreads Book Graph (Wan and McAuley, 2018; Wan et al., 2019) and release per-book predictions for the 89 attributes that pass our validation threshold, two of which are withheld as constant refits (*Withheld constant-refit winners*, below) (Section 4.4). Across all eight release-panel methods, at least one method clears the population diagnostic $\Delta R^2 > 0.01$ for 105 of 301 evaluable attributes (34.9%); stylistic-readability features (38) and GoEmotions (36) have the largest method-specific counts, ahead of ABSA (33) and keyword densities (31). Of those population-diagnostic survivors, 93 fail one or more full release gates and therefore do not enter Tier A. The Tier A keyword-density subset is sensitive to the specific bundle at the magnitude level, though the mechanism-level anchorability relationship transfers across regenerations (Appendix N; §5.4). The ninth comparison method (direct LLM-on-reviews validation reference) covers 31 evaluable units and is not a release candidate (see Limitations, *Method-palette completeness*).

A.1 Format

We release three artifacts, namely per-book predictions organized by winning method (one file per method, indexed by book identifier); a per-attribute release metadata table for the 89 validated winners with winning method, bootstrap CIs, popularity-conditioned incremental ΔR^2 , lexical-anchorability score, popularity-leaning bias-audit flag, and tier; and a full-panel per-attribute table covering all 200 in-scope units so that Tier C units are documented alongside their per-attribute statistics.

A.2 Companion LLM-direct probe (Qwen 2.5-32B, ~10K books)

Alongside the tiered release, we ship a companion LLM-direct annotation set produced by prompting Qwen 2.5-32B-Instruct (AWQ) directly on sampled reviews. We first screen all 228 attributes in the Qwen screening panel against HUMAN-READER ground truth on a frozen 287-book validation set. The 25 attributes clearing partial- $r \geq 0.30$ are then deployed to two Goodreads subsamples: a top-reads 5K covering the most-reviewed titles, and a random 5K drawn uniformly from the 2,018,754-book review pool. The

samples overlap by ten books and yield 9,990 unique books; the shipped predictions parquet includes a per-book `source_subset` column (`topreads_5k / random_5k / both`) so downstream users can reproduce the popularity-vs-random split directly. A further 60 attributes clear $0.15 \leq \rho < 0.30$ in validation, while 143 do not; both groups are documented but not deployed. The probe is a companion cost-tier reference to the main tiered release, not a replacement, since a 9,990-book annotation does not span the 1.7M audit corpus. The Qwen deployment gate is stated as a partial Spearman correlation rather than the main release’s paired-nested-Ridge $\Delta R^2 > 0.05$ because Qwen is a zero-shot per-book prediction with no train/test analogue, and the 287-book frozen validation set is more naturally read as a partial correlation against ground truth than as a trained-model R^2 ; partial- $r \geq 0.30$ corresponds to popularity-conditioned variance-explained of roughly 0.09, which is on the stringent side of the main gate’s $\Delta R^2 > 0.05$ threshold and not directly interchangeable with it.

A.3 Coverage indicators

To make the release’s intended-population scope directly implementable per book, we release a per-book coverage table (`out/release/book_coverage.parquet`) with per-book n_{reviews} (the Goodreads Book Graph review count) joinable to the per-book predictions. Users applying our recommended $n_{\text{reviews}} \geq 10$ filter (see Limitations) can do so directly. The training pipeline applies the historical language filter (§G, `REVIEW_FILTER_AUDIT.md`) to select English-language reviews before feature construction; users needing to reproduce the training-time English-review count per book can regenerate it deterministically from the shipped filter script. We ship the training-set n_{reviews} distribution alongside the coverage table so users can compare their target books against the training regime; predictions on books with per-book coverage far below the training distribution should be interpreted with caution.

A.4 Aggregation gap on per-choice-share targets

A subset of HUMANREADER survey items enter our target matrix only via the per-choice “share of readers picking option X for this book” aggregation of a checkbox or rank-order item, which pro-

duces near-zero-variance targets on rarely-picked options and so fails our predictability filters there. The per-category marginalized-identity indicators (race, ethnicity, disability status, religion, socioeconomic status, gender identity, language, sexual identity, citizenship status, geography, and an open *Other*) enter the survey via two checkboxes, one for the protagonist and one for the ensemble cast, with these eleven options each; a per-protagonist one-hot audit would be more appropriate but is not supported by our current pipeline, and of the 22 resulting units 16 are excluded by the predictability filters while the remaining 6 land in Tier C, so none surfaces with content-driven recovery in our release. The “rank-order the three elements that most influenced how you feel about the book” item splits into ten per-choice units. Four of them (*Something else*, *The emotional impact*, *The intellectual impact*, *The setting(s)*) clear the target gates but fail the degenerate-fit check and are excluded. Five (character, plot or pace, mood, narration, writing style) enter the audit and land in Tier C, and *The message or theme* passes as a Tier B winner under keyword densities. A per-protagonist one-hot audit for the identity items, and a top-1 aggregation for the rank-order item, are both left to future work.

Free-text write-ins in the registry. The 391 registry units are not all closed-form attributes. One Style item offers a free-text *Other* box, and each distinct string a reader typed there becomes its own unit, contributing 65 of the 391 — single words such as *Angsty* or *Bland*, many of them typed once. These are write-ins rather than instrument options, and every one of them fails the predictability filters: 0 enter the 200-attribute analysis panel and none is a winner. The 391 figure therefore describes the registry as expanded, not a count of curated closed-form attributes, and the audited panel is unaffected.

Withheld constant-refit winners. Two Tier B attributes are withheld from the released per-book predictions: the *Epic* mood option won by NMF topics, and the *No* option of the end-state reversal-of-fortune item won by sentence embeddings. In each case the canonically selected Lasso penalty sets every feature coefficient to zero — all 50 topic components and all 384 pooled embedding dimensions respectively — so the refit predicts a single training-set intercept (0.0835 and 0.3662) for every book. Both passed validation on the Spearman route with negative held-out R^2 . A constant column carries no information and would silently

Method	25th	50th	75th	95th	Cat.
Keyword densities	−0.32	−0.11	−0.03	0.10	6%
TF-IDF	−0.41	−0.25	−0.13	−0.01	9%
Sentence embeddings	−0.43	−0.29	−0.16	0.06	3%
Stylistic/readability	−0.06	−0.03	−0.01	0.04	0%
GoEmotions	−0.13	−0.08	−0.02	0.10	0%
ABSA	−0.12	−0.07	−0.03	0.06	0%
NMF topics	−23	−15	−10	−5.63	100%
Meta + ratings	−0.18	−0.11	−0.06	0.01	0%

Table 1: Per-unit popularity-conditioned incremental ΔR^2 percentiles across evaluable attributes for the eight canonical release-panel methods. Cat. is the percentage of method-attribute cells with $\Delta R^2 < -1$. Cells whose combined fit is numerically degenerate are screened out, as in Figure 3 and the diagnostics of §5.1.

propagate into any downstream use, and the predictability filters of §4.4 do not catch it, since they screen the held-out fit rather than the refit used for inference. The attribute’s per-method test statistics are released as for any Tier C attribute, and the release metadata flags it.

B Popularity-Content Confound Distributions by Method

Figure 3 panels the per-attribute distributions underlying §5.1: raw held-out R^2 (panel a) shows every family with a negative median and only LLM-curated keyword densities reaching a positive 95th percentile, and popularity-conditioned incremental ΔR^2 (panel b) shows keyword densities and GoEmotions with the longest positive tails and every other family concentrating at or below zero; per-family counts are reported in §5.1.

C Per-Threshold Winner Staircase

Figure 4 presents two complementary views of the per-attribute winner count. Panel (a) traces raw threshold-exceedance counts as the R^2 threshold varies from 0 to 0.30 (best method per attribute, pre-multiplicity), illustrating how few attributes clear high thresholds. Panel (b) is the post-BH-FDR per-method winner breakdown that composes the 89-winner total: sentence embeddings 25, GoEmotions 27, keyword densities 15, ABSA 11, NMF topics 5, TF-IDF 4, and review meta plus ratings 2; stylistic-readability wins 0.

D Per-Category Pass Rates and Inter-Reader Agreement

Figure 5 presents the per-category pass rates reported in the body against per-category inter-reader

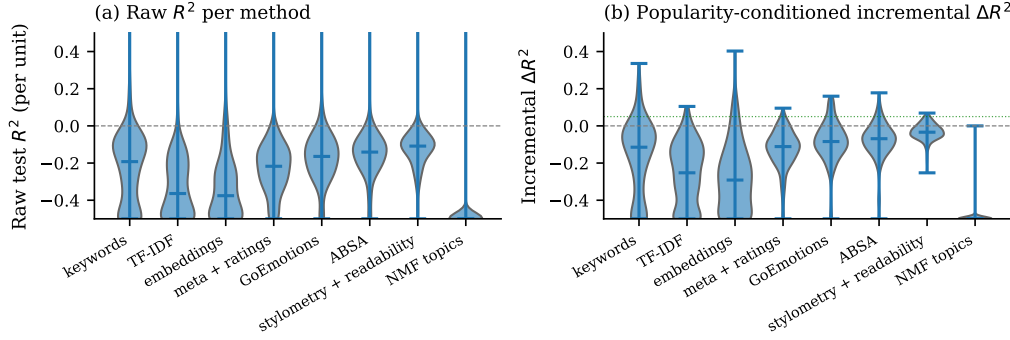


Figure 3: The popularity-content confound dominates the cost-tier comparison. **(a)** Distribution of raw held-out test R^2 per attribute, by method. Every family has a negative median; LLM-curated keyword densities hold the only positive 95th percentile, stylistic-readability features have the median closest to zero, and TF-IDF, sentence embeddings and NMF topics have the heaviest negative tails from overfitting. Keyword densities combine that positive upper tail with a left tail heavier than the embeddings’, so the family is high-variance rather than uniformly better. **(b)** Held-out incremental ΔR^2 from adding the feature family to a $\log(1 + n_{\text{reviews}})$ baseline on the same book splits. Keyword densities and GoEmotions have the longest positive tails and stylistic-readability the median closest to zero, while sentence-embedding, TF-IDF and topic methods incur the largest overfitting penalty. Cells whose combined fit is numerically degenerate are screened out, as in Table 1; bottom tails are clipped at -0.5 for visual range, and the topic method’s distribution falls far below it.

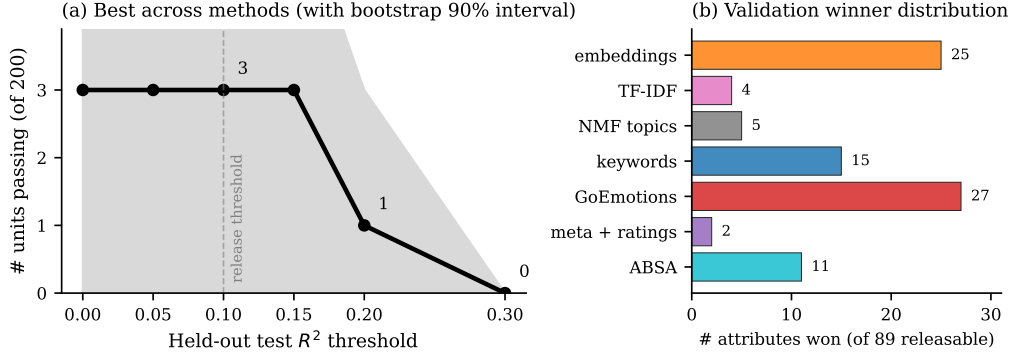


Figure 4: Panel (a): per-unit counts at each held-out R^2 threshold (best across methods by validation R^2). Counts are after BH-FDR adjustment; 3 units pass via the R^2 route. The dashed line marks $R^2 > 0.10$. Panel (b): per-method winner counts after BH-FDR adjustment (the 89-winner total).

agreement (IAA) from the companion dataset paper. Across the seven categories the IAA–pass-rate rank correlation is $\rho = -0.36$ ($p = 0.43$); at $n = 7$ categories the rank correlation carries no information about an IAA–pass-rate relationship.

Residualized companion. The nested statistic of §4.3 is not bounded above by the method’s own raw R^2 , because the popularity-only model has negative held-out R^2 on 98% of evaluable attributes and the paired difference adds that failure back in. We therefore also report a two-stage residualized variant: within each training fold we project $\log(1 + n_{\text{reviews}})$ out of the target and out of every feature column using training-fold coefficients only, refit the method on the residualized data, and score held-out R^2 on the residualized target. Because the popularity-only model then explains essentially none of that target by construction, the resulting value is the incremental content value

directly, and it is bounded by what the method explains of the residual. The residualized variant is correspondingly stricter: 31 attributes clear $\Delta R^2 > 0.01$ against 105 under the nested statistic, and 8 of the 12 Tier A attributes clear the 0.05 gate under residualization. We keep the nested statistic as the release gate because it stays on the raw target’s R^2 scale and adds one fewer fitted stage, and report the residualized variant as the conservative reading. An analogous control is applied as a partial correlation to the direct-LLM comparisons (Appendices Q and R), but it is not the same covariate: both probes condition on the log count of reviews actually shown to the model, which our sampling caps at 20 per book, whereas the release panel conditions on the log of the full Goodreads review count. The probe adjustments therefore remove less popularity variation than the panel’s, and the cost tiers are not adjusted on a common scale.

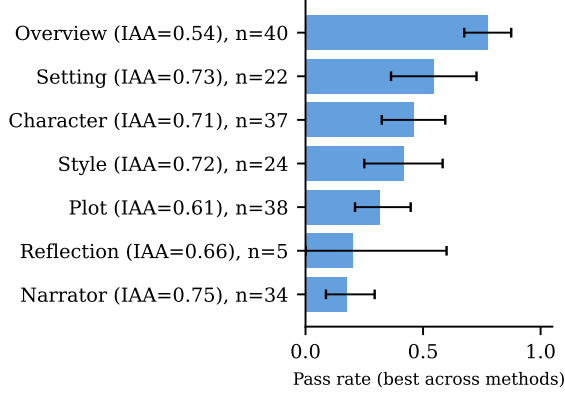


Figure 5: Per-category fraction of (concept, choice) attributes passing the predetermined validation threshold (best across methods), sorted descending. Bootstrap 90% intervals over units (5th/95th percentile). Parenthesized values are the per-category inter-reader agreement (IAA) from the dataset paper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

E Per-Question Variance Ceilings

E.1 Definition

Each per-book target y_b is a sample mean over n_b reader responses. The sampling variance of this estimator is $\sigma_{\text{within}}^2/n_b$, where σ_{within}^2 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

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

which we estimate by substituting $\hat{\mu}_b = y_b$ in the denominator and computing σ_{within}^2 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 the per-book-unit reader count has median 2 and 95th percentile 4, making per-book within-noise estimates too noisy to use as denominators. The ceiling should be interpreted as a population-mean upper bound rather than a per-book bound and may understate the true ceiling for polarizing books.

E.2 Distribution

Across the 391 attributes the variance ceiling has mean 0.47, median 0.44, with the 25th and 75th percentiles at 0.23 and 0.66. 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 2. Per-modeling-unit values are included in the replication package.

Category	Mean	Median	n
Overview	0.46	0.48	54
Plot	0.38	0.37	52
Character	0.50	0.48	86
Narrator	0.81	1.00	52
Setting	0.48	0.52	46
Style	0.32	0.13	96
Reflection	0.40	0.42	5

Table 2: 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.

E.3 Use in the validation threshold

The predetermined predictability filters (Section 4.4) require ceiling > 0.30 before a unit enters the analysis panel. Across the 391 attributes, 191 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; see Appendix T for the full causal-vs-descriptive breakdown). The remaining 200 units form the in-scope analysis panel.

F Keyword Bundle Generation

F.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 a pair of per-book DNF features, a keyword density and a review-coverage density, 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.

F.2 Validation

Before feature extraction, each frozen bundle is lower-cased and normalized with the same tokenizer and preprocessor used on reviews, then re-joined, so a multi-word key becomes exactly the token sequence the vectorizer can emit; empty and within-bundle duplicate entries are removed, as are terms in the global DNF lexicon. The resulting 465 polarity-specific bundles contain 24–30 keywords (median 30). Prompt templates and full keyword bundles are included in the replication package.

F.3 Feature computation

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

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

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 polarity features per (book, unit) cell (ρ_{kw} and ρ_{rev} for both low-end and high-end bundles) and non-monotonic units two, and every cell additionally includes the two global DNF densities, for six and four features in all.

G English-Language Heuristic on Goodreads Reviews

We apply the historical preprocessing heuristic used to construct the analysis corpus from the 15,739,967 reviews in the 2017 Goodreads Book Graph. We report the implemented rule rather than interpreting it as a validated language classifier.

Implemented rule. We retain a review when at least one token from the fixed English indicator list {the, a, and, is, it, in, i, this, that, was, for, of, with, my, book, read, but, to, so, on, not, you} appears among its first six whitespace-separated tokens. This inexpensive rule removes empty reviews, many pure-tag or emoji-only records, and many non-English records, but it is only a heuristic: it can reject English reviews lacking an early indicator and retain non-English or code-switched reviews containing one. No language-detector fallback is in effect: the indicator list above is the whole of the implemented rule, and the replication package records it explicitly so the released results remain reproducible.

The heuristic retains exactly 12,996,191 reviews across 1,707,435 books, rejecting 2,743,776 records (17.4%): 6,938 are empty or non-string texts and 2,736,838 are nonempty texts lacking an early indicator. No JSON line is malformed. We chunk the retained corpus into 44 CSV files for streaming-friendly downstream processing. These exact raw, retained, rejected, book, and reason counts are regenerated in `data_ucsd_chunks_v2/filter_manifest.csv` and `out/corpus_counts.csv`.

As a corpus-definition sensitivity, we separately apply `fast_langdetect` to every nonempty review, retain top-ranked English predictions at confidence at least 0.80, and never mix those outputs with the historical matrices. This version retains 13,147,905 reviews and rejects 2,592,062 (16.5%). The similar net retention masks 1,702,862 changed decisions: 927,288 historical rejections are retained by the detector, while 775,574 historical inclusions are rejected. These changed decisions are classifier disagreements rather than labeled errors because no human language gold standard is available. We rebuild the complete analysis under the detector-based definition in an isolated root and report its result sensitivity below.

The detector-based rebuild leaves the aggregate counts close but not the attribute-level membership. The historical corpus yields 200 in-scope attributes, 89 release-pass attributes, and 12 Tier A attributes, versus 200, 84, and 14 under the corrected definition. The two panels share 58 release-pass attributes and 6 Tier A attributes. So the totals move little while membership moves a lot: barely half of each panel’s Tier A set is shared, and the shared release-pass set is smaller than either panel. The corpus definition therefore constrains which attributes are recoverable more than how many; the package ships the complete paired comparison rather than treating either heuristic as labeled ground truth.

H Feature-Construction Details

TF-IDF family. The vectorizer is fit on a pooled sample of 8,000 reviews per corpus shard, restricted to reviews longer than 30 characters, with `min_df=20`, `max_df=0.95`, unigrams and bigrams, and a top-10,000 vocabulary. The vectorizer sets `use_idf=False` with sublinear term frequencies, so the representation is a sublinear term-frequency vector and not an IDF-weighted one. The 10,000-term matrix is reduced to the top 2,000 terms by global term mass before per-book mean-pooling and L_2 normalization, and the trainer then retains the 200 dimensions with greatest corpus-wide variance. IDF weighting is therefore untested here.

NMF topics. Fit with $K = 50$ components on the reduced 2,000-term matrix above, not on the full 10,000-term vocabulary.

GoEmotions. The checkpoint is a community fine-tune of RoBERTa on the GoEmotions label set. We score at most 50 reviews per book, each truncated to 256 tokens, and mean-pool the 28 logits at the book level. The embedding family uses a larger cap of 200 reviews per book, so the two families are given different amounts of text per book.

Bootstrap depth. Intervals resample the per-seed test values, so they draw on ten values for most methods and five for the full-corpus TF-IDF and NMF topic runs (Appendix I); the attainable p -value grid is correspondingly coarser for those two.

I Bootstrap Intervals

Per (attribute, method) we observe S per-seed test R^2 values $\{r_1, \dots, r_S\}$, 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 $n = 194$ training books per unit. Per-book residual bootstrapping would underestimate the seed-split variance.

J 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 $\{p_1, \dots, p_K\}$. We sort them ascending: $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(K)}$. 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.

K 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, physical-book) pairs with pooled $n_{\text{readers}} \geq 3$. Across the 200 in-scope units, the median number of mapped training books is 86; 199 retain at least 20 and 199 retain at least 50. Under the constant within-book noise model, averaging three rather than two readers reduces sampling variance by one third. The package ships per-unit book counts, response counts, target variance, and target range under both thresholds in `reader_count_sensitivity.csv`; full model re-validation on the smaller subset remains future work.

L Robustness Panel Across Gates

Table 3 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-conditioned incremental ΔR^2 threshold that defines Tier A. At the default settings there are 89 winners and 12 Tier A attributes;

keyword densities, sentence embeddings, ABSA, and GoEmotions account for the Tier A winners.

Variation	Setting	Winners (Tier A)
R^2 threshold	> 0.05	89 (12)
	> 0.10	89 (12)
	> 0.15	89 (12)
	> 0.20	89 (12)
Spearman threshold	> 0.10	96 (12)
	> 0.15	89 (12)
	> 0.20	66 (11)
	> 0.25	48 (11)
BH-FDR q	0.01	78 (12)
	0.05	89 (12)
	0.10	96 (12)
Tier A ΔR^2	> 0.005	89 (14)
	> 0.010	89 (14)
	> 0.030	89 (13)
	> 0.050	89 (12)
	> 0.100	89 (7)

Table 3: Robustness of the 89-winner and 12-Tier-A default counts across alternative validation and popularity-conditioned incremental- ΔR^2 thresholds.

M Lexical Anchorability Scatter

Figure 6 plots per-unit held-out test R^2 against two operationalizations of lexical anchorability across the 200 in-scope attributes with a computed score, coloring points by release-eligibility: (a) the semantic-neighborhood variant, which counts per-book review coverage of the LLM-curated keyword bundle, and (b) a literal-surface-form variant restricted to morphological variants of the response-option text. The semantic-neighborhood variant tracks release-eligibility (Spearman $\rho = 0.20$ on the binary pass indicator, $p = 0.00442$; and $\rho = 0.15$ on the Spearman-route effect size $|\rho_{\text{test}}|$, $p = 0.0371$); the literal variant does not detectably track release-eligibility ($\rho = -0.03$ on binary pass, $p = 0.63$; and $\rho = 0.07$ on $|\rho_{\text{test}}|$, $p = 0.34$), consistent with the mechanism operating through the semantic neighborhood rather than through direct lexical repetition of the option text alone. Grouping the same 200 attributes into anchorability quartiles, 50% of the top quartile clears validation against 22% of the bottom, a ratio of 2.3; the relation is not monotone across the middle quartiles (46% and 60%).

N 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 config-

uration (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 per polarity bundle, so 60 for the monotonic items that have a low and a high bundle and 30 otherwise. Across units, the mean Jaccard similarity between v1 and v2 bundles is 0.149 (median 0.132, range 0–0.429). Tier A units exhibit slightly higher stability (mean Jaccard 0.198). We then rebuild keyword densities on the training-mapped subset under v2 and re-fit. Of the 15 v1 keyword-density winners, 11 remain among attributes passing the relaxed threshold ($R^2 > 0.10$ or $|\rho| > 0.15$ pre-multiplicity) under v2; only 1 retain test $R^2 > 0.05$. The median signed v2-minus-v1 R^2 change is -0.026 , but the median absolute change is 0.148; individual magnitudes are therefore prompt-sensitive despite little median directional shift.

O Permutation Baseline

For each non-embedding validated winner we re-fit the winning feature method and model class on 100 permutations of book-level target labels, repeating the canonical validation-fold regularization selection and averaging R^2 across the same method-specific random 80/10/10 split seeds used for the observed score (five for TF-IDF and NMF topics, ten otherwise). The null distribution of test R^2 per winner spans the range produced by chance label alignments under that 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 nonzero p -value is 0.01; a reported $p = 0$ means no permutation beat the observed.

Of the 64 winners covered, 23 exceed their permutation 95th percentile. Exceedance rates differ sharply by method: ABSA 6/11 and keyword densities 8/15 are highest, meta-plus 1/2 and NMF topics 2/5 sit between on very small denominators, TF-IDF reaches 1/4, and GoEmotions is the clear outlier at 5/27. All 9 Tier A winners in the pass exceed their null at $p \leq 0.01$.

P 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-

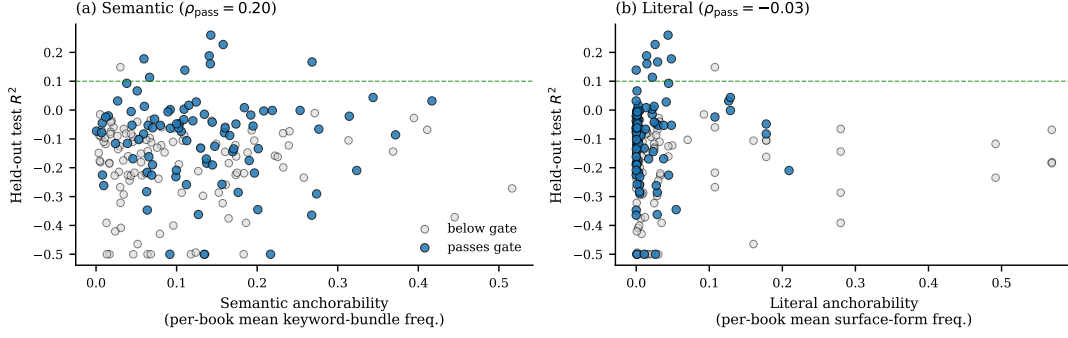


Figure 6: Lexical anchorability operationalizes the recoverability mechanism. **(a)** Per-unit held-out test R^2 versus semantic anchorability (per-book mean fraction of reviews containing any keyword in the LLM-curated semantic neighborhood). **(b)** Per-unit held-out test R^2 versus literal anchorability (per-book mean fraction of reviews containing morphological variants of the response-option text). Semantic anchorability predicts release-eligibility at $\rho = 0.20$ on the binary pass indicator ($p = 0.00442$); the literal surface form does not detectably predict it at $\rho = -0.03$ ($p = 0.63$). Blue points pass the validation threshold, grey points fail it. Outliers $|R^2| > 0.5$ clipped.

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 $\text{stratum_range_z} > 0.5$ and the popularity-conditioned incremental ΔR^2 over the $\log(1 + n_{\text{reviews}})$ baseline is below 0.05 (Figure 7). Of the 87 auditable (method, attribute) winner cells, 27 are flagged, concentrated in ABSA (5/11) and GoEmotions (22/27) winners. We audit every released attribute exactly once, under its own winning method, so the cell count equals the number of released prediction columns. No keyword-density, sentence-embedding, review-meta, TF-IDF or NMF-topic winner is flagged; every flagged cell belongs to one of the two survey-corpus-fit methods, which the coverage caveat below also limits. Per-attribute audit values are released alongside the dataset.

Coverage caveat. ABSA and GoEmotions winners cover only 286 matched books. Because the strata are full-corpus rather than within-subset quintiles, their cells range from 8 to 228 books, versus 339,976–341,487 books for corpus-scale methods. The audit flag does not adjust for this coverage asymmetry, so the ABSA/GoEmotions concentration is consistent with either a real popularity-driven signal or with sampling noise from sparse and highly imbalanced cells.

Q LLM-on-Reviews Per-Concept Recovery

This appendix reports the LLM-on-reviews probe introduced as a validation reference in §4.1, as a complementary direct-extraction comparison at a

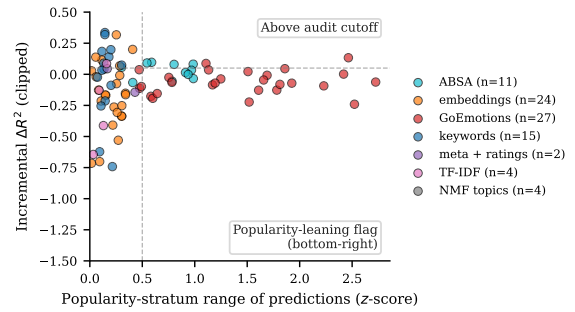


Figure 7: Per-attribute popularity-stratum bias scatter. Each point is a (method, attribute) pair; the x -axis is the across-quintile prediction-mean range (in z -units), and the y -axis is the popularity-conditioned incremental ΔR^2 over $\log(1 + n_{\text{reviews}})$. Points above the dashed $\Delta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

different cost tier from the release-panel methods. Its role is to test whether the release-panel pass rate reflects the scalable method palette rather than an information ceiling on review text; it is not a release candidate, since a per-book LLM call across 1.7M books is prohibitively expensive at release scale.

Setup. On a training subset of 163 books — those with retained reviews among a uniform random sample of 200 mapped training books (§4.1) — we sample up to 20 reviews per book and prompt GPT-5 with the survey question wording and the sampled reviews, asking directly for the item response. After joining the cached predictions to the canonical, reader-weighted physical-book targets, this yields 31 registered (concept, choice) modeling units under the same grain as the main pipeline (§3.1). Four ad hoc free-text “Other” choices are

not survey-registry units and are excluded from this evaluation.

Evaluation. We compute per-unit Spearman between the LLM-on-reviews silver label and the HUMANREADER per-book target, and the corresponding partial correlation controlling for $\log(1 + n_{\text{sampled}})$, the count of reviews actually shown to the model. All 31 canonical units have at least $n \geq 20$ training books and non-degenerate ground-truth variation.

Results. Of the 31 units, 25 reach partial- $r \geq 0.20$ and 16 reach partial- $r \geq 0.30$. The median raw-to-partial drop is 0.002, indicating that the volume-only popularity control does not appreciably reduce the recovered signal for the LLM-on-reviews probe (in contrast to the release-panel methods, which lose substantial signal to the same control). Full per-unit values are included in the replication package.

Interpretation. The LLM-on-reviews probe recovers signal on 25 of 31 evaluable units at partial- $r \geq 0.20$, so reviews contain more attribute signal than the scalable-method palette recovers under the popularity conditioning and release gates (12 of 200 at Tier A on popularity-conditioned incremental $\Delta R^2 > 0.05$). The gap is consistent with a cost-tier limit on the scalable methods rather than an information-content ceiling on review text.

R Qwen 2.5 Subsample Probe as an Open-Weight Companion Release

The Appendix Q GPT-5 probe indicates that the cheap-tier pass rate reflects the scalable method palette rather than a ceiling on review text, but per-book GPT-5 inference does not scale to the 1.7M-book audit corpus. As a companion release at a middle cost tier, we prompt Qwen 2.5-32B-Instruct (AWQ quantization, served locally on a B200 GPU) directly on sampled reviews. We first score all 228 registry (concept, choice) attributes in the Qwen screening panel, a superset of the 200 in-scope for the main release, on the frozen 287-book validation set, then deploy only the 25 attributes clearing the validation threshold to a two-part 10K-book Goodreads subsample: (i) a top-reads 5K, defined as the top 5,000 Goodreads books by the count of nonmissing reviews with at least 30 characters in our separate review-sampling corpus, and (ii) a random 5K uniformly drawn from the same 2,018,754-book pool. This pool is not language-filtered. The two subsamples overlap by ten books

and therefore yield 9,990 unique scored books.

Setup. For each book we sample up to 20 reviews (the same sampling used in Appendix Q) and prompt Qwen 2.5 with an initial batched user message containing the sampled review text and all phase-specific attribute questions, asking for a JSON response with one numeric value per attribute. The validation prompt contains all 228 questions; deployment prompts contain only the 25 frozen validated questions. We use temperature 0, a maximum of 4,096 output tokens, structured JSON parsing, and bounded focused retries for omitted or unparseable keys; a book is excluded if its requested attribute rectangle remains incomplete.

Evaluation against ground truth. The frozen validation set contains 287 books from our HUMANREADER-matched training set (i.e., books appearing in `book_id_map.csv` with a non-null UCSD book id and at least one scorable per-book target). For each attribute we compute Spearman partial- r between the Qwen 2.5 prediction and the HUMANREADER per-book target on its available validation books, controlling for $\log(1 + n_{\text{sampled}})$ as the popularity signal actually available to Qwen (capped at 20 sampled reviews per book, as in the Appendix Q GPT-5 probe; the release panel instead conditions on the full Goodreads count). We label each attribute *validated* at partial- $r \geq 0.30$, *weak* at $0.15 \leq \text{partial-}r < 0.30$, and *not-validated* below 0.15.

Results. Of the 228 attributes, 25 meet the validated threshold and a further 60 meet the weak threshold, for 85 attributes with directionally aligned partial- $r \geq 0.15$. The remaining 143 do not clear. Median partial- r across the 228 audited is 0.106. Keyword-density and sentence-embedding winners of the main tiered release are represented in the validated set, while most ABSA and GoEmotions winners remain below the threshold.

Interpretation. The probe is a companion cost-tier reference to the main release, not a replacement, since 9,990 books does not scale to the 1.7M audit corpus. Users of the companion release should treat the 25 deployed attributes as directly usable. The 60 weak and 143 not-validated attributes are documented in the full validation table but are not deployed in the companion prediction matrix.

S Datasheet for the Released Dataset

Following the framework of Geburu 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 89 validated attributes (12 Tier A content-driven plus 77 Tier B popularity-informed). Of the 51 attributes won by full-corpus methods, 26 (from keyword densities, TF-IDF, review meta plus ratings, or NMF topics) cover all 1,707,435 books, though one of the 26 — the *Epic* mood option won by NMF topics — is withheld as a constant refit (Appendix A), and one of the 25 sentence-embedding winners is withheld for the same reason, so per-book predictions ship for 87 attributes and per-method statistics for all 89, and the released sentence-embedding columns cover 1,704,323 books, a 3,112-book gap from feature-side dropouts: every omitted book has one or two reviews whose text did not yield a usable sentence embedding. 38 attributes won by GoEmotions or ABSA cover 286 matched books only. An additional 111 in-scope attributes (Tier C) will have per-method test statistics released alongside the dataset without per-book predictions. A companion LLM-direct release covers 9,990 books for the 25 of 228 screened attributes directly validated against HUMANREADER ground truth at partial- $r \geq 0.30$ on a frozen 287-book screen; see Appendix R.

Collection process. Ground-truth targets from the HUMANREADER corpus (2024–2025 reader sessions); review text from the 2017 Goodreads Book Graph (Wan and McAuley, 2018); predictions from per-attribute regression models trained 2026.

Preprocessing. Historical first-six-token English-indicator heuristic (Appendix G); title-normalized HumanReader–Goodreads match; predictability filters exclude 191 of 391 attributes. Notable exclusions include 16 of the 22 marginalized-identity checkbox units (the protagonist and ensemble-cast checkboxes, 11 options each) and 6 of the 18 per-time-point character-arc grid slots, excluded for lack of target signal across our 384-book sample and retained in the underlying HumanReader corpus. The narration-switch count (Narrator Q2) and the remaining grid slots are in scope; 11 of the grid slots are released as Tier B. Appendix T lists all 191 excluded items with their per-gate values.

Uses. Recommended for attribute-conditioned

book filtering and computational literary analysis at scale for the 51 attributes with full-corpus coverage (won by sentence embeddings, keyword densities, NMF topics, TF-IDF, or review meta plus ratings); the 38 attributes won by GoEmotions or ABSA cover only 286 matched books. Predictions are clipped to each attribute’s natural range; since all 89 winners pass via the Spearman validation route while only 3 additionally pass via the R^2 route, natural-scale predictions are validated on rank order more broadly than on absolute-scale calibration (the 3 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 27 of 87 auditable releasable-winner cells (31.0%) as popularity-leaning, concentrated in ABSA and GoEmotions winners; the release metadata includes per-attribute incremental ΔR^2 , popularity-leaning, and audibility fields. The two limited methods score 286 matched books, with full-corpus-quintile cells ranging from 8 to 228 books (Appendix P).

Distribution and maintenance. The dataset is released at <https://osf.io/3ncuz/> under a license matching the upstream HumanReader and Goodreads Book Graph terms, and corrections and additions will be issued as versioned updates. The released models are static after publication; subsequent Goodreads snapshots would require re-training the per-attribute models, which is out of scope for this release.

T Full Exclusion Accounting

The 191 attributes excluded by the predictability filter (§4.4, and summarized in the *Preprocessing* entry of Appendix S) are listed exhaustively in Table 4 with their per-gate values. The causal split is 90 units below the operational 16-book modeling floor and 101 units that met the sample floor but were excluded by the remaining gates or downstream fit checks. The mutually exclusive table reasons make that accounting explicit. Among the sample-limited units, 18 fail only the sample gate, 68 also fail the ceiling gate, 3 also fail variance and range, and 1 fails all four gates. Among the evaluated units, 45 reach model evaluation but fail downstream fit checks, 10 fail only variance, 5 fail variance and ceiling, 40 fail only ceiling, and 1 fails

variance, range, and ceiling. All excluded items remain in the underlying HUMANREADER corpus.

Winsorized numeric target. Across 807 raw responses, the “how many named side characters” Character-numeric item has a maximum of 10,000,000, a p99 of 50, and a median of 10; the p99 rule clips 5 responses. Per-reader numeric answers are winsorized at the p99 before both target construction and variance-ceiling estimation. The unit is trained and evaluated by the same pipeline and release gates as every other attribute.

Duplicate canonical books. 2 pairs of survey books map to the same Goodreads book after title normalization: *The Beautiful and Damned* (canonical ids 100050 and 100152; UCSD id 4708) and *A Great Reckoning* (canonical ids 3788 and 100159; UCSD id 30547482). Each pair reflects distinct reader-response sets rather than duplicated rows. Before splitting, the replication pipeline pools each pair into one physical-book target by weighting the two survey means by their respective reader counts and summing those counts. The resulting analysis has one target and one review-feature row per Goodreads book, so a physical title cannot cross the train, validation, and test folds, and neither response set is discarded.

Table 4: Full accounting of the 191 attributes excluded by the predictability filter (§4.4). “Fit” denotes a missing or numerically degenerate best-across-methods fit after the three target-level gates pass; “nan” denotes a unit with no computable target variance, such as a free-text or never-selected option. Question and choice text is truncated upstream; ... marks the cut. Two *Setting* options were additionally split at an internal comma upstream and appear here as a truncated stem and its orphaned tail.

Category	Question	Choice	Var.	Range	Ceiling	Failed gate(s)
Character	Did any members of the ensemble cast belong to any historically marginalized groups - i.e., groups systematically...	Citizenship status	0.002	0.500	1.000	variance
Character	Did any members of the ensemble cast belong to any historically marginalized groups - i.e., groups systematically...	Disability status	0.006	1.000	1.000	fit
Character	Did any members of the ensemble cast belong to any historically marginalized groups - i.e., groups systematically...	Ethnicity	0.004	0.667	1.000	variance
Character	Did any members of the ensemble cast belong to any historically marginalized groups - i.e., groups systematically...	Gender identity	0.007	0.500	1.000	fit
Character	Did any members of the ensemble cast belong to any historically marginalized groups - i.e., groups systematically...	Geography	0.003	0.500	1.000	variance
Character	Did any members of the ensemble cast belong to any historically marginalized groups - i.e., groups systematically...	Language	0.003	0.500	1.000	variance
Character	Did any members of the ensemble cast belong to any historically marginalized groups - i.e., groups systematically...	Other (please list):	0.002	0.500	1.000	variance
Character	Did any members of the ensemble cast belong to any historically marginalized groups - i.e., groups systematically...	Race	0.007	0.667	1.000	fit
Character	Did any members of the ensemble cast belong to any historically marginalized groups - i.e., groups systematically...	Religion	0.001	0.333	1.000	variance
Character	Did any members of the ensemble cast belong to any historically marginalized groups - i.e., groups systematically...	Sexual identity	0.002	0.500	1.000	variance
Character	Did any members of the ensemble cast belong to any historically marginalized groups - i.e., groups systematically...	Socioeconomic status	0.010	0.667	1.000	fit
Character	Does this protagonist belong to any historically marginalized groups - i.e., groups systematically denied access...	Disability status	0.014	1.000	1.000	fit
Character	Does this protagonist belong to any historically marginalized groups - i.e., groups systematically denied access...	Gender identity	0.040	1.000	1.000	fit
Character	Does this protagonist belong to any historically marginalized groups - i.e., groups systematically denied access...	Language	0.010	1.000	1.000	fit
Character	Does this protagonist belong to any historically marginalized groups - i.e., groups systematically denied access...	Religion	0.009	1.000	1.000	fit
Character	Does this protagonist belong to any historically marginalized groups - i.e., groups systematically denied access...	Sexual identity	0.011	0.750	1.000	fit
Character	How proactive was this protagonist? By proactive, we mean motivated and taking steps to reach their goals...	Beginning	0.000	0.000	1.000	sample, variance, range
Character	How proactive was this protagonist? By proactive, we mean motivated and taking steps to reach their goals...	Neither proactive nor passive	0.119	1.000	0.000	sample, ceiling
Character	How proactive was this protagonist? By proactive, we mean motivated and taking steps to reach their goals...	Somewhat passive	0.000	0.000	1.000	sample, variance, range
Character	How proactive was this protagonist? By proactive, we mean motivated and taking steps to reach their goals...	Somewhat proactive	0.037	0.500	0.393	sample
Character	How proactive was this protagonist? By proactive, we mean motivated and taking steps to reach their goals...	Very passive	0.058	0.500	0.000	sample, ceiling

(continued on next page)

(continued from previous page)

Category	Question	Choice	Var.	Range	Ceiling	Failed gate(s)
Character	How proactive was this protagonist? By proactive, we mean motivated and taking steps to reach their goals. . .	Very proactive	0.117	1.000	0.000	sample, ceiling
Character	If you checked any of the above, please specify:	(free-text)	nan	nan	0.000	sample, ceiling
Character	If you checked any of the above, please specify:	(free-text)	nan	nan	0.000	sample, ceiling
Character	If you checked any of the above, please specify:	(free-text)	nan	nan	0.000	sample, ceiling
Character	If you checked any of the above, please specify:	(free-text)	nan	nan	0.000	sample, ceiling
Character	If you checked any of the above, please specify:	(free-text)	nan	nan	0.000	sample, ceiling
Character	If you checked any of the above, please specify:	(free-text)	nan	nan	0.000	sample, ceiling
Character	If you checked any of the above, please specify:	(free-text)	nan	nan	0.000	sample, ceiling
Character	If you checked any of the above, please specify:	(free-text)	nan	nan	0.000	sample, ceiling
Character	In your opinion, how important were the characters to your experience of the book as a whole, relative to other. . .	<scalar>	0.401	3.500	0.221	ceiling
Character	On a scale of 1 (not at all) to 7 (extremely), how relatable did you find this protagonist?	<scalar>	nan	nan	0.000	sample, ceiling
Character	To what extent did the needs and motivations of the ensemble cast reinforce or contradict each other? For example. . .	Completely reinforcing	0.000	0.000	0.437	sample, variance, range
Character	To what extent did the needs and motivations of the ensemble cast reinforce or contradict each other? For example. . .	Equal parts reinforcing and contradicting	0.125	0.500	0.281	sample, ceiling
Character	To what extent did the needs and motivations of the ensemble cast reinforce or contradict each other? For example. . .	Mostly contradicting	0.000	0.000	0.000	sample, variance, range, ceiling
Character	To what extent did the needs and motivations of the ensemble cast reinforce or contradict each other? For example. . .	Mostly reinforcing	0.125	0.500	0.000	sample, ceiling
Character	What primary type of conflict, if any, did this protagonist face in the book? At a high level, conflict can take. . .	Different type of conflict (please describe)	0.014	0.500	0.359	fit
Character	What primary type of conflict, if any, did this protagonist face in the book? At a high level, conflict can take. . .	N/A - Protagonist did not face primary conflict	0.011	0.667	0.485	fit
Character	What primary type of conflict, if any, did this protagonist face in the book? At a high level, conflict can take. . .	Protagonist vs. technology	0.008	1.000	0.213	ceiling
Character	What secondary type(s) of conflict, if any, did this protagonist face in the book? For example, in the first Hunger. . .	Different type of conflict (please describe)	0.010	0.500	0.000	ceiling
Character	What secondary type(s) of conflict, if any, did this protagonist face in the book? For example, in the first Hunger. . .	Protagonist vs. fate/luck	0.045	1.000	0.287	ceiling
Character	What secondary type(s) of conflict, if any, did this protagonist face in the book? For example, in the first Hunger. . .	Protagonist vs. nature	0.024	0.667	0.183	ceiling
Character	What secondary type(s) of conflict, if any, did this protagonist face in the book? For example, in the first Hunger. . .	Protagonist vs. technology	0.007	0.667	0.204	ceiling

(continued on next page)

(continued from previous page)

Category	Question	Choice	Var.	Range	Ceiling	Failed gate(s)
Character	What secondary type(s) of conflict, if any, did this protagonist face in the book? [multi select] For example. . .	N/A - Protagonist did not face secondary conflict	nan	nan	0.000	sample, ceiling
Character	What secondary type(s) of conflict, if any, did this protagonist face in the book? [multi select] For example. . .	Protagonist vs. fate/luck	nan	nan	0.000	sample, ceiling
Character	What secondary type(s) of conflict, if any, did this protagonist face in the book? [multi select] For example. . .	Protagonist vs. other character(s)	nan	nan	0.000	sample, ceiling
Character	What secondary type(s) of conflict, if any, did this protagonist face in the book? [multi select] For example. . .	Protagonist vs. self	nan	nan	0.000	sample, ceiling
Character	What secondary type(s) of conflict, if any, did this protagonist face in the book? [multi select] For example. . .	Protagonist vs. society	nan	nan	0.000	sample, ceiling
Character	What secondary type(s) of conflict, if any, did this protagonist face in the book? [multi select] For example. . .	Protagonist vs. technology	nan	nan	0.000	sample, ceiling
Narrator	If third person narration was used, was it predominantly close or omniscient? In close third person, the narration. . .	Genuinely mixed both	0.052	1.000	0.192	ceiling
Narrator	In your opinion, how important was the narration and POV to your experience of the book as a whole, relative to other. . .	<scalar>	0.468	4.000	0.296	ceiling
Narrator	Was the book narrated primarily in past or present tense? Past tense narration sounds like the story has already. . .	Other (fill in):	0.021	1.000	0.177	ceiling
Narrator	What type of narration was predominantly used? First person narration uses I/WE and comes directly from the character. . .	Genuinely two mixed or more	0.036	1.000	0.173	ceiling
Narrator	What type of narration was predominantly used? First person narration uses I/WE and comes directly from the character. . .	Second person	0.002	0.500	0.000	variance, ceiling
Narrator	What word would you choose to describe this narrator's overall tone? Select all that apply from the alphabetized list. . .	Anxious	0.045	1.000	1.000	fit
Narrator	What word would you choose to describe this narrator's overall tone? Select all that apply from the alphabetized list. . .	Conniving	0.005	0.500	1.000	fit
Narrator	What word would you choose to describe this narrator's overall tone? Select all that apply from the alphabetized list. . .	Crass	0.005	0.500	1.000	fit
Narrator	What word would you choose to describe this narrator's overall tone? Select all that apply from the alphabetized list. . .	Depressed	0.024	1.000	1.000	fit
Narrator	What word would you choose to describe this narrator's overall tone? Select all that apply from the alphabetized list. . .	Devil-may-care	0.004	0.500	1.000	variance
Narrator	What word would you choose to describe this narrator's overall tone? Select all that apply from the alphabetized list. . .	Disingenuous	0.001	0.500	1.000	variance
Narrator	What word would you choose to describe this narrator's overall tone? Select all that apply from the alphabetized list. . .	Hypocritical	0.006	0.500	1.000	fit

(continued on next page)

(continued from previous page)

Category	Question	Choice	Var.	Range	Ceiling	Failed gate(s)
Narrator	What word would you choose to describe this narrator's overall tone? Select all that apply from the alphabetized list...	Irreverent	0.016	1.000	1.000	fit
Narrator	What word would you choose to describe this narrator's overall tone? Select all that apply from the alphabetized list...	Laid-back	0.009	0.667	1.000	fit
Narrator	What word would you choose to describe this narrator's overall tone? Select all that apply from the alphabetized list...	Overwrought	0.005	0.500	1.000	variance
Narrator	What word would you choose to describe this narrator's overall tone? Select all that apply from the alphabetized list...	Pessimistic	0.025	1.000	1.000	fit
Narrator	What word would you choose to describe this narrator's overall tone? Select all that apply from the alphabetized list...	Sarcastic	0.015	1.000	1.000	fit
Narrator	What word would you choose to describe this narrator's overall tone? Select all that apply from the alphabetized list...	Whiny	0.009	1.000	1.000	fit
Overview	How would you describe the mood - or, less formally, the "vibes" - of the book? Select all that apply...	Cozy	0.043	1.000	0.522	fit
Overview	How would you describe the mood - or, less formally, the "vibes" - of the book? Select all that apply...	Energetic	0.027	0.714	0.301	fit
Overview	How would you describe the mood - or, less formally, the "vibes" - of the book? Select all that apply...	Heartwarming	0.081	1.000	0.593	fit
Overview	How would you describe the mood - or, less formally, the "vibes" - of the book? Select all that apply...	Inspirational	0.036	1.000	0.197	ceiling
Overview	How would you describe the mood - or, less formally, the "vibes" - of the book? Select all that apply...	Optimistic	0.059	1.000	0.445	fit
Overview	How would you describe the mood - or, less formally, the "vibes" - of the book? Select all that apply...	Other:	0.020	0.500	0.418	fit
Overview	How would you describe the mood - or, less formally, the "vibes" - of the book? Select all that apply...	Quirky	0.063	1.000	0.486	fit
Overview	Please rank-order the three elements that most influenced how you feel about the book...	Something else	0.011	0.500	0.000	ceiling
Overview	Please rank-order the three elements that most influenced how you feel about the book...	The emotional impact	0.017	0.667	0.300	ceiling
Overview	Please rank-order the three elements that most influenced how you feel about the book...	The intellectual impact	0.011	0.667	0.000	ceiling
Overview	Please rank-order the three elements that most influenced how you feel about the book...	The message or theme	0.028	0.667	0.238	ceiling
Overview	Please rank-order the three elements that most influenced how you feel about the book...	The mood or vibe	0.029	0.667	0.210	ceiling
Overview	Please rank-order the three elements that most influenced how you feel about the book...	The narration	0.036	1.000	0.194	ceiling
Overview	Please rank-order the three elements that most influenced how you feel about the book...	The setting(s)	0.015	0.500	0.169	ceiling

(continued on next page)

(continued from previous page)

Category	Question	Choice	Var.	Range	Ceiling	Failed gate(s)
Plot	Did the ending feature a major emotional release, or other major expression of emotion, on the part of the main. . .	Somewhat	0.092	1.000	0.249	ceiling
Plot	Did the ending feature a major reversal-of-fortune for the main character(s)?	Somewhat	0.103	1.000	0.237	ceiling
Plot	Did you experience a notable "sag" in plot tension at any point in the book? Select all that apply. Please answer. . .	11-20%	0.051	1.000	0.282	ceiling
Plot	Did you experience a notable "sag" in plot tension at any point in the book? Select all that apply. Please answer. . .	31-40%	0.067	1.000	0.236	ceiling
Plot	Did you experience a notable "sag" in plot tension at any point in the book? Select all that apply. Please answer. . .	71-80%	0.045	1.000	0.207	ceiling
Plot	Did you experience a notable "sag" in plot tension at any point in the book? Select all that apply. Please answer. . .	Last 91-100% of the book	0.029	0.667	0.221	ceiling
Plot	How would you characterize its linearity in terms of time? Some books move in a straight sequence with events narrated. . .	<scalar>	0.344	2.500	0.703	fit
Plot	On a percent-of-book basis, please select the point in the novel where any major turning point(s) occurred.	11-20%	0.060	1.000	0.283	ceiling
Plot	On a percent-of-book basis, please select the point in the novel where any major turning point(s) occurred.	21-30%	0.068	1.000	0.205	ceiling
Plot	On a percent-of-book basis, please select the point in the novel where any major turning point(s) occurred.	61-70%	0.065	1.000	0.203	ceiling
Plot	On a percent-of-book basis, please select the point in the novel where any major turning point(s) occurred.	81-90%	0.094	1.000	0.278	ceiling
Plot	On a percent-of-book basis, please select the point in the novel where any major turning point(s) occurred.	Last 91-100% of the book	0.083	1.000	0.206	ceiling
Plot	Which story shape best describes this story's plot, if any?	Oedipus	0.057	1.000	0.257	ceiling
Plot	Which story shape best describes this story's plot, if any?	Rags to riches	0.053	1.000	0.365	fit
Setting	On a percent-of-book basis, at what point(s) was the book set here?	11-20%	0.031	1.000	0.568	fit
Setting	On a percent-of-book basis, at what point(s) was the book set here?	21-30%	0.031	1.000	0.568	fit
Setting	On a percent-of-book basis, at what point(s) was the book set here?	31-40%	0.035	1.000	0.583	fit
Setting	On a percent-of-book basis, at what point(s) was the book set here?	41-50%	0.037	1.000	0.610	fit
Setting	On a percent-of-book basis, at what point(s) was the book set here?	51-60%	0.029	1.000	0.517	fit
Setting	On a percent-of-book basis, at what point(s) was the book set here?	61-70%	0.034	1.000	0.528	fit
Setting	On a percent-of-book basis, at what point(s) was the book set here?	81-90%	0.052	1.000	0.627	fit
Setting	On a percent-of-book basis, at what point(s) was the book set here?	First 1-10% of the book	0.033	1.000	0.490	fit
Setting	On a percent-of-book basis, at what point(s) was the book set here? Please answer on a length basis - and check page. . .	11-20%	0.009	0.667	0.085	ceiling
Setting	On a percent-of-book basis, at what point(s) was the book set here? Please answer on a length basis - and check page. . .	21-30%	0.010	1.000	0.510	fit

(continued on next page)

(continued from previous page)

Category	Question	Choice	Var.	Range	Ceiling	Failed gate(s)
Setting	On a percent-of-book basis, at what point(s) was the book set here? Please answer on a length basis - and check page...	31-40%	0.003	0.500	0.000	variance, ceiling
Setting	On a percent-of-book basis, at what point(s) was the book set here? Please answer on a length basis - and check page...	41-50%	0.006	0.667	0.295	ceiling
Setting	On a percent-of-book basis, at what point(s) was the book set here? Please answer on a length basis - and check page...	51-60%	0.003	0.500	0.000	variance, ceiling
Setting	On a percent-of-book basis, at what point(s) was the book set here? Please answer on a length basis - and check page...	61-70%	0.004	0.500	0.000	variance, ceiling
Setting	On a percent-of-book basis, at what point(s) was the book set here? Please answer on a length basis - and check page...	71-80%	0.000	0.000	0.000	variance, range, ceiling
Setting	On a percent-of-book basis, at what point(s) was the book set here? Please answer on a length basis - and check page...	81-90%	0.001	0.333	0.000	variance, ceiling
Setting	On a percent-of-book basis, at what point(s) was the book set here? Please answer on a length basis - and check page...	First 1-10% of the book	0.020	1.000	0.442	fit
Setting	On a percent-of-book basis, at what point(s) was the book set here? Please answer on a length basis - and check page...	Last 91-100% of the book	0.005	0.500	0.000	ceiling
Setting	When does the book take place? Select all that apply.	In the distant future	0.044	1.000	0.798	fit
Setting	When does the book take place? Select all that apply.	In the near future	0.027	1.000	0.520	fit
Setting	Where does the book take place? Select all that apply.	Kashyyyk)	0.044	1.000	0.633	fit
Setting	Where does the book take place? Select all that apply.	Mars)	0.012	1.000	0.800	fit
Setting	Where does the book take place? Select all that apply.	Somewhere off Earth that is imaginary...	0.044	1.000	0.633	fit
Setting	Where does the book take place? Select all that apply.	Somewhere off Earth that really exists...	0.012	1.000	0.800	fit
Style	How realistic did you find the dialogue in the book?	<scalar>	1.223	5.333	0.268	ceiling
Style	In your opinion, how important was the overall writing style to your experience of the book as a whole, relative...	<scalar>	0.252	2.500	0.065	ceiling
Style	In your opinion, how was the amount of showing and telling throughout the book? - Telling	<scalar>	0.385	3.500	0.290	ceiling
Style	In your opinion, how was the quantity of dialogue throughout the book?	<scalar>	0.208	3.000	0.284	ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Angsty	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Archaic	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Archaic/dated	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Arrogant	nan	nan	1.000	sample

(continued on next page)

(continued from previous page)

Category	Question	Choice	Var.	Range	Ceiling	Failed gate(s)
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	As the plot progression happened mostly through dialogue. . .	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Bittersweet	nan	nan	1.000	sample
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Bland	nan	nan	0.134	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Boring	nan	nan	1.000	sample
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Casual	nan	nan	0.134	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Cinematic	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Clichéd	nan	nan	0.134	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Confusing	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Conversational	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Cutesy	nan	nan	0.134	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Dense	nan	nan	0.134	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Educational(there were lots of footnotes)	nan	nan	1.000	sample
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Exaggerated	nan	nan	1.000	sample
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Fabstastical	nan	nan	0.134	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Florid	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Hopeful	nan	nan	0.134	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	I would combine them and say arch/knowing without being judgmental	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Inconsistent	nan	nan	0.134	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Introspective	nan	nan	1.000	sample
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Linguistic	nan	nan	0.134	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Melodramatic	nan	nan	1.000	sample
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Moody/Descriptive/ Reflective	nan	nan	0.134	sample, ceiling

(continued on next page)

(continued from previous page)

Category	Question	Choice	Var.	Range	Ceiling	Failed gate(s)
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Nostalgic; bittersweet	nan	nan	1.000	sample
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Nuanced	nan	nan	0.134	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Of its time	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Out dated	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Parental	nan	nan	1.000	sample
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Pensive	nan	nan	1.000	sample
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Purple	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Realistic	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Robotic and boring	nan	nan	0.134	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Strong vocabulary	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Superficial	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Un-edited	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Undisciplined	nan	nan	1.000	sample
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Uninspired	nan	nan	0.134	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Unsophisticated	nan	nan	1.000	sample
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Wait	nan	nan	1.000	sample
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Warm	nan	nan	0.134	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Young	nan	nan	1.000	sample
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	Youthful	nan	nan	1.000	sample
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	academic	nan	nan	1.000	sample
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	almanacs	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	as in you could tell it was written in 1950	nan	nan	0.000	sample, ceiling

(continued on next page)

(continued from previous page)

Category	Question	Choice	Var.	Range	Ceiling	Failed gate(s)
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	bad	nan	nan	0.134	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	bleak	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	boring	nan	nan	0.134	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	cringe	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	dreamlike	nan	nan	0.134	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	poems	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	pompous	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	predictable	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	sappy	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	sardonic	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	sinuous	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	specifically nautically so	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	suspenseful	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	the writing style changed based on who was speaking.	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	unimaginative	nan	nan	0.134	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	what (is that a style?)	nan	nan	1.000	sample
Style	Which words best describe the writing style? Select all that apply. - Other: - Text	why deductions	nan	nan	0.000	sample, ceiling
Style	Which words best describe the writing style? Select all that apply. - Selected Choice	Ironic	0.015	1.000	0.271	ceiling
Style	Which words best describe the writing style? Select all that apply. - Selected Choice	Other:	0.021	0.667	0.000	ceiling
Style	Which words best describe the writing style? Select all that apply. - Selected Choice	Sparse (economical with words)	0.014	0.500	0.000	ceiling