HumanReader: A Dataset of Reader Responses to Book-Length Fiction

Anonymous ACL submission

Abstract

Reading a novel produces structured judgments about character, plot, narrator, setting, and style that vary across readers in ways no single answer key resolves. We introduce HUMANREADER, a dataset of structured survey responses from 714 readers across 387 randomly assigned fictional novels. The corpus comprises 809 finished post-reading sessions, with 101 closed-form items in seven subject categories, 55 open-ended items, and per-chapter ratings collected during reading. Mean per-question pairwise inter-reader agreement is 0.68; within-book reader pairs agree more strongly than across-book pairs, and a per-question variance decomposition shows book-level signal varying sharply across categories. Descriptive analyses surface structural correlates of overall enjoyment (e.g. an opening hook, $\rho = +0.48$) and four mood-based story-type clusters (warmth, darkness, melancholic-reflective, adventure) that differ in mean reader enjoyment. We release a 53-title Project Gutenberg subset with full texts and illustrate that pairwise model–human agreement (matching reader responses item-by-item) and cross-book rank correlation (ordering books on each item relative to the average reader) may be distinct evaluation targets for LLMs in this setting: both models we evaluate closely match the inter-reader pairwise baseline, but diverge sharply from each other on cross-book Spearman correlations.

1 Introduction

Large language models are increasingly applied to tasks that require modeling the human experience of book-length narratives, from literary recommendation and analysis to long-form summarization. Existing benchmarks for long-form narrative comprehension, including NarrativeQA (Kočiský et al., 2018), NovelQA (Wang et al., 2025), and NoCha (Karpinska et al., 2024), evaluate models primarily through factual question answering, multi-hop reasoning over plot events, and verifiable claims about characters and settings. These benchmarks measure an important but narrow slice of what it means to understand a novel, since reading is not exhausted by fact recall. Readers experience suspense, find characters likeable or grating, feel that a plot drags or accelerates, and reach the end of a book with judgments about tone, theme, and craft that no canonical answer key can adjudicate. Subbiah et al. (2025) argue that for short-form narrative tasks, this kind of reader disagreement is signal rather than noise, since single-protocol ground truths systematically miss the dimensions on which readers legitimately diverge. We extend this position to book-length narrative experience, where the textual span over which experience accumulates is far longer and the dimensions of reader response are correspondingly richer.

As a concrete illustration, Figure 1 shows three first-time readers of Victor Appleton's *Tom Swift in the Land of Wonders* disagreeing on a single beginning-of-book question, while Llama 4 Maverick and GPT-5 commit to confident single responses. The question admits multiple defensible answers, of the kind HUMANREADER is designed to surface and measure systematically across categories of narrative experience.

In this paper, we introduce HUMANREADER, the largest-to-date dataset of structured reader responses to book-length fiction, comprising 714 readers, 387 randomly assigned novels, and 809 finished post-reading sessions across a three-part instrument (pre-reading, chapter-level, post-reading) with 101 closed-form[1] and 55 open-ended items, each closed-form item annotated as descriptive (verifiable narrative facts) or experiential (subjective reader response). We characterize where book-length narrative readers agree and where they di-

[1] Of 104 administered; 3 Character items are excluded from agreement analysis for insufficient reader coverage.

Q. "Did the book begin with a *hook*—something dramatic, mysterious, or shocking that immediately made the stakes of the story clear?"

Book: Victor Appleton, Tom Swift in the Land of Wonders (1917).

Human readers in our evaluation subset ($n = 3$):

Reader 1	No
Reader 2	No
Reader 3	Arguably—you could make the case, but it's not clear-cut

Models:

Llama 4	**Yes, definitely** — no human reader endorses
GPT-5	**Arguably** — matches the minority human position

Figure 1: Reader disagreement on a Beginning-of-book question: three first-time readers of *Tom Swift in the Land of Wonders* split between "No" and "Arguably" on whether the book opens with a hook. Both models commit to a single confident response where readers themselves disagree (Llama's is endorsed by no reader). The book shown has the minimum reader count ($n = 3$) for the evaluation subset.

verge across seven subject categories, decomposing per-question variance into book-level signal and reader-level idiosyncrasy. We surface descriptive structural correlates of overall enjoyment (an opening hook, major turning points) and four mood-based story-type clusters that differ in mean reader enjoyment. And we release a Project Gutenberg evaluation subset of 53 titles (198 distinct readers, 209 finished sessions) with full texts, responses, and code, alongside an illustrative two-model baseline (Llama 4 Maverick and GPT-5) showing that pairwise model–human agreement and cross-book rank correlation are distinct evaluation targets — though a title-only baseline on this canonical-text corpus shows both metrics largely reflect pretrained literary knowledge rather than text reading.

2 Related Work

Two threads of prior work most directly inform HUMANREADER's position. *Long-context narrative evaluation* has advanced through QA benchmarks (Kočiský et al., 2018; Wang et al., 2025; Karpinska et al., 2024; Xu et al., 2022), summarization benchmarks (Kryściński et al., 2022; Chang et al., 2024), and broader long-context suites and narrative-evaluation frameworks (Bai et al., 2024; Hsieh et al., 2024; Hamilton et al., 2025a,b), each operationalizing understanding as correct answers against canonical per-item targets. We depart from the canonical-target frame by collecting structured subjective responses from first-time readers, with no single ground-truth answer per item.

Rubric-based evaluation of short-story quality (Chen et al., 2022; Chhun et al., 2022) and of LLM-generated short narratives (Harel-Canada et al., 2024) has begun to probe subjective dimensions beyond surface fluency, and Marco et al. (2025) show that reader profiles explain conflicting evaluations of AI-generated creative writing. In parallel, recent work treats annotator disagreement on short-form narratives as task-relevant signal rather than noise (Subbiah et al., 2024, 2025). We extend rubric-style subjective evaluation (building on HANNA's (Chhun et al., 2022) approach) to a seven-category schema designed for book-length narratives, and extend disagreement-as-signal to book length, quantifying the book-level reliability ceiling such disagreement implies for cross-book evaluation. Closest to our scope, Yang and Jin (2025) score 600 books on global aspects derived from reader reviews; we differ in collecting first-read structured responses through a controlled randomized-assignment protocol rather than post-hoc analysis of public reviews.

3 Dataset

The HUMANREADER dataset is generated from a large-scale reading experiment in which habitual readers were randomly assigned a book and surveyed before, during, and after reading. Each session requires a participant to read a full assigned novel, a per-session cost substantially higher than typical NLP labeling tasks; to our knowledge, HUMANREADER is the largest structured reader-response dataset for book-length fiction. The instrument follows four design principles: (i) participants are recruited from active book-discussion communities to ensure a population of habitual readers; (ii) books are randomly assigned within a fixed corpus so that cross-book variation in reader responses reflects the books themselves rather than reader self-selection; (iii) each of the post-reading questionnaire's Narrator, Character, Setting, and Plot sections opens with fact-based items and closes with a "Your Opinion" subsection of subjective items, supporting per-item annotation as descriptive or experiential across the questionnaire; and (iv) the corpus combines a Project Gutenberg sub-

set (released with full book texts to support reproducible LLM evaluation) with a partner-catalog subset of more recent titles (responses-only release). This section describes the corpus and recruitment, the experimental protocol, and the three survey instruments through which responses were collected.

3.1 Corpus and Participants

The corpus comprises 387 distinct books drawn from two sources. Roughly 220 titles come from the catalog of an anonymous private-sector partner, a book-recommendation application, spanning popular recent fiction, award-winning literary fiction, less well-known titles, and older classics such as Austen, Dickens, and Dostoyevsky.[2] The Gutenberg subset was sourced primarily from the top-downloaded novels list, supplemented with selected additional titles, after excluding those too short, non-fiction, or ethically inappropriate to assign. The reproducible evaluation subset comprises the 53 Gutenberg titles for which at least three readers completed the full study (Table 3, Appendix C). The two sources span a wide range of literary periods, genres, and levels of critical recognition, and the same survey instruments were administered to both.

We recruited from book-related subreddits on Reddit, including r/books, r/suggestmeabook, r/52book, and r/bookclub. Recruitment posts described the study as a book reading research project and offered compensation of a free copy of the assigned book and $15 for completing the full study (the book, all chapter-level surveys, and the post-reading questionnaire). All participants provided informed consent prior to enrollment, were free to withdraw at any time, and all data prepared for release are de-identified. We deliberately recruit from active reader communities to ensure a population of habitual readers for whom reading novels is a regular activity, increasing the ecological validity of survey responses. The dataset includes 714 distinct readers across 387 books who completed the full study, producing 809 finished post-reading sessions (some readers completed the protocol and then asked to re-enroll, and so produced data for more than one assigned book).

Each participant was randomly assigned one book from the corpus (per session); participants who had previously read their assigned book were re-randomized to a different title, ensuring all responses reflect first reads. Books were deliberately over-assigned at recruitment time to absorb anticipated attrition, so the number of completed readers per book reflects both random assignment and a non-random completion rate; books that are shorter, more accessible, or better matched to the recruited reader pool may be over-represented among completed responses relative to their original assignment counts (Section 7).

3.2 Survey Instruments

Three types of survey data were collected. The *pre-reading* onboarding questionnaire, administered before the assigned book, captured reading habits (novels read per year, completion rate, preferred genres), familiarity with the assigned author and title, self-assessed reader–book match, and pre-reading expectations for pacing, character relatability, and enjoyment. The *chapter-by-chapter* survey, administered after each chapter during reading, captured current enjoyment (1–7 Likert), surprise, suspense, and uncertainty about future events (1–5 each), and whether a major plot twist occurred (Yes/No with optional free-text description); page numbers were also recorded so ratings can be mapped to normalized book progress rather than raw chapter counts, enabling within-book trajectory analyses such as Figure 2.[3]

The *post-reading* questionnaire, administered after the book, comprised structured items spanning seven categories. The Narrator, Character, Setting, and Plot sections each open with fact-based items and close with a "Your Opinion" subsection of subjective items; the Overview, Style, and Reflection sections interleave both kinds. We use this fact-first-then-opinion structure as a starting basis for our *descriptive/experiential* classification, but apply it per-item across the questionnaire and additionally annotate items as temporally grounded at the beginning, middle, or end of the narrative where applicable. (For example, questions asking

[2] Partner titles were drawn through a combination of broad random sampling of the catalog and selection of algorithmically-flagged top-recommended books. A subset of partner-catalog assignments were paired with experimentally manipulated pre-reading review framings as part of a separate study on review priming; readers in the review-priming conditions are excluded from this dataset and all analyses reported here.

[3] The chapter-level data is included in the dataset but not benchmarked in our later exercise, since within-book temporal items introduce prompt-architecture choices that warrant their own study; similarly, the expectations data are not included in the illustrative benchmark exercise.

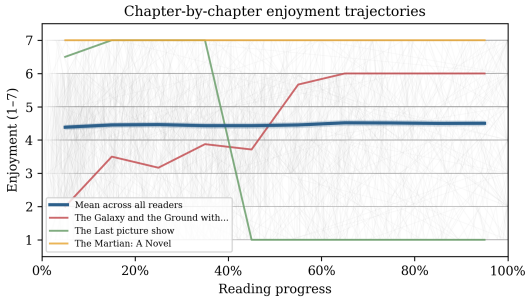

Figure 2: Chapter-by-chapter enjoyment trajectories on three sample books from the corpus. Faint grey lines are individual reader trajectories pooled across all books; the bold blue line is the cross-reader mean (95% CI shaded); three highlighted books illustrate contrasting shapes (rising, falling, flat). Post-reading overall ratings correlate more strongly with late-chapter enjoyment (Spearman $\rho = 0.86$) than with early-chapter enjoyment ($\rho = 0.72$, $n = 577$ reader sessions with both early and late bins), consistent with a recency component in narrative experience. Trajectories will be available for all readers upon dataset release; LLMs are not evaluated on within-book temporal items in this paper.

Sampled post-reading items (one each from seven categories, spanning response formats and the D/E split):

Plot	D, ordinal	"How would you characterize its linearity in terms of time?"
Character	E, 1–7 num, *branched*	"How real did this protagonist feel to you?"
Setting	D, checkbox	"On a percent-of-book basis, at what point(s) was the book set here?"
Overview	E, 1–5 num	"To what extent would you describe the book as plot-driven and character-driven?"
Narrator	D, categorical	"Was the narrator omniscient or an actual character?"
Reflection	E, ordinal	"After reading, how glad are you to have devoted time to this book?"
Style	E, 1–7 num	"How engaging did you find the dialogue?"

Figure 3: One sampled item per post-reading category, with response format and descriptive/experiential (D/E) label. *Branched* items are administered once per identified protagonist, narrator, or setting. Figure 1 shows actual reader and model responses on a single item; the full 156-item inventory (101 closed-form + 55 open-ended) appears in Appendix K.

about character attributes at the beginning versus middle versus end, or plot twists and where they occurred in the narrative.) The instrument also uses branching logic, with the number of narrators, protagonists, and major settings each reader identifies routing them to tailored per-entity items capturing reliability, tone, likability, competence, proactivity, vividness, and immersiveness for each identified setting/narrator/protagonist, respectively. Figure 3 shows a sampling of post-reading items across categories and response formats; the full 156-item inventory appears in Appendix K.

The five formats span checkbox (select all that apply), numeric (integer or scale), ordinal Likert (e.g., "very likable / somewhat likable / neither / somewhat dislikable / very dislikable"), single-choice nominal (e.g., "past tense / present tense"), and open-ended free text. Each closed-form format receives a tailored evaluation metric (Section 5). Table 1 reports the per-category item counts.

The 55 open-ended items are not evaluated in this paper but are released with the dataset. They include plot summaries, character trait descriptions, setting descriptions, turning-point narratives, theme descriptions, and per-section reflections on what worked well, what did not, and what was distinctive about the narration, characters, setting, plot, and writing style — a rich corpus of reader-generated literary micro-criticism that may support future work on open-ended evaluation, reader modeling, and natural-language literary analysis.

3.3 Quality Control and Composition

We retain only sessions in which the reader completed every section of the post-reading questionnaire, dropping incomplete or abandoned sessions from analysis.[4] No further manual screening (free-text engagement checks, duration thresholds, or content-based filtering) was applied. The dataset includes the completion flag, total survey duration, and free-text fields, allowing downstream users to apply additional filtering rules such as duration-based exclusions if their use case requires them. Summary statistics for our analysis dataset are presented in Table 2.

The Project Gutenberg evaluation subset (defined in §3.1) supports reproducible LLM evaluation because its texts can be released alongside the human response data and evaluation pipeline. The non-Gutenberg subset, whose texts cannot be redistributed due to licensing, will be released as human-response-only data, suitable for studying reading

[4] We additionally exclude a small subsample of readers who were randomly assigned to a separate review-priming experimental treatment, since their responses are confounded by the priming manipulation (Appendix E).

Table 1: Number of distinct post-reading questions per category in the analyzed set (101 closed-form questions with sufficient reader coverage; 3 additional Character closed-form items are administered in the instrument but excluded from agreement analysis). Open-ended items are included in the dataset release but not scored in the LLM evaluation here. The instrument applies branching logic to elicit per-protagonist, per-narrator, and per-setting ratings, so the questions in Narrator, Character, and Setting are administered multiple times per reader depending on how many entities the reader identifies.

Category	Closed	Open	Total
Character	29	22	51
Plot	21	4	25
Setting	15	9	24
Narrator	11	12	23
Style	13	3	16
Overview	7	4	11
Reflection	5	1	6
Total	**101**	**55**	**156**

Table 2: Dataset and evaluation-subset summary.

Statistic	Value
Full dataset	
Distinct readers	714
Finished post-reading sessions	809
Books in corpus	387
Pre-reading survey items	11
Closed-form post-reading questions	101
Open-ended post-reading questions	55
Chapter-level numeric dimensions	4 + twist
Evaluation subset (53 PD books, ≥ 3 readers)	
Distinct readers	198
Finished sessions	209

behavior, inter-reader agreement, and reader modeling, but not for direct LLM evaluation without independent access to the texts. Across the 101 closed-form post-reading questions with sufficient reader coverage in the cleaned dataset, mean per-question pairwise inter-reader agreement is 0.68, with substantial heterogeneity across categories that we characterize in Section 4.[5]

4 Characterizing Inter-Reader Agreement

First, we characterize the structure of human responses themselves. Following Subbiah et al. (2025) in considering subjectivity as signal, we document how strongly readers agree on each item, how that agreement varies across the seven subject categories, where on the questionnaire reader disagreement concentrates, and how much of inter-reader variation is stable book-level signal as opposed to reader idiosyncrasy.

4.1 Agreement Metric

We summarize inter-reader agreement using pairwise-agreement: for each (question, book) with at least two readers, we compute the average pairwise agreement across all reader pairs that responded, and aggregate to question-level and category-level means. Different response formats receive metrics tailored to the structure of the response, namely a bounded-difference metric for numeric and ordinal Likert items, Jaccard agreement for checkbox items, and exact-match agreement for single-choice nominal items; formal definitions and the corresponding random-uniform baselines appear in Appendix G.[6]

For branched per-protagonist, per-narrator, and per-setting items, we streamline comparison by focusing just on the first protagonist, narrator, or setting, respectively.[7] The analysis covers all 101 closed-form questions with sufficient reader coverage (i.e., at least one book having two or more reader responses).

4.2 Overall and Per-Category Agreement

Across the 101 closed-form questions with sufficient reader coverage, mean per-question pairwise agreement is 0.68 (per-question distribution: Appendix I, Figure 11); after per-item above-chance normalization that puts the mixed-format aggregate on a single scale (Appendix G), mean above-chance agreement is 0.36. The pattern is consistent with the view in Subbiah et al. (2025) that readers of narrative texts should not be expected to agree perfectly on items that elicit genuinely subjective responses; we extend that view to book-length material. Per-format and per-D/E means appear in

[5] Per-item inter-reader agreement and model–human numbers will accompany the dataset release as a supplementary spreadsheet (Appendix K).

[6] Each reader rated exactly one book under our random-assignment design, so the standard rater-exchangeability framing underlying Krippendorff's α does not apply directly to our data without modification; as such, we focus on the simple metric of average pairwise agreement.

[7] Across 1,544 branched (book, item) cells with ≥ 2 readers on the evaluation subset, 44% of cells have all readers using the same first-entity slot; on the remaining 56% of cells where readers spread across multiple slots, the modal slot captures 62% of readers on average. Different readers' first-named entity need not refer to the same person even when slot positions match, which may understate per-entity reader agreement; we treat the resulting estimates for branched items as conservative.

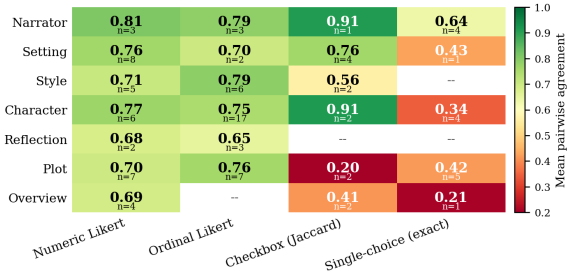

Figure 4: Mean pairwise inter-reader agreement by subject category × response format. Cells show mean and (n) questions; empty cells indicate the combination does not occur.

Appendix G.

Mean agreement varies substantially across the seven subject categories of the post-reading questionnaire. Agreement is highest for Narrator items (0.75, $n = 11$ questions) and Setting items (0.73, $n = 15$), in the middle for Style (0.72, $n = 13$), Character (0.71, $n = 29$), and Reflection (0.66, $n = 5$), and lowest for Overview (0.54, $n = 7$) and Plot (0.61, $n = 21$).

The category ordering plausibly tracks how text-anchored vs. reader-dependent each response is: Narrator, Setting, and Style items concern features readers can anchor to specific passages, while Overview and Plot items elicit gestalt judgments that depend on reader attention and prior expectations.

4.3 The Distribution of Disagreement Across Questions

Per-category means obscure substantial question-level heterogeneity (Figure 11), itself structured along the response-format axis (Figure 4). The high-agreement tail is dominated by counting questions and Likert ratings with clear textual anchors; the low-agreement tail is dominated by interpretive list-selection questions (e.g., the 32-mood-descriptor item at 0.25) and single-choice nominal items.

4.4 Book-Level Signal vs. Reader Idiosyncrasy

Inter-reader agreement aggregates stable book-level signal and reader-level idiosyncrasy. For numeric and ordinal items, we decompose per-question variance via a one-way random-effects ANOVA and report the intraclass correlation ICC_q plus the Spearman–Brown-extended split-half reliability r_{full} of book-level means over books with ≥ 4 readers (Appendix H). On the 72-item shared set used for the cross-book Spearman analysis in Section 5, mean book-mean reliability is $r_{\text{full}} = 0.46$ and the mean per-item Spearman ceiling (the per-item $\sqrt{r_{\text{full},q}}$ averaged across the 71 items with sufficient ≥ 4-reader books) is 0.64. In other words, under the 3–4-reader-per-book counts of the evaluation subset, even an ideal cross-book evaluator — one that always recovers the true population-mean response on each book — could not exceed Spearman $\rho \approx 0.64$ against the observed book-level means, since those observed means are themselves noisy estimates of the underlying book-level signal. Cross-book correlation estimates on this corpus should accordingly be interpreted against this 0.64 ceiling rather than against 1.0.

As a complementary descriptive check, within-book reader pairs agree more strongly than across-book pairs on 93% of items (mean 0.69 vs. 0.62). Branched-item IAA (0.71) runs ∼5 points above unbranched (0.66), consistent with first-entity slot-matching providing additional alignment; a slot-stability sensitivity in App H bounds the contribution of this matching to the gap. We caution that the evaluation subset is weighted toward pre-1925 Anglophone Project Gutenberg titles, so per-category estimates may partly reflect period and genre composition rather than universal narrative properties.

4.5 Structural Correlates of Enjoyment, and Story-Type Clusters

Second, we draw on the dataset to examine which structural features of a book are associated with higher reported overall enjoyment, illustrating one kind of analysis the dataset enables. Among descriptive items (book features rather than evaluative re-statements; halo items in Appendix J), the strongest cross-book correlates with overall enjoyment (Figure 5) are an opening hook ($\rho = +0.48$, $p < 10^{-3}$, $n = 53$) and major turning points ($+0.46$, $p < 10^{-3}$), each associated with higher enjoyment. The pattern is suggestive of reader preference for hook-driven narratives, presented as descriptive cross-book associations.

k-means clustering on book-level mood-selection rates ($k = 4$, the 32-option mood checkbox) yields four interpretable story types: *warmth*, *darkness*, *melancholic-reflective*, and *adventure* (per-cluster mood signatures and exemplar books: Appendix J, Figure 12). Mean overall enjoyment differs across clusters (ANOVA $F = 3.8$, $p = 0.02$): warmth highest (5.4/7),

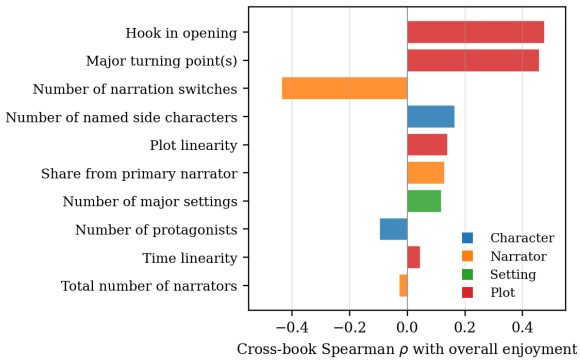

Figure 5: Structural (descriptive) items by cross-book Spearman correlation with overall reader enjoyment on the evaluation subset; bars colored by subject category. Evaluative halo items ("compellingness", "gladness", etc., which correlate more strongly but re-state preference) appear in Appendix J.

melancholic-reflective lowest (3.6/7), adventure and darkness in between (4.5). The cluster structure may reflect both reader experience and the period/genre composition of the Project Gutenberg subset.

5 Illustrative Baselines: Agreement vs. Cross-Book Variation

Third, we present an illustrative benchmarking exercise comparing two contemporary LLMs against reader responses on the evaluation subset, with the goal not of a comprehensive leaderboard but of demonstrating the diagnostic evaluation HUMANREADER enables. We evaluate Llama 4 Maverick (open-source) and GPT-5 (closed-source) on the closed-form questionnaire across the 52-book shared evaluation subset under a single prompting protocol.

On the 52-book shared subset, the two models match the inter-reader pairwise baseline of 0.68 in aggregate (Llama 0.66, GPT-5 0.67). They diverge, however, on cross-book Spearman ρ across the 72 numeric and ordinal items with cross-book coverage: Llama 0.35, GPT-5 0.47 (item-bootstrap 95% gap CI $[0.09, 0.16]$). Per-category breakdowns, the central-tendency baseline, full Spearman figure, and prompting details appear in Appendix F.

That said, supplying both models with only the book's title and author (no text) leaves both metrics essentially unchanged, and, if anything, marginally higher: pairwise agreement 0.67/0.67 for Llama/GPT-5 vs. 0.66/0.67 full-text, and mean per-item cross-book ρ 0.36/0.48 vs. 0.35/0.47 on the matched 72-item set (Appendix F, Table 6). Neither metric isolates text comprehension on this corpus. Agreement and cross-book Spearman score different things, but on a contamination-realistic corpus both are largely driven by pretrained literary-reception knowledge rather than text reading.

6 Discussion and Conclusion

We have introduced HUMANREADER, a dataset of structured reader responses to book-length fiction with a reproducible 53-book Project Gutenberg evaluation subset. Mean per-question pairwise inter-reader agreement is 0.68; book-level signal is strongest for Setting and weakest for Style, and descriptive analyses reveal structural enjoyment correlates and four mood-based story-type clusters.

Beyond the LLM-evaluation illustration in Section 5, HUMANREADER supports a range of downstream uses. The combination of per-question reader-response distributions, descriptive–experiential annotations, and chapter-level reading trajectories opens analyses spanning reader modeling, authorship-style characterization, structural-feature extraction, and rubric-based generation evaluation. One natural extension of the structural-correlates analysis (§4.5) is to develop NLP detectors for structural features such as opening hooks and turning-point timing operating directly on book text, validate them against the corresponding reader-reported labels in this dataset, and then deploy the validated detectors in downstream analyses where reader labels are unavailable. The full release — human responses, chapter-level data, evaluation pipeline, and (for the Gutenberg subset) book texts — will accompany the published version under terms detailed in Appendix A.

7 Limitations

Training-data contamination. Many books in our corpus are likely present in the pretraining data of the models we evaluate; the Project Gutenberg top-downloaded titles from which the evaluation subset is drawn are among the books most likely to appear in pretraining corpora. The title-only baseline reported in Section 5 makes this concrete: supplying both models with only the book's title and author yields pairwise agreement and cross-book Spearman essentially identical to the full-text runs. Neither metric isolates text comprehension on this corpus, and the cross-model Spearman gap should be read as a difference in pretrained literary knowledge rather than in reading the supplied text. Evaluation of text-comprehension specifically would require less-canonical or post-cutoff novels and is left to future work.

Generalization. Participants were recruited from Reddit book-discussion communities, a population of relatively avid Anglophone readers, and the corpus is restricted to English-language fiction drawn from one recommendation-platform catalog and the Project Gutenberg top-downloaded list. The Project Gutenberg evaluation subset is therefore weighted toward older Anglophone fiction (mostly pre-1925, with some later Gutenberg-released titles), and cross-book findings on this subset may not generalize to contemporary novels. The human distributions used as evaluation targets should be interpreted as the judgments of habitual Anglophone genre readers; category-level findings reflect the genres represented in our corpus rather than fiction in general. Because completion is non-random across book length and reader engagement (Appendix E), HUMANREADER responses should be read as describing the population of reader–book completion-likely pairings, not the recruited reader pool as a whole.

Evaluation scope and per-book power. The two-model evaluation uses a single prompt template; conclusions about LLM performance should be replicated across additional models with prompt-sensitivity analysis. For experiential items where readers disagree, a model that predicts the mean response scores well on mean pairwise model–human agreement (the metric in Section 5) while failing to capture the full reader response distribution; richer distributional metrics may be needed to fully characterize model–reader divergence. Most books in the evaluation subset have 3–4 readers (median 3), limiting per-book power for cross-book rank-correlation analyses on individual items; we quantify the resulting attenuation as a per-category Spearman ceiling in Section 4.

8 Ethical Considerations

Participants. This study was approved by the lead institution's Research Ethics Board (protocol details to be provided at de-anonymization). Participants were recruited from public book-discussion subreddits, gave informed consent, were compensated with a free copy of the book and $15 on full completion, and were free to withdraw at any point. No demographic information was retained beyond self-reported reading covariates; all data prepared for release are de-identified. Titles judged so offensive as to violate basic ethics guidelines were excluded from the assignment pool. The flat $15 compensation rate is modest relative to time invested for the longest books in the corpus and likely contributes to differential attrition (Appendix E).

Distribution and downstream use. Release terms (Gutenberg subset full; non-Gutenberg responses-only) are detailed in Appendix A. We discourage applications that would use these data to train systems for automated literary judgment in high-stakes contexts (e.g., publishing acquisitions, content moderation, educational assessment) without substantial additional validation. A user-facing disclaimer covering recruitment population, English-only scope, completion bias, author-assigned category labels, and likely pretraining contamination accompanies the release (Appendix A).

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

References

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

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

Hong Chen, Duc Vo, Hiroya Takamura, Yusuke Miyao, and Hideki Nakayama. 2022. Storyer: Automatic story evaluation via ranking, rating and reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1739–1753.

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836.

Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan K. 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*, 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.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? In *First Conference on Language Modeling*.

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

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

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.

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. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25432–25449, Vienna, Austria. Association for Computational Linguistics.

Oona Rainio, Jarmo Teuho, and Riku Klén. 2024. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086.

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. Association for Computational Linguistics.

Melanie Subbiah, Akankshya Mishra, Grace Kim, Liyan Tang, Greg Durrett, and Kathleen McKeown. 2025. Is the top still spinning? evaluating subjectivity in narrative understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 185–203, Suzhou, China. Association for Computational Linguistics.

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

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun 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 (Volume 1: Long Papers)*, 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 Datasheet for HUMANREADER

We provide a brief datasheet following Gebru et al. (2021), supplemented by data-statement (Bender and Friedman, 2018) and CrowdWorkSheets (Díaz et al., 2022) elements relevant to subjective annotation.

Motivation. The dataset was created to enable empirical research on book-length narrative experience — including literary analysis, narratology, computational social science of reading, reader-modeling tasks, and (as illustrated in this paper) evaluation of LLM alignment against structured reader judgments. It was created by the authors of this paper, with funding and reader recruitment in partnership with an anonymous private-sector book-recommendation platform.

Composition. Each instance is a tuple linking a reader, a book, and one of three survey types (pre-reading, chapter-level, post-reading). The corpus contains 714 distinct readers, 387 books, 809 finished post-reading sessions, 11 pre-reading items, per-chapter ratings on 4 numeric dimensions plus a plot-twist indicator and free-text description, and 101 analyzed closed-form (of 104 administered) plus 55 open-ended post-reading questions. The Project Gutenberg evaluation subset comprises 53 Project Gutenberg titles with 198 distinct readers contributing 209 finished sessions. Reader-level pre-reading attributes captured include novels-per-year, completion rate, preferred genres, and per-book familiarity; no race, gender, age, or geographic data were collected. All reader identifiers are random alphanumeric codes and all data prepared for release are de-identified.

Collection process. Reader recruitment was conducted through public book-discussion subreddits (r/books, r/suggestmeabook, r/52book, r/bookclub) on a rolling basis between January 2025 and January 2026. Books were randomly assigned per Section 3.1; readers who had previously read their assigned book were reassigned. Responses were collected through a Qualtrics-hosted web instrument with three linked surveys (onboarding, chapter-level, post-reading).

Preprocessing and labeling. Released records are at the item level (numeric ratings, checkbox selections, free-text responses). Preprocessing applied before release: (i) exclusion of partial post-reading sessions (we keep only readers who completed every section); (ii) exclusion of the review-priming experimental treatment subset (their responses are confounded by the priming manipulation); (iii) book-identifier deduplication via title-normalized union-find (408 raw book IDs map to 387 consolidated book groups); (iv) reader-identifier normalization (prefix and whitespace stripping, so a reader who used slightly different identifier formats across sessions is recognized as the same person); (v) de-identification of personal identifiers. Per-item D/E (descriptive vs experiential) and temporal labels (beginning/middle/end) were assigned by the authors based on item wording.

Annotator characteristics. Readers are self-selected from book-related subreddits and are accordingly habitual Anglophone readers; we caution against generalizing the response distributions to broader populations. Reading frequency, completion rate, and preferred genres are self-reported in the pre-reading survey and will be included in the released data.

Uses. Intended uses include analysis of reader-response heterogeneity, characterization of subjective reading experience at book length, and (for the Project Gutenberg evaluation subset) evaluation of model alignment with reader responses. We discourage uses involving automated literary judgment in high-stakes contexts without additional validation (see Section 8).

Distribution and maintenance. Upon acceptance, project-created materials (survey responses, instruments, evaluation code, derived analyses) will be released under a permissive license; Gutenberg-subset book texts will be redistributed under their original Project Gutenberg terms (not the permissive license). Non-Gutenberg book texts will not be redistributed; responses will be released as human-response-only data. The public URL will be inserted in the camera-ready version; corrections and additions will be issued as versioned updates.

User-facing disclaimer. Downstream users should account for: (i) self-selected Reddit-recruited habitual readers, not a general reading population; (ii) English-only fiction; (iii) non-random completion by book length and accessibility; (iv) descriptive–experiential labels and seven-category groupings assigned by the authors, not by independent annotators; and (v) likely training-data contamination for many corpus books in current pretrained models.

B Survey Instrument

The post-reading questionnaire was administered through a Qualtrics-hosted web instrument with branching logic on protagonist, narrator, and setting counts (1 through 5 or ensemble for protagonists; 1 through 6 or more for narrators and settings). The full Qualtrics export, including question text, response options, branching conditions, and the underlying logic, will be included in the dataset release for full reproducibility of the instrument. We omit reproductions of the Qualtrics interface screenshots from the printed appendix for length.

C Gutenberg Corpus

Table 3 lists the 53 Project Gutenberg titles in the Project Gutenberg evaluation subset, sorted by reader count.

D Dataset Descriptives

This appendix provides distributional summaries of the full HUMANREADER dataset (387 books, 714 distinct readers, 809 finished post-reading sessions). Figures 6–7 illustrate closed-form response distributions; the open-ended subsection below characterizes the free-text corpus.

Open-Ended Responses

The dataset includes 55 open-ended free-text post-reading items spanning all seven categories (e.g., plot summarization, protagonist description, surprising and disappointing elements, intended audience). Across the 809 finished post-reading sessions, readers provided 23,691 non-empty open-ended responses (median length 10 words, mean 19.4, maximum 3,394), with 32% of responses at ≥ 20 words and 9% at ≥ 50 words. Table 4 lists three illustrative responses to convey the texture of the data; the full open-ended corpus is included in the planned dataset release but is not benchmarked in this paper.

E Data Processing and the Per-Book Reader-Count Distribution

This appendix documents how we move from the raw Qualtrics exports to the cleaned dataset, and why the released per-book reader counts vary as they do.

Readers were assigned books at random from the corpus at enrollment time, with assignments deliberately over-allocated relative to the target completer count, since we anticipated meaningful attrition between book assignment and post-reading completion. Some readers withdrew during reading, some completed the chapter-level surveys but not the post-reading questionnaire, and some completed neither beyond onboarding. The realized per-book completer count therefore reflects both the random assignment and a non-random completion rate that may vary with book length, accessibility, and how well a given title matched the recruited reader pool.

Figure 8 shows the resulting distribution of finished post-reading sessions per book in the cleaned dataset, after the book-identifier deduplication described in the next paragraph. The distribution is heavy on the low end, with 138 books having one finished session and 146 having two, falling off through 67 books with three sessions, 21 with four, and a short tail to a maximum of eight. The threshold of three or more readers, applied to the Gutenberg subset, selects 53 of the deduplicated Gutenberg books, for a total of 209 finished sessions and 198 distinct readers used in the LLM evaluation in Section 5.

Two sources of cross-session ambiguity in the book-identifier fields motivate a deduplication step. The post-reading instrument elicits both a book ID (the admin-assigned identifier Q720) and a free-text book title (Q721). We observe both that the same admin-assigned book ID appears under multiple reader-typed title variants (mostly typos and subtitle variations) and that the same book has been admin-assigned multiple book IDs in different recruitment waves, likely from accidental re-listing of the same title under a new ID. We construct a consolidated book identifier by union-finding book IDs that share any title-normalized form, with light manual aliasing for cases where the same book appears under genuinely different titles (e.g., *Peter Pan* versus *Peter and Wendy*). The resulting key maps 408 raw book IDs to 387 consolidated book groups, with reader sessions pooled within each group for all per-book analyses in this paper.

Beyond the deduplication, the cleaning steps applied to the data prepared for release are minimal. We keep only post-reading sessions in which the reader completed every section of the questionnaire, dropping partial submissions; drop the subset of readers who were assigned to a separate review-priming experimental treatment (positive or negative pre-reading review framing), since their responses are confounded by the priming manipu-

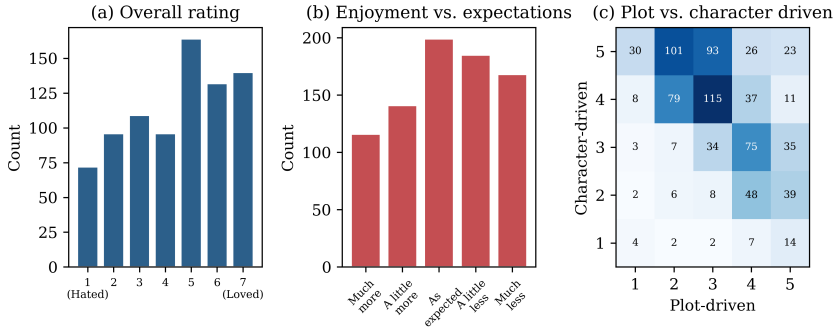

Figure 6: Post-reading response distributions across all 809 finished post-reading sessions (from 714 distinct readers). (a) Overall book ratings on a 1–7 scale, showing a positive skew with a mode at 5. (b) Whether readers enjoyed the book more or less than they expected prior to reading. (c) Heatmap of plot-driven vs. character-driven ratings (1–5 each), showing that readers perceive most books as moderately to highly character-driven regardless of plot intensity.

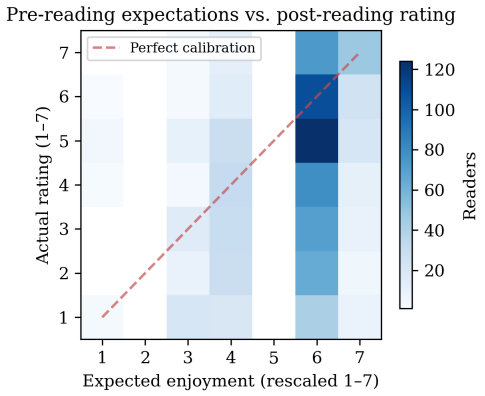

Figure 7: Pre-reading expected enjoyment (from the onboarding survey, rescaled from a 5-point to a 7-point scale) vs. post-reading actual rating. The dashed line indicates perfect calibration. Pre-reading expectations are only weakly correlated with post-reading enjoyment (Spearman $\rho = 0.28$, $n = 798$), and actual ratings are broadly distributed regardless of expectations (34% of post-reading ratings are $\geq 6/7$ and 34% are $\leq 3/7$). The weak calibration confirms the dataset captures realized narrative experience rather than prior reader enthusiasm.

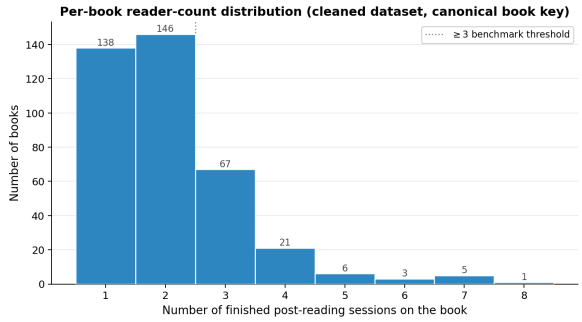

Figure 8: Number of finished post-reading sessions per book in the cleaned dataset (387 books, 809 sessions). The dotted vertical line marks the ≥ 3 threshold applied to the Gutenberg subset to construct the 53-book evaluation subset.

lation and reflect a different population of interest; and normalize reader identifiers, stripping inconsistent prefixes and trailing whitespace so that a reader who used slightly different identifier formats across sessions is correctly recognized as the same person. We do not filter on whether the same reader also completed the onboarding survey, and we do not apply manual quality screening to free-text answers; downstream users who require either filter can apply it using the completion flags, survey-duration field, and free-text columns included in the dataset release.

The full processing pipeline from raw Qualtrics exports through the per-question agreement results in Section 4 and the LLM evaluation in Section 5 will be released alongside the dataset on publication, and is deterministic given the raw data.

Completion descriptives and reader-quality sensitivity. Of 3,780 onboarding entries, 912 (24%) resolve to a finished post-reading session in the cleaned dataset. Completers differ modestly from non-completers on pre-reading covariates collected in onboarding: they report reading more novels per year (43 vs. 37, $p < 0.01$) and slightly higher prior familiarity with their assigned book (4.2 vs. 4.1 on a 1–7 scale, $p < 0.01$); expected enjoyment and self-assessed target-reader fit do not differ meaningfully. As a reader-quality sensitivity, removing the fastest 10% of completed sessions by total questionnaire duration raises headline mean per-question pairwise IAA from 0.68 to 0.69; removing the fastest 20% raises it to 0.69. Reader-quality

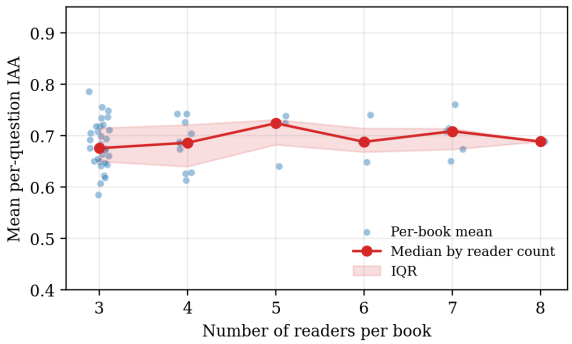

Figure 9: Mean per-question pairwise inter-reader agreement per book in the Project Gutenberg evaluation subset, plotted against the number of readers per book. Per-book means are stable across reader counts (median 0.68–0.72 from $n = 3$ to $n = 8$), indicating that question-level agreement estimates do not drift systematically with per-book reader count in the observed range. Distinct from the noise in book-level mean responses used for the cross-book Spearman analysis, which we characterize via split-half reliability in Appendix H.

filtering does not materially shift the headline number.

F Benchmarking Exercise — Full Results

This appendix reports the full per-category and cross-book Spearman results, the title-only contamination baseline, and implementation details for the benchmarking exercise summarized in Section 5.

Setup. The two models evaluated are Llama 4 Maverick (Meta; mixture-of-experts, approximately 400B total / 17B active parameters) and GPT-5 (OpenAI; parameter count not disclosed), both accessed via OpenRouter at temperature 0, with no fine-tuning or in-context examples. Llama is evaluated on all 53 Project Gutenberg evaluation-subset books; GPT-5 on 52 (Mary Johnston's *The Long Roll*, the longest book in the corpus, requires roughly 370K tokens at the GPT-5 tokenizer and exceeds GPT-5's 400K-token context window when combined with the system prompt; reported model–human and Spearman comparisons therefore use the 52-book intersection, with Llama's *The Long Roll* outputs included in the planned release). The full evaluation comprises ∼6 batches per book, on the order of 600 API calls total; per-call cost is dominated by the long context, and total inference cost is on the order of hundreds of US dollars. For each (book, model) pair, the full book text is placed in the model's system prompt and the closed-form questionnaire is administered in batches of approximately 20 items per API call, with the model returning a structured JSON response per batch. Branched items are administered for the first per-protagonist, per-narrator, and per-setting entity only, matching the per-reader first-slot collapse used in Section 4; model–human agreement uses the same per-question pairwise metric. The system prompt frames the task as follows:

```
You have just finished reading
the novel '{title}' by {author}.
You will now answer a series
of survey questions about the
book as a reader who has just
completed it.  Take the role
of a thoughtful, attentive
reader rather than a literary
scholar.  For experiential
questions (about your feelings,
perceptions, or judgments),
respond as you genuinely would
as a reader.  For descriptive
questions (about verifiable
facts in the text), respond
based on what is in the book.
If a question does not apply to
this book (e.g., asks about a
third protagonist when there are
only two), respond with null for
that item.

=== BEGIN BOOK TEXT ===
{book_text}
=== END BOOK TEXT ===
```

The user-message prompt asks for a structured JSON response over a batch of items. For each item, the prompt specifies the scale or option set. Numeric items request a single integer in the documented range; checkbox items request a JSON object with binary 0/1 values keyed by option label; open-ended items are posed without format constraints. Items that do not apply to the given book may be returned as null.

Per-category pairwise agreement. Table 5 reports per-category model–human pairwise agreement, the inter-reader baseline, and a leave-one-out central-tendency baseline. The two models achieve very similar aggregate pairwise agreement (Llama 0.66, GPT-5 0.67), both within roughly two points of the inter-reader baseline of 0.68. The largest model–human gap is on Setting (Llama −14, GPT-5 −10 pp), followed by Narrator and Reflection. On Plot and Character items, model–human agreement is comparable to or slightly above the inter-reader baseline; we read this as central-tendency approximation, since pairwise agreement flatters that strategy. The leave-one-out central-tendency baseline (rightmost column) scores 0.72 overall, above both the inter-reader level and both model

scores.

Cross-book Spearman. Pairwise agreement does not test whether a model ranks books correctly relative to each other. We therefore compute per-item Spearman ρ between the model's per-book responses and the mean human per-book responses (definition and per-item book counts in Appendix G). The two models diverge sharply on this metric. Across the 72 numeric and ordinal items with sufficient cross-book coverage for both models, mean per-item ρ is 0.47 for GPT-5 (97% positive; 83% significantly positive at $p < 0.05$) and only 0.35 for Llama (94% positive; 65% significantly positive); item-bootstrap 95% CIs are $[0.31, 0.40]$ for Llama and $[0.43, 0.52]$ for GPT-5, with a per-item gap CI of $[0.09, 0.16]$. Figure 10 plots per-item Spearman ρ against mean pairwise agreement: GPT-5 lifts items vertically without substantially changing horizontal position.

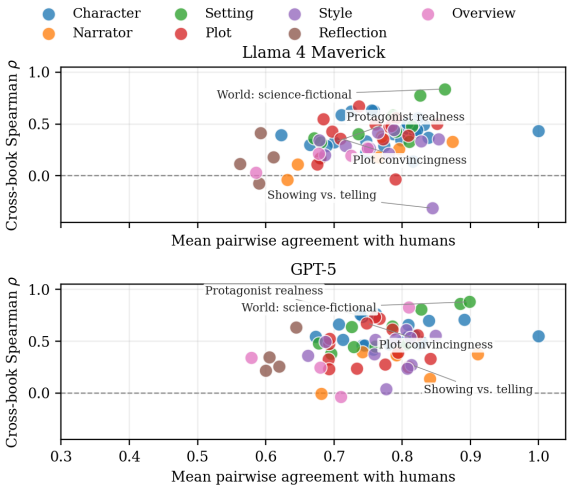

Figure 10: Per-item cross-book Spearman ρ vs. mean pairwise agreement, Llama 4 Maverick (top) and GPT-5 (bottom) on the 72 numeric/ordinal items with cross-book coverage in both models.

Responses on disk and retries. Responses are saved to disk per (book, model) as a structured table after each batched call. On startup, the pipeline skips books for which a completed CSV already exists, allowing resumption after interruptions. All API calls include retry logic for rate-limit and transient network errors.

Reproducibility details. Models are accessed via OpenRouter using the route names `meta-llama/llama-4-maverick` and `openai/gpt-5`; both routes pin a fixed provider snapshot at the time of evaluation, but proprietary model versioning means we cannot guarantee bit-identical replication on future calls. Evaluation is single-run per (book, model): we do not average over decoding seeds because both models run at temperature 0. Batches are administered in fixed questionnaire order with no randomization, and each batch is one independent API call (no cross-batch state). The pipeline retries on rate-limit and parse failures up to three times before recording a null for that item. Nulls (model returns `null`, parse fails after retries, or item is out-of-range) are recorded as missing for that (book, item, reader) cell; they are excluded from agreement and rank-correlation computation rather than counted as disagreement. Null rates are 11.4% of cells for Llama and 5.6% for GPT-5; because per-item null patterns differ across models, per-item cross-book Spearman comparisons (Section 5) are computed on each model's own valid book set, and the 72 shared items used in the cross-model comparison are those where both models have ≥ 3 valid books.

Title-only contamination baseline. For the contamination check in Section 5, we re-ran the full pipeline with an identical prompt template except that the system prompt contains only the book's title and author and no book text. All other settings (temperature, batching, parsing) are unchanged. Table 6 reports the resulting headline numbers alongside the full-text values.

G Formal Metric Definitions and Per-Format Means

This appendix gives precise definitions for the agreement and rank-correlation metrics summarized in Section 4 and Section 5, and reports per-format means and random-uniform baselines. Throughout, $\mathbf{m}$ denotes a model response vector and $\mathbf{h}_i$ the response vector of human respondent i.

G.1 Pairwise Agreement

For checkbox items, responses are sets of selected option labels. Pairwise agreement between two responses M and H is Jaccard, $A(M, H) = |M \cap H|/|M \cup H|$, with the convention $A = 1$ when both sets are empty. Multi-human agreement against a model is the mean over k respondents, $A(M, \mathcal{H}) = \frac{1}{k}\sum_{i=1}^{k} A(M, H_i)$, and pairwise human–human agreement is $A(\mathcal{H}) = \frac{2}{k(k-1)}\sum_{i<j} A(H_i, H_j)$. For numeric items, pairwise agreement is the bounded-difference $\max(0,\ 1 - |m - h|/d_{\max})$, with $d_{\max}$ the scale range ($d_{\max} = 4$ for 1–5 Likert items and 6 for 1–7); this is a normalized MAE-

based metric of the kind surveyed in Rainio et al. (2024). For ordinal Likert items we apply the numeric definition after mapping labels to integer ranks via a fixed encoding. For single-choice nominal items (e.g., narration tense), agreement is exact-match.

G.2 Cross-Book Spearman Rank Correlation

For each numeric or ordinal item, we collect the model's predicted value m_b on each book b and the mean human response $\bar{h}_b = \frac{1}{k_b}\sum_i h_{i,b}$ across the k_b readers of that book, and compute the Spearman rank correlation ρ between $\{m_b\}$ and $\{\bar{h}_b\}$ across the books on which both vectors are defined. Statistical significance is the two-sided p-value of the rank-correlation test. Per-item book counts vary from about 14 (for the most-branched items in Llama; 26 in GPT-5) to the full 52-book evaluation subset, with a median of 51 for both models. For branched per-protagonist, per-narrator, and per-setting items, each reader's response is resolved to the first entity they were routed to in their branch, mirroring the LLM's single-entity prompt; this collapse is applied identically in the inter-reader analysis of Section 4.

G.3 Per-Format Means and Random Baselines

Across the 101 closed-form questions with sufficient reader coverage, mean per-question pairwise inter-reader agreement is 0.73 for numeric ($n = 35$), 0.75 for ordinal ($n = 38$), 0.63 for checkbox-Jaccard ($n = 13$), and 0.44 for single-choice nominal ($n = 15$); descriptive items average 0.66 ($n = 31$) and experiential items average 0.69 ($n = 70$). Random-uniform baselines: 0.62 for 1–7 numeric items (and 0.60 for 1–5); the Jaccard random baseline for the $n = 32$ mood-descriptor checkbox is approximately 0.12 under independent uniform option selection; and exact-match random agreement for k-way nominal items is $1/k$. Per-item above-chance agreement (reported in §4) is computed as $(\text{raw}_q - \text{chance}_q)/(1 - \text{chance}_q)$, using the format-specific baseline above as chance_q, then averaged across the 101 items.

H Book-Level Signal: ICC and Reliability Ceiling

For the numeric and ordinal items used in the Section 5 cross-book Spearman analysis (the 72-item shared set, of which 71 have books with ≥ 4 readers and contribute to the reliability estimates), we compute per-question split-half reliability $r_{\text{half},q}$ over books with ≥ 4 readers (random reader splits, 200 iterations, Spearman across split-mean vectors), Spearman–Brown-extend to full-sample reliability $r_{\text{full},q} = 2r_{\text{half},q}/(1 + r_{\text{half},q})$, and report the per-question ceiling on cross-book Spearman model–human correlation as $\sqrt{r_{\text{full},q}}$ (assuming a deterministic model). Aggregations across items use the mean of per-question values rather than $\sqrt{\ }$ of the mean (these differ by Jensen's inequality); we use the per-item mean since model ρ in Section 5 is reported per item. Table 7 reports per-category means alongside each model's per-category ρ and % of ceiling.

I Per-Question IAA Distribution

J Story-Type Clusters and Attribute Correlations

Pre-reading attribute correlations with overall enjoyment (cross-reader Spearman ρ, $n \approx 790$): pre-reading expected enjoyment $\rho = 0.28$ ($p < 10^{-15}$), target-reader fit $\rho = 0.21$, expected character-relatability $\rho = 0.20$, book-familiarity $\rho = -0.11$ (familiar books rated slightly lower in our sample); pre-reading attributes also correlate with per-category quality outcomes (narrator/character/setting) at $\rho \sim 0.13$–0.24.

Top mood-pair co-occurrences across books (book-level Spearman): Cozy–Heartwarming $\rho = 0.69$; Cozy–Lighthearted 0.63; Sad–Tragic 0.62; Heartwarming–Lighthearted 0.60; Gritty–Raw 0.59; Chilling–Dark 0.59; Quirky–Whimsical 0.57; Funny–Whimsical 0.57.

Evaluative halo items at the top of the overall-enjoyment correlation list (these re-state preference rather than describe book features and are excluded from the main-text Figure 5): gladness-to-have-read ($\rho \approx 0.92$), plot compellingness (0.91), in-the-moment enjoyment (0.79), emotional investment in the protagonist (0.78), writing-style interestingness (0.77). The full ranked list of all 72 items accompanies the dataset release.

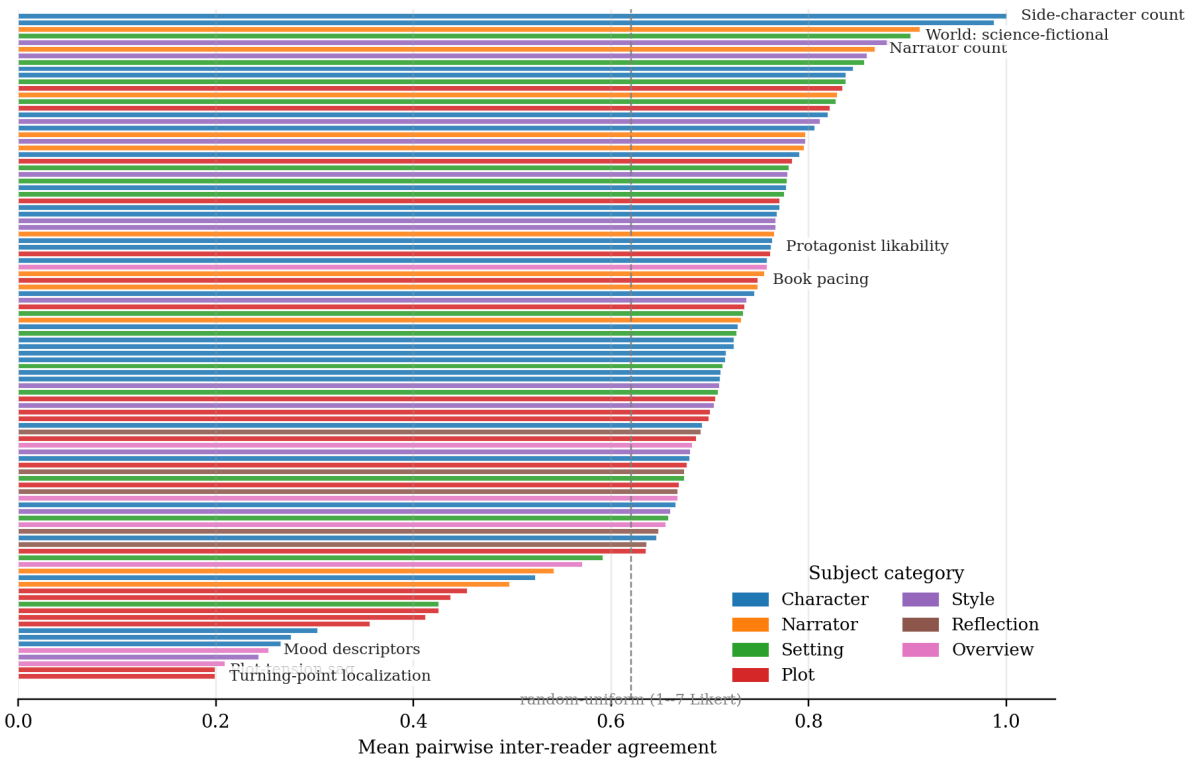

Figure 11: Mean pairwise inter-reader agreement across all 101 closed-form post-reading questions, sorted descending. Bar color encodes subject category; selected questions labeled for orientation. The dashed line marks the random-uniform baseline for 1–7 Likert items (0.62).

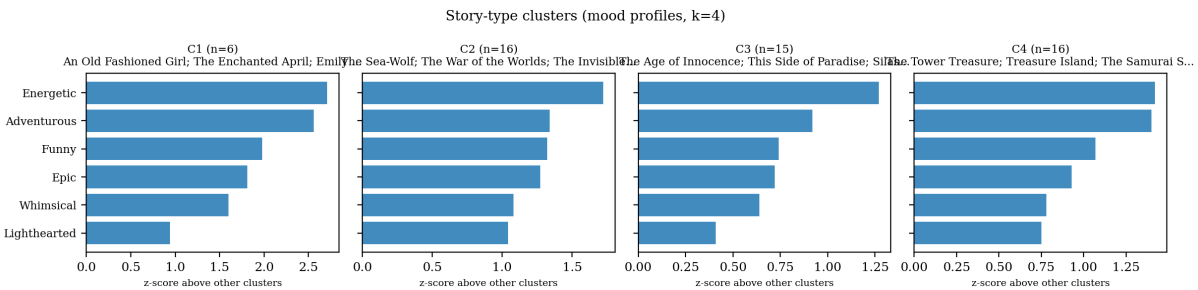

Figure 12: Story-type clusters from k-means ($k = 4$) on the book $\times$ mood-selection-rate matrix (53 books $\times$ 32 mood options). Each panel shows the moods with highest z-score relative to the other clusters and three exemplar books per cluster. Cluster sizes are 6 (warmth), 16 (darkness), 15 (melancholic-reflective), and 16 (adventure).

K Survey Instruments — Full Inventories

This appendix lists the items administered in each of the three surveys: pre-reading (Onboarding), chapter-level (Ancillary), and post-reading. All items are extracted from the source Qualtrics instruments; question text is lightly truncated to fit columns.

K.1 Pre-reading questionnaire (Onboarding, 11 items)

K.2 Chapter-level survey (Ancillary, 9 items per chapter)

K.3 Post-reading questionnaire (101 closed-form items)

Table 8 reports the full per-item table for the 101 closed-form post-reading questions with sufficient reader coverage, grouped by category. D (descriptive) items ask about verifiable narrative facts (number of protagonists, narration tense, presence of a turning point, mood-descriptor checkboxes

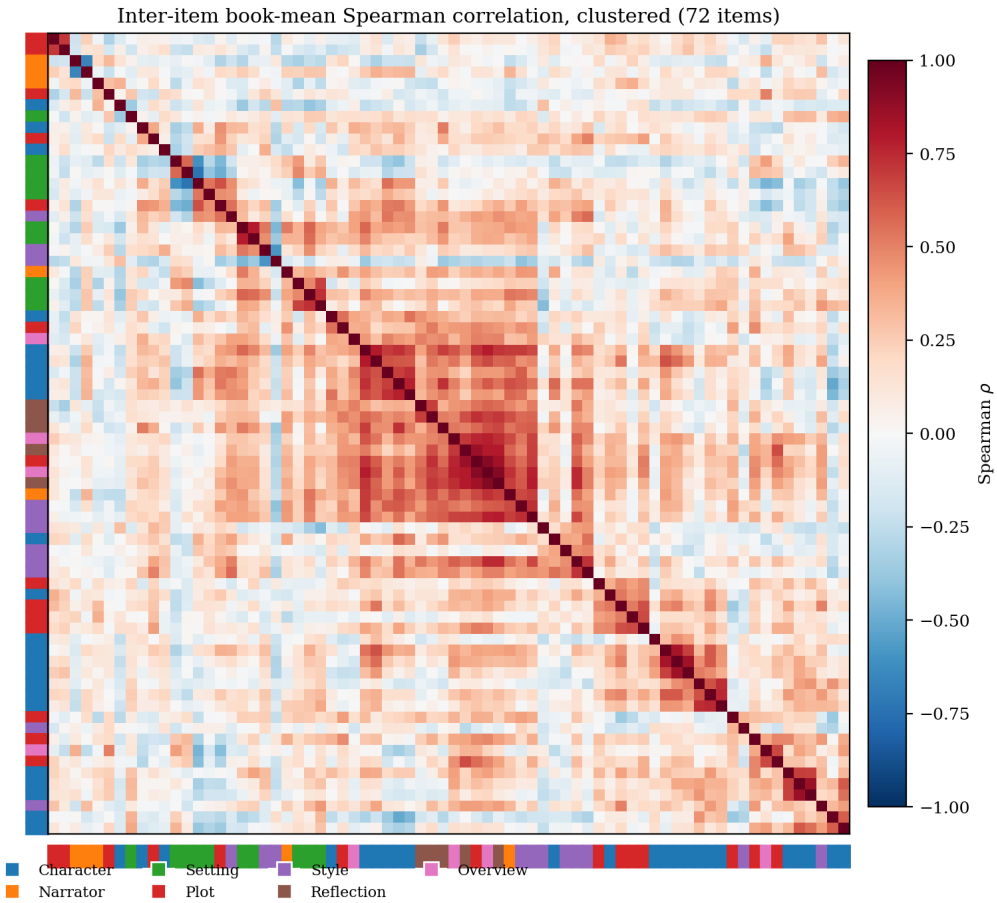

Figure 13: Inter-item Spearman correlation matrix on book-mean responses, 72×72 items reordered by hierarchical clustering (average linkage on $1 - |\rho|$). Side bars color items by subject category. Visible block structure indicates groups of items that co-vary across books; mean $|\rho| = 0.35$.

constrained by the text); E (experiential) items ask about the reader's subjective response (immersiveness, emotional impact, perceived realism, gladness-to-have-read). We caution that the D/E annotation was finalized after questionnaire design: ambiguous items (e.g., genre selection, mood descriptors) were tagged by the author team using majority judgment rather than independent multi-annotator review, and the labels are released as a per-item annotation rather than as a primary organizing axis of the dataset. The 55 open-ended post-reading items (plot summaries, character descriptions, theme reflections, etc.) are released in the dataset but not enumerated here. Question text is lightly truncated to fit the column; the full text, response option sets, branching logic, and per-(book, question) versions covering the 52 shared evaluation-subset books accompany the dataset release as supplementary spreadsheets.

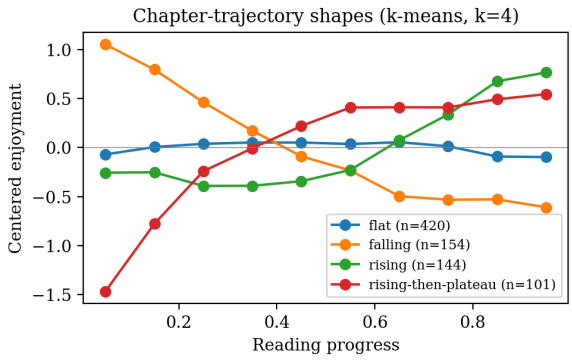

Figure 14: Chapter-by-chapter enjoyment trajectory shapes from k-means ($k = 4$, centered per-reader enjoyment over 10 reading-progress bins, $n = 819$ reader sessions with at least seven bins; this includes some sessions that completed enough chapter-level surveys for trajectory analysis but not the full post-reading questionnaire, hence $n > 809$). Approximately 51% of sessions have flat trajectories; the remainder split between rising, falling, and rising-then-plateau shapes.

Table 3: The 53 Project Gutenberg books in the evaluation subset, defined as all Gutenberg-released titles in the cleaned dataset with three or more readers, sorted by reader count.

Title	Author	n
This Side of Paradise	F. Scott Fitzgerald	8
Peter Pan	J. M. Barrie	7
Treasure Island	Robert Louis Stevenson	7
The Invisible Man	H. G. Wells	7
The Sea-Wolf	Jack London	7
A Princess of Mars	Edgar Rice Burroughs	7
The Tower Treasure	Franklin W. Dixon	6
Silas Marner	George Eliot	6
Carmilla	J. Sheridan Le Fanu	6
Daddy-Long-Legs	Jean Webster	5
The War of the Worlds	H. G. Wells	5
A Simple Story	Elizabeth Inchbald	5
Herland	Charlotte Perkins Gilman	4
Mrs. Dalloway	Virginia Woolf	4
The Castle of Otranto	Horace Walpole	4
Sky Island	L. Frank Baum	4
Jacob's Room	Virginia Woolf	4
The Samurai Strategy	Thomas Hoover	4
The Long Roll	Mary Johnston	4
The Picture of Dorian Gray	Oscar Wilde	4
The Secret Adversary	Agatha Christie	4
An Old-Fashioned Girl	Louisa May Alcott	4
The Enchanted April	Elizabeth von Arnim	3
The Napoleon of Notting Hill	G. K. Chesterton	3
The Jungle	Upton Sinclair	3
A Connecticut Yankee in King Arthur's Court	Mark Twain	3
At the Mountains of Madness	H. P. Lovecraft	3
Twenty Thousand Leagues under the Sea	Jules Verne	3
The Night Land	William Hope Hodgson	3
Emily of New Moon	L. M. Montgomery	3
The Age of Innocence	Edith Wharton	3
Around the World in Eighty Days	Jules Verne	3
Hard Times	Charles Dickens	3
The Call of the Wild	Jack London	3
My Antonia	Willa Cather	3
Lord Jim	Joseph Conrad	3
The Iron Heel	Jack London	3
A Christmas Carol	Charles Dickens	3
The Gadfly	Ethel Lilian Voynich	3
Pride and Prejudice	Jane Austen	3
Father Goriot	Honoré de Balzac	3
The House on the Borderland	William Hope Hodgson	3
Gulliver's Travels	Jonathan Swift	3
The Enchanted Castle	E. Nesbit	3
The Black Tulip	Alexandre Dumas	3
The Valley of the Moon	Jack London	3
Tom Swift in the Land of Wonders	Victor Appleton	3
The Time Machine	H. G. Wells	3
The Gods of Mars	Edgar Rice Burroughs	3
The Queen of Hearts	Wilkie Collins	3
The City of the Sun	Tommaso Campanella	3
Terminal Compromise	Winn Schwartau	3
Frankenstein	Mary Shelley	3

Table 4: Illustrative open-ended responses from the dataset.

Q: "Summarize the plot of the book." "Frederick Elliot is a student at Cambridge. He marries although he isn't in love and takes a position at a school even though he doesn't like it. He meets a half sibling through an Aunt who had ignored him and believes the sibling to be his father's child. He becomes friends with his half sibling and because he feels responsible for his brother, he is killed in an accident."
Q: "In 3–5 words or phrases, describe the key characteristics of this protagonist." "He is loyal, uses formal language, is precise, is unable to hold a normal 'casual' conversation, and is unable to express any form of affection."
Q: "What detracted from the book for you?" "There's quite a lot of racist and sexist elements to the book that I wasn't expecting, and broke the 'flow' of the book for me."

Table 5: Pairwise agreement on the 52 evaluation-subset books shared between the two models. *Llama* and *GPT-5* columns report model–human pairwise agreement. *Human* reports mean per-question pairwise inter-reader agreement on the same 52-book subset so all columns share a denominator. *Central-tend.* is a leave-one-out central-tendency baseline (per-format aggregation; App G).

Category	Llama	GPT-5	Human	Cent.-tend.
All questions	0.66	0.67	0.68	0.72
Character	0.72	0.73	0.72	0.77
Style	0.69	0.67	0.71	0.75
Plot	0.64	0.65	0.60	0.65
Narrator	0.67	0.70	0.75	0.78
Setting	0.59	0.63	0.73	0.76
Reflection	0.60	0.62	0.65	0.71
Overview	0.54	0.55	0.53	0.56

Table 6: Title-only contamination baseline. Both models re-evaluated with only book title and author in the system prompt (no book text), on the same 52-book evaluation subset. *Agree.*: mean per-question pairwise model–human agreement. *Mean* ρ: mean per-item cross-book Spearman ρ; full-text and title-only rows for each model are computed on the same 72-item set (the model's own full-text item set, intersected with its title-only set) so the two rows are directly comparable.

Model	Input	Agree.	Mean ρ	$\rho > 0$	$p < .05$
Llama	Full text	0.66	0.35	94%	65%
Llama	Title only	0.67	0.36	92%	61%
GPT-5	Full text	0.67	0.47	97%	83%
GPT-5	Title only	0.67	0.48	100%	88%

Table 7: Per-category Spearman–Brown-extended split-half reliability (r_{full}) over books with ≥ 4 readers, with implied per-item cross-book Spearman ceiling and observed model ρ vs. ceiling. Reliability and ceiling are computed on the $71/72$-item shared set used in Section 5; per-model ρ is computed on each model's own cross-book-coverage items (Narrator has 5 items for the ceiling vs. 6 for the model ρ, since one branched item meets the cross-book threshold but lacks a ≥ 4-reader book).

Category	n	r_{full}	**Ceil.**	**Llama ρ (%)**	**GPT-5 ρ (%)**
Setting	10	0.60	0.76	0.47 (62%)	0.60 (80%)
Character	23	0.56	0.72	0.42 (58%)	0.56 (77%)
Overview	4	0.51	0.65	0.18 (27%)	0.34 (53%)
Plot	14	0.43	0.62	0.39 (63%)	0.45 (72%)
Reflection	4	0.39	0.54	0.16 (29%)	0.36 (67%)
Narrator	5	0.34	0.55	0.20 (37%)	0.27 (50%)
Style	11	0.24	0.43	0.26 (60%)	0.41 (96%)
All	71	0.46	0.64	0.35 (55%)	0.47 (75%)

Qid	**Type**	**Question**
Q5	free-text (integer)	How many novels do you read in a typical year?
Q12	1–5 ordinal	Do you typically finish every book you start?
Q4	checkbox (genre)	What book genre(s) do you typically read?
Q6	checkbox (genre)	What book genre(s) do you believe this book belongs to?
Q16	1–3 ordinal	How familiar are you with this author?
Q17	1–3 ordinal	How familiar are you with this book?
Q9	1–7 numeric	How much do you believe you are the target reader for this book?
Q10	free-text	In a single sentence, who do you think the target reader for this book is?
Q7	1–5 ordinal	How fast-paced do you expect the book to be?
Q11	1–3 ordinal	How strongly do you expect to relate to the main characters?
Q8	1–5 numeric	How much do you expect to enjoy this book?

Qid	**Type**	**Question**
Q4	free-text	What page are you on?
Q5	1–5 ordinal	How surprising was the preceding section?
Q6	1–5 ordinal	How suspenseful was the preceding section?
Q7	1–5 ordinal	How uncertain do you feel about what's going to happen next?
Q16	yes/no	Were there any major plot twists in the preceding section?
Q17	free-text	If "yes", please briefly describe:
Q8	1–7 numeric	How much are you enjoying the book so far?
Q20	1–7 numeric	How much do you expect to enjoy the rest of the book?
Q21	1–7 numeric	How attentively engaged were you while reading the last chapter?

Table 8: Per-item table for all 101 closed-form post-reading questions. *Cat*: subject category (Over, Plot, Char, Narr, Sett, Styl, Refl). *D/E*: descriptive vs. experiential. *Type*: response format (num/ord/chk/cat). *IAA*: mean per-question pairwise inter-reader agreement on the full corpus. *Llama*/*GPT-5*: model–human pairwise agreement on the 52-book evaluation subset. ρ_L/ρ_G: per-item cross-book Spearman ρ for Llama and GPT-5. "–" indicates the item is not in the Spearman set (checkbox or categorical) or has insufficient cross-book coverage. Rows grouped by category, then by D/E, sorted within group by descending IAA.

Question	**Cat**	**D/E**	**Type**	**IAA**	**Llama**	**GPT-5**	ρ_L	ρ_G
What book genre(s) do you believe this book belongs to? Select all that apply.	Over	D	chk	0.57	0.43	0.50	–	–
To what extent would you describe the book as plot-driven and character-driven? While all books include both eleme…	Over	E	num	0.76	0.75	0.81	0.26	0.83
How closely do you think you fit the author's vision of the "ideal reader" for this book? In other words, how we…	Over	E	num	0.68	0.59	0.58	0.03	0.34
How did you feel about the book?	Over	E	num	0.67	0.68	0.68	0.22	0.25
How much did you like the book relative to your expectations based on the description, blurbs, or anything else yo…	Over	E	num	0.66	0.73	0.71	0.19	-0.04
How would you describe the mood - or, less formally, the "vibes" - of the book? Select all that apply from the a…	Over	E	chk	0.25	0.32	0.32	–	–
Please rank-order the three elements that most influenced how you feel about the book. (If it's really only one or…	Over	E	cat	0.21	0.24	0.24	–	–
How would you characterize its linearity in terms of time? Some books move in a straight sequence with events narr…	Plot	D	ord	0.83	0.85	0.84	0.50	0.33
How would you characterize its linearity in terms of plot? Some books move in a straight sequence with events narr…	Plot	D	ord	0.82	0.79	0.81	-0.03	0.25
Did the book feature one or more major turning points - where critical elements of the protagonists' lives were ne…	Plot	D	ord	0.71	0.76	0.73	0.50	0.24
Did the book begin with a "hook" - something dramatic, mysterious, or shocking that immediately made the stakes…	Plot	D	ord	0.68	0.68	0.69	0.55	0.47
Did the ending feature a major "sudden realization" on the part of the main character(s)?	Plot	D	cat	0.45	0.44	1.00	–	–
How would you characterize the structure of the book in terms of main plots and subplots? Some books have one main…	Plot	D	cat	0.44	0.29	0.23	–	–
Did the ending feature a major reversal-of-fortune for the main character(s)?	Plot	D	cat	0.43	0.42	0.00	–	–
Which story shape best describes this story's plot, if any?	Plot	D	cat	0.36	0.45	0.41	–	–
On a percent-of-book basis, please select the point in the novel where any major turning point(s) occurred.	Plot	D	chk	0.20	0.15	0.22	–	–
In your opinion, how important was the plot/pacing to your experience of the book as a whole, relative to other el…	Plot	E	ord	0.78	0.77	0.77	0.35	0.28
On a scale of 1 (very confusing) to 7 (very easy-to-understand)?	Plot	E	num	0.77	0.81	0.82	0.39	0.56
How related was the sequence of plot events throughout the book? If causal, each event led clearly to the next: A…	Plot	E	ord	0.76	0.78	0.79	0.48	0.39
On the whole, how would you describe the pace of the book? Pace is the speed at which the story develops - how qui…	Plot	E	ord	0.75	0.78	0.79	0.48	0.61
On a scale of 1 (entirely unresolved) to 7 (entirely resolved)?	Plot	E	num	0.74	0.79	0.77	0.56	0.71
On a scale of 1 (very unsatisfying) to 7 (very satisfying)?	Plot	E	num	0.70	0.74	0.76	0.67	0.73
On a scale of 1 (very unsurprising / predictable) to 7 (very surprising)?	Plot	E	num	0.70	0.68	0.69	0.17	0.33
How convincing did you find the plot? By convincing we mean believable.	Plot	E	num	0.69	0.71	0.75	0.36	0.67
How compelling did you find the plot? By compelling we mean interesting and absorbing.	Plot	E	num	0.67	0.70	0.69	0.43	0.52
On a scale of 1 (entirely arbitrary / could have easily happened another way) to 7 (entirely inevitable / it had t…	Plot	E	num	0.63	0.68	0.69	0.11	0.23
Did the ending feature a major emotional release, or other major expression of emotion, on the part of the main ch…	Plot	E	cat	0.41	0.49	0.50	–	–
Did you experience a notable "sag" in plot tension at any point in the book? Select all that apply. Please answe…	Plot	E	chk	0.20	0.38	0.15	–	–
Approximately how many named side characters are in this book? (Feel free to guesstimate)	Char	D	num	1.00	1.00	1.00	0.43	0.55
Did any members of the ensemble cast belong to any historically marginalized groups - i.e., groups systematically…	Char	D	chk	0.99	0.35	0.21	–	–
How many protagonists were there? A protagonist is the person or entity at the heart of the story - the principal…	Char	D	num	0.84	0.83	0.89	0.49	0.71
Does this protagonist belong to any historically marginalized groups - i.e., groups systematically denied access t…	Char	D	chk	0.84	0.53	0.51	–	–
What primary type of conflict, if any, did this protagonist face in the book? At a high level, conflict can take t…	Char	D	cat	0.52	0.51	0.60	–	–
What secondary type(s) of conflict, if any, did this protagonist face in the book? For example, in the first Hunge…	Char	D	cat	0.28	0.29	0.34	–	–
How important would you say the side characters are in this book?	Char	E	ord	0.82	0.82	0.82	0.14	0.47
How competent was this protagonist? By competent, we mean effective at the activities and tasks that matter to the…	Char	E	ord	0.81	0.84	0.83	0.37	0.49
In your opinion, how important were the characters to your experience of the book as a whole, relative to other el…	Char	E	ord	0.79	0.80	0.78	0.35	0.47
How proactive was this protagonist? By proactive, we mean motivated and taking steps to reach their goals. - Overa…	Char	E	ord	0.78	0.82	0.82	0.45	0.56
How proactive was this protagonist? By proactive, we mean motivated and taking steps to reach their goals. - End	Char	E	ord	0.77	0.82	0.80	0.50	0.46
Overall, would you say this protagonist's life or circumstances changed considerably over the course of the book?…	Char	E	ord	0.77	0.81	0.80	0.51	0.49

continued on next page

Table 8 continued from previous page.

Question	Cat	D/E	Type	IAA	Llama	GPT-5	ρ_L	ρ_G
How competent was this protagonist? By competent, we mean effective at the activities and tasks that matter to the…	Char	E	ord	0.76	0.82	0.84	0.42	0.70
How likable was this protagonist? By likable, we mean easy to relate to, empathize with, and root for. - Overall	Char	E	ord	0.76	0.77	0.81	0.29	0.57
How proactive was this protagonist? By proactive, we mean motivated and taking steps to reach their goals. - Middle	Char	E	ord	0.76	0.82	0.81	0.56	0.65
How competent was this protagonist? By competent, we mean effective at the activities and tasks that matter to the…	Char	E	ord	0.74	0.79	0.79	0.40	0.38
How likable was this protagonist? By likable, we mean easy to relate to, empathize with, and root for. - Middle	Char	E	ord	0.73	0.75	0.74	0.33	0.46
How likable was this protagonist? By likable, we mean easy to relate to, empathize with, and root for. - End	Char	E	ord	0.72	0.81	0.81	0.46	0.54
How competent was this protagonist? By competent, we mean effective at the activities and tasks that matter to the…	Char	E	ord	0.72	0.76	0.81	0.63	0.48
In your opinion, how would you rate the quality of the character writing in this book, on a scale of 1-7?	Char	E	num	0.72	0.71	0.74	0.59	0.75
How likable was this protagonist? By likable, we mean easy to relate to, empathize with, and root for. - Beginning	Char	E	ord	0.72	0.75	0.76	0.23	0.42
How real did this protagonist feel to you?	Char	E	num	0.71	0.70	0.74	0.32	0.74
Overall, would you say this protagonist experienced strong personal development over the course of the book? In ot…	Char	E	ord	0.71	0.72	0.76	0.62	0.76
How proactive was this protagonist? By proactive, we mean motivated and taking steps to reach their goals. - Begin…	Char	E	ord	0.69	0.76	0.74	0.63	0.46
How relatable did you find this protagonist?	Char	E	num	0.68	0.69	0.71	0.29	0.51
How emotionally invested were you in this protagonist?	Char	E	num	0.66	0.67	0.67	0.30	0.54
If you checked any of the above, how relevant are these aspect(s) of their identity to their experience in the boo…	Char	E	ord	0.65	0.62	0.71	0.39	0.66
How proactive was this protagonist? By proactive, we mean motivated and taking steps to reach their goals. - Begin…	Char	E	cat	0.30	0.45	0.67	–	–
To what extent did the needs and motivations of the ensemble cast reinforce or contradict each other? For example,…	Char	E	cat	0.27	0.00	0.29	–	–
In total, how many narrators were there? The narrator is the entity or point of view telling the story. Some books…	Narr	D	num	0.87	0.87	0.91	0.33	0.37
Approximately what share of the book was told from this narrator's point of view?	Narr	D	ord	0.80	0.77	0.79	0.38	0.38
How many times did the narration switch? For example, if A narrated the first half and B the second, the narration…	Narr	D	ord	0.80	0.65	0.68	0.11	-0.01
What type of narration was predominantly used? First person narration uses I/WE and comes directly from the charac…	Narr	D	cat	0.76	0.83	0.84	–	–
Was this narrator an omniscient narrator or an actual character featured in the book?	Narr	D	cat	0.75	0.75	0.79	–	–
Was the book narrated primarily in past or present tense? Past tense narration sounds like the story has already h…	Narr	D	cat	0.54	0.71	0.71	–	–
If third person narration was used, was it predominantly close or omniscient? In close third person, the narration…	Narr	D	cat	0.50	0.50	0.54	–	–
What word would you choose to describe this narrator's overall tone? Select all that apply from the alphabetized l…	Narr	E	chk	0.91	0.00	0.00	–	–
How reliable or unreliable was this narrator? A reliable narrator is trustworthy - even if they don't have all the…	Narr	E	num	0.83	0.63	0.84	-0.04	0.14
In your opinion, how important was the narration and POV to your experience of the book as a whole, relative to ot…	Narr	E	ord	0.77	0.79	0.79	0.26	0.37
In your opinion, how would you rate the quality of the narration and POV writing in this book, on a scale of 1-7?	Narr	E	num	0.73	0.76	0.74	0.18	0.40
On a percent-of-book basis, at what point(s) was the book set here? Please answer on a length basis - and check pa…	Sett	D	chk	0.86	0.07	0.01	–	–
Where does the book take place? Select all that apply.	Sett	D	chk	0.83	0.48	0.48	–	–
On a percent-of-book basis, at what point(s) was the book set here?	Sett	D	chk	0.78	0.08	0.06	–	–
How many major settings does the book contain? Here, we're asking you to count at a high-level and focus on major…	Sett	D	num	0.78	0.79	0.81	0.43	0.47
When does the book take place? Select all that apply.	Sett	D	chk	0.59	0.42	0.59	–	–
If more than one major setting, did the book move linearly from one setting to the next, or did it alternate regul…	Sett	D	cat	0.43	0.46	0.55	–	–
How science-fictional was the world of the book? By science-fictional we mean a world of alternative ways of life…	Sett	E	num	0.90	0.86	0.90	0.84	0.88
How fantastical was the world of the book? By fantastical we mean a world that would be characterized as "fantasy…	Sett	E	num	0.84	0.83	0.88	0.78	0.86
How realistic was the world of the book? By realistic we mean plausible - a faithful or true-to-life representatio…	Sett	E	num	0.78	0.79	0.83	0.58	0.80
In your opinion, how would you rate the quality of the setting and world-building writing in this book, on a scale…	Sett	E	num	0.73	0.81	0.73	0.33	0.44
In your opinion, how important was the setting and world-building to your experience of the book as a whole, relat…	Sett	E	ord	0.73	0.81	0.79	0.48	0.64
How vivid did you find this setting? By vivid we mean easy to see in your mind's eye.	Sett	E	num	0.71	0.68	0.70	0.17	0.38
How immersive did you find this setting? By 'immersive' we mean fully realized, with all the details fleshed-out a…	Sett	E	num	0.71	0.68	0.68	0.32	0.48
Overall, how relevant was world-building to the book? World-building is defined as the process of developing a det…	Sett	E	ord	0.67	0.67	0.73	0.37	0.64

continued on next page

Table 8 continued from previous page.

Question	Cat	D/E	Type	IAA	Llama	GPT-5	ρ_L	ρ_G
How familiar was the world of the book? By familiar we mean a setting that you know well, either from your own liv…	Sett	E	num	0.66	0.74	0.76	0.40	0.45
Which words best describe the writing style? Select all that apply. (other: free-text)	Styl	D	chk	0.88	0.24	0.00	–	–
Which words best describe the writing style? Select all that apply.	Styl	D	chk	0.24	0.24	0.24	–	–
In your opinion, how was the quantity of dialogue throughout the book?	Styl	E	ord	0.86	0.85	0.85	0.35	0.55
In your opinion, how was the amount of showing and telling throughout the book? - Showing	Styl	E	ord	0.81	0.84	0.81	-0.31	0.27
In your opinion, how was the amount of showing and telling throughout the book? - Telling	Styl	E	ord	0.80	0.78	0.78	0.21	0.04
How clear did you find the writing style? By clear we mean articulate, coherent, and easy to understand.	Styl	E	num	0.78	0.76	0.81	0.42	0.61
How important was the dialogue to your experience of the book as a whole?	Styl	E	ord	0.77	0.79	0.81	0.44	0.53
Now, let's talk (pun intended) about dialogue! In your opinion, how was the quality of dialogue throughout the boo…	Styl	E	ord	0.77	0.75	0.79	0.27	0.52
In your opinion, how important was the overall writing style to your experience of the book as a whole, relative t…	Styl	E	ord	0.74	0.83	0.81	0.33	0.24
How engaging did you find the dialogue in the book?	Styl	E	num	0.71	0.72	0.76	0.29	0.51
How realistic did you find the dialogue in the book?	Styl	E	num	0.70	0.75	0.76	0.26	0.37
How evocative did you find the writing style? By evocative we mean expressive, eloquent, and memorable.	Styl	E	num	0.68	0.68	0.69	0.34	0.49
How interesting did you find the writing style? By interesting we mean dynamic, varied, and creative.	Styl	E	num	0.66	0.69	0.66	0.20	0.36
To what extent did you feel moved by this book on an emotional basis? An intellectual basis? - Intellectually	Refl	E	num	0.69	0.56	0.61	0.12	0.34
To what extent did you feel moved by this book on an emotional basis? An intellectual basis? - Emotionally	Refl	E	num	0.67	0.59	0.64	0.41	0.63
How interesting do you personally find the book's themes?	Refl	E	ord	0.67	0.61	0.62	0.18	0.26
After having read the book, how glad are you to have devoted time and energy to it? In other words, how worthwhile…	Refl	E	ord	0.65	0.63	0.62	-0.07	–
While you were in the process of reading the book, how enjoyable did you find it? In other words, how much did you…	Refl	E	ord	0.64	0.59	0.60	-0.07	0.22