TEXT AS THE RICHEST PREFERENCE SIGNAL

André F. Cruz[1] **Jon Kleinberg**[2] **Rediet Abebe**[1,3]
[1]Max Planck Institute for Intelligent Systems, Tübingen and Tübingen AI Center
[2]Cornell University [3]ELLIS Institute, Tübingen

ABSTRACT

Preference elicitation algorithms have long relied on structured representations of user preferences: rankings of items, ratings, or simple binary interactions (e.g., views). Over the years, we've slowly become aware of the limitations and biases these representations entail. Users form preferences over items' features rather than items themselves. In this paper, we explore *natural language* as a first-class preference representation, beyond a mere cold-start aid. We study three parallel representations of user preferences: (i) a user-item interaction matrix, (ii) free-form text profiles describing users' preferences, and (iii) interpretable tabular features derived by an LLM from these text profiles. Our findings unfold in three parts. First, text-based predictors substantially outperform collaborative filtering in the cold-start regime and remain competitive as interaction histories grow. Second, most of the predictive signal in text can be retained in a compact, interpretable tabular representation. Third, the three representations are complementary: simple ensembles that combine them consistently achieve the strongest performance. *

1 INTRODUCTION

Human preferences play a central role in many decision-making systems, from allocating public resources and matching students to schools, to recommending content. In all such settings, preferences must be translated into a formal representation that can be stored, compared, and optimized. These representations are rarely neutral: They are shaped by what can be easily elicited at scale and by the algorithms available to process them.

Most large-scale systems rely on *forced-choice* signals: explicit ratings, binary likes, or implicit interaction traces such as clicks and views. These signals are abundant and admit scalable, well-studied algorithms such as collaborative filtering (Resnick et al., 1994; Linden et al., 2003; Hu et al., 2008; Koren et al., 2009; Rendle et al., 2020). However, forced-choice signals are intrinsically lossy: they encode *what* a user chose, but not *why*; they conflate preference with exposure; and they degrade sharply in cold-start settings, where historical data is sparse (Chevaleyre et al., 2008; Robertson & Salehi, 2020). By contrast, users routinely articulate preferences directly in natural language, expressing attributes, trade-offs, and contextual explanations rather than judgments over isolated items. Early ML researchers classified Iris flowers from tabular petal measurements (Fisher, 1936), not because a table of four numbers described a flower better than a photograph, but because models at the time could not process images. Forced-choice interactions occupy a similar role in preference learning: They dominate not because a binary like or a five-star rating captures what users want better than their own words, but because interaction signals are easy to collect and process at scale.

Historically, free-form text has been difficult to operationalize at scale. Earlier systems incorporated language through bag-of-words features, topic models, or hand-engineered pipelines (Wang & Blei, 2011; Lops et al., 2011; McAuley & Leskovec, 2013; Zhang et al., 2014), but these approaches were brittle and task-specific. Even with neural text models, text-only approaches generally failed to outperform interaction-based methods (Rendle et al., 2020; Zangerle & Bauer, 2022). Recent advances in large language models (LLMs) change this picture: Modern LLMs can abstract and categorize user-authored text at scale, producing stable, semantically meaningful representations without task-specific supervision (Gilardi et al., 2023; Ziems et al., 2024). This makes it possible to

[*] Preliminary work, prepared as a non-archival workshop abstract.

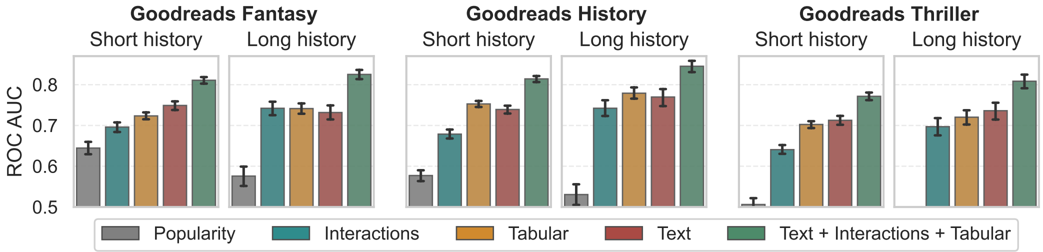

Figure 1: Predictive performance (ROC AUC) on three Goodreads subsets. Free-form text (zero-shot LLM) consistently outperforms or matches collaborative filtering trained on 2–4M domain-specific interactions. Gains are largest for short histories ($k < 10$); long profiles correspond to $k \geq 40$. Error bars: 90% CI.

revisit a longstanding preference elicitation question: When text is available, how informative is it relative to forced-choice interaction data?

1.1 Our Contributions

This paper studies natural language as a first-class source of preference information. We acknowledge a persistent asymmetry: While text is inherently more expressive, interaction data is generated passively at scale. Our experiments preserve this asymmetry: Interaction-based models train on abundant historical data, while text-based representations use off-the-shelf LLMs without fine-tuning.

Our contributions can be summarized by three main findings, showcased by Figure 1:

- **Free-form text is a competitive alternative to abundant interaction data.** Text-based representations outperform collaborative filtering in cold-start regimes and remain competitive as interaction histories grow; CF reaches parity only around $k{\approx}60$ observed interactions.
- **LLM-derived tabular features recover interpretable preference structure.** Categorical features extracted from text reviews capture meaningful, human-interpretable dimensions of preference and often achieve performance close to free-form text.
- **Preference representations carry complementary signal.** Interaction data, free-form text, and tabular features encode distinct aspects of user preferences; simple score-averaging ensembles achieve the strongest overall performance.

2 Related Work

Preference elicitation and representation. The chosen preference language (rankings, cardinal utilities, or restricted reports) is itself a design decision that determines which welfare-relevant trade-offs can be communicated (Chevaleyre et al., 2008; Dütting et al., 2011). Many real preferences cannot be faithfully captured by forced-choice primitives: Binary feedback cannot express conditional or multi-attribute constraints (e.g., "I like political thrillers only when they are not graphic") (Chevaleyre et al., 2008), and elicitation interfaces can introduce systematic biases (Robertson & Salehi, 2020). Interaction logs are typically treated as revealed preferences, but they are entangled with exposure and platform dynamics, complicating inference about true user goals (Kleinberg & Sandler, 2004; Kleinberg et al., 2022). Language-based profiles can be seen as a stated-preference channel that additionally carries explanations and conditional constraints invisible in interaction traces (Robertson & Salehi, 2020; Louviere et al., 2000; Varian, 2012).

Textual preference signals in recommender systems. Text has been used in recommender systems for decades, but early NLP pipelines typically reduced language to sparse tabular signals via TF–IDF (Salton & Buckley, 1988), topic models (e.g., Wang & Blei, 2011), or review-based latent factor models (McAuley & Leskovec, 2013; Zhang et al., 2014). These approaches relied on brittle feature extraction, so text was often most helpful as side information in cold-start scenarios (Lops et al., 2011; Rendle et al., 2020; Zangerle & Bauer, 2022). Recent LLMs enable robust, zero-shot abstraction over user-authored text: Sanner et al. (2023) finds LLMs competitive near the cold-start regime, and Zheng et al. (2024) shows that LLMs can leverage text-rich signals in sequential recommendation. Still, most controlled studies use limited interaction training data and rely on experimentally elicited preference text that is denser than organic free-form inputs.

Interpretable representations of text preference data. Within LLM alignment, a growing body of work extracts structured, interpretable representations from free-form feedback. Preference datasets often admit low-dimensional representations where a small number of features capture much of the signal used by dense reward models (Movva et al., 2025; Laguna et al., 2025). Related work frames alignment as a compression problem, distilling pairwise judgments into concise principles or constitutions (Findeis et al., 2025; Bell et al., 2026), and operationalizes rubric-style evaluation or editable textual profiles for transparency (He et al., 2025; Ramos et al., 2024; Zhou et al., 2024). These directions establish that unconstrained textual feedback can be transformed into compact, interpretable structures, motivating our study of whether similar tabularization preserves preference signal at scale.

Gap and contribution. Prior evidence that text encodes useful preference information comes largely from controlled settings with clean, purpose-collected text (Sanner et al., 2023; Kang et al., 2024; Ramos et al., 2024; Movva et al., 2025; Findeis et al., 2025), and does not address the asymmetry present in deployed systems: Abundant interaction logs versus scarce, noisy, organic preference text. We provide the first head-to-head comparison under these realistic conditions, using large-scale benchmark data from a real-world platform.

3 Experimental Setup

We compare alternative preference representations derived from the same underlying users, items, and observed behaviors, using simple modeling choices to minimize confounds.

3.1 Data

We use a large-scale public dataset from `Goodreads.com` (Wan & McAuley, 2018), the world's largest book cataloging platform, where users track, rate, and review books. We conduct experiments on three genre-specific subsets ("Fantasy & Paranormal", "Mystery, Thriller & Crime", and "History & Biography") with 2.2M–4.2M training interactions per genre. Throughout the paper, an *interaction* corresponds to a book that was both read and rated by a user. Interactions and associated review text form the common substrate from which all preference representations are derived (Figure A.1 in Appendix A). Full split statistics and preprocessing details are in Appendix A.

3.2 Preference representations

The dataset has two preference information sources: User–item interactions and user-written reviews.

Interaction matrix. The standard representation is a sparse user–item matrix $\mathbf{R}$ where $R_{u,i}$ indicates whether user u interacted with item i. This formulation underlies much recommender-system research (Resnick et al., 1994; Linden et al., 2003; Bennett & Lanning, 2007; Hu et al., 2008; He et al., 2017) and provides abundant data for matrix-factorization models, but is not directly interpretable and can reflect exposure effects rather than true preference.

Text preference descriptions. We use a pretrained LLM (GPT-5; Singh et al., 2025) as a zero-shot summarizer to produce a concise description of each user's reading preferences. For each user, we construct a preference context aggregating book synopses and the user's reviews, supplemented with frequency statistics over their most-read authors and sub-genres. The model generates a compact textual profile (at most three paragraphs) describing general preferences; the length constraint acts as an information bottleneck, favoring stable cross-title themes over item-specific detail. Prompt templates are in Appendix C.

Tabular features. We also derive a structured, tabular representation from LLM-generated categorical tags (e.g., `slow-burn`, `female-lead`, `complex-world-building`). Tags are generated without a predefined ontology, then embedded using `Qwen3-8B-Embedding` (Zhang et al., 2025) and clustered into K groups by cosine similarity (K selected on validation data). Each user is represented by a vector over tag clusters; items are placed in the same space by embedding book descriptions and assigning the top l=10 nearest tag clusters. Preference prediction reduces to the dot product $f(u, b) = u^\top \cdot b$, which is fully interpretable, unsupervised, and model-agnostic.

3.3 Models

We pair each representation with a simple, well-established model that maps it to a scalar score for ranking items; the goal is to isolate the predictive value of the underlying signal, not to optimize model architecture.

Interaction-based models. We use collaborative filtering (CF) based on matrix factorization for implicit feedback (Hu et al., 2008; Takács et al., 2011), implemented with the `implicit` library (Frederickson, 2016). At evaluation time, we fold in each test user by solving a regularized least-squares problem for that user's latent vector, holding item factors fixed (He et al., 2016). We also include a popularity baseline that ranks items by total training interactions (user-independent).

Text-based model. We use zero-shot prompting of GPT-5 (Singh et al., 2025). For each user-item pair, the prompt includes the user's text profile, the item description, and up to five community reviews (from the train set). The model outputs a score from 0 to 10. Prompt details are in Appendix C.

Tabular model. The score for a user-item pair is the dot product between their respective feature vectors, without any task-specific training.

3.4 Evaluation

We evaluate how well each representation predicts held-out user–item interactions under a shared ranking task; Figure A.2 (Appendix A) summarizes the data split.

Split in train, validation, and test. We split at the user level. Validation and test each contain 1 000 users sampled uniformly at random among users who (i) interacted with at least 15 items and (ii) wrote at least 5 textual reviews. The training set contains all remaining users.

Profile vs. evaluation interactions. Each validation/test user's interactions are split into a *profile* set and an *evaluation* set. The evaluation set consists of 10 randomly selected interactions (prioritizing those without reviews); the profile set contains the remaining interactions with an associated review. We denote by k the number of profile interactions available for a user; k varies across users, enabling evaluation from cold-start (small k) to information-rich (large k) regimes.

Candidate set and metric. For each user, we form 50 candidates: 10 held-out positives and 40 negatives sampled from unread items (weighted by popularity). Each model scores all 50, and we report mean per-user ROC AUC (additional ranking metrics in Appendix B.1).

4 Results

We now present results organized around our three findings. Throughout, each point on the x-axis corresponds to a cohort of users with k profile interactions; Appendix B contains additional results.

4.1 Text vs. user–item interactions

Figure 2 compares a text-based predictor (zero-shot GPT-5) to collaborative filtering (CF) as a function of profile length k. Across all three datasets, the text-based model substantially outperforms CF in the cold-start regime. This advantage narrows as k increases; only around k≈60 profile interactions does CF reach parity with a zero-shot text model. This reflects how each approach

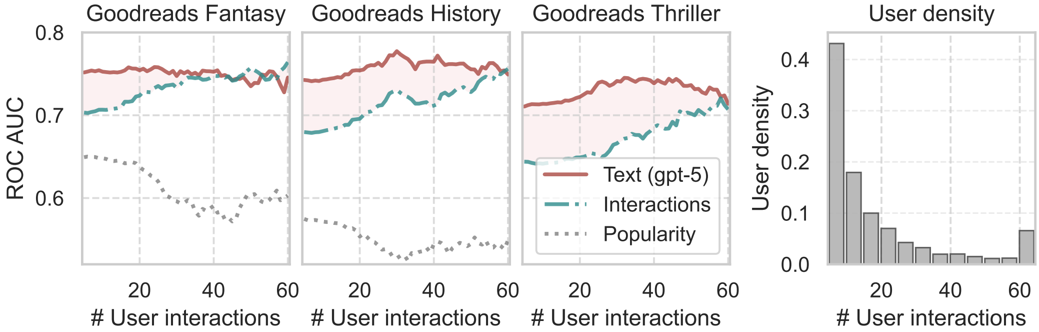

Figure 2: ROC AUC as a function of user profile length (k); rolling window $k \pm 10$. The text-based model (red) outperforms collaborative filtering (teal) at cold start and remains competitive throughout, reaching parity around $k \approx 60$. Right-most plot: user density. Figure B.4 extends to $k = 100$ profiles.

incorporates new information: CF directly optimizes user representations via folding-in on each new interaction, yielding predictable gains; the text-based predictor uses a fixed LLM, so increasing k refines the profile but does not adapt the model itself.

Increasing k also corresponds to a harder prediction problem, as users with longer histories tend to read less popular items (Figure A.3); despite this, the text-based model maintains stable absolute performance, indicating that text profiles grow increasingly informative as more reviews are added.

Takeaway: User-authored text remains competitive with interaction-based models across the entire profile-length spectrum, not just at cold start.

4.2 Tabular features recover interpretable latent categories

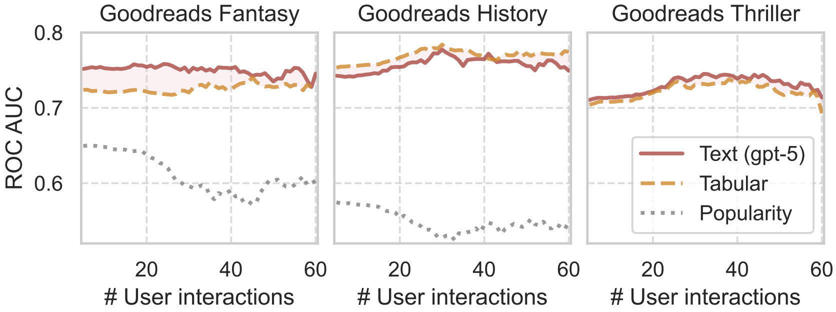

Figure 3: ROC AUC vs. profile length k. Tabular representations closely match text-based profiles (used as LLM inputs). Both are derived from the same user-authored reviews.

The previous comparison establishes text as a strong preference signal. However, the mechanism by which an LLM maps text profiles to scores is not directly interpretable. To test whether text's predictive advantage can be recovered in an explicitly interpretable form, we derive discrete tabular features from the same reviews (Section 3.2); performance peaks around $K \approx 30$ clusters (Appendix B).

Figure 3 shows that the tabular model closely tracks LLM-on-text performance across datasets and user cohorts: An explicitly interpretable, low-dimensional tabularization preserves most of the predictive power of end-to-end text models.

Takeaway: Much of the signal in free-form text can be retained in a compact, interpretable tabular representation.

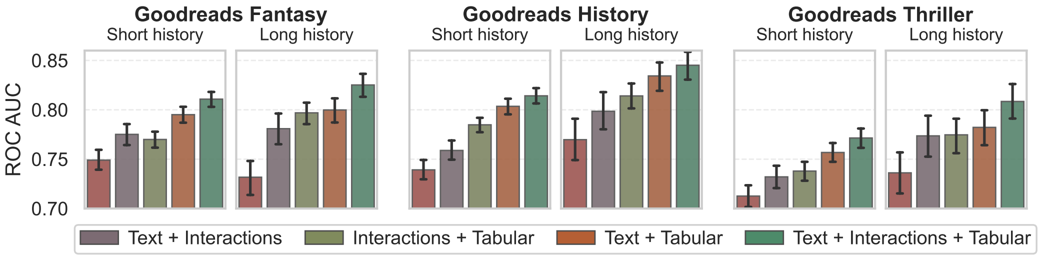

Figure 4: Score-averaging ensembles across preference representations. All pairwise combinations outperform their constituents; the three-way ensemble performs best. The unlabeled left-most bar is the Text-only baseline.

4.3 Complementary preference signals

Having shown that text and tabular representations each carry strong individual signal, we now test whether they capture complementary information. We construct simple ensembles that average predicted scores; Figure 4 reports all two- or three-model combinations with 90% confidence intervals.

Ensembling text and tabular representations yields consistent gains despite both deriving from users' reviews, indicating that each preserves distinct aspects of the preference signal. Similar improvements arise when combining text and interaction-based models. The three-way ensemble achieves the strongest performance across all user cohorts; this pattern holds for Precision@10, NDCG@10, and Recall@10 (Appendix B.1).

Takeaway: Text is the strongest single representation, but all three signals are complementary and can be jointly exploited with simple ensembles.

5 Discussion and Conclusion

This paper revisits a foundational assumption in preference learning: That abundant interaction data is sufficient to infer users' preferences. Our comparison deliberately preserves the data advantage of interaction-based methods: Collaborative filtering trains on millions of interactions, while text-based representations use an off-the-shelf LLM without fine-tuning. Interaction data, though abundant, is incidental: It reflects constrained interface choices rather than deliberate preference expression. Text, by contrast, directly encodes users' goals and trade-offs.

Our results suggest that free-form text is a first-class preference representation, not merely a cold-start fallback. Text-based predictors match or outperform collaborative filtering, compress into interpretable features with little loss of signal, and complement interaction data: Combining all three representations consistently yields the best results.

However, several practical limitations warrant discussion. First, LLM-based scoring is orders of magnitude more costly than dot-product matrix factorization. Second, our evaluation is static, leaving open how richer preference representations interact with strategic behavior and incentive design. Lastly, our evaluation is limited to a single platform (Goodreads), restricting to users who authored reviews. On most platforms today, the volume of user-authored text is likely insufficient to replace interaction-based methods at scale. Our results show, however, that user-authored text is more predictive than interaction data alone. Given that this benefits both users (increased transparency and control) and platforms (higher-quality preference signals), we expect a text-first approach to representing user preferences can be promoted and broadened to a much wider user audience.

Taken in whole, these findings have broad implications for preference elicitation and mechanism design. Rather than inferring intent indirectly from behavioral traces, platforms can elicit preferences directly in natural language, giving users visibility and agency over how they are represented. Our results support a simple design principle: When possible, ask users what they want, and let them answer in their own words.

Acknowledgments

The authors are grateful to Moritz Hardt for invaluable feedback throughout this work. We also thank Andreas Bjerre-Nielsen, Amin Charusaie, Yatong Chen, and Ricardo Dominguez-Olmedo for helpful discussions and comments. We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting André F. Cruz.

References

Henry Bell, Lara Neubauer da Costa Schertel, Bochu Ding, and Brandon Fain. Beyond preferences: Learning alignment principles grounded in human reasons and values. *arXiv preprint arXiv:2601.18760*, 2026.

J. Bennett and S. Lanning. The netflix prize. In *Proceedings of the KDD Cup Workshop 2007*, pp. 3–6, New York, August 2007. ACM. URL `http://www.cs.uic.edu/~liub/KDD-cup-2007/NetflixPrize-description.pdf`.

Yann Chevaleyre, Ulle Endriss, Jérôme Lang, and Nicolas Maudet. Preference handling in combinatorial domains: From ai to social choice. *AI magazine*, 29(4):37–37, 2008.

Paul Dütting, Felix Fischer, and David C Parkes. Simplicity-expressiveness tradeoffs in mechanism design. In *Proceedings of the 12th ACM conference on Electronic commerce*, pp. 341–350, 2011.

Arduin Findeis, Timo Kaufmann, Eyke Hüllermeier, Samuel Albanie, and Robert D. Mullins. Inverse constitutional AI: Compressing preferences into principles. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=9FRwkPw3Cn`.

R. A. Fisher. Iris. UCI Machine Learning Repository, 1936. DOI: https://doi.org/10.24432/C56C76.

Ben Frederickson. implicit: Fast python collaborative filtering for implicit feedback datasets, 2016. URL `https://github.com/benfred/implicit`. Software available from `https://github.com/benfred/implicit`.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023. doi: 10.1073/pnas.2305016120. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2305016120`.

Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. Fast matrix factorization for online recommendation with implicit feedback. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pp. 549–558, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340694. doi: 10.1145/2911451.2911489. URL `https://doi.org/10.1145/2911451.2911489`.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pp. 173–182, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349130. doi: 10.1145/3038912.3052569. URL `https://doi.org/10.1145/3038912.3052569`.

Yun He, Wenzhe Li, Hejia Zhang, Songlin Li, Karishma Mandyam, Sopan Khosla, Yuanhao Xiong, Nanshu Wang, Selina Peng, Beibin Li, et al. Rubric-based benchmarking and reinforcement learning for advancing llm instruction following. *arXiv preprint arXiv:2511.10507*, 2025.

Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 263–272, 2008. doi: 10.1109/ICDM.2008.22.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

Inwon Kang, Sikai Ruan, Tyler Ho, Jui-Chien Lin, Farhad Mohsin, Oshani Seneviratne, and Lirong Xia. Llm-augmented preference learning from natural language. In *Proceedings of the FMGT Workshop at EC 2024*, ACM Conference on Economics and Computation (EC), 2024. URL https://arxiv.org/abs/2310.08523. Workshop on Foundation Models and Game Theory (FMGT).

Jon Kleinberg and Mark Sandler. Using mixture models for collaborative filtering. In *Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '04, pp. 569–578, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138520. doi: 10.1145/1007352.1007439. URL https://doi.org/10.1145/1007352.1007439.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. The challenge of understanding what users want: Inconsistent preferences and engagement optimization. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, pp. 29, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391504. doi: 10.1145/3490486.3538365. URL https://doi.org/10.1145/3490486.3538365.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009. doi: 10.1109/MC.2009.263.

Sonia Laguna, Katarzyna Kobalczyk, Julia E Vogt, and Mihaela Van der Schaar. Interpretable reward modeling with active concept bottlenecks. *arXiv preprint arXiv:2507.04695*, 2025.

Greg Linden, Brent Smith, and Jeremy York. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80, 2003.

Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. *Content-based Recommender Systems: State of the Art and Trends*, pp. 73–105. Springer US, Boston, MA, 2011. ISBN 978-0-387-85820-3. doi: 10.1007/978-0-387-85820-3_3. URL https://doi.org/10.1007/978-0-387-85820-3_3.

Jordan J Louviere, David A Hensher, and Joffre D Swait. *Stated choice methods: analysis and applications*. Cambridge university press, 2000.

Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pp. 165–172, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450324090. doi: 10.1145/2507157.2507163. URL https://doi.org/10.1145/2507157.2507163.

Rajiv Movva, Smitha Milli, Sewon Min, and Emma Pierson. What's in my human feedback? learning interpretable descriptions of preference data. *arXiv preprint arXiv:2510.26202*, 2025.

Jerome Ramos, Hossein A. Rahmani, Xi Wang, Xiao Fu, and Aldo Lipani. Transparent and scrutable recommendations using natural language user profiles. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13971–13984, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.753. URL https://aclanthology.org/2024.acl-long.753/.

Steffen Rendle, Walid Krichene, Li Zhang, and John Anderson. Neural collaborative filtering vs. matrix factorization revisited. In *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys '20, pp. 240–248, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375832. doi: 10.1145/3383313.3412488. URL https://doi.org/10.1145/3383313.3412488.

Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, CSCW '94, pp. 175–186, New York, NY, USA, 1994. Association for Computing Machinery. ISBN 0897916891. doi: 10.1145/192844.192905. URL https://doi.org/10.1145/192844.192905.

Samantha Robertson and Niloufar Salehi. What if I don't like any of the choices? the limits of preference elicitation for participatory algorithm design. *arXiv preprint arXiv:2007.06718*, 2020.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

Scott Sanner, Krisztian Balog, Filip Radlinski, Ben Wedin, and Lucas Dixon. Large language models are competitive near cold-start recommenders for language-and item-based preferences. In *Proceedings of the 17th ACM conference on recommender systems*, pp. 890–896, 2023.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Gábor Takács, István Pilászy, and Domonkos Tikk. Applications of the conjugate gradient method for implicit feedback collaborative filtering. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pp. 297–300, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306836. doi: 10.1145/2043932.2043987. URL https://doi.org/10.1145/2043932.2043987.

Hal R Varian. Revealed preference and its applications. *The Economic Journal*, 122(560):332–338, 2012.

Mengting Wan and Julian McAuley. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, pp. 86–94, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359016. doi: 10.1145/3240323.3240369. URL https://doi.org/10.1145/3240323.3240369.

Chong Wang and David M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pp. 448–456, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450308137. doi: 10.1145/2020408.2020480. URL https://doi.org/10.1145/2020408.2020480.

Eva Zangerle and Christine Bauer. Evaluating recommender systems: Survey and framework. *ACM Comput. Surv.*, 55(8), December 2022. ISSN 0360-0300. doi: 10.1145/3556536. URL https://doi.org/10.1145/3556536.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pp. 83–92, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450322577. doi: 10.1145/2600428.2609579. URL https://doi.org/10.1145/2600428.2609579.

Zhi Zheng, WenShuo Chao, Zhaopeng Qiu, Hengshu Zhu, and Hui Xiong. Harnessing large language models for text-rich sequential recommendation. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pp. 3207–3216, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645358. URL https://doi.org/10.1145/3589334.3645358.

Joyce Zhou, Yijia Dai, and Thorsten Joachims. Language-based user profiles for recommendation. *arXiv preprint arXiv:2402.15623*, 2024.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291, 03 2024. ISSN 0891-2017. doi: 10.1162/coli_a_00502. URL https://doi.org/10.1162/coli_a_00502.

A `Goodreads.com` Dataset

This appendix provides additional details on the `Goodreads.com` dataset (Wan & McAuley, 2018) used throughout the paper. Its purpose is to document dataset composition, preprocessing, and evaluation splits in sufficient detail to support reproducibility, while keeping the main text focused on modeling and empirical comparisons. We also include exploratory analyses that help contextualize the observed performance of different baselines.

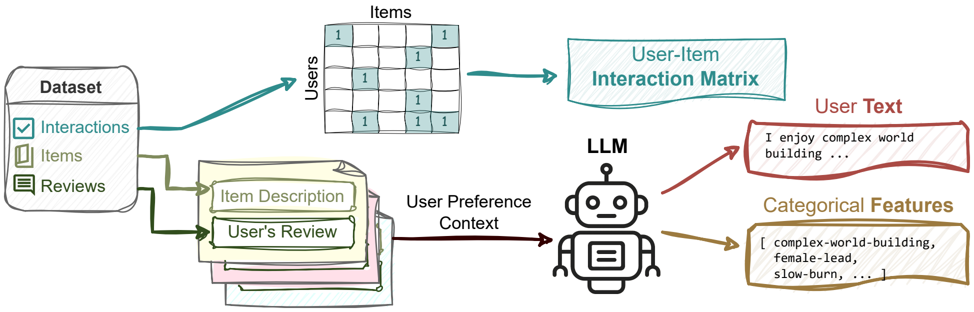

Figure A.1: Deriving parallel preference representations from a shared dataset. User–item interactions form an interaction matrix. Users' reviews and respective item descriptions are processed by an LLM to produce, for each user, a condensed text profile and categorical features.

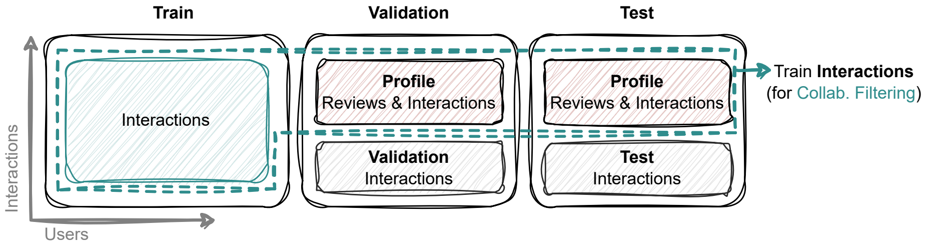

Figure A.2: Diagram of evaluation data split. Interaction-based models are fitted on all *train* interactions and users' *profile* interactions from validation and test. For each user, text-based (or tabular-based) models use their text profile (or tabular features), which are derived from the user's profile reviews.

A.1 Dataset structure and interaction types

The original Goodreads dataset contains multiple forms of user–item interactions, including reading events, ratings, and free-form reviews. These interaction types differ in both coverage and semantic meaning, and raw interaction counts therefore differ from the number of books that were explicitly read or rated. Throughout the paper, we distinguish between raw interaction records and the cleaned interactions used for modeling. Raw records include a read indicator (and, when present, start/end dates), an explicit star rating, and an optional free-form review text. A record may contain any subset of these fields, so review counts are lower than read/rating counts.

A.2 Data preprocessing and filtering

We apply a set of conservative filtering steps to improve data quality and ensure a consistent interpretation of user–item interactions. Specifically, we retain only records with a non-null, strictly positive rating and a positive read indicator. To remove implausible or incomplete reading events, we require both a recorded start date and completion date, with at least two calendar days elapsed between them. For review text, we discard short reviews and retain only reviews exceeding a minimum character-length threshold of 500 characters.

	Train		Test		Validation	
Genre	Users	Interactions	Users	Profile / Eval Int.	Users	Profile / Eval Int.
Fantasy	333K	4.2M	1K	23K / 42K	1K	24K / 43K
Thriller	272K	2.2M	1K	21K / 35K	1K	20K / 36K
History	317K	2.3M	1K	20K / 31K	1K	19K / 29K

Table A.1: Summary of dataset statistics after data cleaning and train/test/validation split. For test and validation users, profile interactions are used to construct preference representations, while evaluation interactions are held out for performance evaluation.

After filtering, each remaining interaction corresponds to a book that was both read and explicitly rated by the user. Unless otherwise stated, all reported interaction statistics and experimental results are computed on this filtered dataset.

A.3 Genre-specific subsets

Experiments are conducted on three genre-specific subsets: *"Fantasy & Paranormal"*, *"Mystery, Thriller & Crime"*, and *"History & Biography"*. Each subset is constructed by restricting books to the corresponding genre labels and retaining all associated filtered interactions and reviews. Table A.1 summarizes the resulting dataset sizes after filtering, as well as the user-level train/validation/test split used in our experiments.

A.4 Predictive power of items' popularity

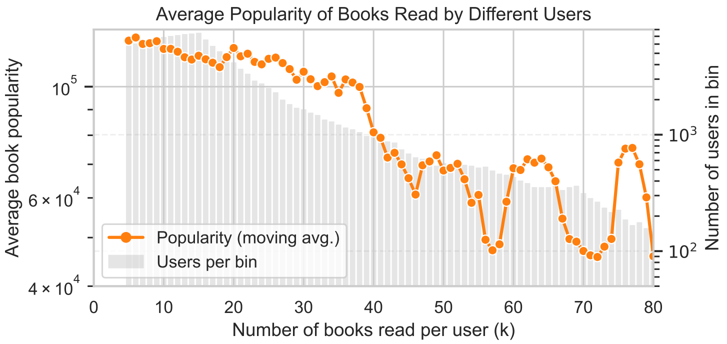

Figure A.3: **[Goodreads-Fantasy, Train]** Moving average popularity of books read by different users, as a function of the number of books each user has read. Computed using a sliding window of users with $k \pm 10$ books read, where k is shown on the x-axis. Book popularity (log-scale) is measured as the total number of ratings a book has received on the platform. *Take-away:* Users with longer reading histories tend to read less popular books.

Figure A.3 shows that users with longer reading histories tend to read, on average, less popular books. This observation helps explain why simple popularity-based baselines perform competitively for users with very short profiles, but deteriorate rapidly as profile length increases. As users become more data-rich, their reading behavior shifts toward the long tail of the catalog, where popularity alone is a weaker predictor of preferences.

B Additional Results and Ablations

This section reports additional results and ablations used to set key modeling choices. Additional results in §B.1 correspond to Test data. Ablation results in §B.2 correspond to validation data: we select a single configuration per method based on these results before reporting final test performance in the main paper.

B.1 Additional Results

This subsection provides two robustness checks for the main results. First, we extend the ROC-AUC analysis to very long user profiles ($k > 60$). Second, we report standard top-K ranking metrics (Precision@10, NDCG@10, Recall@10) using the same short- and long-profile cohorts used in the main text.

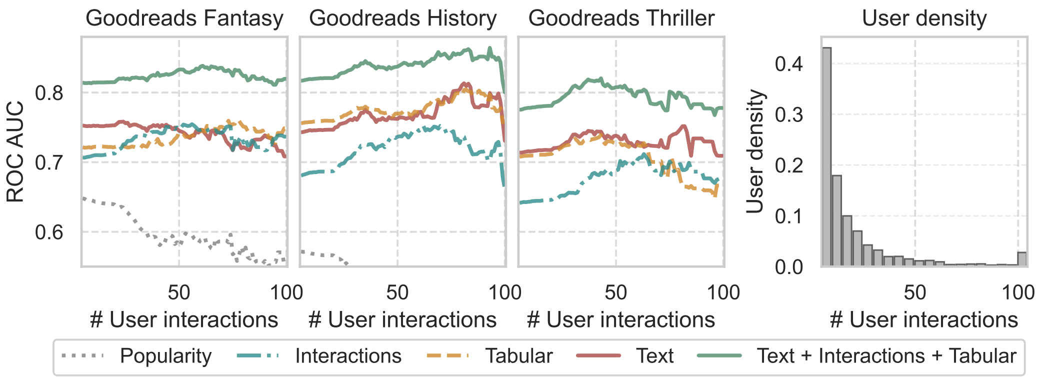

Figure B.4: **[Test]** Performance of all three base models plus a three-way ensemble, up to very long user profiles ($k = 100$). Lines are plotted using a rolling average within $k \pm 15$. The number of users with such a long profile is minuscule, therefore the plotted lines are very unstable at high k. User density per k is plotted on the right, with the last bin representing all users with $k \geq 100$.

Figure B.4 complements Figure 2 in the main text by extending the horizon to $k = 100$ and by including all base models plus the three-way ensemble. However, as k increases, the number of users in each cohort drops sharply, so rolling averages become noisy and should be interpreted as qualitative rather than precise estimates.

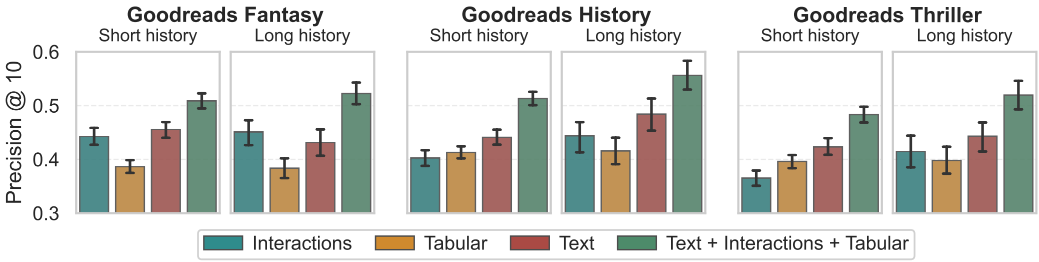

Figure B.5: **[Test]** Performance results using the precision @ 10 metric.

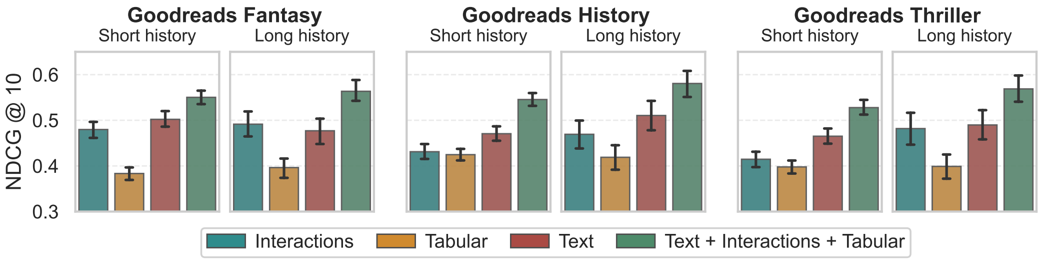

Figure B.6: **[Test]** Performance results using the NDCG@10 metric (Järvelin & Kekäläinen, 2002).

Figures B.5–B.7 report standard top-K ranking metrics that emphasize the quality of the highest-ranked recommendations. These additional metrics serve as a check that conclusions drawn from

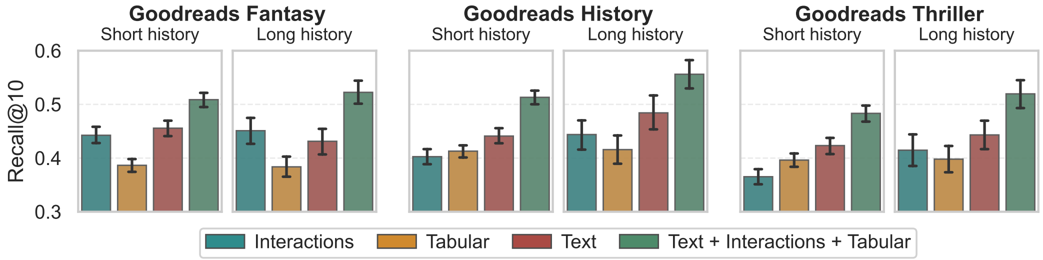

Figure B.7: **[Test]** Performance results using the Recall@10 metric.

ROC AUC (a threshold-free ranking metric) are not an artifact of the particular evaluation measure. As noted in the main text, the qualitative pattern is consistent across these metrics: combining preference representations improves performance over any single channel, and text representations remain strong relative to interaction-only baselines across user cohorts.

B.2 Ablations

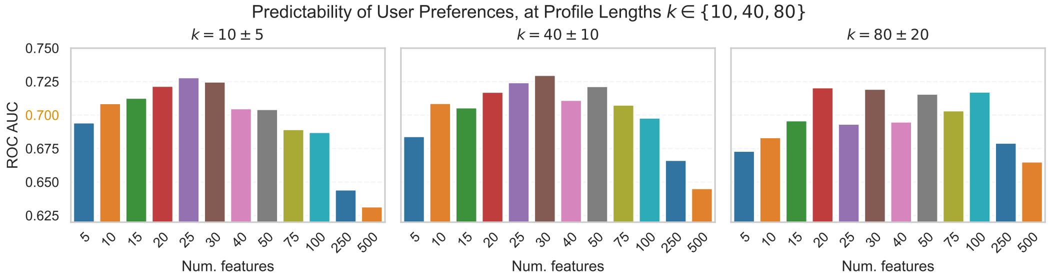

Figure B.8: **[Goodreads-Fantasy, Validation]** The performance of tabular feature recommendations does *not increase* monotonically with higher feature granularity (higher number of tag clusters). The optimal number of clusters of tabular *tags* is approximately 30 throughout the dataset (for both short and long profiles). This contrasts with Collaborative Filtering, where the optimal number of latent factors is generally lower for users with shorter profiles and higher for users with longer profiles.

Figure B.8 illustrates a bias–variance trade-off in the induced tag-cluster feature space. With too few clusters, semantically distinct preferences are merged, reducing the resolution of the representation. With too many clusters, features become sparse and unstable: small lexical variations can split otherwise similar tags into separate clusters, and users/books activate fewer shared dimensions, which weakens similarity-based scoring. This would not be an issue if our tabular model were fitted on domain data, but we score recommendations using a simple dot product between tabular representations. This ensures these representations are easily and meaningfully interpretable. Across profile-length cohorts, we observe a broad optimum around $K \approx 30$, and use this value for tabular baselines in subsequent experiments.

Figure B.9 highlights why a single global rank is a compromise in collaborative filtering. When profile information is limited, higher-rank models overfit and degrade performance, whereas users with richer histories benefit from additional capacity to represent finer-grained taste dimensions. In practice, we therefore choose the CF configuration that performs best on validation overall, while noting that adaptive-capacity variants (not studied here) could further reduce this cohort gap.

Figure B.10 complements this view by showing the joint effect of model capacity (number of latent factors) and regularization, separately for short, medium, and long profiles. The heatmaps make clear that optimal settings are cohort-dependent: for short profiles, stronger regularization and lower-rank models are preferred, while longer profiles benefit from higher capacity and can support weaker regularization.

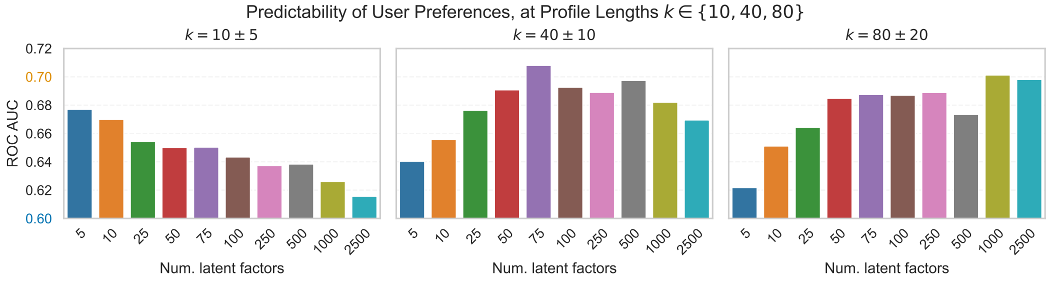

Figure B.9: **[Goodreads-Fantasy, Validation]** Collaborative filtering performance with different number of latent factors, as a function of the number of user profile interactions. Using a higher number of latent factors (higher capacity model) leads to better performance on users with longer profiles, at a detriment to performance on users with shorter profiles.

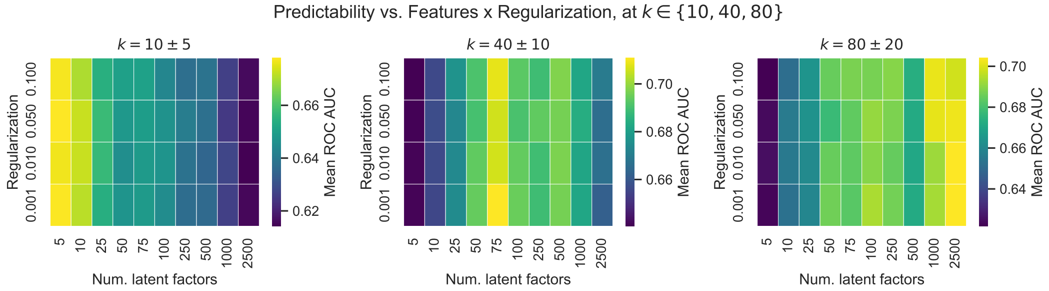

Figure B.10: **[Goodreads-Fantasy, Validation]** Hyperparameter search results for the collaborative filtering model, across different subsections of users (short, medium, or long profiles).

Overall, these ablations show that (i) the tabular representation benefits from an intermediate clustering granularity that is stable across profile lengths, while (ii) collaborative filtering exhibits a more pronounced capacity–data interaction, with optimal rank and regularization shifting as user profiles become longer. We use these validation sweeps to set K for the tag-cluster representation and to select the collaborative filtering hyperparameters used in the main comparisons.

C LLM Prompting Details

This section documents the literal prompt text used in our LLM-based pipelines. We follow a two-message format (system + user) whenever the API supports it, mirroring the OpenAI-style `messages=[{role: system}, {role: user}]` interface. To avoid repetition, prompts may reference reusable sub-templates via placeholders such as `{{book.description}}` and `{{review.description}}`. Unless otherwise stated, placeholders use double braces and dot notation (e.g., `{{user.top_authors}}`).

Leakage controls. To prevent test/validation information from leaking into prompts, all community-review blocks included in item descriptions are drawn from the training split only. In addition, we explicitly exclude any review written by the evaluation user when constructing the candidate item's community-review block.

Prompt 1: [Template] User Reading Profile Generation - System message

```
You are a book analysis expert who synthesizes a user's reading
preferences based on their complete history of book reviews and
ratings.
- Carefully examine the user's past reviews and ratings to identify
recurring themes in what they enjoy and dislike in books.
- Look for specific patterns in favored genres, writing styles,
narrative structures, themes, character types, pacing, or any other
distinctive elements.
- If the user predominantly reads a particular author, book series,
or subgenre, mention it explicitly.
- Write the response in a conversational style across three concise
paragraphs in the first person, as if the user themselves is
summarizing their tastes.  Describe both what they like and what
they dislike (if any).  Avoid starting sentences with phrases like
"Based on my reading history" - instead, use direct and natural
expressions such as "I like..." or "I prefer...".
- At the end, on a separate line, include overarching tags,
attributes, and key characteristics of the user's reading
preferences in a comma-separated list within square brackets, as
in the following format:  "[<tag1>, <tag2>, ...]".
The write-up should be clear and natural, capturing the user's voice
while being informative.  Avoid redundancy and focus on the most
defining aspects of their reading preferences.  This write-up will
later be used to make personalized book recommendations.
```

Prompt 2: [Template] User Reading Profile Generation - User-role message

```
User reading history:
{{book_1_description}}
{{book_1_review}}

{{book_2_description}}
{{book_2_review}}

{{book_3_description}}
{{book_3_review}}
[...]

Most-read authors:  {{user.top_authors}}
Most-read sub-genres:  {{user.top_genres}}
```

Prompt 3: [Template] User-Book Interaction Prediction

```
User reading preferences:
{{user_reading_profile}}

Book description (includes summary, metadata, and up to five
community reviews from the training split):
{{book_description}}

How likely is it that the user will read this book next?  Output
format:  [[RECOMMENDATION_SCORE]]
```

Prompt 4: [Template] Book Description

```
- Title:  {{ book.title }};
- Author(s):  {{ ", ".join(author for author in book.authors) }};
- Pages:  {{ book.pages }};
- Genres:  {{ ", ".join(genre for genre in book.genres) }};

- Summary:
{{ book.summary }};

- Community User Reviews:
{{ "\n\n".join(review for review in book.train_community_reviews) }}
```

Prompt 5: [Example] Book Description

```
- Title:  Need (Need, #1);
- Author(s):  Carrie Jones;
- Pages:  306;
- Genres:  fantasy, young-adult;

- Summary:
 Pain shoots through my head.  Fireworks.  Explosions.  All inside my
 brain.  The white world goes dark and I know what’s about to happen.
 Zara White suspects there’s a freaky guy semi-stalking her.  She’s
 also obsessed with phobias.  And it’s true, she hasn’t exactly been
 herself since her stepfather died.  But exiling her to shivery Maine
 to live with her grandmother?  That seems a bit extreme.  The move
 is supposed to help her stay sane...but Zara’s pretty sure her mom
 just can’t deal with her right now.  She couldn’t be more wrong.
 Turns out the semi-stalker is not a figment of Zara’s overactive
 imagination.  In fact, he’s still following her, leaving behind an
 eerie trail of gold dust.  There’s something not right - not human
- in this sleepy Maine town, and all signs point to Zara.  In this
 creepy, compelling breakout novel, Carrie Jones delivers romance,
 suspense, and a creature you never thought you’d have to fear.;

- Community User Reviews:
 1.  {{ community_review_1_if_exists }}
 2.  {{ community_review_2_if_exists }}
 3.  {{ community_review_3_if_exists }}
```

Prompt 6: [Template] Tag Retrieval Query (Embedding Model)

```
 Instruct:  Given a book description and user reviews, retrieve
 relevant tags that match the book.  Book description:
{{book_description}}
```

Prompt 7: [Template] Book Review Description

```
Review:  {{ review.text }} Review is also on:  {{ review.source }} |
Likes:  {{ review.likes }} | Comments:  {{ review.comments }} | User
rating:  {{ review.rating }} (out of 5).
```

Prompt 8: [Example] Book Review Description

Review: "Need" is one of the first paranormal books I have read in my life and I can tell you that I really enjoyed reading this book! "Need" is a paranormal book written by Carrie Jones and it is about how Zara White realizes that a mysterious stranger has followed her from her hometown Charleston to her new home in Maine. "Need" is a brilliant book that will be a huge hit for paranormal fans. What can I say? Carrie Jones has really brought life to this paranormal romantic story. Carrie Jones has made the book extremely intense yet charming at the same time as Zara White is a truly wonderful character in this book. Zara is not afraid to speak her mind as she is genuinely interested in World Peace and she tells everyone her desire to be a pacifist with such passion that it makes her a truly admirable and courageous character. Another character who is truly memorable in this book is Nick Colt as he seems like a mysterious character, but as the story progresses, he becomes a close friend to Zara as he is willing to protect her from any danger. Carrie Jones brings creativity to this book as each chapter is named after various types of phobias that Zara experiences in her adventures in Maine such as one chapter being called "Didaskaleinophobia - fear of going to school" and another chapter is called "Sitophobia - fear of eating." For adults who do not like bad language in books, this book has a couple of mild profanities, but the profanities in this book are not as strong as some of the books and it might not affect many adults who dislike bad language in books. "Need" is a truly mesmerizing book full of mystery and romance that any hardcore paranormal fan will for ages. Review is also on: Rabbit Ears Book Blog | Likes: 12 | Comments: 2 | User rating: 5 (out of 5).

Prompt 9: System Message - User-Book Interaction Prediction

You are a book recommendation expert. Specifically, you analyze a user’s explicit reading preferences to provide a recommendation score for a given book. Your goal is to maximize user engagement. First, carefully examine the user’s reading preferences, as provided in their own words. Identify key themes, genres, writing styles, and other elements that influence their enjoyment of books. Then, analyze the given book based on its summary, genre, themes, style, and other users’ written reviews. Compare its characteristics to the user’s stated preferences. Use your own knowledge of the given book and the book market to assess how well this book aligns with what the user enjoys. Think step-by-step and reason through your response. Finally, output a recommendation score for the given book between 0 and 10. Your answer must strictly follow the format "[[RECOMMENDATION_SCORE]]", where "[[0]]" means the user is unlikely to read the book at all and "[[10]]" means the user is certain to read the book.

Prompt 10: [Example] User Reading Profile (generated by gpt-5)

I live for fantasy in all its flavors--especially when it skews dark, propulsive, and cinematic. Urban fantasy with a mystery/thriller backbone is my sweet spot (ghosts, mages, angels/demons, shapeshifters, vampires), and I love stories that read like high-octane TV: fast pacing, clean prose, big twists, and memorable ensembles. I’m all in on myth-infused worlds and clear magic systems, whether it’s Greek god mayhem, Maori cosmology, Russian nesting-doll sorcery, or post-apocalyptic clans on the road. YA or adult both work for me as long as the stakes are real and the worldbuilding delivers.

I have a soft spot for strong, capable heroines and reluctant leaders--queens, summoners, thieves-turned-saviors--who grow into their power. Series and authors I keep returning to include Tonya Coffey’s A New World (fae politics with heart), William Stadler’s gritty supernatural action (Dark Connection and Confessions of a Summoner), and J. Thorn’s apocalyptic adventures (Dustfall, plus the mythic chaos of Risen Gods). I also enjoy genre-bending detours: steampunk alt-history in samurai-era Japan, tarot-laced historical mysteries, and thoughtful Christian/“religious paranormal” epics. I don’t mind romance--including queer romance--when it supports the story and keeps the momentum.

What turns me off is cheesy or lightweight tone, meandering plots, or endings that fizzle after a strong start. I generally prefer a grittier edge (I won’t shy away from gore in horror), and I get impatient when books feel long without payoff or dodge the darker implications of their worlds. Crisp editing, steady momentum, and a sense that the story could leap onto a screen go a long way with me.

[dark fantasy, urban fantasy, supernatural mystery, dystopian/ post-apocalyptic, myth and folklore, strong female leads, fast-paced, twisty plots, cinematic feel, shapeshifters/vampires/angels, YA and adult crossover, steampunk alt-history, inclusive romance, gritty with some gore, clean prose, international settings]