Enriching children's stories with LLMs: Delivering multilingual data enrichment for children's books at scale and across markets

Zarah Weiss and Christof Meyer and Mikael Andersson
Nextory AB, Norrtullsgatan 6, 113 29 Stockholm, Sweden

Abstract

This paper presents a user-centered, empirically guided approach to multilingual metadata enrichment for children's books. We combine LLMs with human-in-the-loop quality control in a scalable CI/CD pipeline to curate brand collections that enhance book discovery and engagement for young readers across multiple European markets. Our results demonstrate that this hybrid approach delivers high-quality, child-appropriate labels, improves user experience, and accelerates deployment in real-world production environments. This work offers practical insights for applying generative NLP in the media and publishing industry.

1 Introduction

Enriching book metadata improves content discovery and personalized recommendations (Li et al., 2024; Zhang and Chen, 2018), especially for young readers still developing search strategies (Bilal and Kirby, 2002). Yet, maintaining high-quality annotations in continuously updating data catalogs is resource-intensive and often not feasible. We present a scalable continuous integration and continuous delivery (CI/CD) framework for metadata enrichment with large language models (LLMs) and human-in-the-loop control. We use this to enrich multilingual e-book and audio book data and make our product catalog easier to navigate for children. We focus on curating recognizable *brand collections* to enhance book discovery and engagement for young readers. From the initial proof-of-concept (PoC) through deployment across four European markets, our process has been informed by direct engagement with users—grounded in real-world needs identified through user interviews and iterative feedback.

The main contributions of this paper are:

- A fully integrated CI/CD pipeline for multilingual, LLM-based data enrichment, combining automation with human-in-the-loop control.
- Scalable quality control protocols designed to meet industry standards for deploying LLM-generated labels to sensitive user groups.
- Practical strategies for generating high-quality labels across diverse languages and markets.
- Real-world evidence of user impact, based on live deployment data collected over several weeks across multiple markets.

We discuss related work (Section 2) and our use case definition (Section 3), before reporting our PoC exploration and cross-market expansion (Section 4). We then detail the architecture of our CI/CD framework (Section 5). We report the impact on user experience in Section 6 and conclude with a joined discussion and outlook (Section 7).

2 Related work

LLMs are used in recommendation and retrieval systems to address cold start, interaction sparsity, and generalization challenges (Zhao et al., 2024; Liu et al., 2024), as well as to directly generate personalized content (Li et al., 2024). External tools are often integrated to enhance LLM performance and reduce hallucinations (Li et al., 2024; Wang et al., 2024b). For recent overviews, see Zhao et al. (2024); Li et al. (2024); Lin et al. (2025).

LLMs are also used for automatic data enrichment, improving model performance (Chen et al., 2024; Lyu et al., 2024), recommendation explainability (Li et al., 2024; Zhang and Chen, 2018), and scaling annotation efforts (Tan et al., 2024; Wang et al., 2021). In industry, content annotations remain key for filtering large catalogs, helping reduce latency, costs, and resource demands in recommendation systems (Li et al., 2024), and are especially useful when user-item interactions are sparse (Zhao et al., 2024). However, ensuring label quality is critical. While automatic checks like self-verification, certainty estimates, and consistency evaluations are promising (Madaan et al.,

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 804–812

2023; Xiong et al., 2023; Wang et al., 2023; Lin et al., 2022; Zheng et al., 2023), human-in-the-loop frameworks remain essential for quality control in customer-facing applications (Wang et al., 2024a; Kim et al., 2024; Tan et al., 2024; Madnani et al., 2019). To facilitate this, some hybrid annotation tools have been proposed (e.g., Klie et al., 2018), and there is growing recognition of the need for scalable, cost-efficient LLM data enrichment (Chen et al., 2024; Lyu et al., 2024). Yet, little work addresses integrating LLM-based enrichment with human-in-the-loop control into CI/CD pipelines for continuous, quality-assured deployment. Our work contributes to fill this gap by demonstrating scalable, safe, and cost-efficient LLM-based data enrichment in a production environment.

3 Use case definition

Our use case focused on building an extendable pipeline to automatically enrich e-books and audiobooks with additional meta-information, improving users' ability to navigate our book[1] offers for young readers. We launched a customer-centric discovery process to better understand the needs of our two core user personas in the children's segment: the child and the parent. Previous insights showed that for children to explore the platform independently, parents first need reassurance on safety and trust. We began with a survey of 4,000 customers with active kids' profiles; 200 qualified respondents shared insights on their family's usage. Key findings revealed that children's needs vary significantly by age. Parents of children over six reported more independent use, while younger children required more support. Discovery was a common challenge, especially for children under three and over twelve. Parents rated categorization, search, and navigation lower in the kids' experience than for adults.

To address these issues, we focused on extracting brand labels to curate books into recognizable, age-appropriate brand collections. These include books sharing recurring characters or a common series or (non-)fictional universe. Our initial scope targeted children under 13 across four European markets, focusing on books in their respective dominant languages.[2] Our user research indicated these collections would improve discoverability and engagement (see Section 6). We estimated the opportunity size in terms of (1) substantial market-wise collection uptake and (2) uplift in children's click-to-read ratio (see Section 6). Against this value proposition, we identified three key risks:

Target group suitability Labels had to be harmless[3], accessible, and recognizable.

Limited data access Due to legal uncertainties around processing book content, we restricted data sources to publisher-provided metadata (author, title, series name, descriptions).

Resource drain from LLM exploration Having multiple valid labels[4] complicated evaluation across languages and LLM setups. To stay aligned with the business value, we set strict deadlines: PoC in two languages within one month (170 hours) and a full launch across all four markets within three months (510 hours).

The ideal end state of this use case is defined as:

Content integration Cross-market deployment of brand collections for popular books.

Seamless workflow Full integration of the data enrichment workflow with human expert review into our existing infrastructure.

Scalability Establishing a maintainable and extendable CI/CD framework to support long-term scalability and operational efficiency.

4 Study 1: LLM-based data enrichment with human-in-the-loop control

We structured the use case in two phases: an initial PoC for Market G1 (our largest Germanic-language market) and Market U1 (our Uralic-language market), followed by expansion to Market G2 and Market G3 (the other two Germanic-language markets in our study). The PoC focused on model setup and postprocessing for Market G1, testing generalizability to Market U1 to validate transferability across language families while optimizing for our largest market. We then built a multilingual CI/CD pipeline, refined during the market expansion. This section introduces the datasets, then presents the model setup and results for Market G1, followed by test data from all markets and the final pipeline. While not strictly chronological, this structure highlights our key learnings.

[1] We use the term *book* to refer to a digital book, which may be available in multiple formats, such as e-book or audio-book.

[2] To balance transparency and confidentiality, we anonymized the four markets. They span three Germanic and one Uralic language, testing cross-family generalizability.

[3] Even short brand labels can pose risks if inappropriate, especially in children's products. Therefore, a human-in-the-loop review by content managers was required before release.

[4] E.g., *Peppa Pig*, *Peppa Pig-verse*, *Peppa Pig & friends*.

4.1 Data sets

We created two datasets per market, plus one development set. To streamline the evaluation, we limited annotations to books in each market's dominant language and those linked to a book series.[5]

Development data 1,000 books, sampled from the 100 most prolific series for Market G1.

Test data 1,000 books, sampled from the 100 most relevant series per market (10 book each), selected by content managers based on app popularity and market expertise. For Market G1, we avoided overlap with the development set.

Production data Up to 20 brand collections per age group (0-2, 3-6, 7-9, 10-12, and N/A), combining popular series from the test data with most prolific series in our catalog.

4.2 Development for Market G1

4.2.1 Set-up

We developed a multilingual brand annotation workflow on Market G1 development data, focusing on rapid prototyping for our PoC. To remain within our time constraints, we initially tested on one market, with the option to refine the setup if its performance fell short on Market U1 test data. We framed the task as a book-level classification problem,[6] rather than labeling entire series or clustering books. Book-wise labels support incremental updates as new books are added, align with how metadata is managed in our systems, and allow fine-grained performance evaluation. This makes the approach scalable and maintainable in production, while still enabling brand-level grouping through postprocessing. Figure 1 illustrates the associated prompt and grounding technique.

Our evaluation compared two leading multilingual LLMs available in October 2024: chatgpt-4o-mini-2024-07-18[7] (henceforth ChatGPT) and Gemini-1.5-flash-002[8] (henceforth Gemini). We tested variations in prompt targets (e.g., shared universe vs. protagonist), prompt language (English vs. market language), grounding methods (descriptions vs. web search), and generation settings. Results were tracked using Google's VertexAI Experiments service. For brevity, we report only the best-performing configuration per model.[9]

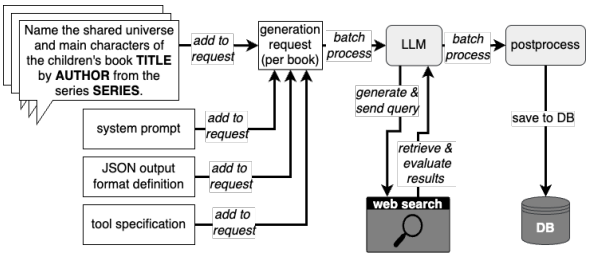

Figure 1: Prompting and grounding example workflow.

LLM	Grounding	SU	main	mix
gemini	web	.54	.69	.75
openai	meta	.37	.52	.53

Table 1: Label accuracy for shared universe (SU), main characters (main), or their combination (mix).

Importantly, this is not a general performance comparison but an evaluation of which model delivered the best out-of-the-box results for our use case, with minimal custom work. Performance across models is only partially comparable: We tested web search grounding exclusively with Gemini, as OpenAI's offering lacked built-in web search support at the time. We postponed custom search development until after evaluating Gemini's performance. Since Gemini with web search grounding delivered sufficient results, no further investment was needed (see 4.2.2). While not a fully controlled comparison, this approach reflects our priority: optimizing for immediate deployment with minimal resource investment in the PoC phase.

4.2.2 Results & discussion

Table 1 summarizes our findings, showing only the top configurations. Gemini with web search grounding outperformed ChatGPT, which was prompted with book metadata and descriptions alone.[10,11] We make three key observations: First,

[5] We focused on books that are parts of series because series information, which is provided to us by publishers, allows us to treat books within the same series as part of the same brand and thus simplify the evaluation process. For example, books that appear in the series *The Ultimate Peppa Pig Collection* belong to the brand *Peppa Pig*. Note that also books from the series *Peppa Pig Bedtime Stories* belong into the *Peppa Pig* brand. Thus, all books in a series belong to the same brand collection, but brand collections can contain many series.

[6] I.e., we annotated each book with a brand label. A brand collection consists of all books with the same brand label.

[7] https://platform.openai.com/docs/models/gpt-4o-mini

[8] https://ai.google.dev/gemini-api/docs/models#gemini-1.5-flash

[9] Model settings: temperature=1; top p=0.95; max output tokens=8,192; frequency penalty=1.9. All configurations performed better when prompting in the market language.

[10] Gemini without web search was comparable to ChatGPT.

[11] It is expected that the use of external tools boosts performance, see also Wang et al. (2024b).

web search grounding outperformed description-based grounding. Publisher descriptions varied in quality—some lacked relevant content—which impacted accuracy. Web grounding also reduced the generation of protagonist lists in favor of more appropriate group labels (e.g., *Avengers* instead of lists of individual names). Second, prompt target selection mattered. Prompts for series protagonists generally outperformed shared universe prompts, but the best results came from choosing between prompt targets case by case. Third, Gemini more reliably detected when no label applied (e.g., classical fairytale collections), allowing us to default to the series name. Yet, both models often returned inaccurate labels instead of "not applicable".

4.2.3 Postprocessing

Neither shared universe nor protagonist prompting consistently outperformed the other, so we developed an automated postprocessing workflow to select the best brand label per book. We computed a label confidence based on four factors: i) string similarity to the series name and the book title (both reinforcing market-specific labels), iii) label length (promoting cross-series groupings, e.g., *Peppa Pig* over *Peppa Pig & friends*), iv) label frequency across a series, and v) a penalty for ambiguous single-name labels (e.g., *Greg* vs. *Greg Heffley*). Similarity was measured using length-normalized longest continuous subsequence (LCS), ignoring case and whitespace. Named entities were identified with stanza (Qi et al., 2020).

We optimized score weights on Market G1 data and applied the highest-scoring label to all books in the series, including unlabeled books, to reduce LLM costs. These became brand collection candidates for manual review. To aid reviewers, we calculated a collection coherence score, flagging risky collections with inconsistent authorship, low series similarity, or low label confidence.

4.3 Results across markets

After finalizing the model setup and postprocessing on development data, we applied it to the test data for Markets G1 and U1.[12] The only adjustment was translating prompts into the market language. We calculated two accuracy metrics: label accuracy (acc_L), the percentage of series labels not renamed by content managers (measured per series), and grouping accuracy (acc_G), the percentage of books remaining in their assigned collection (measured per book). Performance was generally good for both markets (see Table 2), though G1 was notably lower than U1 (discussed below). We expanded to Markets G2 and G3, proceeding to production data after confirming feasibility for G2. We skipped testing for G3 due to consistent results across markets.

	G1	U1	G2
acc_L	.87	.94	.94
acc_G	.94	.99	1.00

Table 2: Labeling with postprocessing on the test data.

Two key observations stand out. First, label accuracy for Market G1 improved on test data compared to development data. This is largely due to our postprocessing. Additionally, the test data contained fewer series that couldn't be reasonably grouped into brands, a major source of errors in the development data. Second, Markets U1 and G2 achieved higher accuracies than G1, despite using G1-optimized weights. This difference was largely due to G1's test data containing brands with common European names (e.g., Klara, Lisa, Anna), which led to misgroupings and required content manager adjustments. This discrepancy likely resulted from the need to sample less common series for G1's test data to avoid overlap with the development data. Ultimately, with grouping accuracies systematically above 90%, we moved to production, as renaming collections involved minimal effort for content managers in our mandatory human-in-the-loop control process.

5 CI/CD framework

Scalable development and seamless integration were two of our three main characteristics for the ideal end state of our use case. To achieve these goals, we designed our system following MLOps principles and included a preliminary CI/CD framework in our PoC, which we finalized within the scope of our product-to-market timeline. Our approach extends CI/CD beyond software deployment to enable end-to-end automation for data pipelines and machine learning workflows. We chose Google Cloud Platform (GCP) services to orchestrate and execute the pipeline. The pipeline versioning follows semantic versioning and is handled through GitLab CI/CD pipelines, which han-

[12] G1 results use label confidence scores fitted to development data; others include scores fitted on G1 development and test data. For now, no market-specific weights are used.

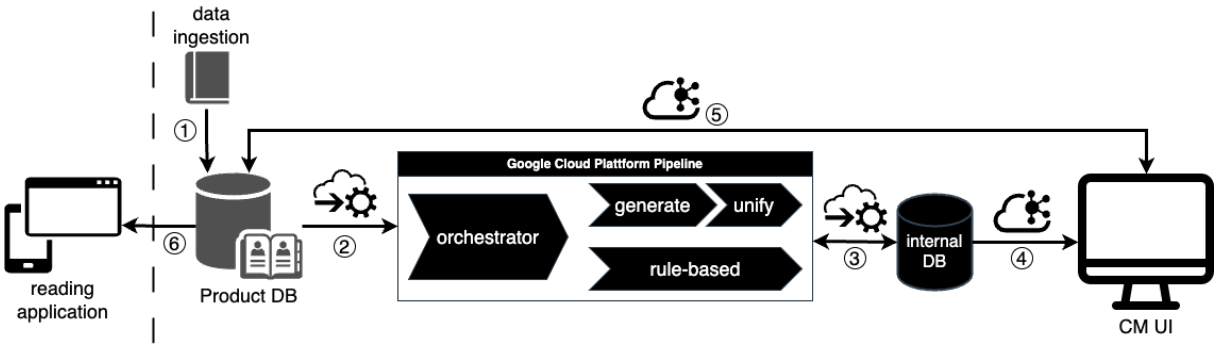

Figure 2: Architecture of our CI/CD framework with human-in-the-loop control.

dle the deployment of new pipeline versions whenever model updates, configuration changes, or code adjustments are required.

The final architecture of our CI/CD framework is shown in Figure 2. New product data is continuously integrated into our product database ①. From there the GCP pipeline, orchestrating the different components, can be triggered manually or automatically based on cron schedules or thresholds for data deltas. Labels for incoming data can be generated rule-based or LLM-based ②. The rule-based component assigns brand labels to products associated with existing brand collections (e.g., books in a series that have already been labeled), reusing previously established knowledge. For other products, the LLM-based component generates labels in batch mode, which are then unified through our custom postprocessing logic. The generated label candidates are stored in a table ③, making them available for further review or downstream processing and ensuring continuous delivery. New brand collections are published via Kafka ④ to our Content Manager user interface. Here, content managers can review, approve, or reject generated collections—integrating human feedback into the automated pipeline. The feedback is automatically synchronized with our product database and metadata is continuously updated ⑤. This enables immediate delivery to the app ⑥.

This set-up allows us to leverage multiple signals to trigger model retraining: Content manager's feedback, such as acceptance and rejection rates, is aggregated to identify degradation in label quality. Additionally we collect qualitative feedback from content managers on specific errors or label inconsistencies, to improve labeling performance. We plan to augment this with automated change rate monitoring, focusing on the difference between submitted and published collections to trigger alerts when significant discrepancies occur. This framework establishes the foundation for scalable, self-improving brand label generation while maintaining human oversight and high-quality standards.

6 Study 2: Effect of brand collections on user experience and engagement

We evaluated the impact of brand collections on our users in two experiments: we conducted user interviews with parents and children and evaluated the engagement that users showed with brand collections. The user interviews were conducted using prototypes and mock data for brand collections to verify the anticipated value proposition prior to investing in the PoC and to speed up development.

6.1 User experience interviews

After the initial survey and discovery phase, we ran qualitative user studies to refine brand collections. During iterative design sprints, we developed prototypes using mock data and tested them with parents. We chose to start with parents due to the low interactivity of the prototypes, which would have frustrated children. Testing with children is most effective when prototypes support natural play (Cantuni, 2020), which ours did not at this stage.

Over two months, we conducted 30–50 minute moderated interviews with 15 parents. This confirmed our survey findings: parents emphasized the need for simplified navigation and recognizable book covers to aid discovery. They expressed a preference for features that reduced repetitive tasks like searching for the same book every night.

In the next phase, we tested an interactive prototype in 30-minute sessions with 12 children (ages 4–11) and their parents. The parents and children invited were our customers and/or employees volunteering for the test in their interest to improve the service as customers and employees in the context

of our work/customer relationship. Sessions took place in our offices and combined observational and think-aloud methods. We paid particular attention to how children discovered and interacted with brand collections, recording their interactions with the screen with a camera. We found that children naturally referred to the brand (e.g., *Disney Cars*) rather than character names, supporting our hypothesis. In cases like *Peppa Pig*, brand and character were the same, consistent with expectations. Children across age groups were primarily drawn to collections with recognizable and appealing cover art. Our implementation of brand collections proved intuitive and aligned with their expectations. When presented with a printed card featuring 30 proposed collections, older children recognized more brands and emphasized the need for age-appropriate groupings. Feedback highlighted the importance of personalized, visually distinct collections and high-fidelity artwork to enhance recognition and appeal. The final cover art is illustrated in Figure 3.

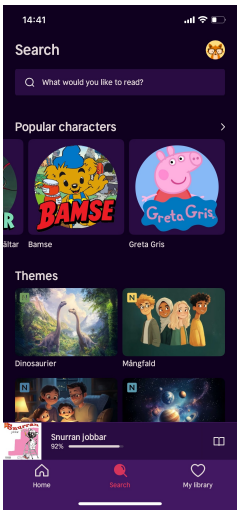

Figure 3: Brand collections shown as "Popular Characters" as a more age-appropriate term for young users.

These insights shaped our final production design, reinforcing the need for age-appropriate, visually distinct brand collections. For this reason, we created brand-specific covers based on images provided by publishers or—if these were not available—by grouping covers from popular books in a collection. We also identified opportunities for further personalization based on user interaction, as children showed clear brand preferences.

6.2 User engagement

We estimated user engagement after five weeks of deployment (February to March 2025) across all four markets (G1, G2, G3, U1). We used two metrics for our evaluation: market-wise collection uptake and the click-to-read ratio.

Collection uptake We measured market-wise collection uptake as brand collection interaction rate (IR_B; see Figure 4). IR_B is the proportion of distinct users interacting with brand collections relative to all distinct app user interactions. This metric reduces bias from highly active users by focusing on unique interactions, with weekly aggregation smoothing out cyclic patterns. To maintain confidentiality, we report IR_B as the difference relative to the stable interaction rate of our most popular discovery screen, direct search (IR_S),[13] using its cross-market mean of the past five months as a reference point (zero). IR_B rises over the first three weeks and stabilizes, indicating sustained engagement. While IR_B is lower than IR_S, it is still close enough to consider brand collections a successful addition given the entrenched use of direct search in our app. Market U1 shows higher IR_B in the first three weeks and a peak in week seven, though the cause is unclear.

Click-to-read ratio We measured the market-wise click-to-read ratio, defined as clicks to read, download, or save to a reading list, normalized by total app interactions and aggregated weekly (C2R; see Figure 5). We compared interactions to the same period in 2024 due to known strong seasonal effects in user activity. On average, C2R was higher in 2025 than in 2024. A Wilcoxon Test showed a statistically significant improvement ($\alpha \leq 0.05$) for all markets ($W = 21, p = .016$). These results suggest brand collections enhance user experience, though the comparison does not isolate this effect from to other changes in our offer. We accepted this limitation, as delaying the rollout for an A/B test was not justified given our previous findings.

7 Discussion & outlook

We aimed to enhance the user experience for children's profiles, addressing both child and parent stakeholders. Our user-centric, data-driven approach, informed by iterative user interviews, demonstrated considerable potential for LLM-based data enrichment but also significant risks, concerning trust and safety, with little margin for error. The limited availability of high-quality in-

[13] Discovery screens are any interfaces for book discovery, including direct search and recommendation lists.

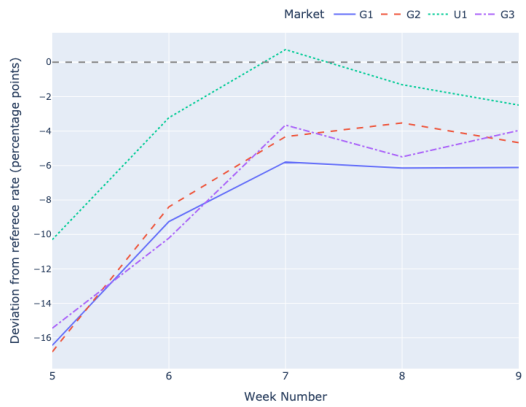

Figure 4: Brand collection uptake as a percentage of interactions, normalized against direct search interactions.

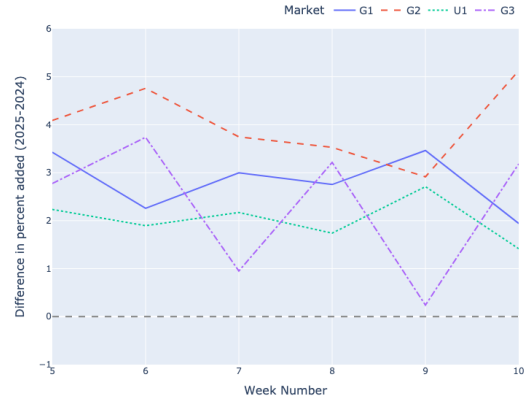

Figure 5: Click-to-read interactions, shown as the percentage point change from 2024 to 2025.

put data posed challenges in label generation. To mitigate these issues, we employed a human-in-the-loop setup, ensuring accurate brand labeling. However, this also introduced complexities in the CI/CD process, necessitating seamless integration with existing workflows for long-term scalability.

We addressed data limitations through search grounding and proposed an automated architecture for data ingestion and labeling, integrating rule-based and stochastic postprocessing with human oversight. Our evaluation involved continuous user feedback and analysis of post-deployment user behavior. While market differences and seasonal effects impacted the cleanliness of results, our findings suggest positive user engagement, consistent with interview insights. We achieved our goals for content integration, efficiency, and scalability.

Future work will include expanding the approach to additional markets and refining postprocessing to improve accuracy and robustness. Another natural direction is adapting the pipeline to other user groups and types of metadata. While this study focused on brand labels for young readers, the underlying CI/CD infrastructure is broadly applicable: Extending it to adult users would require only minor prompt adjustments to ensure age-appropriate labeling. However, given adults' more advanced search capabilities, other forms of metadata—such as themes, tropes, or external references (e.g., adaptations or awards)—may offer greater value. Although our architecture is well suited for generating these types of metadata, it will need to be extended with adapted grounding strategies and postprocessing modules. The modular design of our system supports such extensions with minimal overhead.

Limitations

The study has three core limitations: First, it focuses on markets with similar languages and cultures. Although the inclusion of U1 addresses some linguistic variation, the results may not fully apply to more diverse linguistic or cultural contexts. Second, we concentrated on books from series, which have a strong overlap with brands. Expanding the analysis to include non-series books could provide important additional insights into labeling quality. Third, resource and time constraints during development limited in-depth model comparisons and prevented us to assess the impact of the brand collections on user engagement under ideal conditions, which could affect the robustness of the findings.

Ethical considerations

The user studies involving children were conducted in full compliance with ethical guidelines (Görman, 2023). Informed consent was obtained from parents or legal guardians, and assent was secured from the children themselves, ensuring they understood the purpose and procedures of the study. All participants' privacy and confidentiality were strictly maintained throughout the process. Children were always accompanied by their parents during the sessions, and their well-being was prioritized at all times. Additionally, any video material collected—focused solely on interactions such as children's finger movements on the screen—was permanently deleted after analysis to ensure privacy and data protection.

References

Dania Bilal and Joe Kirby. 2002. Differences and similarities in information seeking: children and adults as web users. In *Information Processing and Management*, volume 38, pages 649–670. Pergamon.

Rubens Cantuni. 2020. *Designing Digital Products for Kids: Deliver User Experiences That Delight Kids, Parents, and Teachers*. apress.

Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, et al. 2024. Exploring the potential of large language models (LLMs) in learning on graphs. *ACM SIGKDD Explorations Newsletter*, 25(2):42–61.

Ulf Görman. 2023. *Guide to the Ethical Review of Research on Humans*. Swedish Ethical Review Authority, PO Box 2110, SE-750 02, Uppsala, Sweden.

Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. MEGAnno+: A human-LLM collaborative annotation system. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–176.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: system demonstrations*, pages 5–9.

Lei Li, Yongfeng Zhang, Dugang Liu, and Li Chen. 2024. Large language models for generative recommendation: A survey and visionary discussions. In *LREC-COLING*, pages 10146–10159. ELRA Language Resource Association.

Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Hao Zhang, Yong Liu, Chuhan Wu, Xiangyang Li, Chenxu Zhu, et al. 2025. How can recommender systems benefit from large language models: A survey. *ACM Transactions on Information Systems*, 43(2):1–47.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Han Liu, Xianfeng Tang, Tianlang Chen, Jiapeng Liu, Indu Indu, Henry Zou, Peng Dai, Roberto Galan, Michael Porter, Dongmei Jia, et al. 2024. Sequential llm framework for fashion recommendation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1276–1285.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, and Jiebo Luo. 2024. Llm-rec: Personalized recommendation via prompting large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 583–612.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Nitin Madnani, Beata Beigman Klebanov, Anastassia Loukina, Binod Gyawali, Patrick L Lange, John Sabatini, and Michael Flor. 2019. My turn to read: An interleaved e-book reading tool for developing and struggling readers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 141–146.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Association for Computational Linguistics (ACL) System Demonstrations*.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024a. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Yanbin Lu, Xiaojiang Huang, and Yingzhen Yang. 2024b. RecMind: Large language model powered agent for recommendation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4351–4364.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations*.

Yongfeng Zhang and Xu Chen. 2018. Explainable recommendation: A survey and new perspectives. In *Foundations and Trends in Information Retrieval*, volume 14, pages 1–85. now.

Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, et al. 2024. Recommender systems in the era of large language models (LLMs). *IEEE Transactions on Knowledge & Data Engineering*, pages 1–20.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.