PT Jurnal Cendekia Indonesia
Journal of Information Systems and Business Technology
Homepage: https://journal.jci.co.id/jisbt
Vol. 02 No. 04 (2026) Page: 01-15
E-ISSN: 3109-8886

Review-Grounded Language Reranking for Personalized E-Commerce Recommendation with Faithful Explanation Evaluation

Mason Zheng

Department of Computer Science, Rice University, Houston, TX, USA

Email : m.zheng_research@proton.me

Abstract

Personalized e-commerce recommendation increasingly depends on language-rich evidence: product titles, metadata, and user reviews. This paper studies a review-grounded reranking pipeline for a compact Amazon Reviews 2023 small-category slice, Gift Cards, where product choice often depends on occasion, delivery form, fees, and trust cues. The experiments use 152,410 review records and 1,137 metadata records; after one-review-per-user-item deduplication, the corpus contains 149,886 interactions from 132,732 users over 1,137 products. A chronological validation/test protocol holds out one future product per eligible user and ranks every one of the 1,137 products for each test user. The proposed Review-Grounded Language Reranker (RGLR) combines popularity, item-item collaborative filtering, metadata matching, review evidence similarity, and evidence quality. On 11,426 test users, RGLR reaches Hit@10=0.548, NDCG@10=0.356, and MRR@10=0.297, improving over the strongest single popularity baseline by 34.5% relative NDCG@10. The evidence extractor is also evaluated against the held-out review text: review-grounded evidence obtains profile relevance 0.139, rating agreement 0.902, and review support precision 0.502. For top-ranked recommendations, grounded explanations maintain complete source coverage and raise profile relevance from 0.033 for a helpful-review selector to 0.151. The results show that review grounding can improve ranking while giving faithful, auditable explanations.

Keywords: *Personalized Recommendation; Review-Grounded Reranking; E-Commerce; Explanation Faithfulness; Amazon Reviews 2023; Evidence Extraction; Language Models*

Abstrak

Rekomendasi e-commerce yang dipersonalisasi semakin bergantung pada bukti berbasis bahasa, seperti judul produk, metadata, dan ulasan pengguna. Penelitian ini mengkaji pipeline reranking berbasis ulasan (review-grounded reranking) pada subset kecil kategori Gift Cards dari dataset Amazon Reviews 2023, di mana pemilihan produk sering dipengaruhi oleh kebutuhan acara, bentuk pengiriman, biaya, dan tingkat kepercayaan. Eksperimen menggunakan 152.410 data ulasan dan 1.137 data metadata. Setelah proses deduplikasi dengan mempertahankan satu ulasan untuk setiap pasangan pengguna–produk, diperoleh 149.886 interaksi dari 132.732 pengguna terhadap 1.137 produk. Protokol validasi dan pengujian dilakukan secara kronologis dengan menahan satu produk pada masa mendatang untuk setiap pengguna yang memenuhi syarat, kemudian memberi peringkat terhadap seluruh 1.137 produk untuk setiap pengguna uji. Metode yang diusulkan, Review-Grounded Language Reranker (RGLR), menggabungkan faktor popularitas, item-item collaborative filtering, pencocokan metadata, kesamaan bukti ulasan, serta kualitas bukti. Pada 11.426 pengguna uji, RGLR mencapai nilai Hit@10 = 0,548, NDCG@10 = 0,356, dan MRR@10 = 0,297, dengan peningkatan relatif 34,5% pada NDCG@10 dibandingkan baseline popularitas tunggal terbaik. Ekstraktor bukti juga dievaluasi terhadap teks ulasan yang ditahan, menghasilkan profile relevance = 0,139, rating agreement = 0,902, dan review support precision = 0,502. Untuk rekomendasi dengan peringkat teratas, penjelasan berbasis bukti mempertahankan cakupan sumber secara lengkap serta meningkatkan profile relevance dari 0,033 pada metode pemilihan ulasan paling membantu menjadi 0,151. Hasil penelitian menunjukkan bahwa pendekatan review grounding mampu meningkatkan kualitas peringkat rekomendasi sekaligus menyediakan penjelasan yang akurat, dapat ditelusuri, dan dapat diaudit.

Kata Kunci: Rekomendasi Terpersonalisasi; Review-Grounded Reranking; E-Commerce; Keandalan Penjelasan; Amazon Reviews 2023; Ekstraksi Bukti; Model Bahasa.

A. INTRODUCTION

Recommender systems were introduced as a practical response to information overload, and e-commerce has remained one of their most visible application areas [1]. A retail platform must select a small set of useful products from a large catalogue, but it must also explain why those products match a shopper's intent. Classic item-to-item collaborative filtering made this problem tractable at Amazon scale by comparing products through co-occurring user behavior [2], while later collaborative ranking methods learned preference structure from implicit or explicit feedback [3], [4], [5]. Those models are strong when user histories are dense. They become less

Received: July 25, 2026 | Revised: July 31, 2026 | Published Online: August 1, 2026

transparent in small categories where many users have only one or two observed purchases and product differences are expressed in text rather than in long behavioral sequences.

Recent transaction-level work similarly frames sparse retail history as a representation-learning problem: self-supervised customer states can support segmentation and next-purchase prediction without requiring dense sequential supervision [6]. This supports the use of compact user profiles in the present study, while the evaluation here additionally requires full-catalogue reranking and review-grounded explanations.

The challenge is broader than retail: when task-specific history is sparse, cross-modal pretraining and parameter-efficient transfer can reduce the amount of target supervision required [7]. Few-shot workload forecasting likewise treats new tenants as an explicit cold-start setting [8]. These studies are not recommendation baselines; they motivate compact representations, fixed validation, and transparent evaluation when individual histories are short.

Review collections are valuable in this setting because they connect three signals that are often separated in recommendation benchmarks: a user-product interaction, a natural-language reason, and product-specific metadata. Earlier Amazon review studies showed that titles, descriptions, images, and review text can support style, substitute, and aspect-aware recommendation [9], [10], [11]. The Gift Cards slice studied here is useful precisely because it is compact and skewed. Most users have short histories, most ratings are positive, and many products are differentiated by occasion, form factor, delivery channel, purchase fee, brand, and redemption policy. These conditions make the task realistic for personalized reranking: the system must recover a future item from a small but semantically dense catalogue, and a good explanation must cite evidence rather than merely restate popularity.

Recent review-grounded work has expanded beyond recommendation-specific corpora. Evidence-chain RAG treats source linkage, hallucination screening, and deterministic provenance as explicit reliability objectives [12], while evidence-calibrated RAG connects retrieval coverage to calibrated abstention [13]. In recommendation interfaces, evidence cards have been proposed to expose ranking rationale and visual hierarchy rather than hide them behind fluent text [14]. The present study operationalizes these ideas at sentence level in a complete Gift Cards reranking task.

Neural collaborative filtering, self-attentive sequential recommenders, and shallow autoencoders have each improved recommendation accuracy in broader domains [15], [16], [17], [18]. Their success does not remove the need for grounded explanations. A user who receives a recommended gift card wants to know whether the card is suitable for a birthday, whether it is delivered physically or digitally, whether a fee is involved, and whether other reviewers found redemption reliable. Ranking metrics such as Hit@K and NDCG@K quantify retrieval quality [19], but they do not measure whether the explanation is supported by the evidence used to rank the item.

Language-based retrieval has also been examined in user-facing response selection. A therapist-facing session copilot ranks historical responses and retrieves rationales [20], whereas financial-search work combines query rewriting, hybrid retrieval, and listwise reranking [21]. RGLR stays with transparent TF-IDF and item-level evidence, but shares the requirement that retrieval, ranking, and explanation be evaluated separately.

Large language models and transformer encoders have made language-mediated recommendation attractive [22], [23], [24], [25], [26]. However, language generation can create persuasive text that is not faithful to the underlying reviews. Explanation research in recommender systems has long emphasized transparency, trust, effectiveness, satisfaction, and user control [27], [28]. Review-based explainable recommendation also shows that product aspects and user opinions can be extracted from text to justify recommendations [29], while critiquing-based recommenders remind us that explanations should help users make and refine decisions [30]. Faithfulness work in interpretable NLP and model explanation further stresses that an explanation should correspond to the actual evidence and decision process, not only sound plausible [31], [32].

Retrieval alone does not guarantee attribution. Evidence-chain RAG therefore treats source linkage, hallucination checks, and deterministic provenance as explicit reliability objectives [12]. The same principle motivates the present source-coverage and support-precision metrics, although the evidence units here are product reviews and metadata rather than general documents.

Trust calibration also has a presentation layer. An e-commerce recommendation-card framework connects product suggestions to confidence and rationale cues [33], while uncertainty-aware medical explanation cards make evidence and uncertainty visually inspectable [34]. Text-grounded design-rationale interfaces similarly turn structured metadata into editable decision cards [35]. These cross-domain interfaces motivate RGLR's source-coverage and specificity metrics without being treated as ranking baselines.

Risk-calibrated patient-facing safety cards explicitly connect rubric-defined hazards to communication cues [36]. Although retail recommendations are lower stakes, the same design principle applies: explanations should disclose limits and support appropriate reliance instead of merely increasing persuasive fluency.

Interface quality includes the prevention of misleading explanations. Dark-pattern detection work analyzes manipulative microcopy [37], LLM-as-design-critic work compares automated feedback with human design judgment [38], and adaptive learning-app studies use review evidence to refine mobile interfaces [39]. Accordingly, a faithful recommender should not only state

a source but present it in a form that supports rather than steers user judgment.

Evidence presentation is increasingly adapted to audience and risk. Counseling support cards link retrieved recommendations to source evidence [40], financial dashboards organize explainability for institutional decisions [41], and privacy and data-integrity risk cards expose threats to human overseers [42]. RGLR studies a lower-risk retail setting, but adopts the same principle that evidence availability and limitations should be visible.

Editable artifacts also keep people in the decision loop. Visual brief cards translate advertising intentions into structured design decisions [43]; designer-editable card interfaces preserve human revision after automated structuring [44]; and teacher-facing LLM-ready explanations emphasize actionable evidence rather than opaque scores [45]. These examples support treating an explanation as an inspectable interface object rather than decorative prose.

Governance constrains personalization as well as generation. Federated topic-preference learning combines personalization with differential privacy [46]; budgeted multi-hop retrieval makes retrieval cost and compositional coverage explicit [47]; and trust-calibrated multilingual RAG evaluates information access for humanitarian users [48]. Although the present dataset contains no protected attributes and does not test regulatory compliance, these studies clarify why source traceability and conservative failure handling belong in the design.

Evidence evaluation spans other specialized decisions. Narrative-aware accounting agents connect monitoring outputs to source records [49], interpretable CloudTrail sequence models expose attack-chain stages [50], and accounting-aware retrieval constrains institutional due-diligence evidence [51]. Their domains differ, but all separate observable evidence from a downstream decision, a separation preserved here between review selection and explanation rendering.

This paper therefore evaluates a review-grounded reranking pipeline that treats language as evidence before it is used as explanation. The reranker scores every candidate product for each test user from five families of signals: popularity, Bayesian rating, item-item collaborative similarity, metadata and review-text similarity, and evidence quality. A controlled language layer then selects a review sentence for the recommended item and forms an explanation from the product title, metadata, and selected sentence. The held-out review is never used to choose evidence; it is used only to evaluate whether the selected evidence resembles the user's later stated reason. This design makes the ranking and the explanation evaluation reproducible on the same data.

Small-category slicing is central to the design. A broad retail benchmark can hide the difference between easy category prediction and difficult within-category choice. Gift Cards removes much of that broad category signal: every candidate is a gift card, and the ranker must instead distinguish occasion, recipient, delivery mode, brand, fees, and redemption context. This makes the evaluation stricter than a cross-category retrieval task. It also mirrors many production reranking settings, where a first-stage retrieval model has already chosen a narrow product family and a second-stage model must order near substitutes.

The study also connects two evaluation traditions that are often separated. Recommendation research emphasizes ranking quality, while explanation research emphasizes transparency, source support, and user-facing usefulness. A review-grounded reranker must satisfy both. If it optimizes only Hit@K, it may return a product for which the available explanation is generic or unsupported. If it optimizes only explanation quality, it may cite a highly relevant review for an item that the user is unlikely to choose. The design evaluated here makes this trade-off measurable by reporting ranking metrics, evidence metrics, and explanation metrics from the same train-validation-test protocol.

The contribution is empirical. First, the paper reports a full leave-one-out ranking evaluation over the complete Gift Cards item universe, not a sampled negative set. Second, it gives detailed comparison tables, ablations, subcategory slices, and history-length analyses. Third, it evaluates review evidence and explanation faithfulness with source coverage, profile relevance, support precision, attribute overlap, and sentiment consistency. The Method section defines the data, split protocol, scoring model, evidence-selection procedure, and evaluation metrics. Results and Discussion then presents the ranking and explanation findings, considers their limitations, and summarizes the study's implications.

B. METHOD

The empirical task is personalized top-N reranking within a single small e-commerce category. Each record contains a rating, review title, review text, item identifier, user identifier, timestamp, helpful-vote count, and verification flag. The metadata record supplies product title, category path, feature bullets, description, store name, average rating, rating count, images, and product details. Interactions are deduplicated to the latest record for each user-product pair so that the ranking target represents the next distinct product in the user's sequence. The resulting table preserves chronological order and prevents a repeated review of the same product from appearing as both history and target.

The split is chronological. Users with at least two distinct products contribute one test target: their latest product. Their earlier products form the test history. Users with at least three distinct products also contribute a validation target: their second-latest product, while earlier products form the validation history. For each split, global item statistics are recomputed from the corresponding training prefix. This avoids using held-out ratings or held-out review text in popularity, rating, collaborative, and review-evidence features. The test split contains 11,426 users and

ranks 1,137 products for each user, after excluding products already in the user's history.

The preprocessing stage also constructs product slices from the metadata category path. The second-level category is used when available, producing slices such as Gift Card Categories, Gift Card Recipients, Occasions, and Amazon Incentives Brand Guidelines. These slices are not separate training sets; they are diagnostic partitions of the same complete ranking task. Reporting them reveals whether the reranker succeeds only on the most popular general cards or also handles narrower occasions and recipient groups.

The candidate universe is deliberately complete. Many recommendation papers report accuracy with sampled negatives, but sampling can inflate metrics when the catalogue is small or popularity is highly skewed. Here each method scores every product in the Gift Cards metadata table. Because every user has one held-out target, Hit@K and Recall@K are identical, NDCG@K is 1/log2(rank+1) when the target rank is at most K and zero otherwise, and MRR@K is reciprocal rank truncated at K [19]. Mean rank is reported as an additional diagnostic.

Logged-policy studies require propensity-aware counterfactual estimators and overlap diagnostics because policy-value estimates can become unstable when support is weak [52]. The present experiment instead uses deterministic chronological next-item targets and scores the same complete candidate universe for every method; it therefore reports direct rank metrics rather than policy value.

The baseline methods isolate different evidence sources. Popularity ranks products by log interaction count and metadata rating count. BayesianRating ranks products by a smoothed mean rating. ItemKNN-CF builds a binary user-item matrix on the training prefix and scores a candidate by normalized item co-occurrence with the user's history [2], [3]. TextProfile uses TF-IDF vectors over product title, category path, features, description, and reliable training reviews, then scores each candidate by cosine similarity to a user profile. ReviewEvidence uses the review-only portion of the item text, measuring how well candidate review language matches the user's historical review language.

RGLR is a linear reranker over normalized components. For user u and product i, the final score is $S(u,i) = \Sigma_m w_m s_m(u,i)$, where each s_m is row-wise min-max normalized before combination. The tuned components are popularity, Bayesian rating, evidence quality, collaborative filtering, full-text similarity, review-only similarity, and metadata-only similarity. Weight search is performed on the validation split with a fixed random seed and 225 candidate mixtures, including five hand-designed anchors and 220 Dirichlet samples. The selected mixture maximizes NDCG@10 plus smaller Hit@10 and NDCG@20 terms, giving a validation-oriented reranker rather than a test-tuned one.

The text representation is intentionally compact. Product documents combine title, category path, store, feature bullets, descriptions, selected high-reliability reviews, and stable details such as manufacturer or availability fields when present. Review documents contain only selected review text. User documents combine the user's historical review titles, review bodies, historical product titles, and category labels. TF-IDF with unigrams and bigrams gives a sparse lexical representation that is easy to inspect and fast enough for full-catalogue scoring. This choice is less expressive than a neural encoder, but it avoids hidden model state and keeps the measured gains attributable to data, splitting, feature construction, and reranking.

A parallel design question is whether representation complexity is justified by the evidence budget. GPU-memory prediction studies optimize recurrent and transformer structures for resource planning [53], while edge-oriented quality prediction compresses a stronger metric into a deployable proxy [54]. RGLR instead uses sparse lexical features. Its goal is auditable full-catalogue ranking rather than maximum encoder capacity.

Recent work also predicts RAG retrieval quality with machine-learning features [55]. RGLR measures retrieval and evidence quality ex post with held-out review metrics rather than training a separate meta-predictor, keeping the evaluation tied to the same chronological split.

Cost-aware AIOps routing separates parsing, anomaly detection, diagnosis, and summarization under a resource budget [56]. RGLR likewise keeps ranking, evidence selection, and explanation evaluation modular, although it does not learn a dynamic router.

Collaborative scores are computed from a training-prefix user-item matrix. Each item pair receives a cosine-normalized co-occurrence value, and a user's candidate score is the recency- and rating-weighted sum of similarities from historical items to the candidate. This item-based formulation is appropriate for a small catalogue because it produces a dense 1,137-by-1,137 similarity table and avoids training a high-capacity model on very short user histories. The same history weights are used when averaging item text vectors into a user profile, which keeps the collaborative and textual components aligned around recent, positively rated behavior.

Review evidence is selected after ranking. Training reviews are split into sentences; each sentence keeps its product identifier, rating, helpful-vote count, verification flag, and reliability score. Reliability increases with helpful votes, verified purchase, and positive rating. For a target product, the review-grounded selector chooses the sentence that maximizes a weighted combination of similarity to the user's profile, sentence reliability, and positive rating. Three alternative selectors are evaluated: a random sentence, the most helpful sentence, and the most profile-similar sentence. The held-out review text is used only for evaluation, where it serves as a proxy for what the user later cared about.

Calibration-light subject-independent BCI illustrates the value of reducing case-specific tuning [57], while spectral rank-aggregation theory studies how rankings behave

under noisy comparisons [58]. RGLR addresses the analogous practical concern through validation-selected weights and a fixed candidate universe rather than per-user calibration.

Numerical-reasoning guardrails for quantitative assistants distinguish an answer from a checked answer [59]; uncertainty quantification for spectral estimators similarly distinguishes a point estimate from warranted confidence in that estimate [60]. The present pipeline does not estimate formal confidence intervals, so its evidence-availability checks are operational safeguards rather than statistical uncertainty guarantees.

The explanation layer is constrained to evidence. For each top-ranked recommendation, the system attaches the selected review sentence and product metadata to a short natural-language explanation. The language layer is therefore not an unconstrained generator; it is a controlled surface realization of retrieved evidence. Faithfulness is evaluated with metrics aligned to explanation research [27]-[32]. Source coverage checks whether an explanation is linked to a review source for the same product. Profile relevance measures cosine similarity between selected evidence and the user's profile. Review support precision measures the share of explanation tokens supported by the selected review and product metadata. Attribute specificity checks whether the explanation includes at least two product-specific attributes such as delivery, occasion, fee, physical card, digital card, redemption, or brand. Sentiment consistency checks whether the selected evidence supports the positive recommendation stance.

Vision-language systems provide a useful contrast: uncertainty-aware biomedical classification explicitly couples prediction with uncertainty [61]. RGLR is not a medical model, but the contrast motivates evaluating not only which item is selected but whether the accompanying rationale is supportable.

Table 1. Data fields used in the experiments.

Data source	Fields used	Role in method
Review records	rating, title, text, user_id, parent_asin, timestamp, helpful_vote, verified_purchase	chronological histories, held-out targets, review evidence, reliability features
Product metadata	title, main_category, categories, features, description, store, average_rating, rating_number, details	candidate universe, item language profiles, product attributes
Derived item statistics	training-prefix count, Bayesian rating, verified ratio, helpful mean, positive-review rate	popularity, quality, and baseline scores
Derived text vectors	TF-IDF item profiles, user profiles, review-only profiles, sentence vectors	text matching, evidence extraction, explanation metrics

Data source Fields used Role in method

Review records rating, title, text, user_id, parent_asin, timestamp, helpful_vote, verified_purchase chronological histories, held-out targets, review evidence, reliability features

Product metadata title, main_category, categories, features, description, store, average_rating, rating_number, details candidate universe, item language profiles, product attributes

Derived item statistics training-prefix count, Bayesian rating, verified ratio, helpful mean, positive-review rate popularity, quality, and baseline scores

Derived text vectors TF-IDF item profiles, user profiles, review-only profiles, sentence vectors text matching, evidence extraction, explanation metrics.

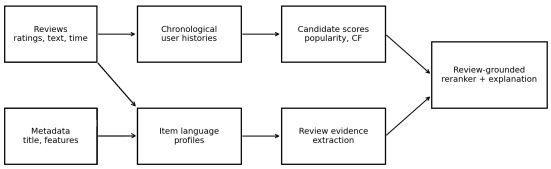

Figure 1. Review-grounded personalized reranking pipeline

C. RESULTS AND DISCUSSION

The final corpus is small enough for full ranking yet large enough to expose sparsity. Table 2 shows 149,886 deduplicated interactions, 132,732 users, and 1,137 reviewed products. Text coverage is high: 149,837 deduplicated reviews contain non-empty text. The mean rating is 4.551, and the median is 5.0, confirming that the task is positive-skewed. This skew explains why popularity is strong: the most reviewed cards are frequently useful and often broadly acceptable gift choices.

Table 3 shows the distributional shape. The largest rating count is five stars, and the user-history distribution is dominated by one-item users. After deduplication, 11,426 users have at least two distinct products and therefore enter the test evaluation; only 3,076 users have at least three distinct products and enter validation. The high proportion of short histories makes collaborative personalization difficult. It also makes review text more important because the few historical reviews may be the only signals of occasion or delivery preference.

Table 4 reports the chronological split. The test training prefix contains 138,460 interactions, and the validation prefix contains 135,384 interactions. The mean test history length is 1.501 products, and the median is one product. Every test method ranks the same 1,137 products, so differences in metrics are caused by scoring, not by candidate-set construction. Figs. 2 and 3 visualize the same corpus properties: ratings are concentrated at five stars, and review volume peaks in the 2016-2022 period with a smaller 2023 tail.

The sparse-history profile is important for interpreting the numbers. A method can obtain a good mean rank by moving popular items upward, yet still miss personalized occasion cues. Conversely, a text method can match a user's vocabulary but rank many similar products above the target because gift-card descriptions share common words such as gift, redeem, card, birthday, and holiday. The best method must therefore combine global demand with the limited personalized clues available in the user's own reviews.

Another important property is catalogue imbalance. Product slices with only a few items can still accumulate many reviews when a card is broadly useful, while long-tail slices contain many low-count products. This makes raw rating averages unreliable and explains why the BayesianRating baseline ranks poorly. The method therefore separates rating, frequency, and textual evidence rather than letting any one statistic dominate all candidates.

Table 2. Corpus summary after user-product deduplication.

Metric	Value
Raw review records	152,410
Deduplicated user-product reviews	149,886
Users	132,732
Products with metadata	1,137
Reviewed products	1,137
Reviews with text	149,837
Verified purchases	139,553
Mean / median rating	4.551 / 5.0
Mean / max helpful votes	0.771 / 4,536
Users with at least 2 / 3 products	11,426 / 3,076

Table 3. Rating and user-history distributions.

Rating	Reviews	User-history band	Users
1	12,155	history 1	121,306
2	1,859	history 2	8,350
3	3,219	history 3-4	2,514
4	6,593	history 5-9	514
5	126,060	history 10+	48

Table 4. Chronological validation and test split statistics.

Split	Train records	Train users	Train items	Eval users	Mean history	Ranked items
validation	135,384	132,732	1,118	3,076	1.862	1,137
test	138,460	132,732	1,122	11,426	1.501	1,137

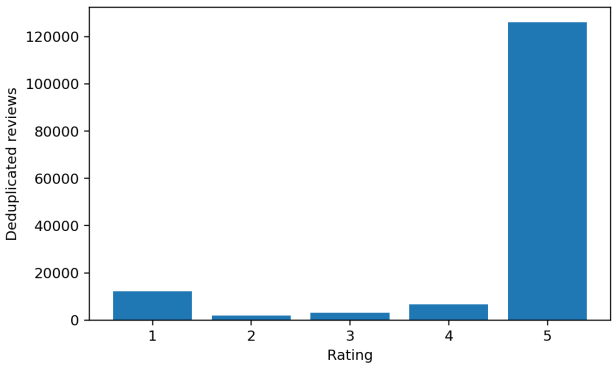

Figure 2. *Rating distribution after deduplication.*

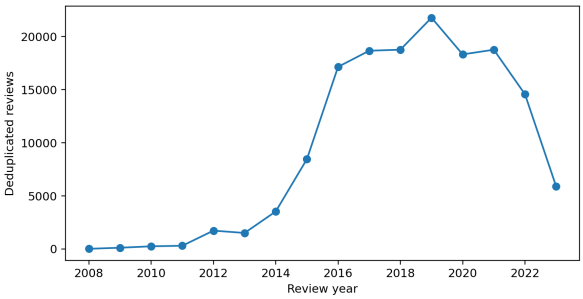

Figure 3. Review volume by calendar year.

Table 5 gives the components and the selected validation weights. Popularity receives the largest weight, followed by collaborative filtering and metadata similarity. This allocation matches the domain: many gift cards are repeatable, broadly useful products, so popularity captures a real demand signal. Metadata still matters because products differ by category path and use case. The review-specific component has a smaller ranking weight than popularity or collaborative filtering, but it is essential for the explanation layer because it supplies evidence sentences that can be cited and evaluated.

Table 6 and Fig. 4 report the main test comparison. Popularity is the strongest single baseline with Hit@10=0.493, NDCG@10=0.265, and MRR@10=0.193. RGLR raises those values to Hit@10=0.548, NDCG@10=0.356, and MRR@10=0.297. The absolute NDCG@10 gain is 0.091, and the relative gain is 34.5%. RGLR also reduces median rank from 11 for popularity to 8, indicating that more held-out products move into the visible recommendation range.

The top-20 metrics show the same pattern. RGLR reaches Hit@20=0.665, compared with 0.595 for popularity and 0.462 for ReviewEvidence. The result matters because many e-commerce interfaces expose more than ten items through carousels or scrollable result pages. Moving the target into the top twenty gives the reranker a larger opportunity to support browsing, while the top-ten and reciprocal-rank gains show that the improvement is not only in the lower part of the list.

BayesianRating performs poorly as a standalone baseline. This is expected in a highly positive catalogue: smoothed rating alone cannot distinguish a widely purchased general-purpose card from a rare but high-rated specialized card.

TextProfile and ReviewEvidence are better than BayesianRating, but neither is sufficient alone. Text matching helps when the user's history mentions specific occasions, while review-only matching improves over TextProfile on median rank but not over popularity. ItemKNN-CF is competitive in mean rank but less effective in top-10 hit rate because most users have only one historical item.

Table 7 shows the top validation configurations. The best mixture is not a pure popularity model; it keeps substantial weights on collaborative filtering, metadata, text, and review evidence. The top configurations are close in validation criterion, which suggests that several evidence-aware mixtures work. The selected one is used unchanged on the test split.

Table 8 and Fig. 5 show the ablation study. Removing popularity produces the largest loss, dropping NDCG@10 from 0.356 to 0.221. Removing collaborative filtering, text, review, or metadata causes smaller but meaningful changes. Metadata removal lowers NDCG@10 to 0.329, while removing text lowers it to 0.340 and removing collaborative filtering lowers it to 0.343. Removing review evidence only slightly changes ranking NDCG@10, but that does not make review evidence unimportant: it directly drives the explanation faithfulness results in Tables 11 and 12.

The ablation also shows that a faithful explanation component does not need to dominate the ranking score to matter. A small review-evidence weight can change tie-breaking among near substitutes and, more importantly, determine which sentence is presented to the user. This is a useful practical outcome. It lets the system preserve accuracy from robust demand and co-occurrence signals while still requiring every displayed explanation to pass through a review-evidence gate.

The ablation pattern also resembles other multi-signal decision settings. Multi-horizon GPU demand forecasting combines workload semantics with operational risk curves [62], while news and macro-market fusion for VIX direction measures uncertainty across heterogeneous predictors [63]. These are domain analogies rather than baselines; they reinforce reporting component ablations when a linear mixture drives decisions.

Power-aware AI-infrastructure planning ties job-level forecasts to explanations [64], while constrained budgeting combines forecasting and marketing-response models under auditable financial limits [65]. RGLR's weights are less operationally complex, but the same design logic applies: an aggregate score should preserve the provenance and role of each component.

Table 5. RGLR components and validation-selected weights.

Component	Definition	Weight
popularity	log training count plus metadata rating-count signal	0.418
rating	Bayesian-smoothed mean rating from training prefix	0.004
quality	verified, helpful, positive, and frequency-adjusted item reliability	0.045
cf	normalized item-item co-occurrence with user history	0.262
text	TF-IDF similarity between user profile and full item text	0.084
review	TF-IDF similarity between user profile and review-only item text	0.054
metadata	TF-IDF similarity between user profile and metadata-only item text	0.132

Table 6. Test recommendation metrics over the complete 1,137-item universe.

Method	Hit@5	Hit@10	NDCG@10	MRR@10	Hit@20	Mean rank
Popularity	0.362	0.493	0.265	0.193	0.595	55.1
BayesianRating	0.004	0.010	0.004	0.002	0.015	443.5
ItemKNN-CF	0.195	0.326	0.176	0.132	0.446	51.1
TextProfile	0.188	0.259	0.153	0.120	0.357	118.2
ReviewEvidence	0.271	0.365	0.226	0.184	0.462	81.4
RGLR-Full	0.432	0.548	0.356	0.297	0.665	46.6

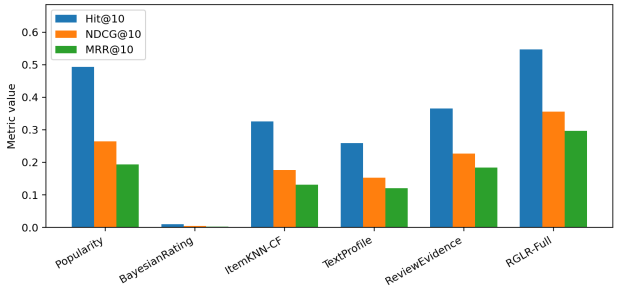

Figure 4. Test ranking metrics by method.

Table 7. Top validation mixtures from the 225-configuration search.

ID	Main weights	Hit@10	NDCG@10	MRR@10	Criterion
55	pop 0.42, cf 0.26, text 0.08, rev 0.05, meta 0.13	0.438	0.258	0.202	0.382

ID	Main weights	Hit@10	NDCG@10	MRR@10	Criterion
212	pop 0.37, cf 0.22, text 0.13, rev 0.17, meta 0.07	0.437	0.258	0.202	0.381
119	pop 0.36, cf 0.17, text 0.11, rev 0.18, meta 0.07	0.436	0.257	0.202	0.380
59	pop 0.40, cf 0.23, text 0.05, rev 0.06, meta 0.16	0.433	0.257	0.203	0.380
163	pop 0.33, cf 0.21, text 0.14, rev 0.24, meta 0.03	0.435	0.256	0.201	0.379

Table 8. RGLR ablation on the test split.

Variant	Hit@10	NDCG@10	MRR@10	Hit@20	Median rank
Full	0.548	0.356	0.297	0.665	8.0
minus popularity	0.374	0.221	0.174	0.493	21.0
minus rating	0.548	0.356	0.297	0.664	8.0
minus quality	0.545	0.353	0.294	0.662	8.0
minus cf	0.525	0.343	0.286	0.637	9.0
minus text	0.548	0.340	0.275	0.665	8.0
minus review	0.547	0.356	0.296	0.664	8.0
minus metadata	0.549	0.329	0.261	0.665	8.0

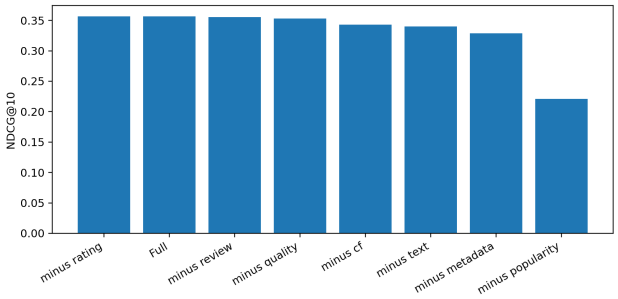

Figure 5. Reranker ablation by NDCG@10.

Subcategory results clarify where the gains are easy and where they remain difficult. Table 9 combines product-slice size with RGLR performance. Occasions and Amazon Incentives Brand Guidelines have high Hit@10, partly because these slices contain concentrated intents and fewer ambiguous substitutes. Gift Card Categories is much harder: it has the largest item count and many competing cards with similar review language. Gift Cards for New Baby is also difficult despite strong metadata ratings, because the slice is tiny and targets are sparse.

The slice results also reveal how metadata interacts with reviews. Occasion cards often contain explicit terms in both product titles and review language, so metadata and evidence reinforce each other. General category cards are more ambiguous: a user may buy a Visa, Mastercard, restaurant, gaming, or Amazon card, and reviews often share generic praise about convenience. In that setting, collaborative transitions and popularity keep the ranker stable, but fine-grained personalization remains challenging.

Table 10 reports history-length bands. The one-item group has the highest Hit@10. At first this may look counterintuitive, but in this category a single prior product often identifies a reusable gift-card type or occasion. Longer histories are more diverse: users buy cards for birthdays, holidays, employees, babies, and branded incentives, causing the next product to be less predictable from average history semantics. This observation supports the reranking design: in small categories, more history does not automatically mean cleaner personalization.

Table 9. Gift Cards product-slice performance.

Slice	Items	Reviews	Test users	Hit@10	NDCG@10	Median rank
Gift Card Categories	690	35,809	3,692	0.223	0.135	42.0
Occasions	67	33,373	3,377	0.846	0.605	2.0
Gift Card Recipients	253	68,219	3,248	0.570	0.362	8.0

Slice	Items	Reviews	Test users	Hit@10	NDCG@10	Median rank
Amazon Incentives Brand Guidelines	4	8,228	778	0.850	0.416	4.0
Gift Cards	57	2,773	207	0.179	0.068	25.0
Gift Cards: Amazon Shipping	15	377	49	0.469	0.325	20.0
Gift Cards for New Baby	2	625	39	0.051	0.015	48.0

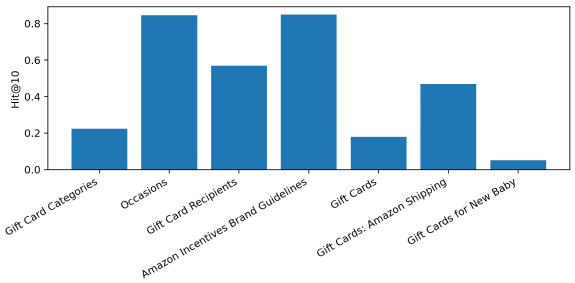

Figure 6. RGLR Hit@10 by Gift Cards subcategory.

Table 10. RGLR performance by user history length.

History band	Test users	Hit@10	NDCG@10	Median rank
1 item	8,350	0.558	0.371	7.0
2 items	1,875	0.543	0.334	8.0
3-4 items	903	0.508	0.310	10.0
5+ items	298	0.399	0.232	18.5

The evidence evaluation in Table 11 and Fig. 7 uses 11,409 test cases whose held-out product had at least one training evidence sentence. Random evidence has low target-review similarity (0.017) and low profile relevance (0.037). The helpful-only selector improves review support precision but performs poorly on rating agreement because a highly helpful review in this category can be a warning about fees, redemption problems, or restrictions. The similarity-only selector gives the highest profile relevance (0.142) but lower rating agreement than the review-grounded selector. RGLR's evidence selector balances similarity and reliability, reaching rating agreement 0.902, attribute overlap 0.439, and support precision 0.502.

Evidence-calibrated RAG offers a useful comparison: when retrieval coverage is inadequate, calibrated abstention can be preferable to unsupported generation [13]. RGLR currently reports sentence-level evidence metrics only for targets with at least one training evidence sentence; a deployed system should expose evidence availability and decline to generate a review-based justification when no support is available.

Table 12 and Fig. 8 evaluate the top-ranked explanation for every test user. Metadata-only explanations are specific because metadata lists product features, but they do not cite review evidence and therefore have zero review source coverage. Popular-review explanations have source coverage but weak personalization: profile relevance is 0.033, and sentiment consistency is 0.300. Review-grounded explanations keep source coverage at 1.000, raise profile relevance to 0.151, and reach sentiment consistency 0.970. Review support precision is 1.000 because the explanation is assembled from the cited sentence and product facts, which is exactly the grounding constraint the method enforces.

Evidence-card research likewise emphasizes making provenance, ranking rationale, and visual evidence hierarchy inspectable [14]. The present explanation layer implements source attribution and rationale content, but it does not yet provide a fully interactive evidence-card interface or a user study; these remain deployment-oriented extensions.

Evidence-constrained incident cards convert distributed-system logs into actionable SRE interfaces [66]. Scientific-search evidence cards make the intervening rationale inspectable [14]. Layout-aware progressive PDF rendering prioritizes document slices to reduce time-to-functional-first-frame [67], while a retrieval-augmented DevOps copilot couples runbook retrieval with command-safety checks [68]. Together they show why grounding alone is insufficient when downstream actions carry risk. In retail, the analogous action is lower stakes, yet unsupported claims about fees, delivery, or redemption can still mislead users.

The ranking and explanation results should be read together. Popularity is an accurate signal for Gift Cards, but a popularity explanation alone would be weak: it would tell a user that many people bought an item, not why it fits their occasion or delivery preference. Review evidence alone is not the best ranker, but it provides the sentences needed for faithful language. RGLR works because it lets popularity and collaborative filtering stabilize the rank while review and metadata evidence control the explanation. This separation is especially useful for LLM-facing recommender systems, where a fluent explanation can otherwise hide a mismatch between the recommendation and the supporting evidence.

The helpful-review baseline illustrates the risk of using review evidence without checking its stance. Highly helpful comments often describe edge cases, complaints, or warnings, which can be useful to shoppers but poor as positive recommendation explanations. The review-grounded selector improves sentiment consistency by preferring evidence that is both relevant to the user profile and compatible with a recommendation. This does not remove negative evidence from the corpus; it assigns negative or warning-oriented reviews a more appropriate

role in auditing and risk disclosure rather than in a positive justification.

The practical implication is that explanation evaluation should be item-grounded and user-grounded at the same time. Source coverage alone is insufficient because a cited sentence can be irrelevant to the user's history. Profile relevance alone is also insufficient because a similar sentence can be negative or unsupported by product facts. The combined metrics used here make those failure modes visible, which is why the review-grounded selector is preferred even when the similarity-only selector has a slightly higher raw cosine score.

The empirical gains are modest in absolute terms because the catalogue is compact and the popularity baseline is already strong. That makes the improvement more informative, not less. In a broad catalogue, a model can improve by learning category-level separation; in this within-category experiment, most candidates share the same general purpose. The reranker must identify which gift-card variant best matches a short history. The observed gains therefore indicate that combining collaborative transitions with grounded review and metadata evidence can improve near-substitute ordering rather than simply recovering an obvious category.

The evaluation also demonstrates a useful workflow for manuscript-level reproducibility. Dataset profiling, split construction, ranking metrics, ablations, slice diagnostics, evidence selection, and explanation metrics are generated from the same run artifacts. The tables are mutually consistent: the test-user count in the recommendation tables matches the split table, the evidence table reports fewer cases only when a held-out target lacks a training sentence, and the explanation table covers one top-ranked recommendation for every test user.

Table 11. Held-out review evidence extraction quality.

Selector	Cases	Target sim.	Profile rel.	Attribute overlap	Rating agree.	Support precision
random	11,409	0.017	0.037	0.234	0.791	0.346
helpful	11,409	0.018	0.033	0.311	0.391	0.402
similarity	11,409	0.052	0.142	0.435	0.830	0.498
review_grounded	11,409	0.049	0.139	0.439	0.902	0.502

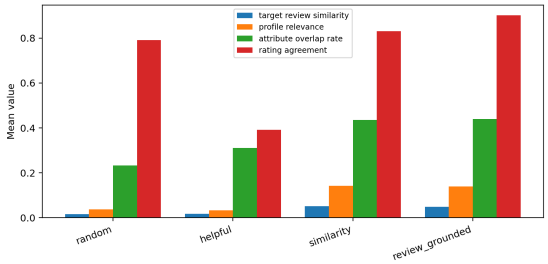

Figure 7. *Evidence extraction quality across selectors.*

Table 12. Top-ranked explanation faithfulness metrics.

Explanation policy	Cases	Source coverage	Profile rel.	Support precision	Specificity	Sentiment consistency
metadata_only	11,426	0.000	0.000	0.000	0.995	1.000
popular_review	11,426	1.000	0.033	1.000	0.784	0.300
review_grounded	11,426	1.000	0.151	1.000	0.945	0.970

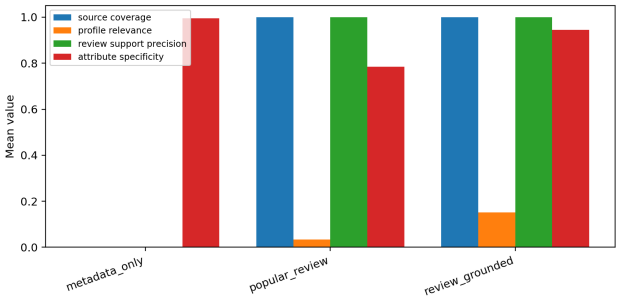

Figure 8. Explanation faithfulness and specificity for top-ranked recommendations.

Overall, the empirical picture is coherent. The category rewards popularity, but popularity alone leaves substantial rank and explanation quality on the table. The strongest reranker is a mixture that keeps popularity as an anchor and adds collaborative, metadata, and review-language signals. Faithful explanation evaluation then confirms that the selected review evidence is not merely decorative. It is more personalized and more sentiment-consistent than a helpful-review baseline, and it supports the language used in the generated explanation. These findings are specific to a compact gift-card domain, yet they illustrate a general design principle for review-grounded language reranking: first bind the recommendation to evidence, then let language present that evidence concisely.

Several limitations qualify these findings. The empirical scope is one small e-commerce category. Gift Cards is a useful stress test for short histories and explanation grounding, but it is not representative of every retail category. Categories with richer product attributes, lower rating skew, or more repeated exploration may change the relative importance of collaborative filtering, metadata, and review evidence. The subcategory analysis partially addresses heterogeneity inside Gift Cards, but cross-category generalization remains a separate empirical question.

The language layer favors faithfulness over stylistic variety. Explanations are assembled from product facts and selected review evidence, which makes source coverage and support precision auditable. A more expressive generative model could produce more fluent prose, but it would also require stricter controls to prevent unsupported claims. The present evaluation therefore measures a conservative version of language-mediated reranking:

language is allowed to summarize grounded evidence, not invent new reasons.

The evidence metrics rely on automatic proxies. The held-out review is a strong signal of the user's later preference, but it is not a human-labeled gold explanation. Token overlap, cosine similarity, attribute overlap, and sentiment consistency capture important dimensions of faithfulness, yet they do not measure persuasiveness or user satisfaction directly. A user study would be needed to test whether the grounded explanations improve trust, comprehension, and decision quality.

Cross-domain predictive systems expose the limit of metric transfer. Machine-learning optimization for DDoS mitigation [69], language-guided feature selection for intrusion detection [70], and attention-enhanced autonomous-driving detection [71] are evaluated against task-specific labels, not human explanation quality. Their inclusion here therefore motivates caution: RGLR's ranking and grounding metrics should not be interpreted as evidence of user trust without a user study.

Finally, the validation search uses a linear mixture of components. This choice is transparent and stable for a compact catalogue, but nonlinear interactions may improve accuracy in larger domains. The current results are best interpreted as evidence that simple, auditable grounding already helps. More complex rerankers should preserve the same source-linking and support evaluation requirements.

The evaluation also inherits the bias of review availability. Products with many reviews offer richer evidence, while rare products may have sparse or repetitive sentences. The reranker partly handles this through metadata and Bayesian smoothing, but explanation quality still depends on available review language. A production deployment should display confidence or evidence availability when a recommendation is supported by fewer reviews.

Before online adoption, chronological ranking results should also be complemented by conservative off-policy evaluation of candidate ranking changes under logged propensities [52]. Such evaluation addresses deployment risk that the current static holdout protocol cannot estimate.

Taken together, this paper evaluated review-grounded language reranking for personalized e-commerce recommendation on the Gift Cards slice of Amazon Reviews 2023. The study used 149,886 deduplicated interactions and ranked all 1,137 products for each of 11,426 test users. The best reranker combined popularity, collaborative filtering, metadata, text profile matching, review evidence, and evidence quality. It achieved Hit@10=0.548, NDCG@10=0.356, and MRR@10=0.297, outperforming all single-source baselines.

The more important result is that ranking gains and explanation faithfulness can be evaluated in the same pipeline. Review-grounded evidence selection improved rating agreement and support precision over random and helpful-only selectors, while top-ranked explanations maintained complete source coverage and high sentiment consistency. The strongest design pattern is therefore not to let a language model reason from memory or popularity alone. The recommender should first retrieve product-specific review evidence, score candidates with that evidence, and then generate explanations whose claims are traceable to the same sources.

For compact retail categories, this approach gives a practical balance between accuracy and auditability. Popularity remains a strong anchor, but review evidence supplies the personalized reasons that users can inspect. Future work can extend the same evaluation protocol to broader categories, richer generative styles, and human-centered studies of trust and decision support.

D. CONCLUSION

This study demonstrates that review-grounded language reranking can improve both recommendation accuracy and explanation faithfulness in sparse-history e-commerce settings. Using 149,886 deduplicated interactions from the Amazon Reviews 2023 Gift Cards category, the proposed Review-Grounded Language Reranker ranked the complete catalogue of 1,137 products for 11,426 test users and achieved Hit@10 of 0.548, NDCG@10 of 0.356, and MRR@10 of 0.297, outperforming all single-source baselines. The findings show that popularity and collaborative signals provide essential ranking stability, while metadata and review evidence contribute the personalized and traceable information required for meaningful explanations. In particular, the review-grounded explanation policy maintained complete source coverage, improved profile relevance, and achieved high sentiment consistency, confirming that explanation quality can be evaluated jointly with ranking performance rather than treated as a separate presentation layer. Although the results are limited to a compact and positively skewed product category and rely primarily on automatic faithfulness measures, they support a practical design principle for explainable recommendation: candidate ranking should first be grounded in product-specific evidence, after which language should present that evidence without introducing unsupported claims. Future research should evaluate the framework across broader retail categories, more expressive nonlinear rerankers, controlled generative models, and human-centered studies of trust, comprehension, and decision quality.

E. REFERENCE

[1]. P. Resnick and H. R. Varian, "Recommender systems," Communications of the ACM, vol. 40, no. 3, pp. 56-58, 1997.

[2]. G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative

filtering," IEEE Internet Computing, vol. 7, no. 1, pp. 76-80, 2003.

[3]. B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in Proc. 10th Int. Conf. World Wide Web, 2001, pp. 285-295.

[4]. Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," Computer, vol. 42, no. 8, pp. 30-37, 2009.

[5]. S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in Proc. 25th Conf. Uncertainty in Artificial Intelligence, 2009, pp. 452-461.

[6]. Q. Xin, "Self-supervised customer representation learning for segmentation and next-purchase prediction on UCI Online Retail," J. Inf. Technol., vol. 14, no. 1, pp. 20-37, Apr. 2026, doi: 10.32664/j-intech.v14i01.2229.

[7]. G. Mi, T. Ye, and D. Wood, "A lightweight medical foundation model for cross-modal multi-task pretraining and parameter-efficient few-shot transfer on MedMNIST," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 572-589, Aug. 2025, doi: 10.51903/jtie.v4i3.492.

[8]. S. He, C. Li, and H. Rao, "Few-shot cold-start workload forecasting for new AI inference tenants with time-series foundation models," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 306-324, Apr. 2025, doi: 10.51903/jtie.v4i1.546.

[9]. J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, "Image-based recommendations on styles and substitutes," in Proc. 38th Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2015, pp. 43-52.

[10]. R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in Proc. 25th Int. Conf. World Wide Web, 2016, pp. 507-517.

[11]. J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing, 2019, pp. 188-197.

[12]. J. Jin, "Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 520-533, Aug. 2025, doi: 10.51903/jtie.v4i2.535.

[13]. Z. S. Zhong, J. Chen, E. Zhong, and X. Sun, "Evidence-calibrated RAG for unanswerable question answering: Retrieval coverage, abstention calibration, and hallucination-proxy analysis on SQuAD 2.0," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 502-520, Aug. 2025, doi: 10.51903/jtie.v4i2.536.

[14]. J. Jin, "LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy," Int. J. Graph. Des., vol. 3, no. 2, pp. 397-414, Oct. 2025, doi: 10.51903/ijgd.v3i2.3698.

[15]. X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in Proc. 26th Int. Conf. World Wide Web, 2017, pp. 173-182.

[16]. W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in Proc. IEEE Int. Conf. Data Mining, 2018, pp. 197-206.

[17]. F. Sun et al., "BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer," in Proc. 28th ACM Int. Conf. Information and Knowledge Management, 2019, pp. 1441-1450.

[18]. H. Steck, "Embarrassingly shallow autoencoders for sparse data," in Proc. World Wide Web Conf., 2019, pp. 3251-3257.

[19]. K. Jarvelin and J. Kekalainen, "Cumulated gain-based evaluation of IR techniques," ACM Transactions on Information Systems, vol. 20, no. 4, pp. 422-446, 2002.

[20]. Yifan Zhang and H. Zhang, "A therapist-facing session copilot for live counseling support: Reasoning-guided retrieval and ranking from multi-turn counseling dialogues," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 464-486, Aug. 2025, doi: 10.51903/jtie.v4i2.547.

[21]. S. Meng, J. Chen, and I. Zheng, "LLM-inspired offline reranking for financial search: Query rewriting, hybrid retrieval, and listwise relevance ranking on FiQA," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 361-378, Apr. 2026, doi: 10.51903/jtie.v5i1.537.

[22]. A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 5998-6008.

[23]. T. B. Brown et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 1877-1901.

[24]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171-4186.

[25]. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-

networks," in Proc. EMNLP-IJCNLP, 2019, pp. 3982-3992.

[26]. P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 9459-9474.

[27]. N. Tintarev and J. Masthoff, "Designing and evaluating explanations for recommender systems," in Recommender Systems Handbook, F. Ricci, L. Rokach, B. Shapira, and P. Kantor, Eds. Boston, MA, USA: Springer, 2011, pp. 479-510.

[28]. Y. Zhang and X. Chen, "Explainable recommendation: A survey and new perspectives," Foundations and Trends in Information Retrieval, vol. 14, no. 1, pp. 1-101, 2020.

[29]. Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma, "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis," in Proc. 37th Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2014, pp. 83-92.

[30]. L. Chen and P. Pu, "Critiquing-based recommenders: Survey and emerging trends," User Modeling and User-Adapted Interaction, vol. 22, no. 1-2, pp. 125-150, 2012.

[31]. A. Jacovi and Y. Goldberg, "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?" in Proc. 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4198-4205.

[32]. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135-1144.

[33]. B. Zhang, Y. Ren, and J. Zou, "LLM-style explainable e-commerce recommendation cards: A UI/UX design framework for trust-calibrated product recommendation," Int. J. Graph. Des., vol. 3, no. 2, pp. 381-396, Oct. 2025, doi: 10.51903/ijgd.v3i2.3697.

[34]. T. Ye, X. Chang, and E. Zhong, "Uncertainty-aware breast ultrasound explanation cards: A visual communication framework for image-based AI diagnostic support using BreastMNIST_224," Int. J. Graph. Des., vol. 3, no. 2, pp. 365-380, Oct. 2025, doi: 10.51903/ijgd.v3i2.3701.

[35]. Q. Wu, S. Meng, and J. Zhao, "Text-grounded LLM-assisted design rationale interfaces: Turning advertising layout metadata into explainable UI/UX decision cards," Int. J. Graph. Des., vol. 3, no. 1, pp. 216-240, May 2025, doi: 10.51903/ijgd.v3i1.3713.

[36]. B. Zhou, C. Li, and L. Liu, "Risk-calibrated patient-facing AI safety cards: A UI/UX design framework for rubric-based medical risk communication," Int. J. Graph. Des., vol. 3, no. 2, pp. 365-380, Oct. 2025, doi: 10.51903/ijgd.v3i2.3696.

[37]. H. Xu, Y. Chen, and A. Med, "Automatic detection and explanation of dark patterns from interface microcopy: Empirical comparison of BERT-style encoders, RoBERTa-style encoders, and LLM-style decoders on the ec-darkpattern dataset," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 590-612, Dec. 2025, doi: 10.51903/jtie.v4i3.491.

[38]. Y. Li, S. Lu, and L. Zhao, "LLM-as-design-critic: Aligning AI-generated UI feedback with human graphic design judgment," Int. J. Graph. Des., vol. 3, no. 1, pp. 196-215, May 2025, doi: 10.51903/ijgd.v3i1.3661.

[39]. J. Zhang, "Adaptive user interface design for volleyball learning apps: Empirical evidence from Google Play reviews and mobile screen analysis," Int. J. Graph. Des., vol. 3, no. 1, pp. 175-195, May 2025, doi: 10.51903/ijgd.v3i1.3618.

[40]. Yifan Zhang and H. Zhang, "Visualizing the right counseling support: Evidence-linked recommendation cards for explainable mental health intake interfaces," Int. J. Graph. Des., vol. 3, no. 1, pp. 214-229, May 2025, doi: 10.51903/ijgd.v3i1.3722.

[41]. Z. Li, S. Zhou, and Z. Zhou, "Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench," Int. J. Graph. Des., vol. 3, no. 1, pp. 196-210, May 2025, doi: 10.51903/ijgd.v3i1.3715.

[42]. W. Su, H. Rao, and E. Ma, "Privacy and data-integrity risk cards for LLM agents: A UI/UX design framework for secure human oversight under prompt-injection attacks," Int. J. Graph. Des., vol. 4, no. 1, pp. 186-191, Apr. 2026, doi: 10.51903/ijgd.v4i1.3699.

[43]. H. Tu, S. Zhao, and A. Zhou, "Visual brief cards for advertising design: A structured UI/UX framework for turning creative intentions into graphic design decisions," Int. J. Graph. Des., vol. 3, no. 1, pp. 210-226, May 2025, doi: 10.51903/ijgd.v3i1.3714.

[44]. J. Mu, Y. Lu, and E. Hwang, "Structured visual brief interfaces for advertising design: A UI/UX framework for turning creative intentions into designer-editable graphic design cards," Int. J. Graph. Des., vol. 4, no. 1, pp. 192-208, Apr. 2026, doi: 10.51903/ijgd.v4i1.3702.

[45]. J. Zhang, "Early warning, grade prediction, and teacher-facing LLM-ready explanations toward an open volleyball course: Reproducible evidence from four public education datasets," J. Technol.

Informatics Eng., vol. 5, no. 2, pp. 20-44, Jun. 2026, doi: 10.51903/jtie.v5i2.525.

[46]. M.-J. Kuo, D. Zheng, and J. Hires, "Federated topic-preference learning for knowledge-grounded chat with differential privacy," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 385-401, Aug. 2025, doi: 10.51903/jtie.v4i2.502.

[47]. W. Su, S. Chen, and C. Zhao, "Budgeted multi-hop retrieval agent for compositional question answering: A retrieval-policy evaluation on the official MultiHop-RAG benchmark," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 649-662, Dec. 2025, doi: 10.51903/jtie.v4i3.543.

[48]. Y. Chen and H. Xu, "Trust-calibrated multilingual RAG for humanitarian information platforms: Empirical evaluation on OMoS-QA for migration information access," Int. J. Graph. Des., vol. 4, no. 1, pp. 141-164, Apr. 2026, doi: 10.51903/ijgd.v4i1.3552.

[49]. S. Zhou, Y. Chen, and K. Lee, "Accounting-aware evidence-constrained agents for disclosure, settlement, and secondary-market risk monitoring in tokenized," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 60-74, Jun. 2026, doi: 10.51903/jtie.v5i2.544.

[50]. J. Bai, S. Chen, D. Zheng, and M.-J. Kuo, "Interpretable attack-chain stage detection from AWS CloudTrail event sequences via linear models and HMM smoothing," Inf. Electr. Electron. Eng., vol. 6, no. 1, pp. 28-43, May 2026, doi: 10.33474/infotron.v6i1.24923.

[51]. Y. Chen, S. Zhou, and E. Lin, "Accounting-aware evidence retrieval for institutional due diligence of tokenized trade receivable RWA," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 649-663, Dec. 2025, doi: 10.51903/jtie.v4i3.542.

[52]. J. Mu, T. Ye, and P. Patel, "Offline counterfactual evaluation for advertising and recommendation slot policies: A reproducible study on the Open Bandit Dataset (Small)," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 521-543, Dec. 2025, doi: 10.51903/jtie.v4i3.500.

[53]. C. Wang, Z. Wen, R. Zhang, P. Xu, and Y. Jiang, "GPU memory requirement prediction for deep learning task based on bidirectional gated recurrent unit optimization transformer," in Proc. 5th Int. Conf. Artificial Intelligence, Virtual Reality and Visualization, Chengdu, China, 2025, doi: 10.1109/AIVRV67401.2025.11350369.

[54]. B. Zhou, H. Wang, and X. Chang, "Distilling VMAF into an edge-deployable quality predictor: A pilot shot-level proxy with LLM-ready quality tokens," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 447-463, Aug. 2025, doi: 10.51903/jtie.v4i2.522.

[55]. R. Zhang, Z. Wen, C. Wang, C. Tang, P. Xu, and Y. Jiang, "Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms," arXiv:2511.19481, 2025, doi: 10.48550/arXiv.2511.19481.

[56]. C. Li, G. Liu, and Z. Zhao, "Cost-aware LLM-style routing for AIOps log analysis: Log parsing, anomaly detection, fault diagnosis, and incident summarization on LogEval task files," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 91-103, Jun. 2026, doi: 10.51903/jtie.v5i2.538.

[57]. Q. Wu, G. Mi, and D. Wood, "Calibration-light subject-independent motor imagery BCI via self-supervised pretraining and Conformer," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 239-262, Apr. 2025, doi: 10.51903/jtie.v5i1.493.

[58]. Z. S. Zhong and S. Ling, "Improved theoretical guarantee for rank aggregation via spectral method," Inf. Inference: J. IMA, vol. 13, no. 3, Art. no. iaae020, Sep. 2024, doi: 10.1093/imaiai/iaae020.

[59]. Z. Li, K. Zhang, and A. Wong, "Numerical-reasoning guardrails for a quant research assistant: A compact reproducible benchmark using SEC and FRED data," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 75-90, Jun. 2026, doi: 10.51903/jtie.v5i2.541.

[60]. Z. S. Zhong and S. Ling, "Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization," arXiv:2408.05944, Aug. 2024.

[61]. [61] S. Lu and T. Zou, "Uncertainty-aware medical vision-language classification on a lightweight MedMNIST-compatible biomedical patch benchmark," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 1-19, Jun. 2026, doi: 10.51903/jtie.v5i2.530.

[62]. [62] S. Zhao, J. Bai, and D. Roberson, "Multi-horizon GPU demand forecasting with workload semantics and operational risk curves: An empirical study on Alibaba Clusterdata GPU Trace," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 544-571, Dec. 2025, doi: 10.51903/jtie.v4i3.498.

[63]. H. Zhou and K. Zhang, "News-based uncertainty and macro-market fusion for VIX direction forecasting: Evidence from 2015-2024 FRED panel," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 487-501, Aug. 2025, doi: 10.51903/jtie.v4i2.540.

[64]. S. He, J. Nie, and C. Li, "Power-aware inventory planning for AI infrastructure using job-level forecasting and LLM workload explanations," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 341-359, Apr. 2026, doi: 10.51903/jtie.v5i1.548.

[65]. Y. Lu, H. Zhou, and Yitian Zhang, "A constrained, data-driven budgeting framework integrating macro demand forecasting and marketing response modeling," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 493-520, Dec. 2025, doi: 10.51903/jtie.v4i3.466.

[66]. J. Nie, G. Liu, C. Li, and T. Zou, "Evidence-constrained incident visualization cards for distributed cloud logs: A UI/UX framework for turning Hadoop, OpenStack, and ZooKeeper logs into actionable SRE design interfaces," Int. J. Graph. Des., vol. 4, no. 1, pp. 179-185, Apr. 2026, doi: 10.51903/ijgd.v4i1.3703.

[67]. H. Wang, Y. Ren, and X. Chang, "Layout-aware progressive PDF rendering: AI prioritization of PDF slices to reduce time-to-functional-first-frame on FUNSD," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 425-446, Aug. 2025, doi: 10.51903/jtie.v4i2.523.

[68]. B. Zhang, X. Sun, G. Liu, and B. Zhou, "LLM-style DevOps copilot for cloud-native troubleshooting: Retrieval-augmented runbook generation and command-safety evaluation," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 104-118, Jun. 2026, doi: 10.51903/jtie.v5i2.534.

[69]. B. Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, "Predictive optimization of DDoS attack mitigation in distributed systems using machine learning," in Proc. 6th Int. Conf. Computing and Data Science, 2024, pp. 89-94.

[70]. Y. Li and S. Lu, "Language-guided feature selection for DDoS and intrusion detection on CICIDS2017," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 284-305, Apr. 2025, doi: 10.51903/jtie.v4i1.531.

[71]. Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, "Optimization of autonomous driving image detection based on RFAConv and triplet attention," in Proc. 2nd Int. Conf. Software Engineering and Machine Learning, 2024.