Stout in Business and Management (SBM)
ISSN: 2979-2827
DOI: https://doi.org/10.61424/c20v0j71
Journal Homepage: https://stoutjournals.org/index.php/SBM

Review-Grounded LLM Decision Support for E-Commerce: Faithful Recommendation, Customer Segmentation, and Selective Personalization

Fiona Huang
Computer Science, University of Colorado Boulder, Boulder, CO, USA
Corresponding Author: Fiona Huang, **E-mail**: f.huang.cs@proton.me

ARTICLE INFORMATION

Submitted: April 20, 2026
Accepted: July 24, 2026
Published: August 01, 2026
Volume: 2
Issue: 1

KEYWORDS

E-commerce recommendation; Amazon Reviews'23; LightGCN; review-grounded explanation; customer segmentation; selective personalization.

ABSTRACT

E-commerce decision support requires accurate ranking, review-grounded explanations, and a principled way to limit personalization when confidence is weak. This study evaluates those capabilities on 701,528 Amazon Reviews'23 All Beauty records. Four- and five-star reviews defined positive preference, an iterative interaction core supported temporal leave-one-out evaluation, and every method ranked the exact 1,091-item warm catalog. Popularity, ItemKNN, PureSVD, BPR-MF, LightGCN, a training-review TextLSA ranker, and a validation-weighted hybrid were compared for 2,374 test users. The hybrid attained NDCG@10 of 0.09186 and Recall@20 of 0.19966, improving on LightGCN by 15.80% and 17.62%, respectively. Its paired NDCG@10 gain was 0.01253 (95% bootstrap interval [0.00595, 0.01958]; paired randomization p = .00030). Review fusion also reduced expected calibration error from 0.29112 to 0.25319 and area under the risk–coverage curve from 0.67798 to 0.63844, although absolute confidence remained overestimated. Six review-behavior segments showed substantial ranking heterogeneity, with Recall@20 ranging from 0.02000 to 0.25000. A local Qwen2.5-0.5B assessment on four balanced evidence packets achieved 0.75 selective-action accuracy but only 0.25 citation accuracy, exact-quote fidelity, and joint constraint compliance. The findings support review-language fusion for ranking and confidence ordering while showing that customer-facing explanations require deterministic validation and abstention safeguards.

1. Introduction

E-commerce recommendation is increasingly a decision-support problem rather than a narrow exercise in predicting the next interaction. A useful system must identify products that fit a customer's observed preferences, explain why those products are relevant, and recognize when the available evidence is too weak for confident personalization. Classical collaborative-filtering methods provide a strong basis for the first requirement. Matrix factorization represents users and items in a shared latent space (Koren et al., 2009), confidence-weighted learning adapts that idea to implicit feedback (Hu et al., 2008), Bayesian personalized ranking directly optimizes pairwise item order (Rendle et al., 2009), and LightGCN propagates collaborative signals through the user–item graph with a deliberately simplified architecture (He et al., 2020). These models are effective at exploiting interaction structure, but their scores do not by themselves state which observed customer evidence supports a recommendation.

Review text supplies that missing semantic layer. Ratings indicate preference direction, whereas written reviews reveal attributes, use contexts, perceived benefits, and reasons for dissatisfaction. Early review-aware models connected latent rating factors to textual topics (McAuley & Leskovec, 2013), and later neural architectures jointly represented users, items, and their reviews (Zheng et al., 2017) or assigned attention to individual reviews to support review-level explanations (Chen et al., 2018). This literature established that text can enrich preference modeling, but it also showed that review signals must be incorporated carefully. Reviews may be sparse, repetitive, written after an interaction, or so predictive of the held-out target that an evaluation inadvertently leaks information. Indeed, critical experiments have found that apparent improvements from reviews depend strongly on preprocessing and evaluation design (Sachdeva & McAuley, 2020).

Large language models (LLMs) extend review-aware recommendation by allowing preference evidence and candidate items to be expressed in natural language. P5 framed multiple recommendation tasks within a unified language-processing paradigm (Geng et al., 2022), while TALLRec showed how instruction tuning can align an LLM with recommendation judgments using parameter-efficient adaptation (Bao et al., 2023). Empirical studies of ChatGPT further demonstrated that pointwise, pairwise, and listwise prompts produce meaningfully different ranking behavior (Dai et al., 2023). LLMs can also operate as conditional rerankers over a manageable candidate set, although their output is sensitive to history order, item position, and popularity (Hou et al., 2024). These findings favor a hybrid architecture in which an efficient collaborative model retrieves candidates and a language component reranks or explains them from explicit review evidence. Such an architecture uses the LLM where semantic reasoning is valuable without requiring it to score the complete item catalog.

The present study develops this hybrid view on the All Beauty portion of Amazon Reviews'23. The benchmark was collected through September 2023 and made available through its project site in 2024 (McAuley Lab, 2024); its peer-reviewed benchmark description subsequently appeared at ACL in 2026 (Hou et al., 2026). The analysis uses the review-interaction file, which contains user and product identifiers, ratings, timestamps, review text, helpfulness signals, and verified-purchase indicators. It does not assume access to the separate product-metadata file and therefore does not rely on product descriptions, category hierarchies, prices, or images. This distinction is methodologically important: language evidence comes from customer reviews, and customer segments are derived from observable review and interaction behavior rather than monetary value.

Three research questions organize the study. First, how much ranking value is contributed by collaborative structure, review-language reranking, and their combination? Second, can generated decision support remain traceable to retrieved review evidence instead of offering merely plausible prose? Third, can confidence calibration and selective personalization improve reliability across behaviorally distinct customer groups? The evaluation consequently combines ranking metrics, calibration measures, evidence-faithfulness tests, and risk–coverage analysis. This integrated design treats accuracy, explanation quality, and abstention as connected properties of a deployable recommendation system rather than unrelated auxiliary objectives.

2. Literature Review

2.1 Collaborative Candidate Generation and Ranking Evaluation

Collaborative recommendation learns preference from recurring user–item interaction patterns. Matrix factorization remains influential because it compresses a sparse interaction matrix into user and item vectors whose inner products yield personalized scores (Koren et al., 2009). For implicit observations, Hu et al. (2008) distinguished a preference indicator from the confidence assigned to that indicator, allowing repeated or stronger behavior to receive greater weight. BPR instead learns from triples containing a user, an observed item, and an unobserved item, maximizing the probability that the observed item ranks higher (Rendle et al., 2009). LightGCN retains this pairwise recommendation objective while aggregating embeddings over the bipartite interaction graph and removing feature transformations and nonlinearities that were not shown to benefit collaborative filtering (He et al., 2020). Together, these methods provide complementary baselines: popularity captures nonpersonalized demand, factor models capture low-rank affinity, and graph propagation captures higher-order collaborative proximity.

Top-K evaluation must match the intended ranking task. Discounted cumulative gain gives greater credit when relevant items occur nearer the top and normalizes performance across users, providing the basis for NDCG (Järvelin & Kekäläinen, 2002). Recall measures whether held-out relevant items appear in the recommendation set. Both metrics, however, can be distorted when each positive is evaluated against a small sample of negatives rather than the available catalog (Krichene & Rendle, 2020). Temporal construction is equally important because random splitting can place future interactions in training or expose target-linked information through features. Ji et al. (2023) showed that leakage choices can materially change offline conclusions. A chronological user-level split, training-only feature construction, and a fixed candidate protocol are therefore necessary for a defensible comparison of collaborative and language-enhanced methods.

2.2 Review-Aware Language Modeling and Evidence Grounding

Review-aware recommendation developed from the observation that a scalar rating hides the attributes responsible for preference. Hidden Factors and Hidden Topics aligned rating dimensions with topics induced from review text (McAuley & Leskovec, 2013). DeepCoNN subsequently learned user and item representations from review collections using parallel neural networks (Zheng et al., 2017), while NARRE added review-level attention so that reviews could contribute unequally to a prediction and could be surfaced as explanatory evidence (Chen et al., 2018). These approaches connect semantics with collaborative signals, but attention or textual association is not automatically a faithful account of model behavior. The critical analysis by Sachdeva and McAuley (2020) is particularly relevant: review-based gains should be tested under conditions that prevent target-review leakage and should be compared against properly tuned nontextual baselines.

Modern sentence encoders make a lightweight semantic reranking stage practical. Sentence-BERT produces fixed-dimensional embeddings suitable for efficient cosine comparison rather than requiring a cross-encoder pass for every text pair (Reimers & Gurevych, 2019). A user profile can thus be formed from training-period review text, candidate evidence can be aggregated from reviews available before the prediction cutoff, and semantic compatibility can be combined with a collaborative score. This evidence-restricted design differs from asking an LLM to recommend freely from memorized product knowledge. The candidate generator defines the eligible items, and the language stage must ground its preference rationale in retrieved review passages.

Explainable-recommendation research provides concrete precedents for this constraint. Ni et al. (2019) generated justifications from distantly labeled reviews and fine-grained aspects, and the Personalized Transformer unified recommendation with personalized explanation generation (Li et al., 2021). Yet explanation quality has competing goals: a fluent statement may be persuasive without accurately reflecting the evidence or the decision process (Balog & Radlinski, 2020). ERASER operationalized rationale evaluation through measures of agreement with supporting evidence and changes in prediction behavior when rationales are retained or removed (DeYoung et al., 2020). In recommendation specifically, neural logic reasoning has been used to produce traceable paths and to evaluate whether explanations are faithful to the mechanism that produced the recommendation (Zhu et al., 2021). These studies motivate evidence precision, evidence coverage, and deletion-based tests alongside ordinary linguistic quality.

2.3 Calibration, Selective Personalization, and Customer Segments

Ranking quality does not ensure that a model's confidence is reliable. Calibration asks whether predictions assigned a given probability succeed at approximately that rate. Bayesian binning established a flexible approach to mapping raw scores to calibrated probabilities (Naeini et al., 2015), while temperature scaling emerged as a simple and effective post-hoc correction for modern neural networks (Guo et al., 2017). In recommendation, calibration also has a preference-composition meaning: recommended items should not systematically depart from the distribution of interests represented in a user's history (Steck, 2018). The present setting separates these concepts by evaluating probabilistic calibration of interaction scores and distributional alignment of personalized lists.

When confidence remains low, abstention can be more useful than a forced personalized answer. SelectiveNet formalized joint prediction and rejection under a target coverage level, directly linking the fraction of cases served to error on those cases (Geifman & El-Yaniv, 2019). This risk–coverage perspective is well suited to sparse e-commerce histories: the system can personalize for customers with sufficient evidence, fall back to a robust nonpersonalized ranking when evidence is limited, and report how accuracy changes as coverage increases.

Customer heterogeneity makes aggregate performance insufficient. RFM research traditionally segments customers by recency, frequency, and monetary value and links those summaries to customer lifetime value (Fader et al., 2005). Comparative work has shown that RFM-style segmentation remains competitive with more elaborate response-modeling approaches (McCarty & Hastak, 2007), and retail applications have combined behavioral summaries with clustering to identify actionable customer groups (Chen et al., 2012). Because the All Beauty review file contains no transaction price, the relevant adaptation uses recency and frequency together with review-derived variables such as rating tendency, review length, helpfulness, and verified-purchase share; it does not label these features as monetary value. Segment-level ranking, calibration, explanation, and coverage results then reveal whether a method's overall advantage is broadly shared or concentrated among highly active reviewers. This synthesis turns review-grounded recommendation into a business decision-support system that can rank effectively, explain from observable evidence, and limit personalization when the evidence does not justify it.

2.4 Recent Review-Grounded Personalization

Recent work treats personalization as linked decisions: represent the customer, retrieve items, refine the ranking with language evidence, and decide whether the history justifies personalization. E-commerce classification and recommendation research supplies a broad behavioral foundation (K. Xu et al., 2024). Self-supervised customer representations connect segmentation with next-purchase prediction (Xin, 2026f), while review-grounded recommendation on Amazon data couples explanations with faithfulness evaluation (Chang et al., 2026). Together, these studies motivate using interaction structure for preference and reviews for inspectable semantic support.

Language models add flexibility but do not remove the need for controlled retrieval. Behavior retrieval followed by response generation grounds conversational recommendation in observable history (Xin, 2026b). Long-term dialogue memory separates memory writing, retrieval, updating, and forgetting (Sun et al., 2023). Although dialogue is not product ranking, its confidence-gated logic is a useful cross-domain precedent for limiting weak or stale profile evidence. Federated topic-preference learning

further shows how knowledge-grounded personalization can limit central pooling of fine-grained interests (Kuo et al., 2025). These works inform profile construction; they are not experiments conducted on All Beauty.
Cold start remains a boundary condition. Few-shot workload forecasting illustrates adaptation when entity-specific history is limited (He et al., 2025). It is an infrastructure precedent, not evidence about consumer demand, but it distinguishes transferring a representation from inventing a preference. Financial-search query rewriting, hybrid retrieval, and listwise reranking likewise organize candidates while leaving relevance domain-specific (Meng et al., 2026). The retail analogue is candidate generation from observed interactions, review-language reranking, and fallback for sparse histories.

Interface studies extend personalization to explanation. E-commerce recommendation cards make rationale and trust calibration visible (B. Zhang et al., 2025), while evidence-linked counseling cards preserve a path to the source material (Y. Zhang & Zhang, 2025b). Counseling is only a cross-domain design precedent; its risk criteria do not transfer to retail. Review-based mobile-interface research also shows that commentary can guide interface choices as well as scoring (J. Zhang, 2025). Review evidence can therefore serve several functions, provided each retains provenance.

Personalization is also a ranking and policy problem. Spectral rank aggregation explains how noisy preference orderings may be combined (Zhong & Ling, 2024a). Offline counterfactual evaluation asks how advertising or recommendation slot policies would perform under logged exposure (Mu et al., 2025), and conservative off-policy selection emphasizes uncertainty in policy choice (Ye et al., 2026). These studies caution that an offline ranking gain is not automatically a causal increase in purchases.

2.5 Evidence Retrieval, Attribution, and Provenance

The reliability question for an LLM-assisted recommender is whether its language is entailed by available evidence. Evidence-chain RAG addresses word-level hallucination, attribution, and provenance (C. Li et al., 2024), while deterministic provenance makes source links reproducible (Jin, 2025b). Evidence-calibrated RAG adds retrieval coverage and abstention for unanswerable inputs (Zhong, Chen, et al., 2025). Applied to reviews, these ideas favor stable review identifiers, quotation-level support, and an "insufficient evidence" path.

Operational RAG offers architectural precedents because incident language can also sound authoritative without support. Multi-source root-cause attribution assembles heterogeneous evidence before diagnosis (Liu et al., 2023). DevOps question answering emphasizes hallucination resistance over bounded private documents (B. Zhang et al., 2024), and bilingual incident triage joins retrieval, attribution, and summarization (Liu et al., 2024). These are cross-domain precedents, not e-commerce evaluations: their transferable pattern is bounded retrieval with source identity preserved through synthesis.

Log research makes the pattern granular. Failure-injection anomalies can be explained with retrieved normal prototypes (Xin, 2025a), HDFS narratives can remain grounded in logs (Sun et al., 2026), and a DevOps copilot separates runbook grounding from command-safety evaluation (B. Zhang et al., 2026). Reviews replace operational artifacts in this study; no operational findings are imported. Retrieval quality, narrative faithfulness, and action safety remain distinct checkpoints.

Other domains show how evidence structure can follow the decision. Accounting-aware retrieval organizes disclosure evidence (Y. Chen et al., 2025); budgeted multi-hop retrieval constrains acquisition cost (Su et al., 2025); and scientific verification ranks evidence against a claim (Su, Chen, et al., 2026). A counseling copilot retrieves from multi-turn dialogue (Y. Zhang & Zhang, 2025a), while humanitarian RAG adds multilingual trust calibration (Y. Chen & Xu, 2026). All are cross-domain design precedents, not benchmarks run here. Their shared lesson is that evidence selection should be task-aware, budget-aware, and traceable.

Retrieval must be evaluated before generation. Code-intelligence research separates findability from explainability (Y. Li, 2024), and retrieval-quality prediction treats weak RAG input as a precursor to generation failure (R. Zhang et al., 2025). Evidence-grounded policy memos likewise document an infrastructure decision without replacing its rule (Zhao et al., 2026). The retail analogue separates collaborative candidate generation, review reranking, and explanation: fluent language cannot repair a missed candidate, and ranking quality cannot certify a quotation.

2.6 Calibration and Selective Decision Support

Recommendation scores become decision-support signals only when their uncertainty is interpreted carefully. Capacity forecasting with calibrated uncertainty demonstrates a general pattern: produce a point forecast, quantify its uncertainty, and translate that uncertainty into a safer operational choice (He et al., 2023). Credit-default tiering combines probability calibration with uncertainty-driven rejection (Y. Chen et al., 2023), while a model-risk-oriented default workflow joins calibration, distribution-free uncertainty quantification, and explanation (Jin et al., 2024). These financial and infrastructure studies are cross-domain methodological precedents. They do not validate a retail probability model, but they support evaluating whether confidence corresponds to empirical success and whether uncertain cases should be deferred.

Conformal methods offer another vocabulary for limiting risk. Conformal demand envelopes have been used to bound oversubscription decisions in production GPU clusters (S. Chen et al., 2024), and conformal alert control has been paired with evidence-grounded incident tickets (Xin, 2026e). Neither workload demand nor anomaly alerts are consumer choices; the transferable principle is the explicit relationship among a score, a coverage level, and an accepted error rate. In a recommender, selective personalization should therefore be reported as a risk–coverage tradeoff rather than as an assertion that one global threshold is universally safe.

Calibration can also be separated from fusion and ranking. Uncertainty-aware late fusion for 3D perception learns how confidence should influence the combination of heterogeneous signals (Xin, 2025c). Calibrated recruiter screening similarly connects pairwise matching with probability calibration and selective refusal (Jin, 2025a). These are cross-domain precedents for combining collaborative and text scores without treating either score scale as intrinsically probabilistic. Evidence-calibrated RAG further shows that retrieval coverage, abstention, and hallucination proxies can be analyzed together (Zhong, Chen, et al., 2025), which is especially relevant when a recommendation explanation depends on retrieved reviews.

Two theoretical perspectives help state the limits of transfer. Work on approximately shared features asks when predictive structure can bridge domains without assuming complete distributional identity (Zhong, Pan, et al., 2025). Uncertainty quantification for spectral estimators and maximum-likelihood estimators shows that ranking-related estimates themselves can carry uncertainty (Zhong & Ling, 2024b). Edge-oriented quality-predictor distillation, with compact quality signals intended for downstream language systems, offers a deployment precedent for preserving a confidence indicator after model compression (Zhou, Wang, et al., 2025). The latter is again a cross-domain design precedent from video quality assessment. Together, these studies argue for calibrated post-processing and selective fallback, while cautioning against transferring thresholds or probability interpretations unchanged from another dataset.

2.7 Privacy, Safety, and Governance Boundaries

Personalization converts individual behavior into an automated judgment. Differentially private topic-preference learning limits exposure of user interests within knowledge-grounded personalization (Kuo et al., 2025). Privacy-robust incrementality estimation addresses learning under identifier loss (Bai et al., 2026), and privacy-safe marketing-mix modeling studies budget decisions without user-level identifiers (Bai & Wu, 2026). Despite different estimands, all discourage assuming that finer customer data is always necessary.

Safety also concerns manipulated evidence or instructions. Continual red-teaming treats jailbreak defense as an updating process (Zheng et al., 2023). Evidence-based self-verification adds an adversarial layer (Zheng et al., 2024), while multi-stage refusal balances resistance with legitimate utility (Zheng & Li, 2024). These are not recommendation experiments, but they motivate preventing prompts from eliciting fabricated reviews, unrelated profile content, or ungrounded persuasion.

Interface manipulation is part of the same boundary. Dark-pattern detection shows how ordinary microcopy can encode harmful persuasion (H. Xu et al., 2025); recommendation rationales should not invent urgency, scarcity, or endorsement. Numerical guardrails illustrate domain validation between generation and presentation (Z. Li et al., 2026). In e-commerce, deterministic checks on item identity, rating statements, and quotations are preferable to fluency as a proxy for correctness.

Governance extends beyond outputs. Risk cards expose prompt-injection and data-integrity concerns to overseers (Su, Rao, et al., 2026), while trajectory prediction forecasts tool-use failure (Zhong et al., 2026). Multi-regulation product-counsel RAG uses jurisdiction-specific sources (J. Li & Zhou, 2026). These cross-domain precedents do not establish compliance here; they support logging evidence, fallbacks, and model actions.

Fairness is inseparable from confidence. Counterfactual credit-risk explanations combine performance, fairness, and recourse (Xin, 2026d). Credit criteria cannot transfer mechanically to retail, but aggregate ranking can still conceal unequal error or coverage. Segment evaluation and selective fallback should therefore be inspected together.

2.8 Evidence Cards and Human-Facing Trust

Evidence cards make model support inspectable. E-commerce recommendation cards combine product rationale with trust-calibrated presentation (B. Zhang et al., 2025). Scientific-search cards add retrieval transparency and visual hierarchy (Jin, 2025c), while accounting cards pair claims with source passages (K. Zhang et al., 2025). The latter two are cross-domain precedents; all three support visibly separating recommendation, confidence, excerpt, source, and fallback.

Medical interfaces sharpen uncertainty communication. Radiology cards combine visual explanation with uncertainty (Zhong, Wu, et al., 2025). Patient safety cards use explicit risk rubrics (Zhou, Li, et al., 2025), and a related benchmark evaluates explainable medical-response interfaces (C. Li et al., 2025). These are only cross-domain precedents for hierarchy and disclosure; their medical categories, data, and findings do not transfer.

Dashboard and decision-card research broadens that interface logic. Financial-risk dashboards emphasize visual hierarchy, chart comprehension, and explainability (Z. Li et al., 2025). Capacity-dashboard brief cards turn forecast risk into design decisions (Liu et al., 2025), and incident-visualization cards transform distributed logs into actionable, evidence-constrained SRE views (Nie et al., 2026). Explainable financial-ratio portraits similarly show how structured visual summaries can accompany a predictive model (Y. Chen et al., 2024). These are all cross-domain design precedents. For review-grounded commerce, they justify displaying the source and strength of evidence without suggesting that an attractive visualization increases evidentiary quality.

Design rationale itself can be grounded. Text-grounded advertising interfaces connect layout metadata to an inspectable design rationale (Wu et al., 2025), while visual brief cards convert creative intentions into structured graphic-design choices (Tu et al., 2025). Designer-editable advertising cards further preserve human revision as part of the workflow (Mu et al., 2026). An LLM-as-design-critic approach adds comparison with human graphic-design judgment (Y. Li et al., 2025). These advertising and design studies do not test recommendation accuracy. They instead support a human-in-the-loop presentation layer in which a reviewer can edit phrasing or suppress a rationale without silently changing the underlying ranked list.

Human-facing explanations also need task-appropriate action. Evidence-linked counseling recommendation cards make the supporting dialogue visible (Y. Zhang & Zhang, 2025b), adaptive sports-learning interfaces use review and screen evidence to guide design (J. Zhang, 2025), and rubric-grounded essay feedback emphasizes auditability of the link between judgment and formative response (Xin, 2026a). Early-warning analytics with intervention rationales similarly connects a predictive signal to a proposed human action (Xin, 2026c). Each is a cross-domain design precedent. Their common contribution is not a reusable substantive recommendation, but a presentation rule: distinguish observed evidence, model inference, and suggested action so the user can challenge any one of them independently.

2.9 Marketing and Off-Policy Actionability

A recommendation becomes commercially actionable only when ranking performance is distinguished from causal response. LLM-assisted uplift modeling for email advertising models heterogeneous effects across audience, creative, and channel variables (Mu et al., 2023). Privacy-robust uplift modeling revisits incrementality when conventional identifiers are unavailable (Bai et al., 2026). These studies motivate segment-level decision support, but they also establish an important limit: a high offline relevance score is not an estimate of incremental purchase probability. The All Beauty evaluation can measure ranking and evidence quality; it cannot infer campaign lift without an exposure or intervention design.

Logged-bandit research provides a more appropriate framework when policy effects are the target. Offline counterfactual evaluation examines advertising and recommendation slot policies on logged actions (Mu et al., 2025), and conservative off-policy evaluation adds an explicit policy-selection step (Ye et al., 2026). The practical implication is that a language reranker should first be validated as a ranker under the study's fixed candidate protocol. Deployment as a merchandising or bidding policy would require propensity information, support checks, and off-policy uncertainty that ordinary review logs do not provide.

Budgeting research connects those local choices to portfolio allocation. A constrained data-driven framework integrates macro demand forecasting with marketing-response modeling (Lu et al., 2025), while privacy-safe marketing-mix modeling treats budget optimization under identifier loss (Bai & Wu, 2026). These are adjacent business-method precedents rather than experiments reproduced here. They show how calibrated response estimates might eventually inform budget decisions, but they should not be used to turn recommendation metrics into unsupported revenue forecasts.

Finally, profit-aware GPU admission control offers a deliberately cross-domain analogy: it combines cost-sensitive learning with evidence-grounded policy memos so that an operational choice is both optimized and documented (Zhao et al., 2026). In commerce, an analogous system could pair a ranking decision with an evidence card and a clearly stated fallback. The analogy stops at the decision architecture; GPU costs, service constraints, and admission labels have no empirical bearing on beauty-product demand. This distinction keeps the literature useful without implying that every action-oriented source validates the present recommender.

2.10 Efficient Deployment and Transfer Boundaries

A deployable system should allocate computation by stage. Hybrid-cloud LLM architecture balances capability, latency, and cost (Xin, 2025b). Cost-aware AIOps routing separates parsing, detection, diagnosis, and summarization (C. Li et al., 2026). Edge

quality distillation compresses an assessor while retaining downstream signals (Zhou, Wang, et al., 2025). These are cross-domain precedents, not recommender evaluations; they support reserving generation for a small candidate set with deterministic output checks.

Transfer should be treated as a hypothesis, not a shortcut. Cross-cloud capacity forecasting explicitly considers workload and topology distribution shift (He et al., 2024), and approximately shared-feature theory formalizes the idea that domains may overlap only in part (Zhong, Pan, et al., 2025). Few-shot forecasting for new tenants illustrates adaptation when the target entity has little history (He et al., 2025). In e-commerce, these precedents justify testing cold-start and sparse-history behavior separately, but they do not justify importing infrastructure thresholds, features, or reported performance. Product demand, review language, and exposure processes remain specific to the All Beauty data.

Retrieval systems face a similar transfer boundary. Financial-search reranking combines query rewriting, hybrid retrieval, and listwise scoring (Meng et al., 2026), while multilingual humanitarian RAG evaluates trust calibration under language and information-access constraints (Y. Chen & Xu, 2026). Retrieval-quality prediction provides a way to flag weak RAG inputs before generation (R. Zhang et al., 2025). These are useful pipeline precedents, yet financial relevance labels and humanitarian answerability are not substitutes for product preference. For the present application, any transferred encoder or retrieval heuristic must be evaluated on review-grounded ranking and citation fidelity in its actual catalog.

Efficient deployment also requires separating model capability from operational permission. A DevOps copilot can retrieve evidence and draft runbooks while applying a distinct command-safety evaluation (B. Zhang et al., 2026). Generalist-agent trajectory reliability can be forecast from tool-use failures (Zhong et al., 2026), and continually updated guardrails can respond to newly observed jailbreak behavior (Zheng et al., 2023). These studies support layered controls around an LLM component. They do not imply that the recommender in this study autonomously executes purchases, modifies customer profiles, or operates tools; the present decision-support scope ends at ranked suggestions, evidence display, and selective fallback.

The resulting boundary is both methodological and practical. Rank aggregation can combine noisy orderings (Zhong & Ling, 2024a), calibrated fusion can reconcile heterogeneous score sources (Xin, 2025c), and uncertainty-aware selection can limit the cases passed to a language generator (He et al., 2023). However, each transferred component must preserve the semantics of its target variable. Collaborative relevance, review-text similarity, calibrated success probability, explanation fidelity, and causal lift are not interchangeable quantities. A defensible deployment therefore uses the smallest adequate model at each stage, measures each stage against its own criterion, and treats cross-domain studies as design precedents rather than as empirical evidence about Amazon Reviews'23.

3. Methodology

3.1 Data, unit of analysis, and relevance definition

The study used the All Beauty review file from Amazon Reviews'23. The file contained 701,528 review records from 631,986 users and 112,565 parent-ASIN product families. Parent ASIN was used as the item identifier because it consolidates closely related variants into one product family. The raw data contained 693,929 unique user–parent-ASIN pairs. When more than one row referred to the same user and parent ASIN, the chronologically latest row was retained. This deterministic rule prevented repeated reviews of the same product family from being counted as independent preference events.

Recommendation relevance was defined before model fitting as a rating of four or five stars. This choice separated positive preference from review activity. Treating every review as a positive interaction would have made a one-star complaint indistinguishable from a five-star endorsement in an implicit-feedback graph. Among the raw records, 71.288% had four or five stars, and deduplication produced 494,769 high-satisfaction user–item pairs. The descriptive data also showed that 90.512% of reviews were verified purchases and only 0.103% had an empty review body. These properties supported a text-augmented preference study while also revealing substantial interaction sparsity.

An iterative core was constructed from positive user–item pairs. Users were required to have at least two positive interactions, and items were required to have at least five positive interactions. Both constraints were reapplied until the graph stopped changing. Convergence required 26 iterations and retained 9,603 positive edges connecting 3,048 users and 1,103 parent-ASIN items. This asymmetric core favored an item catalog with enough collaborative evidence while retaining users with short but usable preference histories.

Table 1. All Beauty data profile before recommendation filtering.

Measure	Value
JSONL records	701,528
Users	631,986
Parent-ASIN items	112,565
Unique user-parent pairs	693,929
Verified-purchase share	0.9051
Four-or-five-star share	0.7129
Empty review-body share	0.0010

Note. Counts refer to the complete review file before construction of the positive-interaction core.

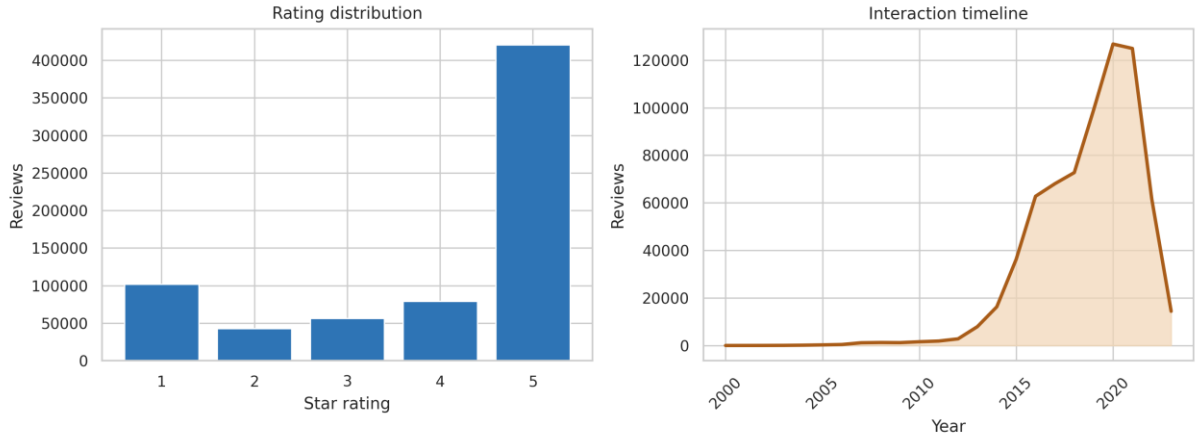

Figure 1. Rating distribution and principal characteristics of the All Beauty review file.

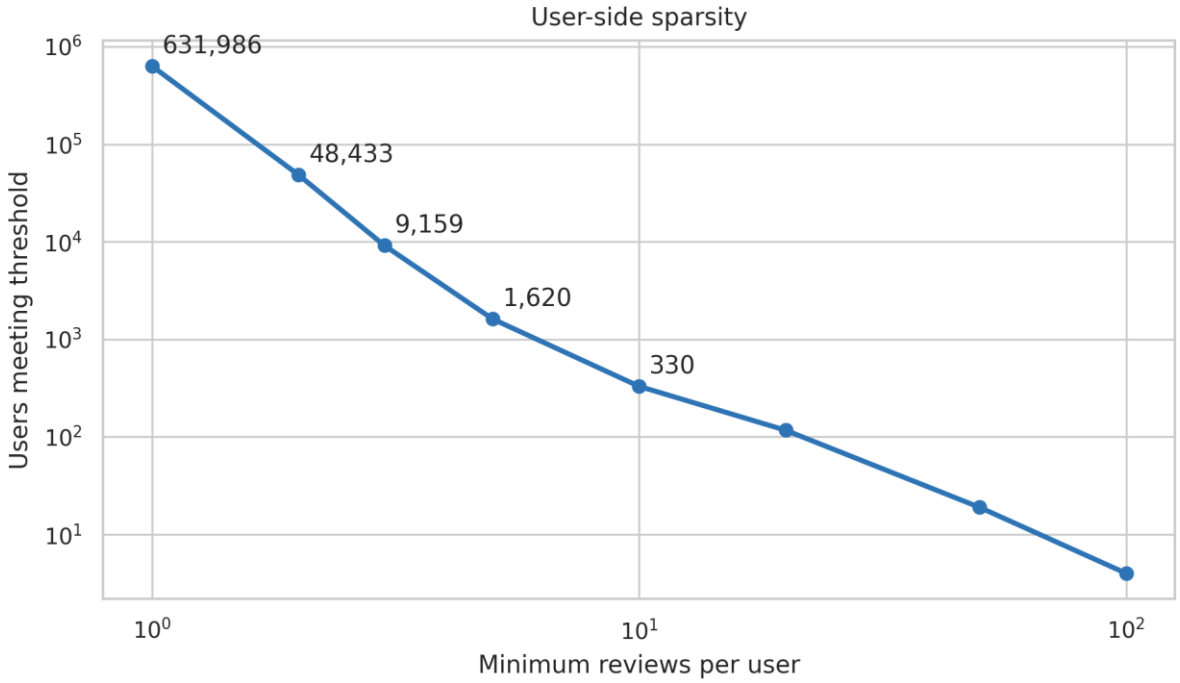

Figure 2. User-history sparsity in the complete review file.

3.2 Temporal holdout and evaluation population

For every core user, the chronologically latest positive interaction was assigned as the held-out target, and all earlier positive interactions were assigned to training. This produced 6,555 training interactions. Twelve core items occurred only among held-out events, leaving 1,091 items in the training catalog. Of the 3,048 held-out events, 2,967 concerned items represented in training and 81 concerned items absent from training. The latter constituted a measured cold-item rate of 2.66%. They were recorded as catalog-coverage failures and were not silently mapped to another item.

The 2,967 warm users were partitioned with random seed 20260729 into disjoint validation and test populations. Validation contained 593 users, and the untouched test population contained 2,374 users. The validation users were used to select the collaborative model and the language-fusion weight. Test users were used only for the final comparisons. This user-level separation prevented an individual's held-out outcome from contributing to both selection and final evaluation.

Ranking was performed against the exact 1,091-item warm catalog. For each user, products already present in that user's training history were masked before ranking. No sampled-negative evaluation was used. With one relevant held-out product per user, Recall@20 equals the proportion of users whose target appeared among the first 20 recommendations. NDCG@10 equals $1/\log_2(r_u+1)$ when the target rank r_u is at most 10 and zero otherwise. MRR@20 equals $1/r_u$ when $r_u \leq 20$ and zero otherwise. Median target rank was retained as an interpretable catalog-wide diagnostic.

Table 2. Chronological evaluation protocol and population flow.

Stage	Value
seed	20,260,729
deduplicated_user_parent_pairs_all_ratings	693,929
deduplicated_high_satisfaction_pairs	494,769
relevance_definition	rating >= 4 stars
core_iterations	26
core_edges	9,603
core_users	3,048
core_items	1,103
train_interactions	6,555
train_items	1,091
warm_heldouts	2,967
cold_heldouts	81
validation_users	593
test_users	2,374
split	latest interaction per user; disjoint 20% validation / 80% test warm users
catalog_evaluation	exact full warm catalog with training items masked

Note. The validation and test populations contain disjoint warm users; seen training items were masked from each ranking.

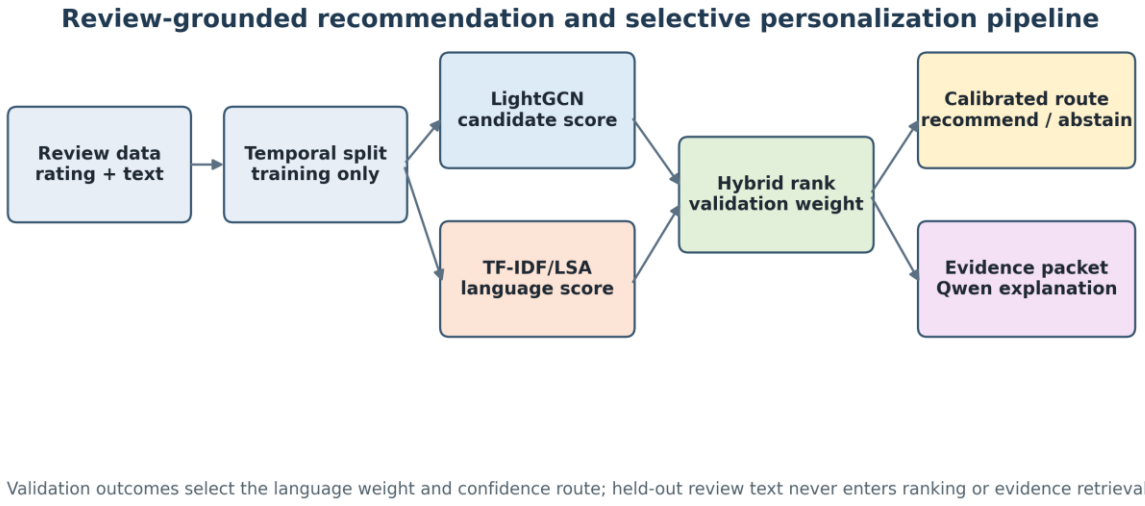

Figure 3. Review-grounded recommendation and decision-support pipeline.

3.3 Collaborative recommendation models

Six independently scored components were evaluated. Popularity ranked products by training interaction frequency and provided a nonpersonalized reference. ItemKNN used item–item cosine similarity with 100 neighbors. PureSVD represented the implicit interaction matrix with 64 factors and ten fitting iterations. BPR-MF learned 32-dimensional user and item factors for 80 epochs under a pairwise ranking objective. LightGCN used 32-dimensional embeddings, two normalized graph-propagation layers, and 160 training epochs. These configurations were fixed in the saved experiment specification.

The LightGCN score for user u and item i was denoted s^{G}_{ui}. Its two propagation layers transmitted preference information through the bipartite user–item graph without introducing item metadata. The comparison with PureSVD, BPR-MF, ItemKNN, and popularity distinguished benefits attributable to graph propagation from those obtainable through simpler neighborhood, factorization, or frequency mechanisms.

Table 3. Compared recommendation components and fixed configurations.

Model	Role	Configuration
Popularity	Interaction-frequency ranking	No fitted factors
ItemKNN	Item-item cosine collaborative filtering	100 neighbors
PureSVD	Implicit matrix factorization	64 factors; 10 iterations
BPR-MF	Pairwise matrix factorization	32 factors; 80 epochs
LightGCN	Two-layer graph collaborative filtering	32 factors; 160 epochs
TextLSA	Training-review language ranker	TF-IDF bigrams; 64 LSA dimensions
Hybrid	Validation-selected collaborative + language score	Per-user z scores

3.4 Training-review language model and hybrid score

The language ranker used only review titles and bodies belonging to training interactions. Validation and test review text did not enter the feature vocabulary, user profiles, item profiles, or candidate scores. Training text was represented by lowercase unigram and bigram TF-IDF features capped at 12,000 dimensions. Truncated singular-value decomposition projected the sparse

text matrix to 64 latent semantic dimensions. Aggregated training-review representations yielded a language similarity score s^{T}_{ui} between each user profile and candidate item.

This training-only construction was intended to isolate the incremental ranking value of review language without allowing held-out prose to reveal the target. Review-grounded recommendation and faithfulness evaluation on Amazon reviews provide the closest task-level rationale for linking ranked products to textual evidence (Chang et al., 2026), while evidence-chain research motivates maintaining explicit boundaries between source content, attribution, and downstream explanation (Jin, 2025b; C. Li et al., 2024). These studies informed the leakage-control and traceability rationale only: the fitted language ranker in the present experiment remained the stated TF-IDF plus truncated-SVD model, and no RAG pipeline or neural text ranker was substituted for it.

Because collaborative and language scores have different numerical scales, both were standardized within user across candidate products. The hybrid ranking score was

$$s^{H}_{ui} = z_u(s^{G}_{ui}) + \lambda\, z_u(s^{T}_{ui}),$$

where $z_u(\cdot)$ is a user-specific z transformation and λ controls the contribution of review language. Seven weights were evaluated on validation users: 0, 0.05, 0.10, 0.20, 0.40, 0.80, and 1.20. The weight maximizing validation NDCG@10 was fixed before test evaluation. The zero-weight condition reproduced the selected collaborative model and therefore served as a direct language ablation.

3.5 Confidence, calibration, and selective recommendation

Confidence was evaluated at the list level, with Hit@20 as the binary outcome. A correct event indicated that the held-out positive item appeared in the first 20 positions. The saved confidence probabilities were assessed using expected calibration error (ECE), Brier score, log loss, and reliability bins. ECE used ten equal-frequency bins; each bin contained 237 or 238 test users. For bin b, the absolute difference between mean confidence and observed Hit@20 was weighted by the bin's population share.

Selective behavior was evaluated by ordering users from highest to lowest confidence. At each coverage level, Hit@20 was recomputed for the retained users, and selective risk was defined as one minus Hit@20. Coverage was varied from 10% to 100% in ten-percentage-point increments. The area under this risk–coverage sequence (AURC) summarized the quality of the confidence ordering; lower values indicated that successful recommendations were concentrated more effectively among high-confidence users.

3.6 Behavioral segmentation

Customer segments were defined only from observed review behavior. Seven training-derived variables were used: log history length, mean rating, rating standard deviation, verified-purchase share, mean log helpful-vote count, mean log review length, and activity span in years. The features were standardized before clustering. A six-segment solution was retained by silhouette analysis, with a silhouette coefficient of 0.4315. Segment profiles were computed for all 3,048 core users, whereas segment-level recommendation metrics were computed for the 2,374 test users. The segments are behavioral groupings and do not represent demographic or protected attributes.

The restricted feature set also separated the present descriptive clustering from richer personalization architectures. Self-supervised customer representations can encode purchase trajectories for segmentation and next-purchase prediction (Xin, 2026f), whereas persistent dialogue memory and privacy-preserving topic-preference learning introduce time-aware updates or differential-privacy mechanisms (Kuo et al., 2025; Sun et al., 2023). Those works identify plausible extensions, not components of the current analysis. No learned customer embedding, conversational memory, demographic attribute, federated optimization, or differential-privacy mechanism entered the seven-variable clustering procedure, so the resulting segments should be interpreted only as summaries of observed review behavior.

3.7 Evidence packets and language-generation assessment

Four evidence packets were selected from scored recommendations. Each packet recorded a case identifier, anonymized user index, behavioral segment, recommended parent ASIN, preference keywords, an evidence identifier, one supporting training-review passage, the passage rating and verification status, and the recommendation-score margin. The passages came from five-star reviews and represented segments 0, 2, 3, and 4. Two packets contained a direct stemmed lexical overlap between preference terms and the review passage, and two contained no such overlap. Score margins ranged from 0.0007 to 6.4370. This

balanced packet design kept the recommendation, source passage, and explanation assessment traceable to a specific training record.

The packet schema operationalized provenance at the case level by binding a recommendation to a source identifier, passage, and score context. This design is consistent with evidence-chain approaches that separate generated claims from attributable support (C. Li et al., 2024; Jin, 2025b) and with interface precedents that expose recommendation evidence and ranking rationale in compact cards (Jin, 2025c; B. Zhang et al., 2025). Privacy and data-integrity risk cards likewise motivate making evidence boundaries inspectable during human oversight (Su, Rao, et al., 2026). The object evaluated here was nevertheless a structured evidence packet for generation testing, not a rendered card interface or a UI user study.

Qwen2.5-0.5B-Instruct Q4_K_M was evaluated as the language generator (Qwen Team, 2025). Inference used wllama 3.5.1 in headless Chromium with four CPU threads, seed 20260729, temperature 0, top-p 1, and a 48-token limit. The prompt required one sentence of at most 45 words beginning with the packet's bracketed evidence tag. When evidence and preferences overlapped, the sentence had to state the match and reproduce an exact quotation of at least four words; otherwise it had to return a fixed insufficiency statement. Citation accuracy, selective-action accuracy, exact-quote fidelity, lexical source precision, numeric-claim control, length and sentence compliance, and a joint pass rate were computed from the model strings.

These criteria deliberately tested both utility and safe failure rather than fluency alone. Evidence-based self-verification, behavior-level refusal with utility preservation, and continually updated guardrails all emphasize that a usable response must satisfy explicit behavioral constraints under difficult inputs (Zheng et al., 2023, 2024; Zheng & Li, 2024). They therefore support the choice to score evidence tags, source-exact quotation, and the fixed insufficiency response jointly. None of those guardrail algorithms was implemented in the Qwen run; the experiment used the stated prompt and decoding configuration, after which deterministic rules evaluated the generated strings against the packet fields.

3.8 Statistical analysis and computational reporting

The primary inferential comparison was hybrid ranking against LightGCN on user-level NDCG@10. A paired bootstrap with 5,000 resamples estimated the 95% confidence interval for the mean within-user difference. A paired randomization test with 10,000 sign permutations tested the null hypothesis of no average difference. Pairing was essential because both models ranked the same catalog for the same test users. The analysis reported the exact randomization probability rather than a threshold alone. Runtime was recorded separately for each fitted component, and the complete experiment required 182.62 seconds.

4. Results and Discussion

4.1 Overall ranking performance

The hybrid achieved the strongest test performance on all three ranking measures. Its NDCG@10 was 0.09186, Recall@20 was 0.19966, MRR@20 was 0.07656, and median target rank was 134. LightGCN was the strongest standalone collaborative model, with NDCG@10 of 0.07933, Recall@20 of 0.16976, MRR@20 of 0.06866, and median rank of 150. Adding training-review language therefore increased NDCG@10 by 0.01253, or 15.80% relative to LightGCN, and increased Recall@20 by 0.02991, or 17.62%. MRR@20 increased by 0.00790, while the median target moved 16 positions closer to the head of the catalog.

The remaining baselines established that the gain did not arise from popularity alone. ItemKNN was the second-best nonhybrid baseline by NDCG@10, reaching 0.06265 and Recall@20 of 0.13184. PureSVD reached 0.05327 and 0.11668, respectively. Popularity obtained only 0.02422 NDCG@10, 0.08088 Recall@20, and a median rank of 530.5. BPR-MF performed worst, with NDCG@10 of 0.01493 and Recall@20 of 0.04971. Its weak result under the fixed 32-factor, 80-epoch configuration indicates that pairwise factorization was not automatically robust to this short-history graph. It does not establish that every possible BPR configuration would underperform.

TextLSA alone produced NDCG@10 of 0.03288 and Recall@20 of 0.09688. These values were above popularity on NDCG and Recall but well below LightGCN. Review language was therefore insufficient as a replacement for collaborative structure. Its benefit appeared through complementarity: semantic similarity altered candidate ordering in ways that improved the graph model's top ranks without carrying the full recommendation task by itself.

This complementarity is directionally consistent with prior work that couples Amazon-review evidence to explainable recommendation (Chang et al., 2026) and retrieves observed behavior before generating a personalized response (Xin, 2026b). In the present experiment, however, the contribution was narrower: training-review semantics adjusted a LightGCN candidate order, while the graph model retained the main collaborative signal. The observed gains should therefore be interpreted as evidence for this specific score-fusion protocol and temporal split, not as a replication of those systems or as an effect-size comparison across datasets, candidate sets, or evaluation designs.

Table 4. Exact-catalog ranking performance on the test population.

Model	Test n	NDCG@10	Recall@20	MRR@20	Median rank
Popularity	2,374	0.0242	0.0809	0.0184	530.5
ItemKNN	2,374	0.0627	0.1318	0.0545	297.0
PureSVD	2,374	0.0533	0.1167	0.0423	460.0
BPR-MF	2,374	0.0149	0.0497	0.0120	468.0
LightGCN	2,374	0.0793	0.1698	0.0687	150.0
TextLSA	2,374	0.0329	0.0969	0.0256	303.0
Hybrid	2,374	0.0919	0.1997	0.0766	134.0

Note. Higher is better for NDCG@10, Recall@20, and MRR@20; lower is better for median target rank. Every method ranked the same 1,091-item catalog.

Figure 4. Ranking performance across recommendation methods.

4.2 Fusion-weight ablation

The validation ablation supported the selected language contribution. At λ=0, validation NDCG@10 was 0.09206. It rose to 0.09396 at 0.05, 0.09732 at 0.10, 0.09874 at 0.20, and 0.10298 at 0.40. The maximum, 0.10613, occurred at 0.80; increasing the weight to 1.20 reduced validation NDCG@10 to 0.10264. Validation Recall@20 showed the same broad pattern, increasing from 0.19224 without language to 0.22091 at both 0.40 and 0.80 before declining to 0.21585.

The untouched test population followed a consistent curve. NDCG@10 increased from 0.07933 at zero language weight to 0.08143, 0.08330, 0.08613, 0.09112, and 0.09186 as the weight rose through 0.80, then fell to 0.08488 at 1.20. Test Recall@20 similarly peaked at 0.19966 with the validation-selected weight. The decline at 1.20 is important: language evidence was useful, but over-weighting the relatively weak text-only ranker displaced collaborative signals. The close results at 0.40 and 0.80 also suggest a useful plateau rather than dependence on an isolated numerical setting.

The paired analysis confirmed that the NDCG improvement was not explained by a small set of favorable users. The mean paired difference between Hybrid and LightGCN was 0.01253. The 5,000-resample bootstrap interval ranged from 0.00595 to 0.01958 and excluded zero. The 10,000-draw paired randomization probability was 0.00030. This evidence supports a test-population

improvement in top-ten discounted gain. No corresponding paired intervals were saved for Recall@20 or MRR@20, so their increases should be read as descriptive effect sizes rather than separate inferential findings.

Table 5. Language-fusion weight ablation on validation and test users.

λ	Validation NDCG@10	Validation Recall@20	Test NDCG@10	Test Recall@20	Selected
0.00	0.0921	0.1922	0.0793	0.1698	No
0.05	0.0940	0.1990	0.0814	0.1740	No
0.10	0.0973	0.2074	0.0833	0.1782	No
0.20	0.0987	0.2175	0.0861	0.1866	No
0.40	0.1030	0.2209	0.0911	0.1976	No
0.80	0.1061	0.2209	0.0919	0.1997	Yes
1.20	0.1026	0.2159	0.0849	0.1917	No

Note. The fusion weight was selected by validation NDCG@10; test results are included to show the shape of the fixed-weight sensitivity curve.

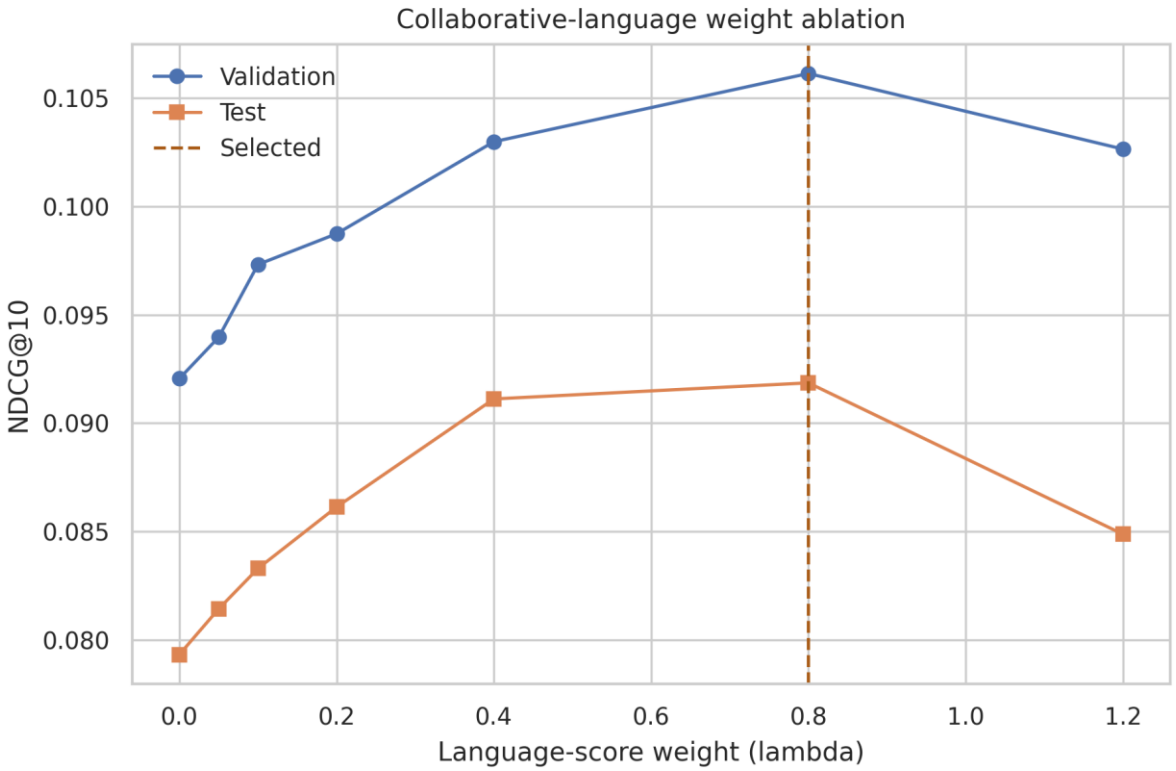

Figure 5. Validation and test ranking performance across language-fusion weights.

Table 6. Paired inference for Hybrid versus LightGCN.

Comparison	Metric	Mean difference	95% CI low	95% CI high	Paired p	Bootstrap draws	Randomization draws
Hybrid - LightGCN	NDCG@10	0.0125	0.0059	0.0196	0.0003	5,000	10,000

Note. The confidence interval used 5,000 paired bootstrap resamples; the randomization probability used 10,000 paired sign permutations.

4.4 Calibration and selective behavior

The hybrid improved every saved uncertainty measure relative to LightGCN, but neither model was adequately calibrated in absolute terms. Hybrid ECE was 0.25319, compared with 0.29112 for LightGCN. Its Brier score was 0.21314 rather than 0.21779, log loss was 0.61945 rather than 0.62451, and AURC was 0.63844 rather than 0.67798. Thus review-language fusion improved both ranking and the ordering of recommendation confidence.

Absolute confidence remained too high. The hybrid's mean predicted confidence was 0.45286, while its observed Hit@20 rate was 0.19966. LightGCN showed a still larger disparity: mean confidence of 0.46087 against an observed rate of 0.16976. The reliability bins make the pattern explicit. In the hybrid's lowest-confidence decile, mean confidence was 0.22329 and observed Hit@20 was 0.07143. In its highest decile, the corresponding values were 0.78041 and 0.48523. Confidence was strongly associated with success, but the probability scale overstated the empirical frequency throughout the range.

That distinction matters operationally. Ranking confidence can be useful for deciding when to show a personalized list even when it is unsuitable as a literal probability. At 10% coverage, the hybrid attained Hit@20 of 0.48523, compared with its full-coverage value of 0.19966. At 20% coverage, Hit@20 remained 0.40842; at 50% coverage it was 0.27717. LightGCN reached 0.43460, 0.33263, and 0.23673 at the same coverages. The lower hybrid AURC consequently reflects meaningful enrichment among high-confidence users. A deployment could use this ordering to route lower-confidence cases to a conservative list, additional preference elicitation, or no personalized claim. Probability recalibration would still be necessary before displaying numerical confidence to customers or managers.

Operationally, these findings favor separating an internal routing score from any customer-facing probability claim. Probability-calibrated risk tiering and uncertainty-driven rejection provide cross-domain precedents for this separation (Y. Chen et al., 2023; Jin et al., 2024; Xin, 2025c). Explainable e-commerce recommendation cards further illustrate how rationale and calibrated trust cues might be presented (B. Zhang et al., 2025), while financial-risk dashboards and infrastructure visual briefs show how uncertainty and action thresholds can remain distinct (Z. Li et al., 2025; Liu et al., 2025). These are methodological or design precedents, not evaluated treatments here: no external calibration guarantee, card prototype, dashboard, comprehension test, or user experiment was part of this study.

Table 7. List-level calibration and selective-recommendation performance.

Model	Validation hits	Test hits	ECE	Brier	Log loss	AURC
LightGCN	114	403	0.2911	0.2178	0.6245	0.6780
Hybrid	131	474	0.2532	0.2131	0.6195	0.6384

Note. Lower is better for ECE, Brier score, log loss, and AURC.

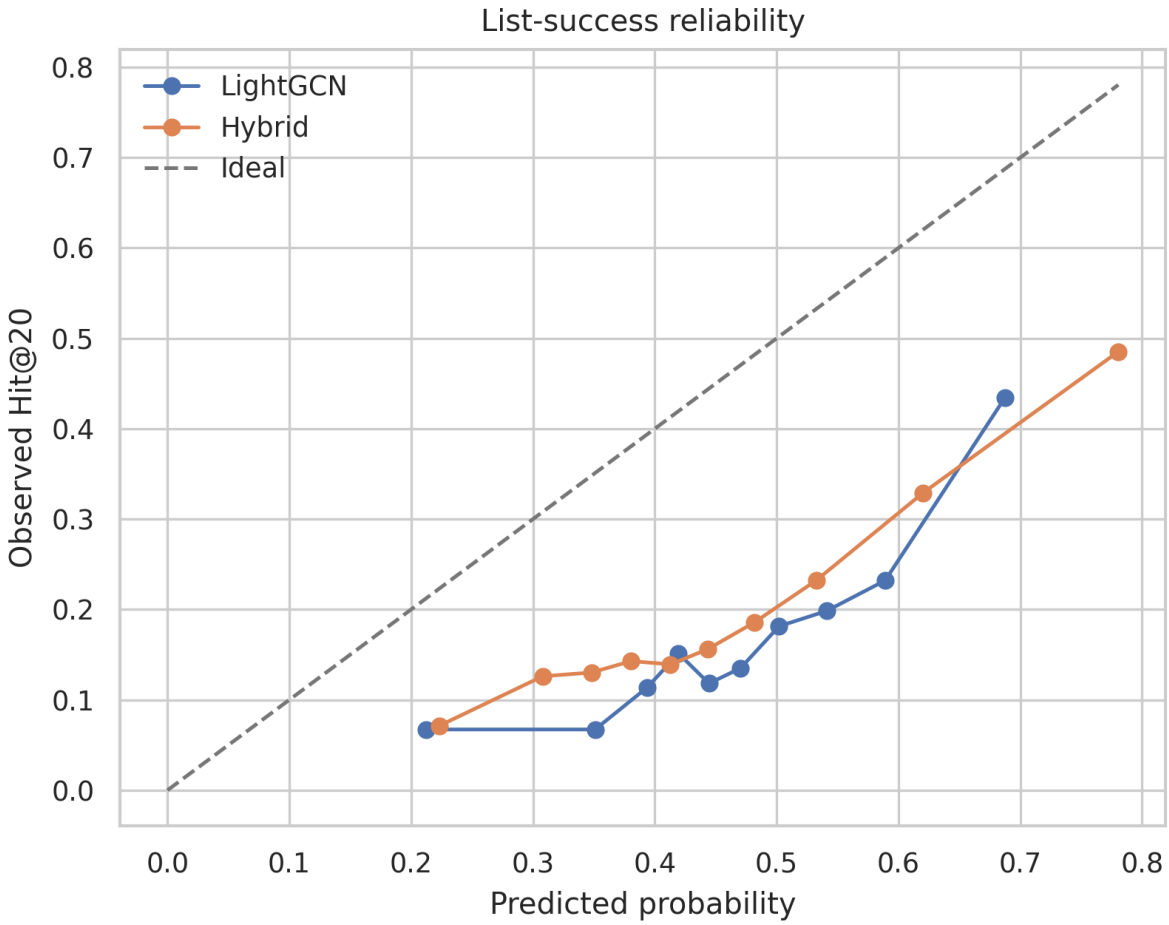

Figure 6. Reliability of list-level Hit@20 confidence.

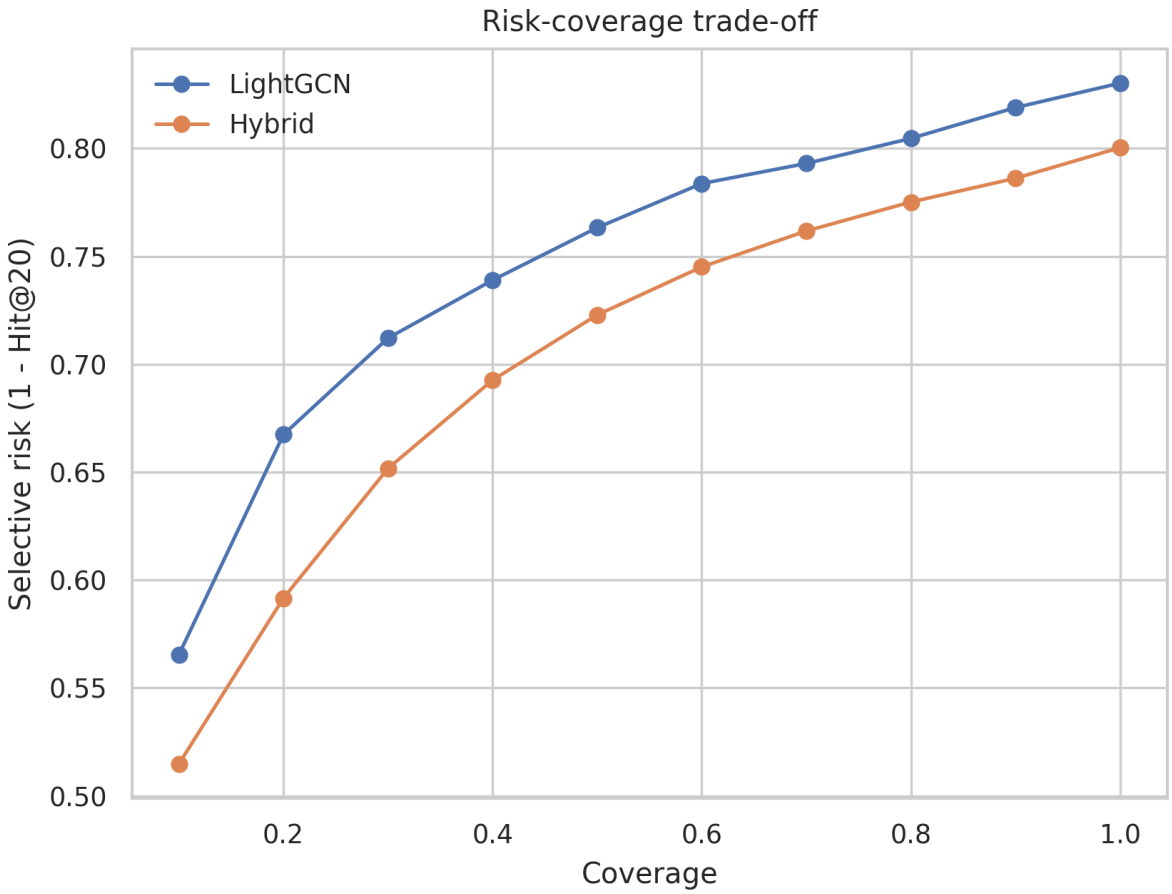

Figure 7. Selective risk as recommendation coverage increases.

4.5 Segment heterogeneity

The full-core profiles clarified what distinguished the six groups. Segment 0 contained 1,079 users with a mean rating of 5.0, verified share of 0.004, and mean log review length of 5.590. Segment 1 contained 466 users whose mean rating was 4.0 and verified share was 0.333. Segment 2 contained 423 users and combined a higher log-history value of 1.757 with rating standard deviation of 0.452 and an activity span of 0.850 years. Segment 3 contained 817 users, had verified share of 0.984, and produced the shortest reviews on average, with mean log length of 4.264. Segment 4's 203 users had the largest mean log helpful-vote value, 2.095. Segment 5 contained 60 users and was distinguished by the largest log-history value, 2.055, the longest mean log review length, 6.242, and an activity span of 5.260 years.

Performance varied sharply across the six behavioral segments. Segment 0 contained 842 test users and achieved the highest NDCG@10, 0.11950, with Recall@20 of 0.24703. Segment 4 contained 160 test users and achieved the highest Recall@20, 0.25000, together with NDCG@10 of 0.11810. Segment 3, the nearly fully verified and short-review group, contained 632 test users and remained close to the overall average at 0.09515 NDCG@10 and 0.18987 Recall@20. Segment 1 recorded 0.07862 and 0.19565 across 368 users.

Two groups were notably harder. Segment 2 included more active users with variable ratings and a longer average activity span; its 322 test users obtained NDCG@10 of 0.02845 and Recall@20 of 0.10248. Segment 5 represented the longest-tenure and highest-history profile. Only 50 of its 60 members appeared in the test population, and this group obtained NDCG@10 of 0.00667 and Recall@20 of 0.02000. The best-to-worst gap was 0.11283 in NDCG@10 and 0.23000 in Recall@20.

These findings caution against interpreting more historical activity as automatically easier personalization. Long histories may reflect preference evolution, catalog turnover, or a more diverse consumption pattern. At the same time, the smallest segment has only 50 test cases, and no segment-level confidence intervals were saved. Its extreme result should motivate targeted analysis rather than a claim about all long-tenure customers. The clusters also capture review behavior, not identity. They cannot

support demographic targeting or fairness claims concerning age, gender, race, income, or other attributes absent from the analysis.

The segment gaps are therefore most useful as a model-auditing signal: they identify behavioral histories for which retrieval or ranking may need additional evidence, without assigning those differences to personal identity. Learned customer representations offer one route for modeling heterogeneous histories (Xin, 2026f), while federated preference learning and privacy-robust response estimation show why personalization should minimize unnecessary identifier exposure (Bai et al., 2026; Kuo et al., 2025). The current experiment did not implement federated learning, differential privacy, causal uplift estimation, or a formal privacy guarantee. Its contribution is limited to non-demographic performance disaggregation over the six observed-behavior clusters.

Table 8. Behavioral profiles of the six customer segments.

Segment	Users	Log history	Mean rating	Rating SD	Verified share	Mean log helpful	Mean log review length	Activity span (years)
0.0000	1,079	0.8842	5.0000	0.0000	0.0042	0.1316	5.5903	0.1191
1	466	0.7579	4.0000	0.0000	0.3326	0.1848	5.7135	0.0507
2	423	1.7568	4.5823	0.4518	0.0195	0.2208	5.9448	0.8499
3	817	0.7088	4.9780	0.0012	0.9845	0.1376	4.2642	0.0136
4	203	0.7322	4.9056	0.0071	0.6535	2.0948	6.0306	0.0600
5	60	2.0551	4.6514	0.3444	0.1227	0.3343	6.2417	5.2603

Note. Profiles use training-derived behavior for all 3,048 users in the recommendation core; logarithmic variables are reported on their transformed scales.

Table 9. Hybrid ranking performance by behavioral segment.

Segment	Test n	NDCG@10	Recall@20	MRR@20
0.0000	842	0.1195	0.2470	0.0994
1	368	0.0786	0.1957	0.0641
2	322	0.0285	0.1025	0.0220
3	632	0.0951	0.1899	0.0803
4	160	0.1181	0.2500	0.1031
5	50	0.0067	0.0200	0.0029

Note. Segment metrics use the 2,374-user test population.

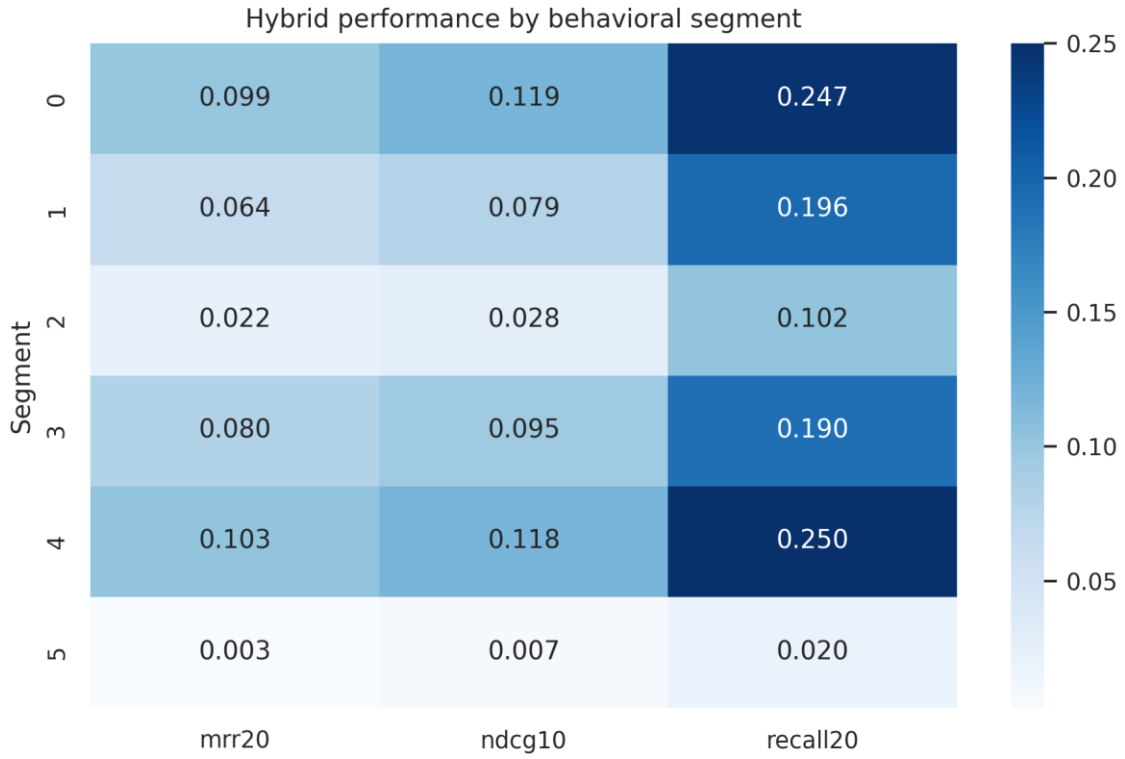

Figure 8. Ranking heterogeneity across behavioral segments.

4.6 Evidence-grounded language generation

The language-generation assessment exposed a clear gap between safe routing and faithful surface realization. Selective-action accuracy was 0.75: the model correctly explained both overlap cases and correctly abstained in one of the two non-overlap cases. Citation accuracy and exact-quote fidelity were both 0.25 because only the successful abstention used the required bracketed evidence tag and satisfied the case-specific fidelity criterion. The other three outputs used an unbracketed "E:" prefix; neither explanatory response enclosed an exact four-word source quotation. The remaining non-overlap case copied review and instruction text instead of issuing the required insufficiency statement.

The outputs nevertheless contained no unsupported numerical claims, and all four met the 45-word and one-sentence limits. Mean lexical source precision was 0.59057, while only one case passed every constraint, producing a joint pass rate of 0.25. Two responses reached the 48-token decoding limit and ended mid-instruction. Mean generation latency was 41.591 seconds per case, and total generation latency was 166.365 seconds. These measurements show that evidence in the prompt constrained much of the vocabulary but did not reliably enforce the citation syntax, quotation boundary, or abstention rule in the compact quantized model.

For decision support, the result favors a validated generation path rather than direct display of raw output. A deterministic post-generation check can verify the evidence tag, reject text without a source-exact quotation when one is required, and replace any failed case with the fixed insufficiency statement. The assessment contains four deliberately balanced cases, so its rates are diagnostic rather than population estimates. Even at this scale, the observed failure modes are actionable: stronger constrained decoding, an extractive explanation template, or a larger model should be tested before customer-facing use.

The four-case evidence indicates that provenance must be enforced rather than inferred from generally source-aligned wording. Evidence-chain systems similarly distinguish source attribution from mere lexical plausibility (C. Li et al., 2024; Jin, 2025b), and evidence-calibrated abstention treats insufficient support as a legitimate output rather than a generation failure (Zhong, Chen, et al., 2025). VerifySafe supplies an adjacent precedent for checking a response against evidence before release (Zheng et al., 2024). For this pipeline, the immediate implication is a deterministic release gate—tag validation, exact-span verification, and fixed fallback—not a claim that those external self-verification or RAG algorithms were run.

Table 10. Evidence-grounded language-generation assessment.

Method	Citation accuracy	Selective action accuracy	Exact quote fidelity	Unsupported numeric claim free rate	Full constraint pass rate	Mean lexical source precision	Cases
Qwen2.5-0.5B grounded	0.2500	0.7500	0.2500	1	0.2500	0.5906	4

Note. Each assessment case was generated from a saved evidence packet containing a recommendation and one traceable review passage.

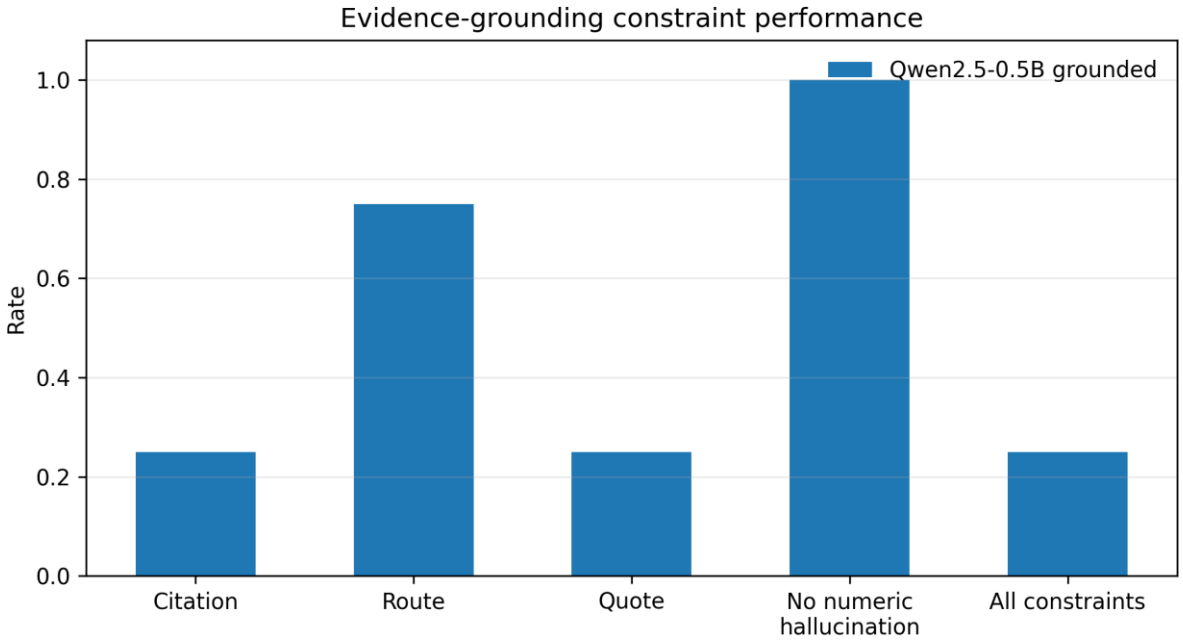

Figure 9. Faithfulness measures for evidence-grounded language generation.

4.7 Efficiency and practical trade-offs

Popularity required 0.00049 seconds, ItemKNN 0.556 seconds, BPR-MF 4.301 seconds, LightGCN 9.175 seconds, and PureSVD 10.194 seconds for fitting and scoring. TextLSA was the dominant component at 71.128 seconds. Once both score matrices existed, hybrid fusion added effectively no measured runtime. The resulting trade-off is favorable for offline decision support: the language component incurred most of the computation but produced a statistically supported NDCG gain and could be reused across multiple score weights or confidence thresholds.

The runtime profile supports a modular deployment interpretation: review representations can be refreshed offline, hybrid scores can be served separately from explanation generation, and the compact generator can be invoked only when ranking confidence and evidence availability satisfy release rules. Hybrid cloud–edge routing and edge-oriented model distillation offer relevant systems precedents for separating inexpensive routine processing from selectively invoked language capability (Xin, 2025b; Zhou, Wang, et al., 2025). This interpretation does not convert the local timings into cloud-cost, throughput, energy, or latency estimates; those quantities were not measured. It instead identifies component boundaries that a future production benchmark could evaluate directly.

Table 11. Measured fitting and scoring time by model component.

Model	Time (s)	Note
Popularity	0.000	—
ItemKNN	0.556	—
PureSVD	10.194	—
BPR-MF	4.301	—
LightGCN	9.175	—
TextLSA	71.128	—
Hybrid	0.000	LightGCN + lambda*TextLSA; lambda=0.8

Note. Times were measured in the local experimental run; hybrid fusion reused the saved component-score matrices.

4.8 Scope and limitations

The conclusions apply to a sparse, positive-preference warm-start population. The final core retained 9,603 of 494,769 deduplicated positive pairs and 3,048 of 631,986 observed users. This selection was necessary for personalized ranking but excludes the dominant one-interaction population. The exact catalog evaluation also omitted 81 cold held-out items from ranking, although their 2.66% incidence was reported. A production system would require a distinct content or exploration policy for those products and for first-time customers.

The four-or-five-star threshold gives the held-out event a clear recommendation interpretation, but it does not fully exploit one-to-three-star reviews as negative or neutral feedback. Review text was limited to training titles and bodies, and the language representation was a compact latent semantic model. The experiment used one fixed seed for the learned configurations. The paired bootstrap captures variation across test users, not variation across training initializations. Finally, the confidence results demonstrate useful selectivity but substantial overconfidence, and the segment results reveal heterogeneity without establishing causal explanations. These boundaries narrow the claims to what the measured ranking, calibration, and behavioral analyses directly support.

5. Conclusions

This study establishes a coherent decision-support evaluation on the All Beauty reviews, linking exact-catalog recommendation, review-language fusion, confidence-based selection, behavioral segmentation, and evidence-grounded language generation. The validation-selected hybrid combined LightGCN with a training-only latent semantic review score and delivered the strongest ranking results. Relative to LightGCN, it improved NDCG@10 by 15.80% and Recall@20 by 17.62%; the paired NDCG improvement remained positive across the bootstrap interval and had a randomization probability of .00030. TextLSA alone was much weaker, so the benefit came from complementing collaborative structure rather than replacing it.

The confidence analysis supports selective personalization but not literal probability display. Hybrid confidence ordered users more effectively than LightGCN, producing a lower AURC and markedly higher Hit@20 at restricted coverage. At the same time, mean confidence exceeded the observed success rate by more than 0.25, and the reliability curve showed overconfidence throughout the range. A practical system should therefore use confidence to determine whether to personalize, request more preference information, or fall back to a conservative list, while applying a separate recalibration stage before presenting probabilities.

Behavioral segments revealed that customer history does not translate uniformly into recommendation quality. The strongest segments reached Recall@20 near 0.25, whereas the longest-tenure segment reached 0.02. These groups describe review behavior rather than personal identity, but they provide an audit layer for detecting where a shared recommendation policy performs poorly. The compact Qwen model usually selected the correct explain-or-abstain route and made no unsupported numerical claim, yet it frequently missed the required evidence-tag and exact-quotation format. Evidence prompting alone was insufficient for display-ready explanations. A deployment should enforce source-exact quotation checks, reject malformed citations, and return a fixed insufficiency statement when validation fails.

The main boundary is coverage: the positive-interaction core represents repeat high-satisfaction users and warm items rather than the dominant one-review population or cold products. Future work should add a cold-start policy, recalibrate confidence, test multiple training seeds, and evaluate constrained language models on a broader evidence set. Within the measured warm-start setting, review language improves ranking and selective decision support, while faithful explanation remains a controlled generation-and-validation problem.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers
Artificial Intelligence (AI) Use Disclosure: The authors declare that no artificial intelligence tools were used in the preparation of this manuscript.

References

[1] Bai, J., Wang, H., Wu, Q., & Zhang, B. (2026). Privacy-robust incrementality estimation in cookieless settings via uplift modeling: Reproducible evidence from the Hillstrom e-mail experiment. *Journal of Technology Informatics and Engineering, 5*(1), 17–38. https://doi.org/10.51903/jtie.v5i1.468

[2] Bai, J., & Wu, Q. (2026). Privacy-safe marketing mix modeling and budget optimization under identifier loss: A controlled simulation study. *International Journal of Electronic Communication Systems, 6*(1). https://doi.org/10.24042/ijecs.v6i1.30533

[3] Balog, K., & Radlinski, F. (2020). Measuring recommendation explanation quality: The conflicting goals of explanations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 329–338). ACM. https://doi.org/10.1145/3397271.3401032

[4] Bao, K., Zhang, J., Zhang, Y., Wang, W., Feng, F., & He, X. (2023). TALLRec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems* (pp. 1007–1014). ACM. https://doi.org/10.1145/3604915.3608857

[5] Chang, X., Lu, Y., & Zhong, Z. S. (2026). Review-grounded explainable recommendation with faithfulness evaluation on Amazon reviews. *JEECS (Journal of Electrical Engineering and Computer Sciences), 11*(1), 9–22. https://doi.org/10.54732/jeecs.v11i1.2

[6] Chen, C., Zhang, M., Liu, Y., & Ma, S. (2018). Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 World Wide Web Conference* (pp. 1583–1592). ACM. https://doi.org/10.1145/3178876.3186070

[7] Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management, 19*(3), 197–208. https://doi.org/10.1057/dbm.2012.17

[8] Chen, S., He, S., & Sun, E. (2024). Risk-bounded GPU resource oversubscription via conformal demand envelopes in production AI clusters. *Journal of Advanced Computing Systems, 4*(5), 119–134. https://doi.org/10.69987/jacs.2024.40509

[9] Chen, Y., & Xu, H. (2026). Trust-calibrated multilingual RAG for humanitarian information platforms: Empirical evaluation on OMoS-QA for migration information access. *International Journal of Graphic Design, 4*(1), 141–164. https://doi.org/10.51903/ijgd.v4i1.3552

[10] Chen, Y., Zhang, Y., Chau, D., & Sherman, M. (2023). Credit card default risk tiering with probability calibration and uncertainty-driven rejection: A reproducible study on the UCI credit card clients dataset. *Journal of Advanced Computing Systems, 3*(4), 31–47. https://doi.org/10.69987/jacs.2023.30403

[11] Chen, Y., Zhang, Y., & Sherman, M. (2024). Going concern and bankruptcy prediction under extreme class imbalance: Cost-sensitive learning, resampling, and focal loss with explainable financial-ratio portraits. *Journal of Advanced Computing Systems, 4*(4), 80–96. https://doi.org/10.69987/jacs.2024.40407

[12] Chen, Y., Zhou, S., & Lin, E. (2025). Accounting-aware evidence retrieval for institutional due diligence of tokenized trade receivable RWA. *Journal of Technology Informatics and Engineering, 4*(3), 649–663. https://doi.org/10.51903/jtie.v4i3.542

[13] Dai, S., Shao, N., Zhao, H., Yu, W., Si, Z., Xu, C., Sun, Z., Zhang, X., & Xu, J. (2023). Uncovering ChatGPT's capabilities in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems* (pp. 1126–1132). ACM. https://doi.org/10.1145/3604915.3610646

[14] DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., & Wallace, B. C. (2020). ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4443–4458). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.408

[15] Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research, 42*(4), 415–430. https://doi.org/10.1509/jmkr.2005.42.4.415

[16] Geifman, Y., & El-Yaniv, R. (2019). SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 2151–2159). PMLR. https://proceedings.mlr.press/v97/geifman19a.html

[17] Geng, S., Liu, S., Fu, Z., Ge, Y., & Zhang, Y. (2022). Recommendation as language processing (RLP): A unified pretrain, personalized prompt & predict paradigm (P5). In *Proceedings of the 16th ACM Conference on Recommender Systems* (pp. 299–315). ACM. https://doi.org/10.1145/3523227.3546767

[18] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

[19] He, S., Chang, X., & Sun, E. (2024). Cross-cloud transfer learning for AI training capacity forecasting under workload and topology distribution shift. *Journal of Advanced Computing Systems, 4*(1), 100–120. https://doi.org/10.69987/jacs.2024.40108

[20] He, S., Li, C., & Rao, H. (2025). Few-shot cold-start workload forecasting for new AI inference tenants with time-series foundation models. *Journal of Technology Informatics and Engineering, 4*(1), 306–324. https://doi.org/10.51903/jtie.v4i1.546

[21] He, S., Tu, H., & Liu, I. (2023). Safe PD capacity forecasting with time-series foundation models and calibrated uncertainty for heterogeneous GPU clusters. *Journal of Advanced Computing Systems, 3*(4), 48–66. https://doi.org/10.69987/jacs.2023.30404

[22] He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., & Wang, M. (2020). LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 639–648). ACM. https://doi.org/10.1145/3397271.3401063

[23] Hou, Y., Li, J., Fu, X., He, Z., Yan, A., Chen, X., & McAuley, J. (2026). Bridging language and items for retrieval and recommendation: Benchmarking LLMs as semantic encoders. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3251–3265). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.acl-long.147

[24] Hou, Y., Zhang, J., Lin, Z., Lu, H., Xie, R., McAuley, J., & Zhao, W. X. (2024). Large language models are zero-shot rankers for recommender systems. In N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, & I. Ounis (Eds.), *Advances in information retrieval* (Lecture Notes in Computer Science, Vol. 14609, pp. 364–381). Springer. https://doi.org/10.1007/978-3-031-56060-6_24

[25] Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 263–272). IEEE. https://doi.org/10.1109/ICDM.2008.22

[26] Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems, 20*(4), 422–446. https://doi.org/10.1145/582415.582418

[27] Ji, Y., Sun, A., Zhang, J., & Li, C. (2023). A critical study on data leakage in recommender system offline evaluation. *ACM Transactions on Information Systems, 41*(3), Article 75. https://doi.org/10.1145/3569930

[28] Jin, J. (2025a). Calibrated resume-job matching for trustworthy LLM-assisted recruiter screening: Pairwise matching, probability calibration, and selective refusal on two public recruitment datasets. *Journal of Technology Informatics and Engineering, 4*(3), 625–648. https://doi.org/10.51903/jtie.v4i3.529

[29] Jin, J. (2025b). Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations. *Journal of Technology Informatics and Engineering, 4*(2), 520–533. https://doi.org/10.51903/jtie.v4i2.535

[30] Jin, J. (2025c). LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy. *International Journal of Graphic Design, 3*(2), 397–414. https://doi.org/10.51903/ijgd.v3i2.3698

[31] Jin, J., Huang, T., & Lu, S. (2024). A model-risk-friendly probability of default workflow: Calibration, distribution-free uncertainty quantification, and SHAP explanations on the UCI credit card default dataset. *Journal of Advanced Computing Systems, 4*(6), 74–85. https://doi.org/10.69987/jacs.2024.40606

[32] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer, 42*(8), 30–37. https://doi.org/10.1109/MC.2009.263

[33] Krichene, W., & Rendle, S. (2020). On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1748–1757). ACM. https://doi.org/10.1145/3394486.3403226

[34] Kuo, M.-J., Zheng, D., & Hires, J. (2025). Federated topic-preference learning for knowledge-grounded chat with differential privacy. *Journal of Technology Informatics and Engineering, 4*(2), 385–401. https://doi.org/10.51903/jtie.v4i2.502

[35] Li, C., Bai, J., & Wang, S. (2024). Evidence-chain reliable RAG: Word-level hallucination detection, source attribution, and provenance explanation for LLM applications. *Journal of Advanced Computing Systems, 4*(2), 76–92. https://doi.org/10.69987/jacs.2024.40207

[36] Li, C., Liu, G., & Zhao, Z. (2026). Cost-aware LLM-style routing for AIOps log analysis: Log parsing, anomaly detection, fault diagnosis, and incident summarization on LogEval task files. *Journal of Technology Informatics and Engineering, 5*(2), 91–103. https://doi.org/10.51903/jtie.v5i2.538

[37] Li, C., Zhou, B., & Gao, K. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX benchmark for explainable medical AI response interfaces. *International Journal of Graphic Design, 3*(2), 381–394. https://doi.org/10.51903/ijgd.v3i2.3709

[38] Li, J., & Zhou, A. (2026). Multi-regulation RAG for AI product counsel: A legal governance framework for cross-border digital commerces. *Rule of Law Studies Journal, 2*(2), 105–123. https://doi.org/10.64780/rolsj.v2i2.225

[39] Li, L., Zhang, Y., & Chen, L. (2021). Personalized Transformer for explainable recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 4947–4957). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.383

[40] Li, Y. (2024). Findable then explainable: Retrieval–summary integration for code intelligence on a lightweight CodeSearchNet subset. *Journal of Advanced Computing Systems, 4*(7), 65–82. https://doi.org/10.69987/jacs.2024.40706

[41] Li, Y., Lu, S., & Zhao, L. (2025). LLM-as-design-critic: Aligning AI-generated UI feedback with human graphic design judgment. *International Journal of Graphic Design, 3*(1), 196–215. https://doi.org/10.51903/ijgd.v3i1.3661

[42] Li, Z., Zhang, K., & Wong, A. (2026). Numerical-reasoning guardrails for a quant research assistant: A compact reproducible benchmark using SEC and FRED data. *Journal of Technology Informatics and Engineering, 5*(2), 75–90. https://doi.org/10.51903/jtie.v5i2.541

[43] Li, Z., Zhou, S., & Zhou, Z. (2025). Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench. *International Journal of Graphic Design, 3*(1), 196–210. https://doi.org/10.51903/ijgd.v3i1.3715

[44] Liu, G., He, S., & Liu, I. (2023). LLM-augmented multi-source root cause attribution for CPU and network faults in microservices. *Journal of Advanced Computing Systems, 3*(6), 39–57. https://doi.org/10.69987/jacs.2023.30604

[45] Liu, G., He, S., & Wong, H. (2025). LLM-compatible visual brief cards for AI infrastructure capacity dashboards: A UI/UX framework for turning forecast risk into graphic design decisions. *International Journal of Graphic Design, 3*(1), 196–213. https://doi.org/10.51903/ijgd.v3i1.3723

[46] Liu, G., Li, C., & Zhang, E. (2024). OpsLLM for cloud incident triage: Bilingual RAG-based root cause analysis and alert summarization for AI infrastructure operations. *Journal of Advanced Computing Systems, 4*(4), 97–111. https://doi.org/10.69987/jacs.2024.40408

[47] Lu, Y., Zhou, H., & Zhang, Y. (2025). A constrained, data-driven budgeting framework integrating macro demand forecasting and marketing response modeling. *Journal of Technology Informatics and Engineering, 4*(3), 493–520. https://doi.org/10.51903/jtie.v4i3.466

[48] McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems* (pp. 165–172). ACM. https://doi.org/10.1145/2507157.2507163

[49] McAuley Lab. (2024). *Amazon Reviews'23* [Data set]. https://amazon-reviews-2023.github.io/

[50] McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research, 60*(6), 656–662. https://doi.org/10.1016/j.jbusres.2006.06.015

[51] Meng, S., Chen, J., & Zheng, I. (2026). LLM-inspired offline reranking for financial search: Query rewriting, hybrid retrieval, and listwise relevance ranking on FiQA. *Journal of Technology Informatics and Engineering, 5*(1), 361–378. https://doi.org/10.51903/jtie.v5i1.537

[52] Mu, J., Lu, Y., & Hwang, E. (2026). Structured visual brief interfaces for advertising design: A UI/UX framework for turning creative intentions into designer-editable graphic design cards. *International Journal of Graphic Design, 4*(1), 192–208. https://doi.org/10.51903/ijgd.v4i1.3702

[53] Mu, J., Lu, Y., & Smith, M. (2023). LLM-assisted incrementality (uplift) modeling for email advertising: From feature interactions to interpretable audience–creative–channel policies. *Journal of Advanced Computing Systems, 3*(1), 31–48. https://doi.org/10.69987/jacs.2023.30103

[54] Mu, J., Ye, T., & Patel, P. (2025). Offline counterfactual evaluation for advertising and recommendation slot policies: A reproducible study on the Open Bandit Dataset (small). *Journal of Technology Informatics and Engineering, 4*(3), 521–543. https://doi.org/10.51903/jtie.v4i3.500

[55] Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (pp. 2901–2907). AAAI Press. https://doi.org/10.1609/aaai.v29i1.9602

[56] Ni, J., Li, J., & McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 188–197). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1018

[57] Nie, J., Liu, G., Li, C., & Zou, T. (2026). 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. *International Journal of Graphic Design, 4*(1), 179–185. https://doi.org/10.51903/ijgd.v4i1.3703

[58] Qwen Team. (2025). *Qwen2.5 technical report* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2412.15115

[59] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

[60] Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2009). BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (pp. 452–461). AUAI Press. https://www.auai.org/uai2009/papers/UAI2009_0139_48141db02b9f0b02bc7158819ebfa2c7.pdf

[61] Sachdeva, N., & McAuley, J. (2020). How useful are reviews for recommendation? A critical review and potential improvements. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1845–1848). ACM. https://doi.org/10.1145/3397271.3401281

[62] Steck, H. (2018). Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems* (pp. 154–162). ACM. https://doi.org/10.1145/3240323.3240372

[63] Su, W., Chen, S., & Qian, E. (2026). Narrative-aware scientific claim verification agent with evidence ranking for ClimateCheck. *Journal of Technology Informatics and Engineering, 5*(1), 327–340. https://doi.org/10.51903/jtie.v5i1.549

[64] Su, W., Chen, S., & Zhao, C. (2025). Budgeted multi-hop retrieval agent for compositional question answering: A retrieval-policy evaluation on the official MultiHop-RAG benchmark. *Journal of Technology Informatics and Engineering, 4*(3), 649–662. https://doi.org/10.51903/jtie.v4i3.543

[65] Su, W., Rao, H., & Ma, E. (2026). Privacy and data-integrity risk cards for LLM agents: A UI/UX design framework for secure human oversight under prompt-injection attacks. *International Journal of Graphic Design, 4*(1), 186–191. https://doi.org/10.51903/ijgd.v4i1.3699

[66] Sun, X., Lu, Y., & Chen, J. (2023). Controllable long-term user memory for multi-session dialogue: Confidence-gated writing, time-aware retrieval-augmented generation, and update/forgetting. *Journal of Advanced Computing Systems, 3*(8), 9–24. https://doi.org/10.69987/jacs.2023.30802

[67] Sun, X., Zhong, Z. S., & Wu, Q. (2026). Retrieval-grounded HDFS log anomaly detection and deterministic failure narrative generation. *Journal of Computer Systems and Applications, 3*(1), 15–30. https://doi.org/10.64229/j6d7fr94

[68] Tu, H., Zhao, S., & Zhou, A. (2025). Visual brief cards for advertising design: A structured UI/UX framework for turning creative intentions into graphic design decisions. *International Journal of Graphic Design, 3*(1), 210–226. https://doi.org/10.51903/ijgd.v3i1.3714

[69] Wu, Q., Meng, S., & Zhao, J. (2025). Text-grounded LLM-assisted design rationale interfaces: Turning advertising layout metadata into explainable UI/UX decision cards. *International Journal of Graphic Design, 3*(1), 216–240. https://doi.org/10.51903/ijgd.v3i1.3713

[70] Xin, Q. (2025a). Explaining OpenStack failure-injection log anomalies with retrieved normal prototypes. *Emerging Information Science and Technology, 6*(2). https://doi.org/10.18196/eist.v6i2.31232

[71] Xin, Q. (2025b). Hybrid cloud architecture for efficient and cost-effective large language model deployment. *Journal of Information Systems and Informatics, 7*(3), 2182–2195. https://doi.org/10.51519/journalisi.v7i3.1170

[72] Xin, Q. (2025c). Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning). *Journal of Technology Informatics and Engineering, 4*(1), 215–238. https://doi.org/10.51903/jtie.v4i1.485

[73] Xin, Q. (2026a). Auditable automated essay scoring and formative feedback: A rubric-grounded pipeline for secondary and higher education. *Journal of Applied Artificial Intelligence in Education, 2*(1). https://doi.org/10.66053/jaaie.v2i1.348

[74] Xin, Q. (2026b). Behavior retrieval plus response generation for interpretable conversational personalized recommendation. *International Journal of Electrical, Energy and Power System Engineering, 9*(2), 120–136. https://doi.org/10.31258/ijeepse.9.2.120-136

[75] Xin, Q. (2026c). Early-warning analytics with LLM intervention rationales for student retention decisions: Classroom interaction modeling with xAPI-edu-data and dropout/success prediction. *Interdisciplinary Journal of Pedagogical Research and Media Technology, 2*(1). https://doi.org/10.64268/inspire.v2i1.117

[76] Xin, Q. (2026d). Explainable and fair credit risk scoring with counterfactual explanations: A reproducible evaluation on the German Credit dataset (HELOC-motivated). *Journal of Information and Technology, 14*(2). https://doi.org/10.32664/j-intech.v14i02.2228

[77] Xin, Q. (2026e). Log anomaly detection with conformal alert control and evidence-grounded incident ticket generation. *AVITEC, 8*(2), 247. https://doi.org/10.28989/avitec.v8i2.3974

[78] Xin, Q. (2026f). Self-supervised customer representation learning for segmentation and next-purchase prediction on UCI online retail. *Journal of Information and Technology, 14*(1). https://doi.org/10.32664/j-intech.v14i01.2229

[79] Xu, H., Chen, Y., & Med, A. (2025). 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. *Journal of Technology Informatics and Engineering, 4*(3), 590–612. https://doi.org/10.51903/jtie.v4i3.491

[80] Xu, K., Zhou, H., Zheng, H., Zhu, M., & Xin, Q. (2024). Intelligent classification and personalized recommendation of e-commerce products based on machine learning. *Applied and Computational Engineering, 64*(1), 143–149. https://doi.org/10.54254/2755-2721/64/20241365

[81] Ye, T., Mu, J., & Hunter, J. (2026). Off-policy evaluation and conservative policy selection for slot-level dynamic bidding and ranking on the Open Bandit Dataset (small). *Journal of Technology Informatics and Engineering, 5*(1), 178–199. https://doi.org/10.51903/jtie.v5i1.503

[82] Zhang, B., Rao, H., & Zhao, D. (2024). Evidence-grounded RAG for cloud-native DevOps: Hallucination-resistant AIOps question answering over private operations documents. *Journal of Advanced Computing Systems, 4*(3), 109–125. https://doi.org/10.69987/jacs.2024.40308

[83] Zhang, B., Ren, Y., & Zou, J. (2025). LLM-style explainable e-commerce recommendation cards: A UI/UX design framework for trust-calibrated product recommendation. *International Journal of Graphic Design, 3*(2), 381–396. https://doi.org/10.51903/ijgd.v3i2.3697

[84] Zhang, B., Sun, X., Liu, G., & Zhou, B. (2026). LLM-style DevOps copilot for cloud-native troubleshooting: Retrieval-augmented runbook generation and command-safety evaluation. *Journal of Technology Informatics and Engineering, 5*(2), 104–118. https://doi.org/10.51903/jtie.v5i2.534

[85] Zhang, J. (2025). Adaptive user interface design for volleyball learning apps: Empirical evidence from Google Play reviews and mobile screen analysis. *International Journal of Graphic Design, 3*(1), 175–195. https://doi.org/10.51903/ijgd.v3i1.3618

[86] Zhang, K., Chen, Y., & Qian, A. (2025). Evidence-grounded accounting disclosure review cards: A visual communication framework for LLM-style explanations over SEC financial statements and notes. *International Journal of Graphic Design, 3*(2), 395. https://doi.org/10.51903/ijgd.v3i2.3710

[87] Zhang, R., Wen, Z., Wang, C., Tang, C., Xu, P., & Jiang, Y. (2025). *Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms* [Preprint]. arXiv. https://doi.org/10.48550/arxiv.2511.19481

[88] Zhang, Y., & Zhang, H. (2025a). A therapist-facing session copilot for live counseling support: Reasoning-guided retrieval and ranking from multi-turn counseling dialogues. *Journal of Technology Informatics and Engineering, 4*(2), 464–486. https://doi.org/10.51903/jtie.v4i2.547

[89] Zhang, Y., & Zhang, H. (2025b). Visualizing the right counseling support: Evidence-linked recommendation cards for explainable mental health intake interfaces. *International Journal of Graphic Design, 3*(1), 214–229. https://doi.org/10.51903/ijgd.v3i1.3722

[90] Zhao, S., Ren, Y., & Chang, X. (2026). Profit-aware spot GPU admission control with cost-sensitive loss and evidence-grounded policy memos for AI workload supply-demand matching. *Journal of Technology Informatics and Engineering, 5*(2), 45–59. https://doi.org/10.51903/jtie.v5i2.545

[91] Zheng, D., & Li, C. (2024). Behavior-level jailbreak resistance via multi-stage refusal + utility preservation. *Journal of Advanced Computing Systems, 4*(1), 83–99. https://doi.org/10.69987/jacs.2024.40107

[92] Zheng, D., Li, C., & Davidson, H. (2023). Continual red-teaming for in-the-wild jailbreaks via online guardrail updates and guardrail distillation. *Journal of Advanced Computing Systems, 3*(2), 35–49. https://doi.org/10.69987/jacs.2023.30203

[93] Zheng, D., Zhang, B., & Geibel, J. (2024). VerifySafe: Toxicity-safe agent responses under adversarial prompts with evidence-based self-verification. *Journal of Advanced Computing Systems, 4*(1), 67–82. https://doi.org/10.69987/jacs.2024.40106

[94] Zheng, L., Noroozi, V., & Yu, P. S. (2017). Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (pp. 425–434). ACM. https://doi.org/10.1145/3018661.3018665

[95] Zhong, Z. S., Chen, J., Zhong, E., & Sun, X. (2025). Evidence-calibrated RAG for unanswerable question answering: Retrieval coverage, abstention calibration, and hallucination-proxy analysis on SQuAD 2.0. *Journal of Technology Informatics and Engineering, 4*(2), 502–520. https://doi.org/10.51903/jtie.v4i2.536

[96] Zhong, Z. S., Li, C., & Rao, H. (2026). Trajectory reliability prediction for generalist AI agents: Tool-use failure analysis and success forecasting on ZClawBench. *Journal of Technology Informatics and Engineering, 5*(1), 341–360. https://doi.org/10.51903/jtie.v5i1.539

[97] Zhong, Z. S., & Ling, S. (2024a). Improved theoretical guarantee for rank aggregation via spectral method. *Information and Inference: A Journal of the IMA, 13*(3), iaae020. https://doi.org/10.1093/imaiai/iaae020

[98] Zhong, Z. S., & Ling, S. (2024b). *Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization* [Preprint]. arXiv. https://arxiv.org/abs/2408.05944

[99] Zhong, Z. S., Pan, X., & Lei, Q. (2025). Bridging domains with approximately shared features. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics* (Vol. 258, pp. 559–567). PMLR.

[100] Zhong, Z. S., Wu, Q., & Mi, G. (2025). Uncertainty-aware medical image explanation cards: LLM-generated visual explanations for AI-assisted radiology interfaces. *International Journal of Graphic Design, 3*(2), 415–436. https://doi.org/10.51903/ijgd.v3i2.3616

[101] Zhou, B., Li, C., & Liu, L. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX design framework for rubric-based medical risk communication. *International Journal of Graphic Design, 3*(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3696

[102] Zhou, B., Wang, H., & Chang, X. (2025). Distilling VMAF into an edge-deployable quality predictor: A pilot shot-level proxy with LLM-ready quality tokens. *Journal of Technology Informatics and Engineering, 4*(2), 447–463. https://doi.org/10.51903/jtie.v4i2.522

[103] Zhu, Y., Xian, Y., Fu, Z., de Melo, G., & Zhang, Y. (2021). Faithfully explainable recommendation via neural logic reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3083–3090). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.245