PT Jurnal Cendekia Indonesia
Journal of Information Systems and Business Technology
Homepage: https://journal.jci.co.id/jisbt
Vol. 02 No. 04 (2026) Page: 79-89
E-ISSN: 3109-8886

Review-Grounded Explainable Recommendation under Extreme User Sparsity: Self-Supervised Reconstruction, Behavior Retrieval, Client-Partitioned Preference Learning, and Evidence-Support Evaluation

Joshua Baker[1], Yan Wang[2], Bradley Cook[3], Xiaoyu Liu[4]

[1]Department of Computer Science, Texas A&M University, College Station, TX, USA
[2]Department of Computer Science, Arizona State University, Tempe, AZ, USA
[3]Department of Computer Science, North Carolina State University, Raleigh, NC, USA
[4]Department of Computer Science, University of Virginia, Charlottesville, VA, USA

Email : [1*]jbaker92@proton.me

Abstract

Amazon Reviews'23 Gift Cards tests whether review text helps when users are sparse and products popularity-dominated. Its 152,410 reviews cover 132,732 users, 1,137 parent products, and complete metadata. Collapsing repeat user-product pairs yielded 125,704 verified positives rated at least four stars. Leave-two-out retained 120,154 training interactions and evaluated 2,769 active users over all 1,009 items with one frozen pre-validation history. Ten recommenders covered popularity, transitions, item/user retrieval, low-rank and corrupted-view reconstruction, BPR, graph propagation, review retrieval, and gated fusion. Separate branches tested first-observation cold start, clipped-noise client learning, and post-hoc evidence attribution. Markov achieved NDCG@10 0.2346 and HR@10 0.4056; gating gave it all weight, so reviews neither improved warm ranking nor caused recommendations. In the 2021 cold-start proxy, metadata TF-IDF reached micro NDCG@10 0.4613 over 113 candidates, but item-macro and dominant-target-excluded NDCG@10 scores were 0.0627 and 0.0605. Across 500 cases, the cited template reached 0.998 lexical support and 1.000 source localization; shuffling preserved support but reduced user-profile cosine from 0.1683 to 0.0392. Thus ranking, support, alignment, and causal faithfulness diverge.

Keywords: Explainable Recommendation, Amazon Reviews'23, Review Retrieval, Self-Supervised Reconstruction; Cold Start; Client-Partitioned Recommendation; Faithfulness; Evidence Attribution; Temporal Recommendation.

Abstrak

Dataset Amazon Reviews'23 Gift Cards menguji apakah teks ulasan bermanfaat dalam situasi di mana jumlah pengguna sedikit dan popularitas produk sangat mendominasi. Dataset ini mencakup 152.410 ulasan dari 132.732 pengguna dan 1.137 produk induk, serta dilengkapi dengan metadata lengkap. Penggabungan pasangan pengguna-produk yang berulang menghasilkan 125.704 interaksi positif terverifikasi dengan peringkat minimal empat bintang. Metode leave-two-out menyisakan 120.154 interaksi pelatihan dan mengevaluasi 2.769 pengguna aktif terhadap seluruh 1.009 item, dengan satu riwayat pra-validasi yang dibekukan. Sepuluh sistem rekomendasi diuji, mencakup pendekatan popularitas, transisi, pengambilan item/pengguna, rekonstruksi low-rank dan corrupted-view, BPR, propagasi graf, pengambilan ulasan, dan gated fusion. Cabang pengujian terpisah mengevaluasi skenario cold start pada observasi pertama, pembelajaran klien dengan clipped-noise, serta atribusi bukti post-hoc. Model Markov mencapai skor NDCG@10 sebesar 0,2346 dan HR@10 sebesar 0,4056; mekanisme gating memberikan seluruh bobot pada metode ini, sehingga ulasan tidak meningkatkan peringkat pada kondisi warm maupun memengaruhi rekomendasi yang dihasilkan. Dalam proksi cold-start tahun 2021, metadata TF-IDF mencapai skor micro NDCG@10 sebesar 0,4613 dari 113 kandidat, namun skor item-macro NDCG@10 dan skor NDCG@10 dengan pengecualian target dominan masing-masing hanya sebesar 0,0627 dan 0,0605. Pada 500 kasus, templat yang digunakan mencapai dukungan leksikal 0,998 dan lokalisasi sumber 1,000; pengacakan (shuffling) mempertahankan tingkat dukungan namun menurunkan nilai kosinus profil pengguna dari 0,1683 menjadi 0,0392. Dengan demikian, aspek peringkat, dukungan, penyelarasan, dan kesetiaan kausal menunjukkan perbedaan hasil.

Kata kunci: Rekomendasi yang Dapat Dijelaskan (Explainable Recommendation); Amazon Reviews'23; Pengambilan Ulasan; Rekonstruksi Self-Supervised; Cold Start; Rekomendasi dengan Partisi Klien; Kesetiaan (Faithfulness); Atribusi Bukti; Rekomendasi Temporal.

A. INTRODUCTION

Diterima Redaksi: xx-xx-2022 | Selesai Revisi: xx-xx-2022 | Diterbitkan Online: xx-xx-2022

Amazon Reviews'23 links review text, ratings, timestamps, metadata, users, and items [1]. Gift Cards is a sparse, compact, repetitive stress test of whether richer representations and natural-language evidence actually improve ranking and explanation.

Collaborative recommendation spans matrix factorization [2], pairwise implicit ranking [3], and LightGCN [4]. Review-aware models use joint encoders [5], attention [6], review graphs [7], and contrastive objectives [8], but text helps only when it adds information beyond popularity [9]. Recent e-commerce work combines BERT with nearest-neighbor retrieval [10] and jointly evaluates review-grounded ranking and explanation faithfulness [11]. With 84.15% five-star reviews, Gift Cards is framed as full-catalog ranking from incomplete positives rather than rating prediction.

Temporal and self-supervised alternatives include SASRec [12], BERT4Rec [13], S3-Rec [14], and perturbed graph views [15]. Transaction self-supervision can support next-purchase prediction [16], while long-term personalization motivates confidence-gated writing, time-aware retrieval, and forgetting [17]. Gift Cards histories are too short to presume that capacity wins: the last reviewed format may carry more signal than a representation estimated from one or two events.

Language-based recommendation uses unified pretraining [18], review evidence [19], personalized Transformers [20], and aspect factors [21]. E-commerce cards expose ranking, review evidence, and confidence [22]; counseling cards pair suggestions with evidence and review controls [23]. Search evidence cards expose hierarchy [24], while AI critique can diverge from human priorities [25]. Cross-domain risk work separates accuracy, calibration, fairness, and recourse [26]. Fluent factual generation still requires traceable support [27], [28].

Federated averaging [29], collaborative filtering [30], secure matrix factorization [31], and graph personalization [32] motivate decentralization. Topic-preference learning can avoid centralizing raw dialogue [33], while formal differential privacy needs calibrated noise [34] and sacrifices utility [35]. Our logical clients and Gaussian perturbations are only a robustness test, not an epsilon guarantee.

Evaluation goals differ: confidence calibration [36], cold start [37], unseen-entity transfer [38], and list calibration [39]. Selective text matching separates calibrated scores from refusal [40]; multilingual RAG joins answerability, abstention, and visible evidence [41]. Late-fusion calibration [42] and shared-feature transfer [43] address other risks. Propensity-based OPE [44] requires logged actions; because sampled negatives can reverse model orderings [45], all 1,009 items were ranked. Uplift [46] additionally needs randomized exposure absent here.

The study decomposes where temporal, collaborative, self-supervised, and review signals work; whether gating rejects harmful channels; whether post-hoc evidence is supported and aligned; and how client-partitioned learning responds to noise.

B. METHODS

Data, Identity, And Preprocessing

We used Amazon Reviews'23 Gift Cards reviews and metadata [1], joining all 1,137 parents by parent_asin because child variants lack one-to-one metadata. Reviews span August 2008-September 2023; 2024 is the benchmark release year. Table 1 summarizes the raw data.

Table 1 Dataset Profile Before Recommendation Filtering

Measure	Value	Unit
Review records	152,410	rows
Distinct users	132,732	users
Distinct parent items	1,137	items
Metadata records	1,137	rows
Review-item metadata coverage	100.00	%
Verified-purchase share	93.14	%
Mean rating	4.553	stars
Median review length	11	words
Mean review length	18.279	words
Reviews with helpful votes	14.32	%
Repeated user-item records	2,524	rows
Earliest interaction	2008-08-06	date
Latest interaction	2023-09-06	date
Review file size	47.90	MiB
Metadata file size	1.94	MiB

Rows were ordered by timestamp and source line; the latest repeated user-parent event was retained. Positive implicit preference required a verified rating of at least four, yielding 125,704 pairs. Snapshot aggregate ratings were excluded. Item text combined title, store, category, features, description, and nonaggregate details.

Primary temporal protocol

Users with at least three positives contributed penultimate validation and final test targets; earlier events formed history. Shorter histories remained in training. Warm targets yielded 2,769 users, 120,154 training interactions, 110,302 training users, and 1,009 products. Both targets used the same frozen history; validation was masked at test, not appended.

The user-level split is chronological, not globally causal. Evidence had to precede each query, while cold start used a global cutoff. Training items were masked in both splits and validation targets at test; all remaining items were ranked. Table 2 quantifies filtering loss from 2-core through 5-core.

Table 2 Sparsity Sensitivity And Primary Temporal Split

Protocol	Interactions	Users	Items	Density (%)	Median user	Median item
2-core	25,402	10,159	551	0.4538	2.0	9.0
3-core	10,160	2,643	354	1.0859	3.0	8.0
4-core	4,723	929	202	2.5168	4.0	11.0
5-core	2,067	324	116	5.4997	6.0	11.5

Protocol	Interactions	Users	Items	Density (%)	Median user	Median item
Primary train	120,154	2,769 eval	1,009	0.1842	-	-

Recommendation Models

MostPop used log training frequency. Markov estimated smoothed consecutive-item transitions with a 0.10 popularity prior. ItemKNN kept 40 cosine item neighbors. Behavior-Retrieval aggregated the 50 nearest nonself histories among 110,302 training users. Conversational retrieval ranks responses alongside retrieved rationales [47], while behavior-retrieval generation has been proposed for interpretable personalization [48].

PureSVD used 48 components. SSL-Dropout-SVD trained a shared 48-component decoder on two 20%-edge-dropout views to reconstruct intact histories; this is corrupted-view reconstruction, not contrastive sequence learning. BPR-MF learned 48-dimensional factors for 100 epochs. BPR+LightGCN applied two normalized propagation layers and averaged representations, testing [4]'s aggregation without end-to-end LightGCN.

Review-SVD encoded training reviews and static metadata as at most 14,000 sublinear TF-IDF unigrams and bigrams, reduced them to 64 dimensions, and ranked by cosine to prior-review queries. Held-out reviews entered no profile, candidate document, vocabulary, or evidence store.

RG-Hybrid standardized temporal, graph-propagated, and review scores. A 0.10-step validation grid selected nonnegative weights summing to one and could reject any channel. Unlike spectral rank aggregation over noisy pairwise orderings [49], this was score-level gating that directly tested incremental ranking value.

Table 3 Model And Evaluation Settings

Component	Setting
Temporal protocol	All verified ratings >=4; active-user last-two-out; inactive users retained in training
Candidate protocol	Frozen training-history query; full catalog; validation item masked at test
ItemKNN	Cosine similarity; 40 neighbors
Behavior retrieval	User cosine similarity; 50 neighbors
PureSVD	48 components; 10 power iterations
SSL-Dropout-SVD	Two views; 20% interaction dropout; 48 components
BPR-MF	48 dimensions; 100 epochs; lr 0.035; L2 0.002
BPR+LightGCN Prop.	BPR initialization; 2 post-hoc normalized propagation layers; layer mean
Review-SVD	TF-IDF 1-2 grams; max 14,000 terms; 64 latent dimensions
RG-Hybrid	Temporal/graph/review fusion; weights selected on validation grid of 0.1
Client-partition stress test	32 dimensions; 80 rounds; 5% clients/round; clip 0.10; no privacy accountant
Cold-start	2021-01-01 raw first-observation cutoff; static metadata candidate encoding
Bootstrap	5,000 paired user resamples; seed 2024; model seeds averaged; Holm correction

Figure 1 separates the four evaluation branches, while Table 3 fixes their model and statistical settings.

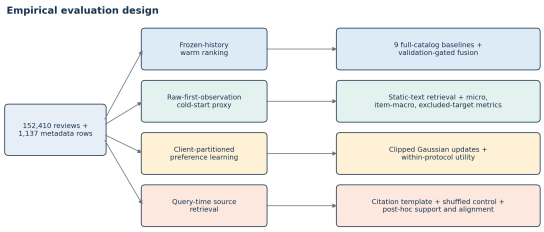

Figure 1 Empirical Design Comprising Frozen-History Warm Ranking, Raw-First-Observation Cold Start, A Client-Partitioned Noise Stress Test, And Post-Hoc Citation-Controlled Evidence Evaluation.

Cold-Start And Client-Partitioned Simulations

Cold start used a January 1, 2021 cutoff. Candidates had no prior raw review; positive histories and first targets required verified ratings of at least four. This yielded 388 queries, 24 targets, and 113 candidates. One target covered 84.28% of queries; median/maximum horizons were 396/869 days. Because first review is not launch and metadata lacks availability dates, we report query-micro, target-macro, and dominant-target-excluded metrics. Post-cut reviews were excluded. Baselines were Random, Metadata-TFIDF, Review-to-Metadata, Semantic-SVD, and Text Hybrid.

The client-partitioned simulation used 2,769 users, 5,147 frozen-history edges, local user vectors, and shared item vectors. Across 80 rounds, 5% of clients were sampled; item updates were clipped at L2 0.10 and aggregated with Gaussian noise. Multipliers 0-2 used three seeds. Because budgets differ, results are not comparable with centralized BPR; without secure aggregation or an accountant, they are not a privacy guarantee.

Grounded Explanation Protocol

Explanation evaluation sampled 500 test users (seed 2024). Retrieval-summary [50], multi-hop [51], and evidence-reranking [52] pipelines motivate separating selection from verbalization; here one prior sentence was selected by TF-IDF. Because RG-Hybrid gave reviews zero weight and reproduced Markov, the sentence is post-hoc support, not a causal account. Sources were static metadata or reviews preceding the query.

Five methods were evaluated: uncited Generic Popularity, high-weight Aspect Template, nearest-sentence Extractive Evidence, a Citation-Controlled Evidence Template, and a Shuffled Evidence Control drawn from another product. The deterministic template isolates retrieval and attribution from fluency; no language model was used in the experiment.

Evidence-chain RAG separates retrieval, attribution, and provenance [53]; evidence-calibrated RAG separates coverage from abstention [54]. Lexical support was content-token overlap with a cited source; uncited output scored zero. Aspect recall covered four top item terms, user relevance was prior-profile cosine, and source localization required an exact ID. Evidence-replacement change was

Jaccard distance to another sentence, not score sensitivity; Distinct-2 measured bigram diversity. A shuffled citation can be supported yet misaligned [27], [28]; rubric-grounded scoring likewise keeps evidence anchors upstream of verbalization [55].

Metrics And Statistical Analysis

Ranking used HR@5/10, NDCG@10, MRR@10, coverage@10, and novelty@10. Stochastic models used seeds 17, 29, and 43; deterministic pipelines ran once. User-level comparisons averaged seeds before 5,000 paired bootstrap resamples, used plus-one two-sided p-values, and Holm-corrected nine tests. Efficiency reports wall time, stored-numeric-state proxy, score memory, peak memory, software/data/code digests, and protocol ID.

C. RESULTS AND DISCCUSSION

Sparsity And Observation Structure

The 152,410 reviews came from 132,732 users (mean 1.148). Verified purchases were 93.14%; five-star and at-least-four-star shares were 84.15% and 88.54%. These properties make random splits prone to future-history leakage and ordinary rating classifiers majority-dominated. Figure 2 shows review volume rising through the late 2010s and a long, thin user-activity tail.

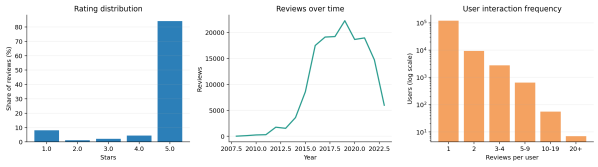

Figure 2 Rating, Temporal, And User-Frequency Distributions Computed From All 152,410 Reviews.

Core filtering would change the task substantially: the positive-set 3-core has 10,160 interactions and 354 products, while the 5-core has 2,067 and 116. The active-user protocol preserved more than eleven times the 3-core evidence while keeping ordered held-out targets.

Full-Catalog Ranking

Table 4 and Figure 3 show First-Order Markov at NDCG@10 0.2346, MRR@10 0.1821, and HR@10 0.4056. MostPop had higher HR@10 0.4381, nearly equal NDCG@10 0.2335, and lower MRR@10 0.1694: popularity retrieved more top-ten targets, while Markov ranked successes closer to one.

Table 4 Full-Catalog Next-Item Ranking; Stochastic Models Report Three-Seed Mean And Standard Deviation.

Model	Seeds	HR@5	HR@10	NDCG@10	MRR@10	Coverage@10	Novelty@10
RG-Hybrid	3	0.2839 ± 0.0000	0.4056 ± 0.0000	0.2346 ± 0.0000	0.1821 ± 0.0000	0.2597	6.3619
First-Order Markov	1	0.2839	0.4056	0.2346	0.1821	0.2597	6.3619
MostPop	1	0.3164	0.4381	0.2335	0.1694	0.0178	4.8006
BPR+LightGCN Prop.	3	0.2089 ± 0.0028	0.2800 ± 0.0054	0.1657 ± 0.0040	0.1304 ± 0.0059	0.3515	7.2020
ItemKNN	1	0.1939	0.2571	0.1573	0.1267	0.5808	10.3021
BPR-MF	3	0.0934 ± 0.0042	0.1440 ± 0.0023	0.0798 ± 0.0019	0.0604 ± 0.0018	0.4549	8.8345
SSL-Dropout-SVD	3	0.0679 ± 0.0019	0.1157 ± 0.0023	0.0589 ± 0.0012	0.0419 ± 0.0009	0.2223	9.1937
PureSVD	1	0.0679	0.1123	0.0570	0.0404	0.2180	9.1688
Behavior-Retrieval	1	0.0191	0.0242	0.0159	0.0134	0.4133	13.8332
Review-SVD	1	0.0051	0.0116	0.0049	0.0030	0.7344	13.2125

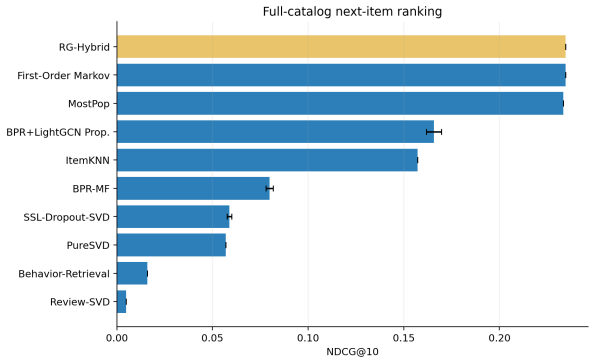

Figure 3 NDCG@10 Across All Ranking Models; Error Bars Show Seed Standard Deviation Where Applicable.

Graph-propagated BPR reached NDCG@10 0.1657, ahead of ItemKNN 0.1573 but below Markov. BPR-MF, SSL-Dropout-SVD, PureSVD, Behavior-Retrieval, and Review-SVD reached 0.0798, 0.0589, 0.0570, 0.0159, and 0.0049. Thus corrupted-view reconstruction barely improved SVD, and higher retrieval coverage/novelty did not offset weak ranks. This supports [9]: short generic reviews may add semantic coverage without discriminative signal. Financial-search reranking has likewise lost to BM25 [56], underscoring strong lexical baselines.

Validation gating assigned weights of 0.0 to graph, 1.0 to temporal, and 0.0 to review in every seed. RG-Hybrid consequently reproduced the Markov list exactly. This is a useful negative result rather than a fusion success: a gate protected test quality by rejecting unhelpful channels, but it did not establish complementary review signal.

History, Popularity, Ablation, And Significance

Table 5 Empirical design comprising frozen-history warm ranking, raw-first-observation cold start, a client-partitioned noise stress test, and post-hoc citation-controlled evidence evaluation. and Figure 4 show both simple baselines weakening with more history. RG-Hybrid NDCG@10 was 0.2470 for one training event, 0.2331 for two, and 0.1954 for three or more; MostPop showed the same decline. Longer-history users may buy specialized cards that the catalog head poorly predicts.

Table 5 Performance By Available User-History Length

Model	Cohort	Users	HR@10	NDCG@10
BPR+LightGCN Prop.	1 interaction	1,672	0.2911	0.1728
BPR+LightGCN Prop.	2 interactions	592	0.2748	0.1660
BPR+LightGCN Prop.	3+ interactions	505	0.2495	0.1418
Behavior-Retrieval	1 interaction	1,672	0.0233	0.0142
Behavior-Retrieval	2 interactions	592	0.0287	0.0200
Behavior-Retrieval	3+ interactions	505	0.0218	0.0168
MostPop	1 interaction	1,672	0.4629	0.2489
MostPop	2 interactions	592	0.4358	0.2303
MostPop	3+ interactions	505	0.3584	0.1860
RG-Hybrid	1 interaction	1,672	0.4270	0.2470
RG-Hybrid	2 interactions	592	0.4122	0.2331
RG-Hybrid	3+ interactions	505	0.3267	0.1954
Review-SVD	1 interaction	1,672	0.0096	0.0045
Review-SVD	2 interactions	592	0.0152	0.0061
Review-SVD	3+ interactions	505	0.0139	0.0050

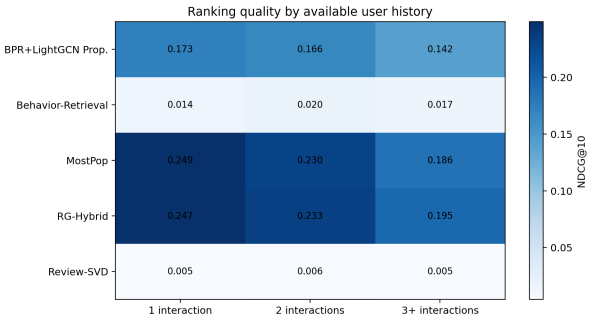

Figure 4 NDCG@10 By Model And Training-History Length

Among 2,769 test targets, 2,689 had more than 20 training interactions, 65 were torso, and 15 tail (Table VI). RG-Hybrid hit 1.54% of torso targets; Review-SVD hit 4.62%, and other reported models hit none. No model hit a tail target. Small denominators limit generalization but locate the regime where content deserves further study.

Table 6 Ranking By Target-Item Training Popularity

Variant	Val. NDCG	Test HR	Test NDCG	w_graph	w_temp	w_review
Temporal only	0.2276	0.4056	0.2346	0.0	1.0	0.0
Temporal + Review	0.2276	0.4056	0.2346	0.0	1.0	0.0
Graph + Temporal	0.2276	0.4056	0.2346	0.0	1.0	0.0
Graph + Temporal + Review	0.2276	0.4056	0.2346	0.0	1.0	0.0
Graph only	0.1549	0.2831	0.1614	1.0	0.0	0.0
Graph + Review	0.1549	0.2831	0.1614	1.0	0.0	0.0
Review only	0.0054	0.0116	0.0049	0.0	0.0	1.0

Table 7 confirms the gate result. Temporal-only and every combination containing temporal reached the same test NDCG@10 of 0.2346 because validation chose temporal weight one. Graph-propagation-only reached 0.1614 for seed 17, while review-only reached 0.0049.

Table 7 Validation-Selected Fusion Ablations For Seed 17

Variant	**Val. NDCG**	**Test HR**	**Test NDCG**	**w_graph**	**w_temp**	**w_review**
Temporal only	0.2276	0.4056	0.2346	0.0	1.0	0.0
Temporal + Review	0.2276	0.4056	0.2346	0.0	1.0	0.0
Graph + Temporal	0.2276	0.4056	0.2346	0.0	1.0	0.0
Graph + Temporal + Review	0.2276	0.4056	0.2346	0.0	1.0	0.0
Graph only	0.1549	0.2831	0.1614	1.0	0.0	0.0
Graph + Review	0.1549	0.2831	0.1614	1.0	0.0	0.0
Review only	0.0054	0.0116	0.0049	0.0	0.0	1.0

Table 8 found no RG-Hybrid/Markov difference because their rankings were identical. Markov-MostPop delta NDCG@10 was 0.0012 (95% CI [-0.0098, 0.0116], Holm p = 1.000). Markov HR@10 was 0.0325 lower (CI [-0.0488, -0.0166], p = 0.0048), while its MRR advantage did

not survive correction. Against seed-averaged BPR+LightGCN, Markov improved NDCG@10 by 0.0689 (CI [0.0578, 0.0796], p = 0.0036).

Table 8 Seed-Averaged User-Level Paired Bootstrap Comparisons With Holm Correction.

Comparison	Metric	Mean Δ	95% CI	Holm p
RG-Hybrid vs. First-Order Markov	NDCG@10	0.0000	[0.0000, 0.0000]	1.0000
RG-Hybrid vs. First-Order Markov	HR@10	0.0000	[0.0000, 0.0000]	1.0000
RG-Hybrid vs. First-Order Markov	MRR@10	0.0000	[0.0000, 0.0000]	1.0000
First-Order Markov vs. MostPop	NDCG@10	0.0012	[-0.0098, 0.0116]	1.0000
First-Order Markov vs. MostPop	HR@10	-0.0325	[-0.0488, -0.0166]	0.0048
First-Order Markov vs. MostPop	MRR@10	0.0128	[0.0018, 0.0234]	0.1080
First-Order Markov vs. BPR+LightGCN Prop.	NDCG@10	0.0689	[0.0578, 0.0796]	0.0036
First-Order Markov vs. BPR+LightGCN Prop.	HR@10	0.1256	[0.1086, 0.1422]	0.0036
First-Order Markov vs. BPR+LightGCN Prop.	MRR@10	0.0518	[0.0412, 0.0621]	0.0036

Long-Horizon First-Observation Item Cold Start

Table 9 and Figure 5 show high query-micro scores that did not generalize under target-balanced views. Metadata-TFIDF achieved NDCG@10 of 0.4613, MRR@10 of 0.3602, and HR@10 of 0.7526 over 113 candidates. The standardized Cold-Start Hybrid raised HR@10 to 0.7706 and coverage to 1.000 while reaching NDCG@10 of 0.4595. Review-to-Metadata reached 0.2048 NDCG@10, Semantic-SVD reached 0.0819, and Random reached 0.0433.

Table 9 Long-Horizon Raw-First-Observation Proxy With Query-Micro, Target-Item-Macro, And Dominant-Target-Excluded Results

Model	Seeds	HR@10	NDCG@10	Item-macro NDCG	Excl. dominant NDCG	Coverage
Metadata-TFIDF	1	0.7526	0.4613	0.0627	0.0605	0.6372
Cold-Start Hybrid	1	0.7706	0.4595	0.1118	0.0704	1.0000
Review-to-Metadata	1	0.3866	0.2048	0.1012	0.0591	0.9823
Semantic-SVD	1	0.1753	0.0819	0.0189	0.0189	0.9558
Random	3	0.0945	0.0432	0.0475	0.0502	1.0000

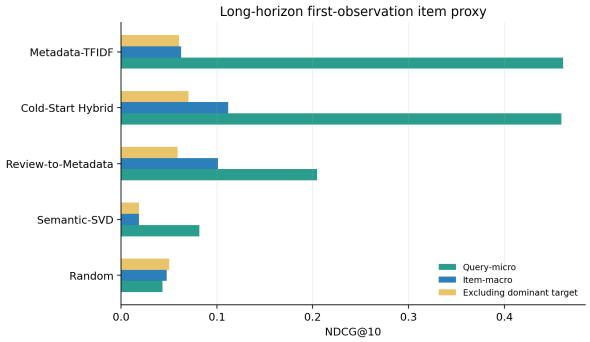

Figure 5 Query-Micro, Target-Item-Macro, And Dominant-Target-Excluded NDCG@10 In The First-Observation Proxy

Micro results are Gift-Cards-specific: one target covered 84.28% of queries. Metadata-TFIDF fell from 0.4613 micro NDCG@10 to 0.0627 under equal target weighting and 0.0605 after removing the dominant target; the hybrid reached 0.1118 and 0.0704. Static text therefore offers limited transfer, while micro strength mainly reflects target skew. This is a long-horizon first-observation proxy, not product launch or strict next-item prediction [37], [38].

Post-Hoc Evidence Support And Alignment

Table 10 and Figure 6 show Extractive and Citation-Controlled Evidence at 0.998 lexical support and 1.000 source localization. Extractive Evidence reached user-profile cosine 0.3398; the longer template reached 0.1683. Aspect Template covered 54.05% of top item terms but had no sentence identifier. Uncited Generic Popularity received zero support and only 0.0138 cosine.

Table 10 Post-Hoc Lexical Support, Source Localization, User Alignment, Evidence-Replacement Change, And Diversity On 500 Recommendations.

Method	N	Lexical support	Aspect	Profile	Localized	Evidence change	Distinct-2
Aspect template	500	0.9300	0.5405	0.2584	0.0000	0.0000	0.1756
Citation-controlled evidence template	500	0.9980	0.0180	0.1683	1.0000	0.5778	0.1705
Extractive evidence	500	0.9980	0.0170	0.3398	1.0000	0.5775	0.4977

Method	N	Lexical support	Aspect	Profile	Localized	Evidence change	Distinct-2
Generic popularity	500	0.0000	0.0050	0.0138	0.0000	0.0000	0.0020
Shuffled evidence control	500	1.0000	0.0030	0.0392	1.0000	0.8335	0.0972

Deterministic provenance separates retrieval from attribution [57], and evidence-grounded RAG can filter unsupported citations [58]. The shuffled control shows why lexical support is insufficient: support stayed 1.000, but aspect recall was 0.0030 and user-profile cosine fell to 0.0392. Correct attribution reached 0.0180 and 0.1683. A single sentence rarely covered four dominant document terms, explaining low absolute recall; nevertheless, the gap shows retrieved evidence aligned with the user's query more than an arbitrary product sentence.

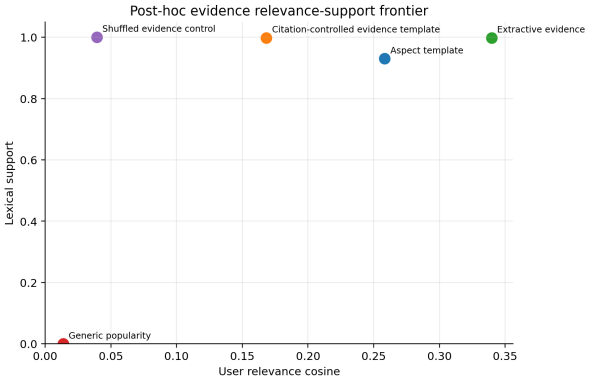

Figure 6 Post-Hoc User-Relevance Cosine Versus Lexical Support For Five Evidence Strategies

Table 11 Grounded Explanation Cases With Exact Evidence Identifiers.

Item	Evidence ID	Evidence	Explanation
B00BXLSJXA	review-line-22536	Easy Christmas gift	A cited source states, "Easy Christmas gift" It was selected because its wording is closest to the user's prior review profile.
B00ADR2LV6	review-line-136796	Wonderful way to give a gift card	A cited source states, "Wonderful way to give a gift card" It was selected because its wording is closest to the user's prior review profile.
B01MSBQB1P	review-line-110397	I was able to use it anywhere	A cited source states, "I was able to use it anywhere" It was selected because its wording is closest to the user's prior review profile.
B015WY0DOQ	metadata-B015WY0DOQ-8	Gift Card Categories.	A cited source states, "Gift Card Categories." It was selected because its wording is closest to the user's prior review profile.
B00JFBLZ90	review-line-26380	Great gift..fast delivery	A cited source states, "Great gift..fast delivery" It was selected because its wording is closest to the user's prior review profile.

Table 11 uses short sources and explicit IDs. Scientific evidence cards [24], counseling recommendation cards [23], and grounded rationale interfaces [59] motivate this hierarchy but do not replace user studies. Wording states only what a source says and how it was selected. Query-time filtering, IDs, and the shuffled control separate attribution from linguistic quality [19], [27], [28]. Because reviews had zero fusion weight, these are post-hoc attribution and alignment metrics, not causal faithfulness.

Client-Partitioned Preference Learning, Efficiency, And Limitations

Within the client-partitioned stress test, NDCG@10 ranged 0.0043-0.0047 across noise multipliers 0-2 (Table 12, Figure 7), below three-seed variability. The slight rise with noise is not evidence of regularization. Because its users, edges, and budget differ from centralized BPR, this supports only a within-protocol robustness observation.

Table 12 Client-Partitioned BPR Stress Test Under Clipped Gaussian Update Noise; No Privacy Budget Was Computed

Noise σ	HR@10	NDCG@10	Coverage	Median update norm	Time (s)
0.00	0.0099 ± 0.0018	0.0043 ± 0.0008	0.9828	0.0157	0.625
0.25	0.0102 ± 0.0019	0.0045 ± 0.0009	0.9818	0.0157	0.660

Noise σ	HR@10	NDCG@10	Coverage	Median update norm	Time (s)
0.50	0.0104 ± 0.0017	0.0045 ± 0.0009	0.9818	0.0157	0.610
1.00	0.0106 ± 0.0019	0.0046 ± 0.0009	0.9815	0.0157	0.582
2.00	0.0107 ± 0.0020	0.0047 ± 0.0008	0.9805	0.0157	0.620

Figure 7 Client-Partitioned BPR NDCG@10 Across Gaussian Update-Noise Multipliers.

Table 13 reports MostPop/Markov training in 0.014/0.006 s and ItemKNN in 0.158 s. BPR-MF, propagated BPR, Review-SVD, and gated fusion averaged 81.96, 82.21, 58.74, and 140.95 s. Each score matrix used 10.66 MiB; the full run took 497.2 s and peaked at 1,020.1 MiB.

Table 13 CPU Training Time, Stored Numeric Values, And Score-Matrix Memory.

Model	Runs	Time (s)	State values (proxy)	Score matrix (MiB)
BPR+LightGCN Prop.	3	82.209	5,342,928	10.66
BPR-MF	3	81.956	5,342,928	10.66
Behavior-Retrieval	1	3.595	120,154	10.66
First-Order Markov	1	0.006	1,018,081	10.66
ItemKNN	1	0.158	1,018,081	10.66
MostPop	1	0.014	1,009	10.66
PureSVD	1	13.843	48,432	10.66
RG-Hybrid	3	140.952	7,335,588	10.66
Review-SVD	1	58.743	974,576	10.66
SSL-Dropout-SVD	3	13.916	48,432	10.66

Post-transaction reviews lack impressions, and unobserved items are not dislikes. Without propensities, OPE [60] and counterfactual LTR [61] are unavailable; without randomized exposure, uplift [62] is unidentified. The main split is within-user chronological; cold start uses a global cutoff. Simulated clients have no privacy budget, secure aggregation, demographics, or fairness labels. Results support ranking, first-observation cold start, and post-hoc attribution, not deployed privacy, causal explanations, or general retail.

D. CONCLUSIONS

This CPU-scale evaluation found Markov NDCG@10 0.2346, nearly tied with MostPop 0.2335, while MostPop had significantly higher HR@10. Both beat graph propagation, factors, behavioral neighbors, corrupted-view reconstruction, and review-semantic baselines on principal metrics. Validation gating removed graph/review channels, matching Markov and avoiding forced-fusion loss. In cold start, Metadata-TFIDF achieved micro NDCG@10 0.4613 but target-macro 0.0627, exposing dominant-target skew.

Review/metadata retrieval produced exact, time-aware source IDs and near-perfect lexical support, but shuffling showed that support does not ensure user/product alignment. Because evidence was retrieved after a zero-review-weight Markov decision, it is not a causal explanation. Evaluation must separate ranking, support, alignment, and decision faithfulness. Client-partitioned noise multipliers 0-2 showed no meaningful utility trend and no privacy guarantee.

For Gift Cards, benchmark temporal against popularity, use static text cautiously for first-observation cold start, and label retrieved sentences as post-hoc evidence, not causal explanations. A later verbalizer should retain source IDs and self-verify before release [63]. Deployed systems can combine confidence-based local/cloud routing [64] with drift monitoring, human review, and new impression-level evaluation.

E. REFERENCES

[1] Y. Hou, J. Li, Z. He, A. Yan, X. Chen, and J. McAuley, "Bridging language and items for retrieval and recommendation," arXiv:2403.03952, 2024, doi: 10.48550/arXiv.2403.03952.

[2] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," Computer, vol. 42, no. 8, pp. 30-37, Aug. 2009, doi: 10.1109/MC.2009.263.

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

[4] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "LightGCN: Simplifying and powering graph convolution network for recommendation," in Proc. 43rd Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2020, pp. 639-648, doi: 10.1145/3397271.3401063.

[5] L. Zheng, V. Noroozi, and P. S. Yu, "Joint deep modeling of users and items using reviews for recommendation," in Proc. 10th ACM Int. Conf. Web Search and Data Mining, 2017, pp. 425-434, doi: 10.1145/3018661.3018665.

[6] C. Chen, M. Zhang, Y. Liu, and S. Ma, "Neural attentional rating regression with review-level explanations," in Proc. 2018 World Wide Web Conf., 2018, pp. 1583-1592, doi: 10.1145/3178876.3186070.

[7] C. Wu, F. Wu, T. Qi, S. Ge, Y. Huang, and X. Xie, "Reviews meet graphs: Enhancing user and item representations for recommendation with hierarchical attentive graph neural network," in Proc. EMNLP-IJCNLP, 2019, pp. 4884-4893, doi: 10.18653/v1/D19-1494.

[8] J. Shuai, K. Zhang, L. Wu, P. Sun, R. Hong, M. Wang, and Y. Li, "A review-aware graph contrastive learning framework for recommendation," in Proc. 45th Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2022, pp. 1283-1293, doi: 10.1145/3477495.3531927.

[9] N. Sachdeva and J. McAuley, "How useful are reviews for recommendation? A critical review and potential improvements," in Proc. 43rd Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2020, pp. 1845-1848, doi: 10.1145/3397271.3401281.

[10] K. Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, "Intelligent classification and personalized recommendation of e-commerce products based on machine learning," Applied and Computational Engineering, vol. 64, pp. 147-153, 2024, doi: 10.54254/2755-2721/64/20241365.

[11] X. Chang, Y. Lu, and Z. S. Zhong, "Review-grounded explainable recommendation with faithfulness evaluation on Amazon Reviews," J. Electr. Eng. Comput. Sci., vol. 11, no. 1, pp. 9-22, Jun. 2026, doi: 10.54732/jeecs.v11i1.2.

[12] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in Proc. IEEE Int. Conf. Data Mining, 2018, pp. 197-206, doi: 10.1109/ICDM.2018.00035.

[13] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "BERT4Rec: Sequential recommendation with bidirectional encoder representations from Transformer," in Proc. 28th ACM Int. Conf. Information and Knowledge Management, 2019, pp. 1441-1450, doi: 10.1145/3357384.3357895.

[14] [14] K. Zhou, H. Wang, W. X. Zhao, Y. Zhu, S. Wang, F. Zhang, Z. Wang, and J.-R. Wen, "S3-Rec: Self-supervised learning for sequential recommendation with mutual information maximization," in Proc. 29th ACM Int. Conf. Information and Knowledge Management, 2020, pp. 1893-1902, doi: 10.1145/3340531.3411954.

[15] J. Wu, X. Wang, F. Feng, X. He, L. Chen, J. Lian, and X. Xie, "Self-supervised graph learning for recommendation," in Proc. 44th Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2021, pp. 726-735, doi: 10.1145/3404835.3462862.

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

[17] X. Sun, Y. Lu, and J. Chen, "Controllable long-term user memory for multi-session dialogue: Confidence-gated writing, time-aware retrieval-augmented generation, and update/forgetting," J. Adv. Comput. Syst., vol. 3, no. 8, pp. 9-24, Aug. 2023, doi: 10.69987/JACS.2023.30802.

[18] S. Geng, S. Liu, Z. Fu, Y. Ge, and Y. Zhang, "Recommendation as language processing: A unified pretrain, personalized prompt and predict paradigm," in Proc. 16th ACM Conf. Recommender Systems, 2022, pp. 299-315, doi: 10.1145/3523227.3546767.

[19] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in Proc. EMNLP-IJCNLP, 2019, pp. 188-197, doi: 10.18653/v1/D19-1018.

[20] L. Li, Y. Zhang, and L. Chen, "Personalized Transformer for explainable recommendation," in Proc. 59th Annual Meeting of the Association for Computational Linguistics, 2021, pp. 4947-4957, doi: 10.18653/v1/2021.acl-long.383.

[21] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma, "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis," in Proc. 37th Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2014, pp. 83-92, doi: 10.1145/2600428.2609579.

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

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

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

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

[26] Q. Xin, "Explainable and fair credit risk scoring with counterfactual explanations: A reproducible evaluation on the German Credit dataset (HELOC-motivated)," J-INTECH, vol. 14, no. 2, pp. 215-231, Jun. 2026, doi: 10.32664/j-intech.v14i02.2228.

[27] [27] Z. Xie, S. Singh, J. McAuley, and B. P. Majumder, "Factual and informative review generation for explainable recommendation," Proc. AAAI Conf. Artificial Intelligence, vol. 37, no. 11,

pp. 13816-13824, 2023, doi: 10.1609/aaai.v37i11.26618.

[28] A. Jacovi and Y. Goldberg, “Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?” in Proc. 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4198-4205, doi: 10.18653/v1/2020.acl-main.386.

[29] [29] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in Proc. 20th Int. Conf. Artificial Intelligence and Statistics, vol. 54, 2017, pp. 1273-1282.

[30] M. Ammad-ud-din, E. Ivannikova, S. A. Khan, W. Oyomno, Q. Fu, K. E. Tan, and A. Flanagan, “Federated collaborative filtering for privacy-preserving personalized recommendation system,” arXiv:1901.09888, 2019, doi: 10.48550/arXiv.1901.09888.

[31] D. Chai, L. Wang, K. Chen, and Q. Yang, “Secure federated matrix factorization,” IEEE Intelligent Systems, vol. 36, no. 5, pp. 11-20, Sep.-Oct. 2021, doi: 10.1109/MIS.2020.3014880.

[32] C. Wu, F. Wu, L. Lyu, T. Qi, Y. Huang, and X. Xie, “A federated graph neural network framework for privacy-preserving personalization,” Nature Communications, vol. 13, Art. no. 3091, 2022, doi: 10.1038/s41467-022-30714-9.

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

[34] C. Dwork, F. McSherry, K. Nissim, and A. D. Smith, “Calibrating noise to sensitivity in private data analysis,” in Theory of Cryptography, LNCS 3876. Berlin, Germany: Springer, 2006, pp. 265-284, doi: 10.1007/11681878_14.

[35] F. McSherry and I. Mironov, “Differentially private recommender systems: Building privacy into the Netflix Prize contenders,” in Proc. 15th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2009, pp. 627-636, doi: 10.1145/1557019.1557090.

[36] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in Proc. 34th Int. Conf. Machine Learning, vol. 70, 2017, pp. 1321-1330.

[37] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, “Methods and metrics for cold-start recommendations,” in Proc. 25th Annual Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2002, pp. 253-260, doi: 10.1145/564376.564421.

[38] M. Volkovs, G. Yu, and T. Poutanen, “DropoutNet: Addressing cold start in recommender systems,” in Advances in Neural Information Processing Systems 30, 2017, pp. 4957-4966.

[39] H. Steck, “Calibrated recommendations,” in Proc. 12th ACM Conf. Recommender Systems, 2018, pp. 154-162, doi: 10.1145/3240323.3240372.

[40] J. Jin, "Calibrated resume-job matching for trustworthy LLM-assisted recruiter screening: Pairwise matching, probability calibration, and selective refusal on two public recruitment datasets," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 625-648, Dec. 2025, doi: 10.51903/jtie.v4i3.529.

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

[42] Q. Xin, “Uncertainty-aware late fusion for 3D perception: Confidence calibration and fusion rule learning,” J. Technol. Informatics Eng., vol. 4, no. 1, pp. 215-238, Apr. 2025, doi: 10.51903/jtie.v4i1.485.

[43] Z. S. Zhong, X. Pan, and Q. Lei, “Bridging domains with approximately shared features,” in Proc. 28th Int. Conf. Artificial Intelligence and Statistics, PMLR 258, 2025, pp. 559-567.

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

[45] W. Krichene and S. Rendle, “On sampled metrics for item recommendation,” in Proc. 26th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2020, pp. 1748-1757, doi: 10.1145/3394486.3403226.

[46] J. Mu, Y. Lu, and M. Smith, “LLM-assisted incrementality (uplift) modeling for email advertising: From feature interactions to interpretable audience-creative-channel policies,” J. Adv. Comput. Syst., vol. 3, no. 1, pp. 31-48, Jan. 2023, doi: 10.69987/JACS.2023.30103.

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

[48] Q. Xin, “Behavior retrieval plus response generation for interpretable conversational personalized recommendation,” IJEEPSE, vol. 9, no. 2, pp. 120-136, Jul. 2026, doi: 10.31258/ijeepse.9.2.120-136.

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

[50] Y. Li, “Findable then explainable: Retrieval-summary integration for code intelligence on a lightweight CodeSearchNet subset,” J. Adv.

Comput. Syst., vol. 4, no. 7, pp. 65-82, Jul. 2024, doi: 10.69987/JACS.2024.40706.

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

[52] W. Su, S. Chen, and E. Qian, "Narrative-aware scientific claim verification agent with evidence ranking for ClimateCheck," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 327-340, Apr. 2026, doi: 10.51903/jtie.v5i1.549.

[53] C. Li, J. Bai, and S. Wang, “Evidence-chain reliable RAG: Word-level hallucination detection, source attribution, and provenance explanation for LLM applications,” J. Adv. Comput. Syst., vol. 4, no. 2, pp. 76-92, Feb. 2024, doi: 10.69987/JACS.2024.40207.

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

[55] Q. Xin, “Auditable automated essay scoring and formative feedback: A rubric-grounded pipeline for secondary and higher education,” J. Appl. Artif. Intell. Educ., vol. 2, no. 1, pp. 1-19, Jul. 2026, doi: 10.66053/jaaie.v2i1.348.

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

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

[58] B. Zhang, H. Rao, and D. Zhao, “Evidence-grounded RAG for cloud-native DevOps: Hallucination-resistant AIOps question answering over private operations documents,” J. Adv. Comput. Syst., vol. 4, no. 3, pp. 109-125, Mar. 2024, doi: 10.69987/JACS.2024.40308.

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

[60] T. Ye, J. Mu, and J. Hunter, “Off-policy evaluation and conservative policy selection for slot-level dynamic bidding and ranking on the Open Bandit Dataset (Small),” J. Technol. Informatics Eng., vol. 5, no. 1, pp. 178-199, Apr. 2026, doi: 10.51903/jtie.v5i1.503.

[61] H. Zhang, “Counterfactual learning-to-rank for ads: Off-policy evaluation on the Open Bandit Dataset,” J. Adv. Comput. Syst., vol. 5, no. 12, pp. 1-11, 2025.

[62] J. Bai, H. Wang, Q. Wu, and B. Zhang, “Privacy-robust incrementality estimation in cookieless settings via uplift modeling: Reproducible evidence from the Hillstrom e-mail experiment,” J. Technol. Informatics Eng., vol. 5, no. 1, pp. 17-38, Apr. 2026, doi: 10.51903/jtie.v5i1.468.

[63] D. Zheng, B. Zhang, and J. Geibel, "VerifySafe: Toxicity-safe agent responses under adversarial prompts with evidence-based self-verification," J. Adv. Comput. Syst., vol. 4, no. 1, pp. 67-82, Jan. 2024, doi: 10.69987/JACS.2024.40106.

[64] Q. Xin, “Hybrid cloud architecture for efficient and cost-effective large language model deployment,” J. Inf. Syst. Informatics, vol. 7, no. 3, pp. 2182-2195, Sep. 2025, doi: 10.51519/journalisi.v7i3.1170.