International Journal of Graphic Design (IJGD)
Volume 3 No. 2 October 2025
E-ISSN: 2987-9434; P-ISSN : 2988-0343, Page 381-396
DOI: https://doi.org/10.51903/ijgd.v3i2.3697
Available online at: https://journal.stekom.ac.id/index.php/ijgd

LLM-Style Explainable E-Commerce Recommendation Cards: A UI/UX Design Framework for Trust-Calibrated Product Recommendation

Boning Zhang[1], Yuxuan Ren*[2], Jocelyn Zou[3]
[1]Computer Science, Georgetown University, DC, USA
[2]Chemical Engineering, University of Washington, WA, USA
[3]Information Experience Design, Pratt Institute, NY, USA
Email Address: boningzhang819@gmail.com

Abstract. *This paper presents and empirically evaluates a UI/UX design framework for explainable e-commerce recommendation cards. The framework addresses a practical visual-communication problem: product lists can be useful but opaque, while explanation-heavy cards can create unwarranted confidence when the system has weak evidence. The revised study therefore uses the term LLM-style for the language condition and treats it as a grounded card-generation and confidence-display policy rather than as evidence of a black-box large-language-model recommender. Experiments were conducted on Amazon Reviews'23 All_Beauty raw reviews and item metadata, together with the Beauty_and_Personal_Care 5-core benchmark split as a larger same-domain warm-user check. The All_Beauty review file contains 701,528 review records from 631,986 users and 112,565 parent items, and the metadata file contains 112,590 parent items with near-complete title and image coverage. On the sparse All_Beauty all-user test, Recall@10 remained low for all methods, with the LLM-style reciprocal-rank reranker reaching 0.007945. On the All_Beauty warm-user slice, the same reranker reached Recall@10 of 0.008079. On the larger Beauty_and_Personal_Care 5-core test, it reached Recall@10 of 0.021463, improving over popularity and last-item co-history baselines but still indicating modest recommendation effectiveness. Card-level evaluation on All_Beauty shows that the LLM-style explanation plus confidence card achieved the highest confidence-discrimination AUC (0.700), while the review-evidence card offered a simpler evidence-forward alternative. The results support an interface-oriented conclusion: recommendation cards should separate ranking quality, grounded evidence, and confidence display, and UI/UX claims should be framed as proxy-based evidence until validated with a controlled user study.*

Keywords *: E-Commerce, Explainable Recommendation, Human-Centered AI, Trust Calibration, Visual Communication.*

INTRODUCTION

Recommendation interfaces are often evaluated by ranking metrics, but customers experience recommendations as visual objects: product cards, lists, carousels, badges, ratings, thumbnails, and short explanatory phrases. In e-commerce, a recommendation card must help a shopper compare products quickly while also communicating why the recommendation deserves attention. A visually polished explanation can improve perceived transparency, but it can also encourage over-trust if the card hides weak evidence or presents uncertain predictions as confident advice.

This paper treats recommendation cards as a design problem at the boundary of recommender systems, visual communication, and human-centered AI. Prior work on explanations in recommender systems has argued that explanations can support transparency, trust, satisfaction, scrutability, and decision effectiveness (Tintarev & Masthoff, 2007, 2012;

Received: July 2025; Revised: August 2025; Accepted: September 2025; Published: October 2025
*Corresponding author, boningzhang819@gmail.com

Zhang & Chen, 2020). Work on trust in automation further shows that appropriate reliance depends on calibration: users should trust a system in proportion to what it can reliably do (Lee & See, 2004). The design challenge is therefore not simply to make users trust recommendations more. A better goal is trust calibration: the interface should help users understand when a recommendation is well supported and when it is speculative.

Language-model-style explanations make this issue more visible because fluent natural language can sound more certain than the available evidence justifies. The revised framing of this study therefore uses LLM-style recommendation cards rather than claiming that an online LLM improves the recommender. The language condition summarizes multiple grounded signals into a compact rationale and adds a confidence cue, but the ranking and evaluation remain deterministic and auditable. This narrower framing keeps the empirical claim aligned with the method.

The empirical contribution is an evaluation on Amazon Reviews'23 beauty-domain data. Amazon Reviews'23 includes user reviews, item metadata, links, and benchmark processing, making it suitable for studying both recommendation behavior and product-card evidence (Hou et al., 2024). The study uses All_Beauty raw reviews and metadata to ground product titles, review availability, metadata tags, and visual card slots, and it adds the larger Beauty_and_Personal_Care 5-core split to reduce the weakness of drawing practical conclusions from the very sparse All_Beauty setting alone.

The study asks a focused research question: when recommendation-card explanations are constrained by available interaction, review, and metadata evidence, do richer LLM-style cards provide better confidence discrimination and visual evidence communication than simpler product-list or evidence-card alternatives? The answer is intentionally limited. The experiments provide ranking results and computational UI/UX proxies; they do not claim to prove user trust, comprehension, or purchase decision quality without a human-subject study. Three contributions follow. First, the paper proposes a card-level UI/UX framework that separates item rank, evidence support, and confidence display. Second, it reports recommendation performance on a sparse All_Beauty split and a larger Beauty_and_Personal_Care 5-core split. Third, it evaluates four card designs using proxy metrics for grounded support, confidence calibration, readability, and information density.

LITERATURE REVIEW

Recommender systems have long balanced predictive performance against user experience. Matrix factorization methods made rating prediction and top-N recommendation scalable by

learning user and item latent factors (Koren et al., 2009), while implicit-feedback methods reframed observed interactions as preference signals rather than explicit satisfaction (Hu et al., 2008). Pairwise ranking methods such as Bayesian Personalized Ranking directly optimized the ordering of observed positives above unobserved negatives (Rendle et al., 2009). These algorithms are useful for e-commerce because they rank products efficiently, but they do not automatically explain why a particular product is shown.

A further difficulty is that offline ranking performance and interface quality do not move in lockstep. A model may produce a useful ranking but offer little reason that can be displayed in a product card. Conversely, a fluent natural-language explanation may be disconnected from the model evidence that produced the ranking. For design-oriented recommendation research, this creates a measurement gap: the researcher must report both whether the system retrieves relevant items and whether the displayed card truthfully represents the evidence available to the system.

Textual and retrieval-based methods add another path for grounded explanation. BM25 remains a strong lexical retrieval baseline because it combines term frequency, inverse document frequency, and document length normalization (Robertson & Zaragoza, 2009). In product recommendation, item titles, descriptions, and reviews can be represented by lexical or neural encoders. Sentence-BERT showed that transformer encoders can produce effective sentence embeddings for semantic search (Reimers & Gurevych, 2019), and BLaIR positions Amazon Reviews 2023 as a benchmark for recommendation, collaborative filtering, and product search with language-item representations (Hou et al., 2024).

Explainable recommendation research studies how systems answer the user's question of why this item. Sinha and Swearingen (2002) showed early evidence that transparency affects recommender acceptance. Tintarev and Masthoff (2007) organized explanation benefits around transparency, scrutability, trust, effectiveness, persuasiveness, efficiency, and satisfaction. Later empirical work found that explanation style matters: not all explanations improve decision making, and some explanation types can increase persuasion without improving understanding (Gedikli et al., 2014; Tintarev & Masthoff, 2012). Zhang and Chen (2020) further distinguished model-intrinsic and post-hoc explainable recommenders.

Human-centered AI research (Kuhn et al., 2024) adds design guidance for how explanations should be displayed. Amershi et al. (2019) proposed guidelines for human-AI interaction (Chen & Chan, 2023), including making system status clear, supporting appropriate trust, and enabling user control. Liao and Varshney (2021) argued that explainable AI should be designed around stakeholders' questions rather than only around algorithmic outputs. From a graphic design perspective, this means the explanation card should make the evidential structure

visible: what signal supports the recommendation, how strong the signal is, and what uncertainty remains.

Trust calibration is the central concept linking these literatures. Over-trust occurs when persuasive explanation design makes uncertain recommendations appear more reliable than they are. Under-trust occurs when a strong recommendation lacks enough evidence for the user to accept it. E-commerce cards need additional visual design rules because users encounter explanations while scanning products, not while auditing a model. This paper therefore treats card text, evidence cues, density, and confidence displays as measurable design variables rather than decorative additions.

METHODS

A. Dataset and splits

Experiments used Amazon Reviews'23 All_Beauty raw review records and item metadata, plus the Beauty_and_Personal_Care 5-core benchmark split. The experimental unit is the parent_asin item identifier because it consistently links interactions to item metadata. Table 1 summarizes the data sources used in the revised evaluation. Table 2 reports split statistics, and Table 3 reports warm-user and target-item observability rates for the evaluation splits.

Table 1. Dataset files used in the revised evaluation

File role	File name	Rows	Users	Items	Fields / use
All_Beauty raw reviews	All_Beauty.jsonl	701,528	631,986	112,565	rating, text, helpful votes, verified flag, timestamp, parent asin
All_Beauty item metadata	meta_All_Beauty.jsonl	112,590	-	112,590	title, description/features, price, images, store, details
BPC 5-core train	BPC.train.csv.gz	5,165,289	729,576	207,385	user_id, parent_asin, rating, timestamp, history
BPC 5-core valid	BPC.valid.csv.gz	729,576	729,576	137,448	standard validation split
BPC 5-core test	BPC.test.csv.gz	729,576	729,576	131,428	standard test split

Note: All_Beauty train, validation, and test splits were derived temporally from the raw review records. BPC abbreviates Beauty_and_Personal_Care.

For All_Beauty, a temporal last-out protocol was derived from the raw review records. For each user, the most recent interaction was assigned to test. Users with at least two interactions contributed their second most recent interaction to validation. Earlier interactions formed the training history. This produces a sparse realistic setting: every user has a test case, but only users with earlier history can receive personalized recommendations. The raw review and metadata files also make it possible to ground visible product-card slots in real item titles, metadata availability, image availability, review text availability, verified-purchase indicators, and review counts.

Table 2. Split statistics for All_Beauty and Beauty_and_Personal_Care

Dataset	Split	Rows	Users	Items	Mean rating	Rating >= 4	Verified rate	Review-text rate
All_Beauty	raw_reviews	701,528	631,986	112,565	3.960	0.713	0.905	1.000
All_Beauty	train	21,109	9,159	11,146	4.185	0.783	0.507	1.000
All_Beauty	valid	48,433	48,433	25,934	4.050	0.736	0.866	1.000
All_Beauty	test	631,986	631,986	106,456	3.946	0.709	0.921	1.000
BPC 5-core	train	5,165,289	729,576	207,385	4.259	0.795	-	-
BPC 5-core	valid	729,576	729,576	137,448	4.188	0.771	-	-
BPC 5-core	test	729,576	729,576	131,428	4.107	0.748	-	-

Note: Verified-purchase and review-text rates apply to All_Beauty raw review-derived splits.

The Beauty_and_Personal_Care 5-core split is used as a larger same-domain warm-user check. Its train, validation, and test files contain 5,165,289, 729,576, and 729,576 rows respectively, and all validation and test users appear in training. This setting does not replace the sparse All_Beauty analysis; instead, it tests whether the same ranking and evidence-display logic remains meaningful when every user has a history and nearly every target item is observable in training.

Table 3. Warm-user and target-item observability rates

Evaluation split	Split	Rows	Warm-user rate	Target item in train	Warm users
All_Beauty	valid	48,433	0.189	0.422	9,159
All_Beauty	test	631,986	0.014	0.383	9,159
BPC 5-core	valid	729,576	1.000	0.999	729,576
BPC 5-core	test	729,576	1.000	0.999	729,576

Note: Low All_Beauty warm-user rates explain why all-user recommendation metrics remain small. Beauty_and_Personal_Care 5-core is a dense warm-user check.

B. Recommendation task and models

The task is next-interaction top-10 recommendation. For each validation or test row, the system ranks candidate parent_asin items; a case is counted as a hit if the held-out item appears in the top 10. Metrics are Recall@10, NDCG@10, MRR@10, catalog coverage, and novelty. Items already present in the user's training history are masked before ranking. In the sparse All_Beauty test, users without training history receive the same popularity fallback, while warm-user results are reported separately.

Table 4 summarizes the implemented models. POP ranks items by training popularity. BM25 and TF-IDF build item-context documents from co-review histories in All_Beauty training data: items that occur in the same user's prior history become contextual evidence for one another. The Beauty_and_Personal_Care scalable baseline uses last-item co-history, which ranks likely next items from the user's most recent training item. LLM-style RR is a deterministic reciprocal-

rank reranker that fuses retrieval or co-history candidates with a popularity prior and exposes the resulting agreement, review support, and metadata support to the card layer. The language condition is therefore reported as LLM-style rather than as an online LLM recommender.

Table 4. Model configurations and roles in the framework

Model	Role	Configuration	Reason in framework
POP	Popularity baseline	Ranks by log-transformed training item frequency	Cold-user fallback and baseline
BM25	All_Beauty co-review retrieval	BM25 over item co-history documents; user query is prior-history items	Evidence source for history-based cards
TF-IDF	All_Beauty co-review retrieval	TF-IDF item co-history vectors with cosine scoring	Lexical embedding-style baseline
Last-item co-history	Beauty_and_Personal_Care sequential baseline	Ranks next items from observed transitions from the user's latest history item	Scalable dense warm-user baseline
LLM-style RR	Explanation-aware reranker	Deterministic reciprocal-rank fusion plus popularity, metadata, and review-support signals	Supplies confidence and language-card features

Note: The LLM-style condition is a deterministic card-generation and confidence-display policy; it does not claim black-box LLM generation or evaluation.

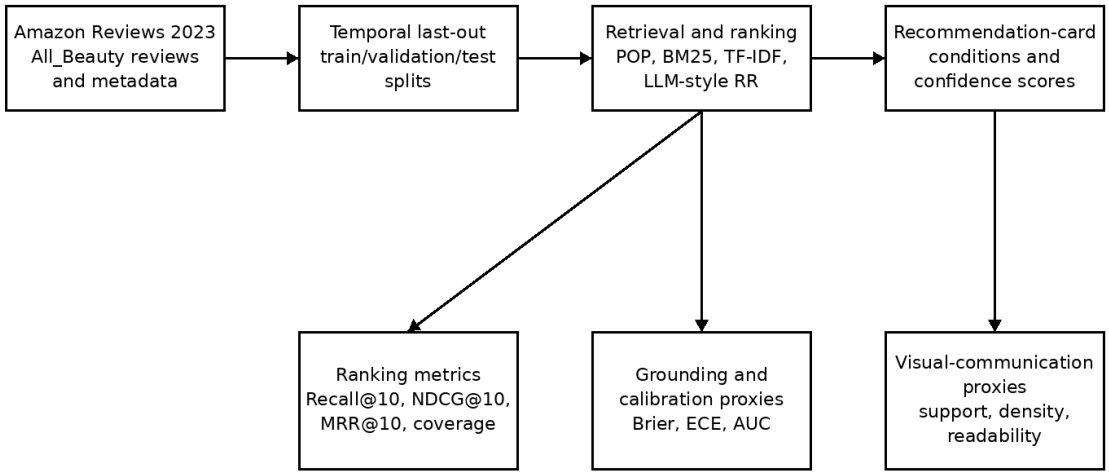

Figure 1. Workflow for grounded recommendation-card evaluation

C. Recommendation-card conditions and proxy metrics

Four UI card conditions were evaluated. The plain product list shows a product block with rating and price slots but no explanation. The metadata-tag card adds compact tags derived from item-level metadata, rating strength, review volume, and product details. The review-evidence card exposes observed review and co-history evidence, such as whether similar histories reviewed the item and whether the item has review-text support. The LLM-style explanation plus confidence card summarizes the same grounded signals into short natural-language copy and displays a calibrated confidence cue.

Card evaluation uses the All_Beauty warm-user validation and test slices. For each warm user, the LLM-style RR top-10 cards are labeled positive when the card item equals the held-out

next interaction. Confidence scores are calibrated on validation and evaluated on test. The reported card metrics are mean displayed confidence, Brier score, Expected Calibration Error with 10 bins, and ROC-AUC. Visual-communication proxies include grounded evidence support, mean grounded signals per card, Flesch Reading Ease, information density, and visible confidence indicator. These metrics are pre-user-study checks; they are not treated as direct evidence of user trust, comprehension, or purchase quality.

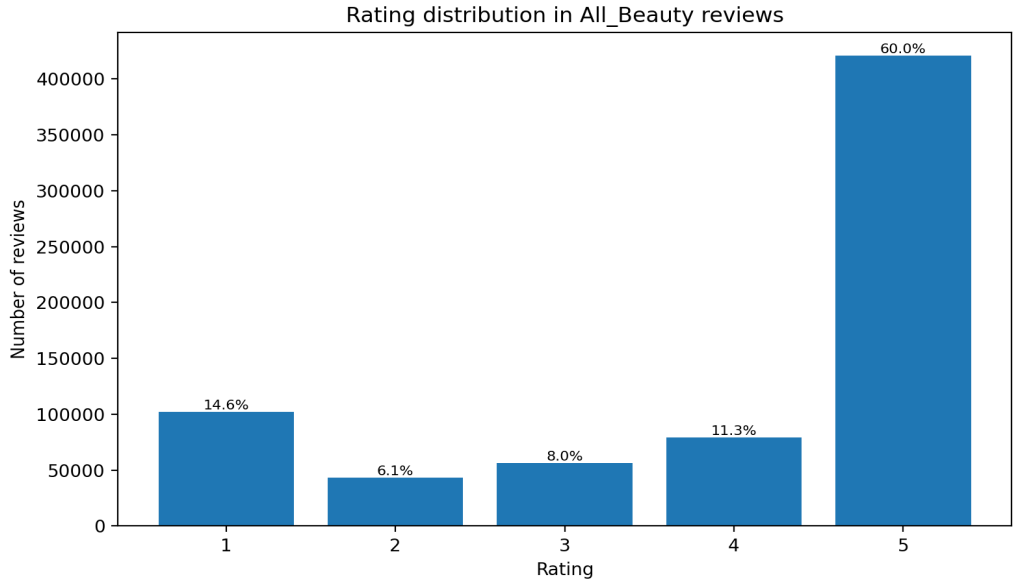

Figure 2. Rating distribution in the All_Beauty raw review file

RESULTS

A. Data distribution and ranking performance

Figures 2 and 3 summarize the All_Beauty empirical distribution. Ratings are strongly skewed toward high scores, and item review counts follow a long-tail pattern. This data structure matters for UI design because a confidence cue can be grounded in abundant evidence for some products but only weak aggregate signals for many others.

Table 5. All_Beauty all-user test results

Model	N	Target in train	Recall@10	NDCG@10	MRR@10	Novelty	Coverage
POP	631,986	0.383	0.007904	0.004385	0.003290	9.811	0.001
BM25	631,986	0.383	0.007935	0.004404	0.003306	9.827	0.428
TFIDF	631,986	0.383	0.007940	0.004401	0.003301	9.827	0.445
LLM-style RR	631,986	0.383	0.007945	0.004406	0.003306	9.826	0.437

Note: All-user results include cold users. Users absent from training receive the popularity fallback.

Table 5 reports All_Beauty all-user test results. Because only 1.45% of test users have training history, all-user Recall@10 is low across methods. The LLM-style RR condition has the highest Recall@10 in this setting, but the absolute value is only 0.007945. This result directly limits the practical recommendation claim: the sparse All_Beauty setting is useful for studying evidence display and cold-start communication, not for claiming strong recommendation effectiveness.

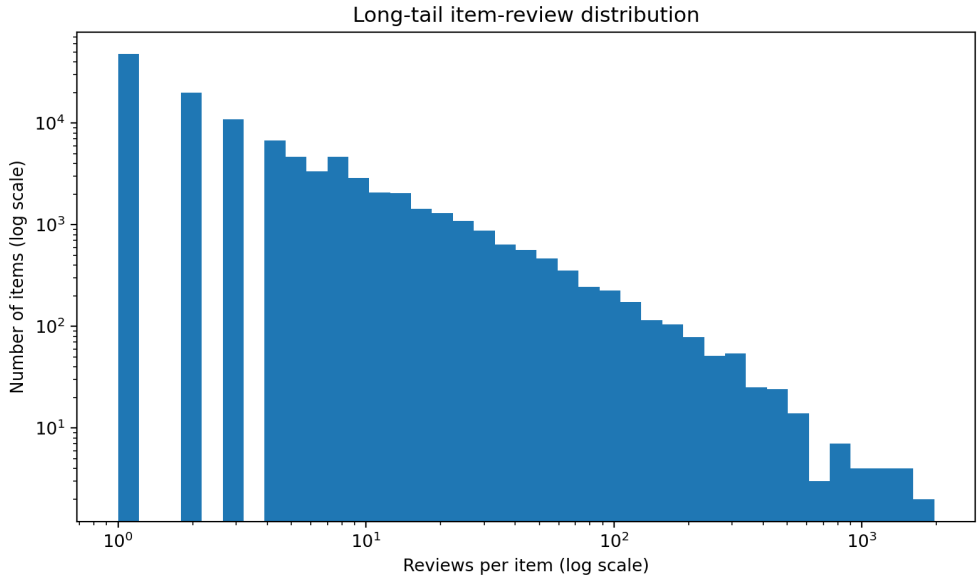

Figure 3. Long-tail item-review distribution in the All_Beauty raw review file

Table 6 isolates the All_Beauty warm-user slice. This slice is more informative for personalization because every evaluated user has a training history. LLM-style RR reaches Recall@10 of 0.008079, followed by TF-IDF at 0.007752 and BM25 at 0.007424. Figure 4 visualizes the warm-user comparison. The gains are directionally useful for the interface framework, but the absolute recommendation accuracy remains modest.

Table 6. All_Beauty warm-user test results

Model	N	Target in train	Recall@10	NDCG@10	MRR@10	Novelty	Coverage
POP	9,159	0.455	0.005241	0.002714	0.001928	9.812	0.001
BM25	9,159	0.455	0.007424	0.004028	0.003006	10.890	0.428
TFIDF	9,159	0.455	0.007752	0.003850	0.002667	10.893	0.445
LLM-style RR	9,159	0.455	0.008079	0.004201	0.003041	10.834	0.437

Note: Warm-user results evaluate the 9,159 test users with training history.

Table 7 reports the larger Beauty_and_Personal_Care 5-core test. In this dense setting, all test users appear in training and 99.86% of target items are observed in training. LLM-style RR reaches Recall@10 of 0.021463, improving over POP and last-item co-history. This check

strengthens the empirical basis of the paper relative to the small All_Beauty 5-core setting, while still supporting a moderate interpretation: the recommendation layer is useful as an evidence source for card design, but it is not presented as a high-performance production recommender.

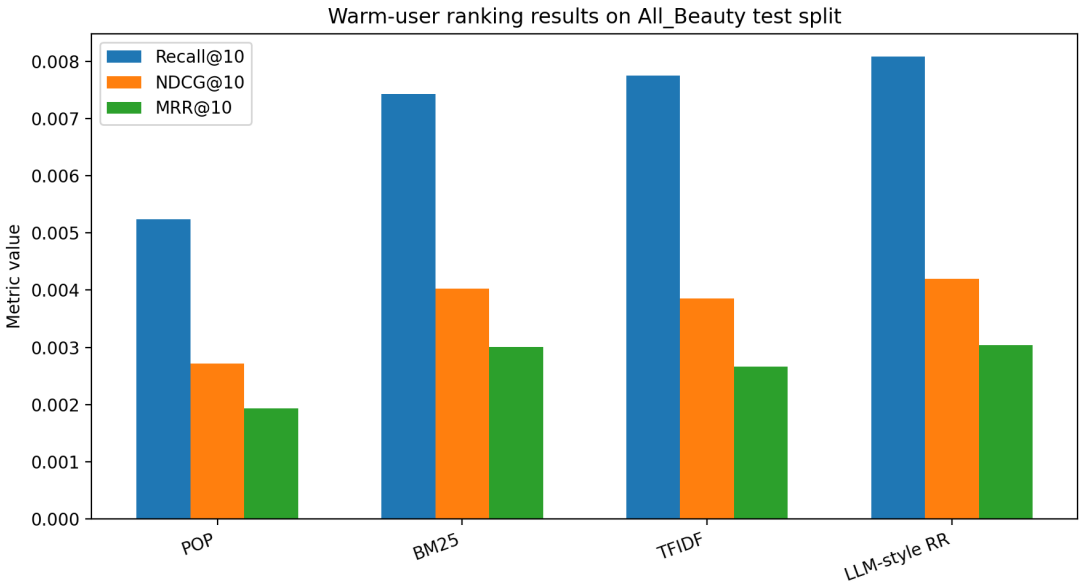

Figure 4. Warm-user ranking results on the All_Beauty test split

Table 7. BPC 5-core test results

Model	N	Target in train	Recall@10	NDCG@10	MRR@10	Novelty	Coverage
POP	729,576	0.999	0.008759	0.005055	0.003926	10.313	0.000
Last-item co-history	729,576	0.999	0.019725	0.013077	0.011028	14.952	0.858
LLM-style RR	729,576	0.999	0.021463	0.013957	0.011648	13.536	0.668

Note: The 5-core split is used as a larger warm-user robustness check.

B. Card confidence and visual-communication proxies

Table 8 evaluates card-level confidence on All_Beauty warm-user test cards. The LLM-style explanation plus confidence card achieves the highest ROC-AUC (0.700), meaning that its confidence score better separates correct from incorrect displayed cards than the simpler variants. The review-evidence card is the closest alternative, with ROC-AUC of 0.664. Brier scores remain small because positives are rare at the displayed-card level. Figure 5 shows the reliability curves for the same confidence estimates.

Figure 6 shows the four card prototypes used to organize the visual-communication conditions. Table 9 then reports the proxy metrics for these card variants. The LLM-style card

has full grounded support and the most grounded signals per card, but it also has the highest information density. Figure 7 visualizes the support-density trade-off.

Table 8. Card confidence calibration on All_Beauty warm-user test cards

Card variant	Test cards	Test positives	Mean confidence	Positive-card confidence	Brier	ECE-10	ROC-AUC
Plain product list	91,590	74	0.001441	0.001441	0.000808	0.000633	0.500
Metadata-tag card	91,590	74	0.001505	0.001483	0.000808	0.000697	0.494
Review-evidence card	91,590	74	0.001459	0.002598	0.000808	0.000651	0.664
LLM-style explanation + confidence card	91,590	74	0.001450	0.002828	0.000809	0.000642	0.700

Note: Scores are calibrated on validation cards and evaluated on 91,590 warm-user test cards.

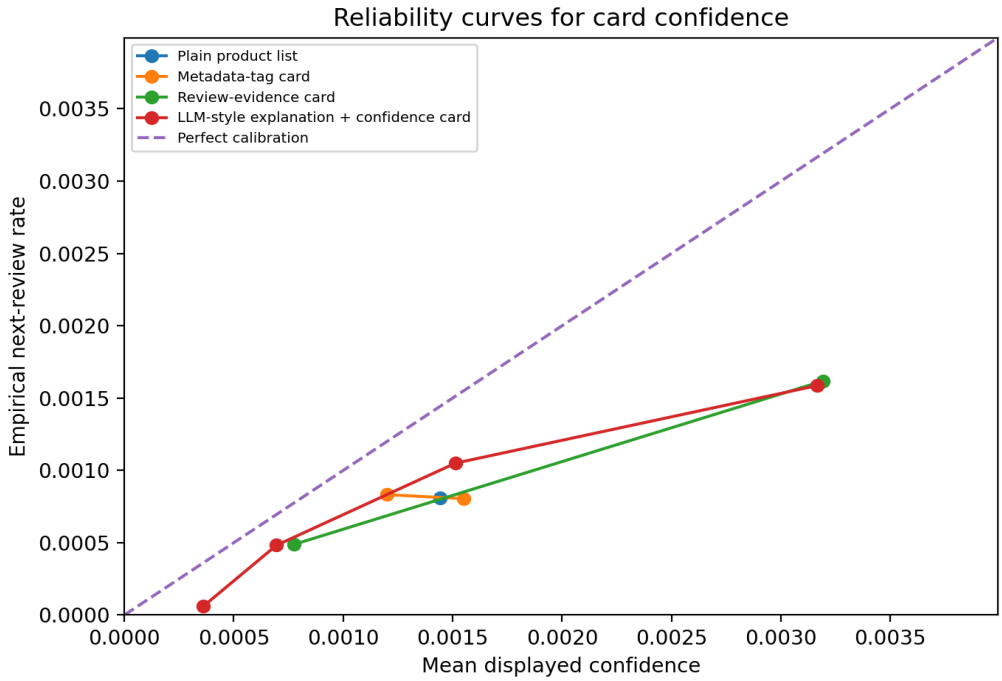

Figure 5. Reliability curves for card confidence on All_Beauty warm-user test cards

Table 10 translates the empirical pattern into design rules. The main design result is not that every product card should contain generated prose. Rather, explanation type should match evidence strength: plain lists are appropriate as a low-density control, metadata tags support quick scanning, review-evidence cards expose observed signals with moderate density, and LLM-style explanations should be reserved for cases where multiple grounded signals and a calibrated confidence cue are available.

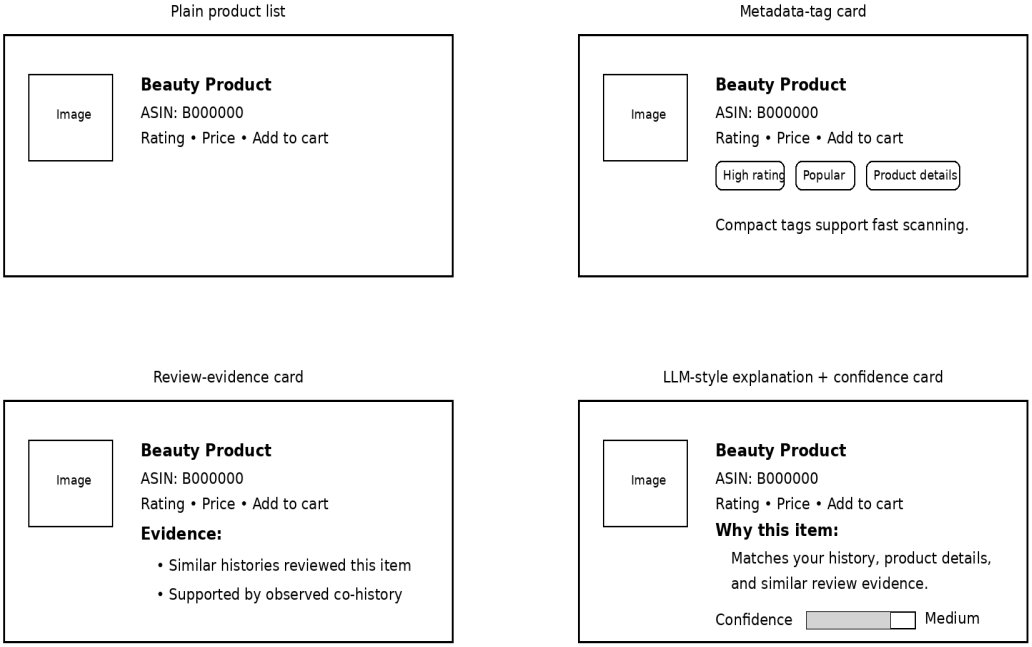

Figure 6. Wireframes for the four recommendation-card conditions

DISCUSSION

The revised results narrow the paper's empirical claim. In the All_Beauty all-user test, most users are cold, so all models are constrained by fallback behavior and Recall@10 remains below 0.008. This makes the practical value of the recommendation layer limited in the sparse setting. The paper therefore treats the recommender primarily as a source of card-level evidence and confidence, not as a strong standalone e-commerce recommendation system.

Table 9. UI/UX card proxy metrics

Card variant	Evidence support	Signals / card	Words / card	Flesch ease	Info density	Confidence visible
Plain product list	0.000	0.000	7	83.610	0.088	0
Metadata-tag card	1.000	2.000	7	30.530	0.088	0
Review-evidence card	1.000	1.000	8	8.365	0.101	0
LLM-style explanation + confidence card	1.000	4.000	14	48.659	0.177	1

Note: Information density is measured as words per 1,000 pixels on a fixed 360 x 220 pixel card canvas.

The Beauty_and_Personal_Care 5-core check partly addresses this limitation. When every user has a history and almost every target item is observed in training, LLM-style RR improves over POP and last-item co-history. Even here, the absolute Recall@10 is 0.021463, so the interpretation remains cautious. The stronger claim is not that the recommender is highly accurate; it is that the interface should expose evidence strength and confidence separately so that weak recommendations are not visually over-presented.

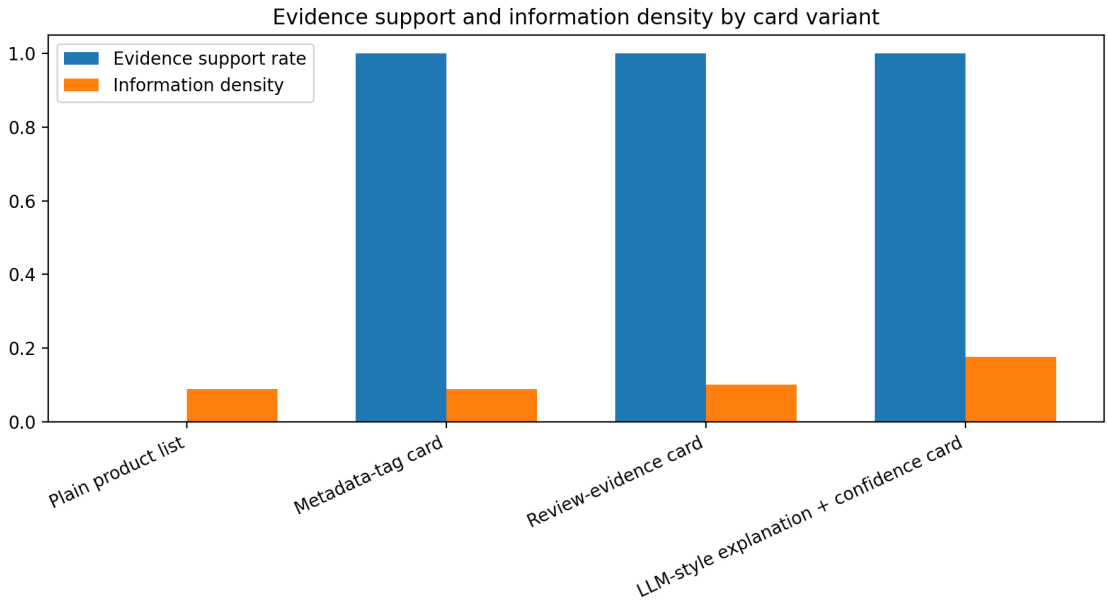

Figure 7. Evidence support and information density by card variant

The card results clarify the contribution. The LLM-style explanation plus confidence card achieves the best confidence discrimination, while the review-evidence card provides a simpler and less dense alternative. This trade-off is important for graphic design and UI/UX practice. Richer language can combine multiple signals into a readable reason, but it also consumes more visual space and may imply more certainty than the model deserves. A responsible product grid should therefore use progressive disclosure: tags by default, evidence bullets when observed support exists, and LLM-style explanation with confidence only when multiple grounded signals agree.

Table 10. Recommendation-card design rules derived from the evaluation

Card	Visible components	Grounding rule	Confidence display	UI/UX purpose
Plain product list	Item block, rating/price slot	No explicit explanation	No displayed confidence	Lowest-density control
Metadata-tag card	Tags for review volume, rating strength, product detail availability	Item metadata and aggregate review statistics	None	Fast comparison and scanning
Review-evidence card	Short evidence bullets	Review text availability, verified-purchase signals, and co-history evidence	None	Evidence-forward explanation without narrative
LLM-style explanation + confidence card	Short rationale plus confidence cue	Ranking agreement, metadata support, review support, and popularity	Calibrated confidence label	Trust-calibration proxy condition

Note: Rules specify when each visual explanation style should be used.

These findings complement, rather than replace, human-subject evaluation. Confidence AUC, calibration error, evidence support, readability, and information density are useful pre-user-study checks because they reveal unsupported claims and excessive density before participant testing. They do not prove that users will trust the system appropriately, understand the explanation, or make better purchase decisions. Stronger claims about trust calibration would require a controlled study measuring perceived transparency, reliance, comprehension, task time, and choice quality.

Limitations

First, the LLM-style condition is deterministic. It follows the explanation pattern associated with LLM interfaces, but the reported experiment does not evaluate a stochastic online LLM generator or black-box LLM reranker. This choice keeps the ranking and card metrics auditable and addresses the present UI/UX research question, but it limits claims about LLM model performance.

Second, the UI/UX evidence is proxy-based. Calibration, confidence discrimination, grounded support, readability, and density are necessary checks for an explainable interface, but they do not directly measure user trust, comprehension, or purchase decision quality. A controlled user study is needed before stronger behavioral claims can be made.

Third, All_Beauty is extremely sparse under the all-user temporal last-out protocol. Only 1.45% of test users have training history, so all-user ranking metrics are dominated by cold-start fallback. The Beauty_and_Personal_Care 5-core experiment provides a larger warm-user check, but future work should evaluate additional categories and richer models.

Fourth, the card framework uses available review and metadata fields, not full product-page context. Some fields such as price and description/features are incomplete for All_Beauty metadata. A production interface would need additional safeguards for missing, stale, or inconsistent product information.

CONCLUSION

This paper developed and evaluated a UI/UX design framework for LLM-style explainable e-commerce recommendation cards. The empirical evaluation on Amazon Reviews'23 beauty-domain data shows that sparse all-user recommendation remains weak, while larger warm-user data provides a more informative test of ranking and evidence-display behavior. The LLM-style explanation plus confidence card provides the strongest confidence discrimination among the tested card variants, but it also increases information density. The review-evidence card offers a simpler calibrated alternative. The main conclusion is that trust-calibrated recommendation cards

should separate three design layers: ranking quality, grounded explanation evidence, and confidence display. Keeping these layers distinct allows language-style explanations to support understanding without turning uncertain recommendations into unconditional persuasion.

REFERENCES

Amershi, S., Weld, D., Vorrell, M., Lee, B., Kapoor, A., Fourney, A., Nushi, B., & Horvitz, E. (2019). Guidelines for human-AI interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1-13. https://doi.org/10.1145/3290605.3300233

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

Chen, X., Zhang, Y., & Wen, J. (2022). Measuring why in recommender systems: A comprehensive survey on the evaluation of explainable recommendation. arXiv. https://arxiv.org/abs/2202.06466

Gedikli, F., Jannach, D., & Ge, M. (2014). How should I explain? A comparison of different explanation types for recommender systems. International Journal of Human-Computer Studies, 72(4), 367-382. https://doi.org/10.1016/j.ijhcs.2013.12.007

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T.-S. (2017). Neural collaborative filtering. Proceedings of the 26th International Conference on World Wide Web, 173-182. https://doi.org/10.1145/3038912.3052569

Hou, Y., Li, J., He, Z., Yan, A., Chen, X., & McAuley, J. (2024). Bridging language and items for retrieval and recommendation. arXiv. https://arxiv.org/abs/2403.03952

Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. Proceedings of the 2008 IEEE International Conference on Data Mining, 263-272. https://doi.org/10.1109/ICDM.2008.22

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). Recommender systems: An introduction. Cambridge University Press.

Jason Kuhn, Yushan Chen, & Evelyn Chan. (2024). AI-Driven Mobile UI Pattern Recognition and Design Topic Mining on RICO: Semantic Clustering and Screenshot-Based Topic Classification. Journal of Advanced Computing Systems , 4(5), 67-83. https://doi.org/10.69987/JACS.2024.40506

Konstan, J. A., & Riedl, J. (2012). Recommender systems: From algorithms to user experience. User Modeling and User-Adapted Interaction, 22, 101-123. https://doi.org/10.1007/s11257-011-9112-x

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

Kula, M. (2015). Metadata embeddings for user and item cold-start recommendations. Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender Systems, 14-21.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Liao, Q. V., & Varshney, K. R. (2021). Human-centered explainable AI (XAI): From algorithms to user experiences. arXiv. https://arxiv.org/abs/2110.10790

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

Nielsen, J. (1994). Usability engineering. Morgan Kaufmann.

Norman, D. A. (2013). The design of everyday things: Revised and expanded edition. Basic Books.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 3982-3992. https://doi.org/10.18653/v1/D19-1410

Rendle, S. (2010). Factorization machines. Proceedings of the 2010 IEEE International Conference on Data Mining, 995-1000. https://doi.org/10.1109/ICDM.2010.127

Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2009). BPR: Bayesian personalized ranking from implicit feedback. Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, 452-461.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. https://doi.org/10.1145/2939672.2939778

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4), 333-389. https://doi.org/10.1561/1500000019

Sinha, R., & Swearingen, K. (2002). The role of transparency in recommender systems. CHI '02 Extended Abstracts on Human Factors in Computing Systems, 830-831. https://doi.org/10.1145/506443.506619

Tintarev, N., & Masthoff, J. (2007). A survey of explanations in recommender systems. 2007 IEEE 23rd International Conference on Data Engineering Workshop, 801-810. https://doi.org/10.1109/ICDEW.2007.4401070

Tintarev, N., & Masthoff, J. (2012). Evaluating the effectiveness of explanations for recommender systems. User Modeling and User-Adapted Interaction, 22, 399-439. https://doi.org/10.1007/s11257-011-9117-5

Yushan Chen, & Evelyn Chan. (2023). Multimodal UI Representation Learning: Ablation of Screenshot, Wireframe, and View-Hierarchy Proxies on an Uploaded 168-Screen Dataset. Journal of Advanced Computing Systems , 3(1), 1-15. https://doi.org/10.69987/JACS.2023.30101

Zhang, Y., & Chen, X. (2020). Explainable recommendation: A survey and new perspectives. Foundations and Trends in Information Retrieval, 14(1), 1-101. https://doi.org/10.1561/1500000066