
Explainable Generative AI for Personalized E-Commerce Decision Support: Integrating Customer Behaviour, Trust and Purchase Prediction

Original Research Article | Volume 1 | Issue 5 | 2026 | Article Number: 101

Accepted: 04 October 2026 | Published: 10 October 2026 | ISSN: 2979-8582 (Online)



Crossref : <https://doi.org/10.67693/BJCR-PYRRLVRI>



Dr. S Afsar Ahmed

Assistant Professor, Department of Commerce, Mazharul Uloom College(Autonomous), Ambur, Tamil Nadu, India

Correspondence: Dr. S Afsar Ahmed, deptofbcmuc@gmail.com

Abstract

E-commerce systems increasingly predict which products a customer may click or purchase, but a ranked list rarely explains why an item fits a person's needs or which trade-offs deserve attention. Generative artificial intelligence (GenAI) creates opportunities for conversational product comparison, review summarization, and personalized guidance; however, fluent output can contain unsupported claims. This paper proposes a conceptual decision-support framework integrating customer-behaviour modelling, purchase-propensity prediction, explainable machine learning, and retrieval-grounded generation. The framework combines searches, views, cart activity, purchases, returns, and review interactions with current product attributes, price, availability, delivery, and return-policy information. A predictive model estimates purchase propensity, an explanation layer identifies relevant evidence, and a generative layer converts it into concise answers with source links and uncertainty disclosures. Trust is treated as an outcome to evaluate—not an automatic benefit of AI—and is supported through transparency, privacy safeguards, user correction, and control over personalization. The paper presents an architecture, research questions, an experimental protocol, and metrics for prediction, calibration, explanation faithfulness, factual grounding, and user understanding. No dataset or user study is claimed; this is a research framework, not a report of measured performance. Its goal is to help shoppers make informed decisions rather than simply maximize conversion.

Keywords: Generative AI, Explainable AI, E-commerce, Recommender Systems, Purchase Prediction, Customer Behaviour, Trust, Retrieval-augmented Generation

I. INTRODUCTION

Online shopping offers extensive choice but often leaves customers to reconcile product specifications, reviews, prices, shipping terms, and competing recommendations. Recommender systems narrow the field using interaction histories, item features, and collaborative patterns [1]–[5]. Yet personalization alone does not guarantee decision quality. A customer may care about compatibility, durability, fit, delivery timing, repairability, or budget—factors a click-oriented ranking objective may not fully capture.

Explainability is therefore important in recommendation research [6]–[8]. Generative AI adds a natural-language interface that can explain trade-offs and answer follow-up questions. Still, language models may hallucinate or present uncertain claims confidently [9], [10]. Retrieval-augmented generation (RAG) can ground responses in product records and policy documents, but retrieval does not guarantee that every generated statement is accurate [11]. A dependable design must connect prediction, explanation, retrieval, validation, and customer control rather than treat a chatbot as a stand-alone solution.

This paper proposes a conceptual framework for explainable GenAI decision support in e-commerce. It addresses behaviour-aware prediction, faithful explanations, verifiable product information, and trust-supporting interaction design. Its contribution is an implementable architecture and evaluation plan; it does not claim that the system has already been tested.

II. BACKGROUND AND RELATED WORK

A. Recommendation and Customer Behaviour

Collaborative filtering identifies preference patterns across users or items, while matrix factorization learns compact representations of interactions [1], [2]. Content-based methods use product attributes, and hybrid systems combine evidence sources [3], [4]. Neural and sequential recommenders model nonlinear relationships and action order [12], [13]. In commerce, behaviour remains an imperfect proxy for preference: a view may reflect curiosity, a cart addition may be tentative, and a purchase may be followed by cancellation or return. Event timing, session context, stock, discounts, and alternatives affect interpretation.

B. Purchase Prediction

Purchase prediction estimates the probability of a purchase within a defined future window. Candidate features include recency and frequency of views, search-to-product transitions, cart actions, prior purchases, price changes, product attributes, and session context. The prediction timestamp must be fixed to prevent future-event leakage. A score is a model estimate, not certain knowledge of intention; calibration and monitoring matter alongside ranking quality [14], [15].

C. Explainable AI

Local methods such as LIME and SHAP estimate which features contributed to a prediction [16], [17]. Counterfactual explanations describe how a result might change under altered conditions [18]. These methods should not be mistaken for causal accounts of a person's motives. Explanations can be unstable, incomplete, or difficult for nontechnical users; faithfulness, stability, clarity, and usefulness should be assessed.

D. Generative AI, Grounding and Trust

Large language models support flexible dialogue but can generate plausible yet incorrect statements [9], [10]. RAG combines a language model with retrieved evidence [11], [19], allowing answers to draw on product specifications, inventory, policies, and review passages. Trust in automation varies with task, experience, transparency, and the ability to correct errors [20], [21]. Privacy and fairness also matter:

personalization should use proportionate data and ranking systems should be monitored for effects on product visibility [22]–[25].

III. PROBLEM STATEMENT AND RESEARCH QUESTIONS

The problem is to provide personalized product guidance that uses behaviour without over-interpreting it, predicts purchase propensity without overstating certainty, and explains recommendations through evidence customers can inspect.

RQ1: How can customer events and product context estimate purchase propensity while distinguishing tentative browsing from stronger signals? RQ2: How can model explanations be translated into natural language without inventing motives or misrepresenting model behaviour? RQ3: How do evidence-linked explanations, uncertainty disclosures, and preference controls affect understanding and calibrated trust? RQ4: Which metrics jointly assess prediction, explanation faithfulness, grounding, privacy, and decision usefulness?

IV. PROPOSED FRAMEWORK

A. Behaviour and Context Layer

The input layer processes relevant events such as searches, impressions, clicks, dwell-time bands, comparisons, cart additions/removals, purchases, cancellations, returns, and review interactions. Events should be timestamped and associated with product/session identifiers. Direct identifiers should be separated from analytical features, access restricted, and retention limited to the stated purpose. The system must distinguish user-stated constraints from interests inferred from behaviour.

The context store contains verified attributes, current price, stock, delivery estimates, return conditions, and review evidence. Dynamic facts should be fetched from authoritative sources at response time. If current data are unavailable, the system should say so rather than rely silently on stale model knowledge.

B. Purchase-Propensity Layer

A supervised model estimates purchase probability within a specified horizon. Logistic regression and gradient-boosted trees can serve as baselines; neural or sequence-aware models may be evaluated when data and validation design justify them [12]–[14]. The target, prediction timestamp, and outcome window must be documented before training.

Let x represent behaviour and context features and y indicate purchase within the target window:

$$\hat{p} = P(y = 1 \mid x) \quad (1)$$

The score may support ranking but must not be presented as certainty or as a complete measure of suitability. Calibration can be examined with Brier score, log loss, and reliability plots [15]. Sparse or stale histories should lead to less confident claims.

C. Explanation Layer

The explanation layer selects relevant reasons from feature attributions and verifiable product evidence. Examples include “matches the budget you selected,” “fits the weight limit you specified,” or “reviews frequently mention battery life,” provided each statement is supported. If comparison activity influenced the score, the system may describe that observed signal; it should not claim to know why the customer acted.

Reasons should be identified as user-provided requirements, observed platform behaviour, model-derived contributions, or product facts. This lets shoppers correct assumptions. A counterfactual may show how

changing a budget or feature requirement affects the shortlist [18].

D. Retrieval-Grounded Generative Layer

The generator receives candidate items, scores, explanation features, current product records, review passages, and source identifiers. It should answer from this evidence, link factual claims to sources, disclose missing information, and avoid inventing specifications or review consensus. A validator checks price, stock, delivery, compatibility, warranty, and return claims. If validation fails, the system should narrow its answer, ask a clarifying question, or abstain.

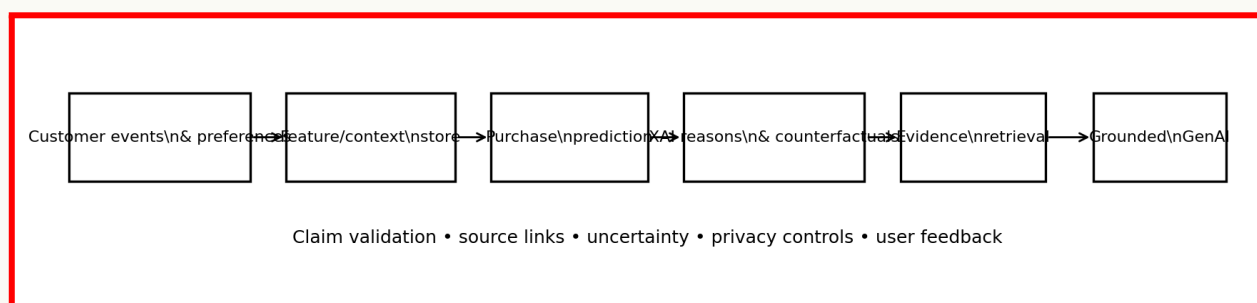
RAG offers a practical retrieval-before-generation pattern [11], [19]. Production safeguards should include freshness checks, access control, prompt-injection testing, and treating seller descriptions and reviews as untrusted data. Retrieval reduces dependence on model memory but does not eliminate verification.

E. Trust and User Control

The interface should explain whether it uses browsing history, purchase history, stated preferences, or only the current session. Customers should be able to edit inferred preferences, compare alternatives, and opt out of personalization. Uncertainty should be visible when evidence is weak, and urgency claims should be based on actual inventory or policy data. These controls aim to encourage appropriate reliance rather than blind acceptance.

V. SYSTEM ARCHITECTURE AND WORKFLOW

The pipeline is: event ingestion → feature/context preparation → purchase prediction → explanation extraction → evidence retrieval → grounded generation → validation and display → feedback and monitoring. This modular design helps diagnose distinct failures: a wrong score, an unfaithful explanation, stale retrieval, or distorted generation.



PLEASE REVIEW THE IMAGE, AS IT IS NOT DISPLAYING CLEARLY. IF YOU WISH TO REPLACE IT, KINDLY REPLY TO THIS EMAIL WITH THE UPDATED IMAGE ATTACHED.

Fig. 1. Proposed architecture for explainable, evidence-grounded e-commerce decision support.

A useful customer-facing answer should be concise but auditable: compare products against stated requirements, identify trade-offs, and link factual claims to specifications or review evidence. The interface should distinguish likelihood of purchase from suitability for the shopper's stated needs.

VI. PROPOSED EXPERIMENTAL METHODOLOGY

A. Data and Protocol

An offline study could use timestamped interactions, impressions, clicks, cart events, purchases, product metadata, and—where permitted—returns. Public product/review and session datasets may support initial experiments, but their coverage and licensing must be checked [26]–[28]. Review-only data may lack reliable impression logs or purchase labels and should not be represented as a complete behavioural dataset

without verification.

Use chronological training, validation, and test periods. Prevent future-event leakage and separate users or sessions where the research question requires generalization. Compare popularity ranking, collaborative filtering, content-based recommendation, and supervised purchase prediction. Evaluate prediction separately from text generation.

B. Metrics

Predictive metrics may include ROC-AUC, precision-recall AUC, precision/recall at k, NDCG@k, Brier score, and log loss. Report results across relevant categories and user groups, with uncertainty estimates where possible. Explanation assessment should cover faithfulness, stability, clarity, and evidence coverage. Human annotators can audit generated claims against retrieved records.

$$\text{Grounding rate} = \text{supported factual claims} / \text{all factual claims} \quad (2)$$

Pair this metric with severity-weighted error analysis: an unsupported colour claim differs from an invented compatibility or safety claim. Trust can be assessed with a pre-registered questionnaire and behavioural measures such as evidence inspection, correction rate, and appropriate reliance. High self-reported trust is not necessarily desirable if the model is wrong.

C. User Study

A controlled study could compare (A) recommendations without explanations, (B) template-based explanations, and (C) evidence-linked generative explanations with uncertainty disclosure. Participants could complete shopping tasks, such as choosing a laptop under a fixed budget with explicit weight and battery requirements. Randomize task order and disclose AI use.

Measure decision accuracy against stated constraints, time to decision, recall of trade-offs, transparency, calibrated trust, and satisfaction. Determine sample size through power analysis, obtain informed consent, minimize personal data, and allow withdrawal. No completed experiment or statistical significance is claimed in this paper.

VII. GOVERNANCE AND DESIGN PRINCIPLES

A. Evidence Before Fluency

Polished language can make weak claims sound authoritative. Concrete statements should be traceable to current specifications, policies, or identified review evidence. Conflicting or sparse evidence should be disclosed rather than converted into confident summaries.

B. Privacy, Fairness and Security

Use only signals relevant to the decision and explain their use in plain language. Data minimization, purpose limitation, access controls, and retention limits reduce unnecessary surveillance [22], [23]. Historical interaction data can overexpose popular products and underexpose new sellers or niche items; monitor coverage and exposure distribution [24], [25]. Treat seller text and reviews as untrusted input and test against prompt injection.

C. Lifecycle Oversight

Assign ownership for product-data quality, model updates, privacy requests, and incident response. Changes to the model or retrieval pipeline should trigger regression tests. High-impact product claims should be verified by deterministic services. Risk-management guidance such as the NIST AI Risk Management Framework can inform lifecycle governance [29].

VIII. DISCUSSION AND LIMITATIONS

Separating prediction, explanation, retrieval, and generation makes the system easier to audit than one in which a language model invents both a recommendation and its rationale. It clarifies that purchase propensity, product suitability, and trust are distinct. A high score may reflect repeated exposure or discount sensitivity; it does not prove that an item best satisfies the customer's needs.

Limitations remain: behaviour logs are incomplete proxies for preference, public datasets may not represent current shopping contexts, feature attribution is not a causal explanation of human motivation, and generation introduces additional error modes. Trust varies by task and experience. The architecture remains conceptual and requires implementation, source-alignment audits, and a preregistered user study.

IX. CONCLUSION

This paper presented a conceptual framework for explainable GenAI in personalized e-commerce decision support. It connects customer behaviour and product context to purchase-propensity estimation, local explanations, retrieval-grounded generation, claim validation, and user controls. Trust should be measured and protected, not assumed. Future evaluation should report prediction, calibration, explanation faithfulness, grounding, user understanding, privacy impacts, and appropriate reliance—not only conversion-oriented metrics. The objective is to help customers understand alternatives and choose according to their own needs.

References

- [1] J. B. Schafer, D. Frankowski, J. L. Herlocker, and S. Sen, "Collaborative filtering recommender systems," in *The Adaptive Web*. Springer, 2007, pp. 291–324.
- [2] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.
- [3] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," in *The Adaptive Web*. Springer, 2007, pp. 325–341.
- [4] R. Burke, "Hybrid recommender systems: Survey and experiments," *User Modeling and User-Adapted Interaction*, vol. 12, pp. 331–370, 2002.
- [5] A. Adomavicius and G. Tuzhilin, "Toward the next generation of recommender systems," *IEEE Trans. Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.
- [6] J. Vig, S. Sen, and J. Riedl, "Tagsplanations: Explaining recommendations using tags," in *Proc. ACM IUI*, 2009, pp. 47–56.
- [7] Y. Zhang and X. Chen, "Explainable recommendation: A survey and new perspectives," *Foundations and Trends in Information Retrieval*, vol. 14, no. 1, pp. 1–101, 2020.
- [8] F. Ricci, L. Rokach, and B. Shapira, eds., *Recommender Systems Handbook*, 3rd ed. Springer, 2022.
- [9] T. Brown et al., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.
- [10] J. Ji et al., "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, 2023, Art. no. 248.
- [11] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.

-
- [12] X. Wang et al., “Neural collaborative filtering,” in Proc. WWW, 2017, pp. 173–182.
- [13] F. Sun et al., “BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer,” in Proc. CIKM, 2019, pp. 1441–1450.
- [14] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [15] J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.
- [16] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’ Explaining the predictions of any classifier,” in Proc. ACM SIGKDD, 2016, pp. 1135–1144.
- [17] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [18] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box,” *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018.
- [19] R. Bommasani et al., “On the opportunities and risks of foundation models,” arXiv:2108.07258, 2021.
- [20] B. J. Dietvorst, J. P. Simmons, and C. Massey, “Algorithm aversion: People erroneously avoid algorithms after seeing them err,” *Journal of Experimental Psychology: General*, vol. 144, no. 1, pp. 114–126, 2015.
- [21] J. A. C. de Melo et al., “Algorithm appreciation: People prefer algorithmic to human judgment,” *Organizational Behavior and Human Decision Processes*, vol. 151, pp. 90–103, 2019.
- [22] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.
- [23] European Union, “Regulation (EU) 2016/679 (General Data Protection Regulation),” *Official Journal of the European Union*, 2016.
- [24] R. Burke, “Multisided fairness for recommendation,” in Proc. FATREC Workshop, 2017.
- [25] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, “Evaluating collaborative filtering recommender systems,” *ACM Trans. Information Systems*, vol. 22, no. 1, pp. 5–53, 2004.
- [26] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, “Image-based recommendations on styles and substitutes,” in Proc. SIGIR, 2015, pp. 43–52.
- [27] R. He and J. McAuley, “Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering,” in Proc. WWW, 2016, pp. 507–517.
- [28] J. Leskovec and A. Krevl, “SNAP Datasets: Stanford Large Network Dataset Collection,” Stanford University, 2014. [Online]. Available: <https://snap.stanford.edu/data/>
- [29] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, 2023.
- [30] OECD, “OECD principles on artificial intelligence,” OECD, 2019.
- [31] Y. Gao et al., “Retrieval-augmented generation for large language models: A survey,” arXiv:2312.10997, 2023.
- [32] Y. Koren, “Collaborative filtering with temporal dynamics,” in Proc. ACM KDD, 2009, pp. 447–456.

- [33] J. A. Konstan and J. Riedl, "Recommender systems: From algorithms to user experience," *User Modeling and User-Adapted Interaction*, vol. 22, pp. 101–123, 2012.
- [34] Y. K. Dwivedi et al., "So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI," *International Journal of Information Management*, vol. 71, 2023, Art. no. 102642.
- [35] ISO/IEC, "ISO/IEC 23894:2023 Information technology—Artificial intelligence—Guidance on risk management," 2023.



© 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the **Creative Commons Attribution (CC BY 4.0) licence** (<https://creativecommons.org/licenses/by/4.0/>)