International Journal of Graphic Design (IJGD)
Volume 4 No. 1 April 2026
E-ISSN: 2987-9434; P-ISSN : 2988-0343, Page 293-310
DOI: https://doi.org/10.51903/ijgd.v4i1.3992
Available online at: https://journal.stekom.ac.id/index.php/ijgd

Evidence-Grounded Visual Explanation Design for E-commerce Recommendation Cards: Complementary and Substitute Relations, Image Availability, and Interface Hierarchy

Hailin Zhou*[1]
[1]Applied Analytics, Columbia University, NY, USA
Email Address: hailin.zhou1668@yahoo.com

Abstract. *E-commerce recommendation cards must explain why an item is related while maintaining a compact, consistent visual hierarchy. This study presents an evidence-grounded card framework and evaluates its retrieval inputs with the final TREC Product Recommendation 2025 topics, a 48,616-product corpus, 7,633 NIST judgments for 47 topics, and the SQID supplementary image-URL table. Fixed lexical methods generated separate top-10 complement and substitute rankings over the full corpus before assessor judgments were applied. Under the official scoring procedure, the relation-aware lexical method achieved an average nDCG@10 of 0.1840, with a complement nDCG@10 of 0.0562 and a substitute nDCG@10 of 0.3119. Its average difference from full-text TF–IDF was not statistically reliable (difference = 0.0111, 95% CI [-0.0090, 0.0321]). The complement-minus-substitute gap was -0.2557 (95% CI [-0.3151, -0.1937]), confirming that complementary recommendation remains the weaker relation. Requiring a confirmed supplementary image URL increased confirmed URL coverage among displayed items to 100% but reduced average nDCG@10 to 0.0138. The findings support relation-qualified reasons, confirmed-image and text-led fallback states, and conservative language for complementary suggestions. The framework specifies evidence and layout behavior; shopper responses and different verbalization methods require direct comparative evaluation.*

Keywords: *E-commerce recommendation; visual explanation; complementary products; substitute products; interface hierarchy*

INTRODUCTION

E-commerce interfaces increasingly present related products in compact cards rather than undifferentiated ranked lists (B. Zhang et al., 2025; Go & Mothelsang, 2024). A useful card must do more than identify a potentially relevant item: it must indicate how that item relates to the product currently being viewed, expose enough evidence for the recommendation to be intelligible, and remain visually usable within a small display area. These requirements are especially important for related-item recommendation, where two products may be substitutes that serve the same function or complements that perform connected functions (K. Xu et al., 2024). Hybrid recommender systems have long combined heterogeneous signals to improve recommendation quality (Burke, 2002), while broader recommender-system research has shown that different prediction settings require different representations of item and contextual evidence (Ricci et al., 2015; Xin, 2026b). The interface problem considered here begins after candidate evidence has been assembled: the system must rank related products and communicate the basis of the relation without overstating what the data establish.

Complementary and substitute relations create distinct ranking and explanation problems. A substitute is normally similar in function to the reference product, so shared titles, brands, attributes, and descriptions can provide useful lexical evidence. A complement may be

Received: Januari 2026; Revised: February 2026; Accepted: March 2026; Published: April 2026
*Corresponding author, hailin.zhou1668@yahoo.com

functionally connected while sharing little vocabulary with the reference product. The Shopping Queries Dataset formalized product-search relevance with Exact, Substitute, Complement, and Irrelevant labels (Reddy et al., 2022), and the TREC Product Search and Recommendations Track subsequently defined separate short-ranking tasks for complementary and substitute products (TREC Product Search Track Organizers, 2025). This separation matters for card design because a generic "similar item" label can be appropriate for a substitute but misleading for a complement (H. Xu et al., 2025). Relation-aware cards therefore require both relation-specific ranking evidence and language that preserves the distinction between similarity and functional pairing.

Showing evidence is not equivalent to showing every model feature. Explainable-recommendation research treats transparency, effectiveness, scrutability, and related user-centered qualities as design and evaluation goals rather than automatic consequences of adding text (Jin, 2025c; Tintarev & Masthoff, 2015; Y. Zhang & Chen, 2020). Miller (2019) further argues that human explanations are typically selective and contrastive. For a recommendation card, this suggests a short reason tied to a visible product attribute or relation (Chang et al., 2026), not an exhaustive technical account of the ranking model. The distinction is important: the present study evaluates ranking behavior and defines field-traceability requirements, while shopper comprehension, confidence, and behavior require direct user evaluation.

The visual structure of the card imposes an additional constraint. Usability principles emphasize visibility of system status, consistency, and recognition rather than recall (Nielsen, 1994), while Norman (2013) connects understandable action to legible system states and signifiers. Cognitive-load theory cautions that additional information can impede processing when it is redundant or poorly organized (Sweller, 1988). Tufte's (2001) account of information density similarly favors meaningful visual content over decoration. Applied to related-product cards, these principles motivate a stable hierarchy: confirmed-image state or a text-led fallback, relation label, concise evidence phrase, and action control (Mu et al., 2026). This hierarchy is a design proposition grounded in established principles; its comparative usability remains a question for controlled evaluation (Zhou, Li, & Liu, 2025).

Image availability is treated as a display condition rather than evidence of semantic relevance. The Shopping Queries Image Dataset (SQID) extends the ESCI setting with product image information and multimodal features (Al Ghossein et al., 2024). Its official repository distributes both a primary image-URL table and a supplementary URL table (Crossing Minds, 2026). The analysis uses only URL availability. It does not inspect image pixels or use image embeddings to infer relatedness. This boundary permits a direct audit of whether a ranked item has a confirmed image URL without turning that audit into a multimodal-effectiveness claim.

The empirical analysis combines the official TREC 2025 product corpus with the final recommendation topics and NIST assessor judgments. The product corpus supplies titles, descriptions, brands, and bullet points for lexical representation (TREC Product Search Track, 2025). The final topics and recommendation qrels provide external evaluation of complementary and substitute rankings (National Institute of Standards and Technology [NIST], 2026). Ranking is assessed separately by relation because an average can conceal weak complement retrieval behind stronger substitute results. Image-URL coverage is joined by product identifier and analyzed separately as a card-readiness property.

The study addresses three questions. First, how effectively do fixed lexical and relation-aware methods rank complementary and substitute products under the official TREC evaluation? Second, what relevance cost accompanies a hard requirement for a confirmed supplementary image URL? Third, how can relation evidence and image state be translated into a compact interface hierarchy without implying effects that were not measured? The contribution is an evidence-grounded recommendation-card framework that connects retrieval evaluation to explicit display decisions, reports image availability separately from relevance, and qualifies complementary suggestions in proportion to the observed evidence.

LITERATURE REVIEW

Related-Product Recommendation and Retrieval

Collaborative and latent-factor recommenders infer preferences from interaction patterns; matrix factorization is a prominent representation of user–item structure (Koren et al., 2009). Hybrid systems combine collaborative, content-based, demographic, or knowledge-based components to address weaknesses in a single evidence source (Burke, 2002). Related-product recommendation can instead be posed without individual user history: given a reference item, the system ranks products according to a functional relation and distinguishes items that replace the reference from items that work with it.

Titles, descriptions, brands, and bullet points encode functional and categorical information. Manning et al. (2008) describe term weighting, vector-space retrieval, and cosine comparison, while Robertson and Zaragoza (2009) formalize probabilistic relevance and the BM25 family. These methods are interpretable baselines, but their limitations differ by relation. Substitutes often share vocabulary because they occupy similar categories; complements may be connected through use or compatibility despite low term overlap. Relation-specific evaluation is therefore necessary before lexical similarity becomes a visible reason.

The ESCI benchmark supplies an important data lineage for this problem. Reddy et al. (2022) introduced a large multilingual collection of shopping queries, product results, and graded

relevance judgments organized around Exact, Substitute, Complement, and Irrelevant labels. The 2025 TREC recommendation task adapts related-product evidence into a seed-product setting: systems receive a reference product and produce distinct complement and substitute rankings (TREC Product Search Track Organizers, 2025). The official track corpus provides the product text used to represent seeds and candidates (TREC Product Search Track, 2025), while the final NIST topics and qrels provide assessor judgments for evaluation (NIST, 2026).

Ranking Evaluation and Evidence-Grounded Explanations

Related-product cards display only a few items, making rank position consequential (Mu et al., 2025). Cumulative-gain measures reward relevant results near the top of a ranking and accommodate graded judgments (Järvelin & Kekäläinen, 2002). NDCG is therefore well suited to the TREC task. Reporting complement and substitute NDCG separately is essential because a single average can obscure whether a method behaves consistently across both relations. Supplementary measures such as reciprocal rank and recall can describe first-relevant placement and candidate coverage, but they should not replace the official relation-specific metric (Su et al., 2025). The implementation uses TF–IDF vectorization and sparse matrix operations from scikit-learn (Pedregosa et al., 2011).

Explainable recommendation includes model- and user-oriented questions (Y. Zhang & H. Zhang, 2025a). Y. Zhang and Chen (2020) organize the field by mechanism, information source, presentation form, and evaluation objective. Tintarev and Masthoff (2015) identify goals including transparency, scrutability, effectiveness, trust, and satisfaction, each requiring suitable evidence. An offline relevance score establishes ranking behavior; it cannot show that a shopper understood an explanation or changed a decision. The narrower term evidence-grounded is used here because each reason must resolve to an available relation label or product field (Chen & Xu, 2026).

This approach differs from post hoc explanation of an opaque predictor. Ribeiro et al. (2016) use local surrogate behavior to explain individual classifier predictions, and Lundberg and Lee (2017) provide a unified feature-attribution framework based on Shapley values. These methods demonstrate the value of associating an output with identifiable local evidence (Jin, 2025b). A recommendation card, however, has less space and a different communicative purpose than a diagnostic explanation panel. It must select one or two decision-relevant cues while preserving the semantic meaning of the recommendation.

Miller's (2019) synthesis of social-science research helps explain why selection matters (Y. Zhang & H. Zhang, 2025b). People tend to seek contrastive answers—why this item rather than another—and prefer causes relevant to the immediate question. For substitutes, an evidence phrase can point to shared function or an overlapping attribute. For complements, the phrase

should describe a use relationship or compatibility signal when that evidence is present; low lexical overlap alone is not a meaningful reason. If the data establish only a broad complementary category, the card should avoid precise compatibility language.

Image Availability and Interface Hierarchy

Visual product information can support retrieval and presentation, but those roles should not be conflated. Al Ghossein et al. (2024) introduced SQID as an image-enriched extension of the Shopping Queries Dataset, including image URLs and multimodal features. The repository documents the primary and supplementary product-image URL resources (Crossing Minds, 2026.). A non-null URL indicates that an image-led card may be renderable; it does not establish image quality (Zhou, Wang, & Chang, 2025), semantic usefulness, currency, or causal benefit. Nor does URL availability constitute multimodal ranking.

Interface-design theory translates this distinction into a card state. Nielsen's (1994) visibility and consistency principles support predictable confirmed-image and text-led fallback layouts. Norman (2013) emphasizes that controls and system states should be perceptible and interpretable. Sweller (1988) shows that unnecessary cognitive load can interfere with processing, and Tufte (2001) argues that visual elements should carry substantive information. Together, these principles support a small set of ordered fields: relation, evidence, image state, and action. They provide a testable design basis, not proof of improved comprehension.

METHODS

Research Design

The research design combines a full-corpus recommendation evaluation with an image-readiness audit and a card-schema analysis. Figure 1 summarizes the sequence. Seed products and corpus records are represented first; separate complement and substitute lists are then generated; NIST judgments are applied only for evaluation; and the resulting evidence is mapped to card fields. This order keeps the final judgments outside candidate generation and model scoring.

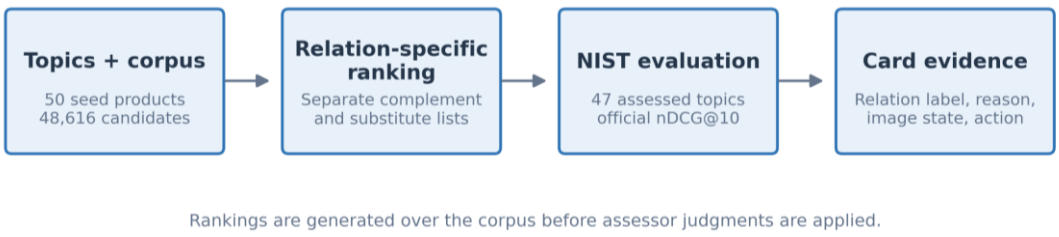

Figure 1. Experimental pipeline from seed-product retrieval to card evidence.

Data Sources and Coverage

The official recommendation topics contain 50 distinct seed products. All 50 appear in the product corpus, which contains 48,616 unique records. The fields used for ranking are title, description, brand, and product bullet points. The final qrels contain 7,633 unique topic–product

judgments for 47 topics; the three remaining topics have no final judgments and are excluded from effectiveness averages. The qrels are used only after a ranking has been produced (NIST, 2026). Table 1 lists the data sources and their analytic roles.

Table 1. Data sources and analytic roles.

Source	Records	Analytic role
TREC recommendation topics	50 seed products	Reference products and released topic descriptions
TREC evaluation corpus	48,616 products	Full-corpus candidate universe and product text
NIST final qrels	7,633 judgments; 47 topics	Official complement and substitute scoring only
SQID supplementary URLs	27,139 identifiers; 26,229 non-null URLs	Confirmed image-URL state for exact ID matches

Note. The recommendation corpus and final evaluation files are distributed for the TREC 2025 Product Search Track; the URL file is the SQID supplement.

The SQID supplementary table contains 27,139 product identifiers, including 26,229 with a non-null image URL. An exact identifier join is used to mark confirmed URL availability. The join is intentionally narrow: an item absent from the supplementary table is not classified as image-less. Figure 2 shows the separation between TREC relation evidence and the SQID display state, which meet only in the card contract (Al Ghossein et al., 2024; Crossing Minds, 2026.).

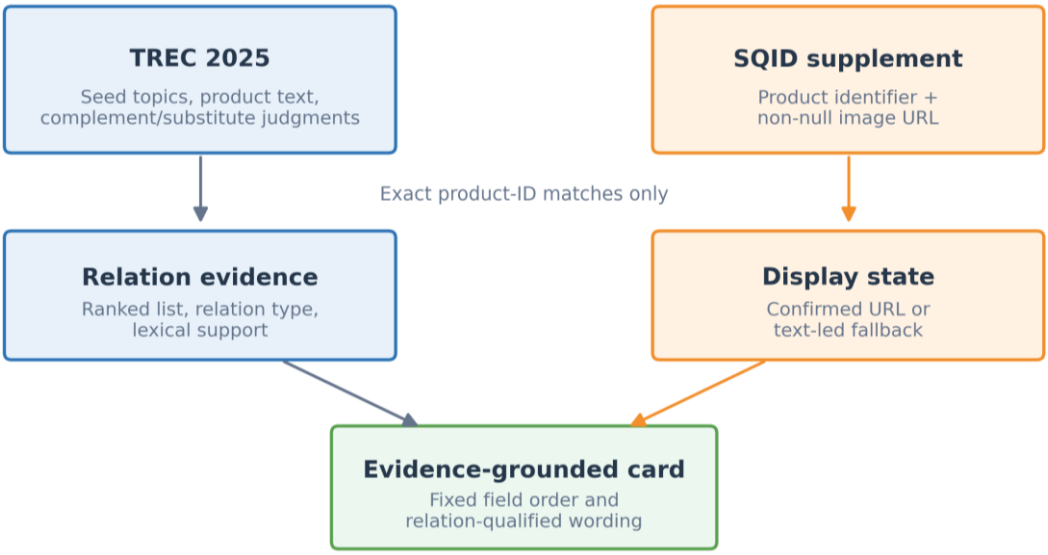

Figure 2. Separation of TREC relation evidence and the SQID display state.

Table 2 reports the evaluation accounting. The corpus contains 5,540 of the 7,103 unique assessed candidate identifiers. Every candidate with a C0–C2 or S0–S2 judgment is present; the missing identifiers carry only NR or UA labels. Title is populated for every corpus record, while description, brand, and bullet-point coverage are 54.58%, 95.76%, and 99.45%, respectively. Eight assessed topics have no positive complement after the seed is excluded, and four have no positive substitute. Those topics remain in the 47-topic averages; a ranking with no positive target beyond the seed contributes zero.

Table 2. Dataset coverage and evaluation accounting.

Item	Value
Released topics	50
Topics with final qrels	47
Topics without final qrels	3
Unique qrels pairs	7,633
Unique assessed candidate IDs	7,103
Assessed candidate IDs present in corpus	5,540
Seed products present in corpus	50 of 50
Title coverage	48,616 (100.00%)
Description coverage	26,535 (54.58%)
Brand coverage	46,557 (95.76%)
Bullet-point coverage	48,351 (99.45%)
Topics without positive complement	8 of 47
Topics without positive substitute after seed removal	4 of 47

Note. All candidates carrying a C0–C2 or S0–S2 judgment are present in the corpus; the absent assessed identifiers are labeled NR or UA.

Ranking Methods

Text is normalized by lowercasing, Unicode accent stripping, whitespace cleanup, and unigram–bigram tokenization. TF–IDF uses sublinear term frequency, L2 normalization, a minimum document frequency of two, and a maximum document frequency of 0.98. Title TF–IDF compares seed and candidate titles. Full-text TF–IDF compares concatenated title, description, brand, and bullet-point fields. Topic TF–IDF compares the released topic description and product title with each candidate's full text. These operations use scikit-learn (Pedregosa et al., 2011).

Table 3. Ranking runs and evaluation purpose.

Run	Score or restriction	Purpose
Deterministic random	Fixed-seed random corpus order	Scale reference
Title TF–IDF	Seed-title to candidate-title cosine	Narrow lexical baseline
Full-text TF–IDF	Cosine over title, description, brand, and bullets	Broad lexical baseline
Topic TF–IDF	Released topic text to candidate full text	Topic-context baseline
Relation-aware lexical	Fixed asymmetric complement/substitute formulas	Relation-specific lexical run
Image-constrained	Relation-aware scores restricted to confirmed URL IDs	Display-readiness stress test

Note. All methods rank the corpus without using the final qrels for fitting, candidate selection, or weight adjustment.

The relation-aware lexical method applies fixed asymmetric weights. Let t be seed–candidate title similarity, f full-text similarity, u the similarity between the topic title and the candidate's supporting fields, v the shared-title similarity in the same vocabulary, and c an indicator for a generic relation cue such as "compatible with," "fits," "designed for," or "works with." The substitute score is $0.75t + 0.25f - 0.25cu$. The complement score is $0.50f + 0.50\max(u - v, 0) + 0.25cu$. The weights are not fitted to the final judgments. A deterministic random order provides a scale reference. Table 3 summarizes the six evaluated runs, including the image-constrained stress test.

Evaluation Protocol

Every run ranks the full 48,616-product corpus and excludes the seed product from its own recommendation list. Each submitted relation list contains 10 items, while the official scorer discounts all positive qrels in the ideal ranking. It also retains the seed's S2 judgment in the substitute ideal ranking; the primary nDCG values follow that procedure. Gains are 2 for C2 or S2, 1 for C1 or S1, and 0 for C0, S0, the opposite relation, NR, UA, and unjudged items. Complement and substitute nDCG@10 are averaged to obtain the relation-averaged score, following the task definition (Järvelin & Kekäläinen, 2002; TREC Product Search Track Organizers, 2025). A diagnostic score also removes the seed from the ideal ranking.

Uncertainty is estimated at the topic level. Mean nDCG intervals use 20,000 percentile bootstrap resamples. Paired model differences and complement-minus-substitute gaps use 100,000 two-sided sign-flip randomizations; Holm adjustment is applied within each reported family. Precision@10, recall@10, MRR@10, qrels-match rate, effective-assessment rate, and confirmed-URL rate describe retrieval behavior alongside the official nDCG. Because the judgments are pooled, unjudged retrieved items receive zero gain and can depress estimates for methods that retrieve outside the assessment pool (Ye et al., 2026).

Card Schema

The card schema receives a ranked item, its list relation, product fields, and the confirmed-URL state. It fixes the display order as image state, relation label, one concise evidence phrase, and action control. Complement wording is qualified unless compatibility is independently established; substitute wording emphasizes alternative function. The reason can be rendered by a deterministic template or another constrained text component, but the renderer may not create new product facts or alter the relation label (Su, Rao, & Ma, 2026).

Table 4. Official nDCG@10 on the 47 assessed topics.

Run	Complement nDCG@10 [95% CI]	Substitute nDCG@10 [95% CI]	Average nDCG@10 [95% CI]
Random	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]
Title TF–IDF	0.0354 [0.0093, 0.0698]	0.3035 [0.2478, 0.3597]	0.1695 [0.1360, 0.2035]
Full-text TF–IDF	0.0581 [0.0238, 0.0992]	0.2878 [0.2286, 0.3481]	0.1730 [0.1349, 0.2135]
Topic TF–IDF	0.0628 [0.0240, 0.1099]	0.2739 [0.2183, 0.3304]	0.1683 [0.1339, 0.2039]
Relation-aware lexical	0.0562 [0.0262, 0.0922]	0.3119 [0.2547, 0.3678]	0.1840 [0.1502, 0.2194]
Image-constrained stress test	0.0030 [0.0000, 0.0083]	0.0246 [0.0092, 0.0425]	0.0138 [0.0060, 0.0229]

Note. Intervals are 20,000-resample topic-bootstrap percentile intervals. The average is the mean of complement and substitute nDCG for each topic. The official IDCG discounts all positive qrels rather than only the first 10.

RESULTS

Overall Ranking Effectiveness

Table 4 presents official nDCG@10 for all evaluated runs. The relation-aware lexical method is numerically highest on the average metric at 0.1840, followed by full-text TF–IDF at

0.1730. The bootstrap intervals overlap substantially. Figure 3 displays the unconstrained relation-averaged results and their 95% topic-bootstrap intervals. The relation-aware method does not establish a reliable improvement over full-text TF–IDF. Their average difference is 0.0111, with a 95% confidence interval of [-0.0090, 0.0321] and a Holm-adjusted p-value of 1.000. The appropriate interpretation is comparable aggregate performance among the lexical methods, not a general superiority claim.

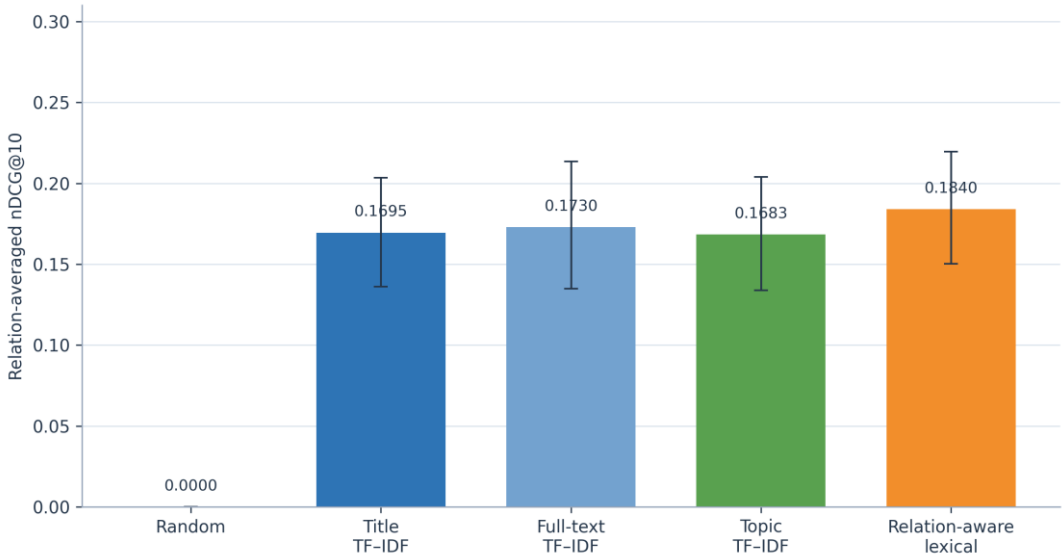

Figure 3. Relation-averaged nDCG@10 with 95% topic-bootstrap intervals.

Complement and Substitute Performance

Relation-specific values in Table 4 and Figure 4 reveal a much larger and more consistent pattern. For the relation-aware lexical method, complement nDCG@10 is 0.0562, whereas substitute nDCG@10 is 0.3119. The pattern also appears for title, full-text, and topic TF–IDF. Complement performance is therefore modest under every lexical formulation. Table 5 adds top-10 diagnostics for the relation-aware method. Complement precision@10 is 0.0383, recall@10 is 0.0790, and MRR@10 is 0.0826. The corresponding substitute values are 0.3745, 0.3598, and 0.7540. Figure 5 makes the first relevant placement gap especially visible.

Table 5. Relation-aware lexical retrieval diagnostics.

Relation	Precision@10	Recall@10	MRR@10	Qrels match@10	Confirmed URL@10
Complement	0.0383	0.0790	0.0826	57.87%	0.85%
Substitute	0.3745	0.3598	0.7540	79.36%	1.06%

Note. Positive labels are C1/C2 for complements and S1/S2 for substitutes. Unjudged retrieved items do not count as hits.

Paired Comparisons and Relation Gap

Table 6 reports paired model comparisons. None of the title, topic, or relation-aware average differences from full-text TF–IDF are significant after Holm adjustment. The image-

constrained run is different: its average nDCG@10 is lower than the unconstrained relation-aware run by 0.1702, with a 95% confidence interval for the signed difference of [-0.2083, -0.1335] and Holm-adjusted p < .001.

Table 6. Paired nDCG@10 comparisons.

Difference	Relation	Mean	95% CI	Holm-adjusted p
Title − Full-text	Average	-0.0035	[-0.0229, 0.0148]	1.000
Topic − Full-text	Average	-0.0046	[-0.0249, 0.0143]	1.000
Relation-aware − Full-text	Complement	-0.0019	[-0.0278, 0.0249]	1.000
Relation-aware − Full-text	Substitute	0.0240	[-0.0025, 0.0523]	.862
Relation-aware − Full-text	Average	0.0111	[-0.0090, 0.0321]	1.000
Image-constrained − Relation-aware	Average	-0.1702	[-0.2083, -0.1335]	< .001

Note. Differences are challenger minus reference. Confidence intervals are paired topic-bootstrap intervals; p-values use 100,000 sign-flip randomizations.

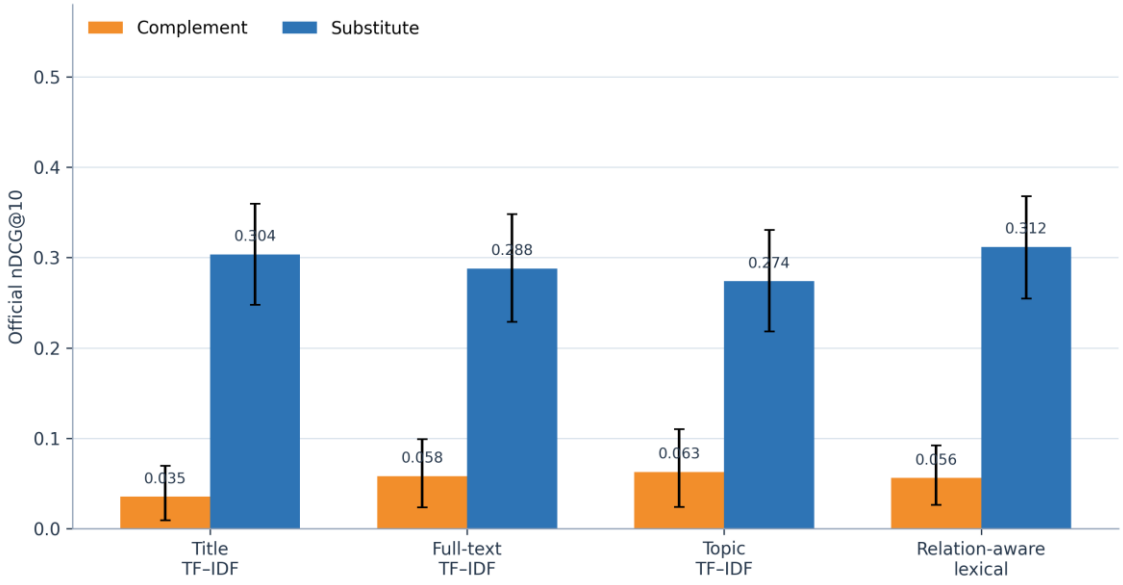

Figure 4. Complement and substitute nDCG@10 with 95% topic-bootstrap intervals.

The complement-minus-substitute tests in Table 7 reach the same substantive conclusion for each non-random lexical method. For the relation-aware run, the mean gap is -0.2557, with a 95% confidence interval of [-0.3151, -0.1937] and Holm-adjusted p < .001. Figure 6 orders all 47 topic-level gaps. Only four topics favor complement ranking; most differences are negative, and the lowest is −0.6131.

Table 7. Complement-minus-substitute nDCG@10 gaps.

Run	Mean gap	95% CI	Holm-adjusted p
Title TF–IDF	-0.2681	[-0.3278, -0.2083]	< .001
Full-text TF–IDF	-0.2297	[-0.2932, -0.1668]	< .001
Topic TF–IDF	-0.2111	[-0.2817, -0.1385]	< .001
Relation-aware lexical	-0.2557	[-0.3151, -0.1937]	< .001
Image-constrained	-0.0217	[-0.0405, -0.0053]	.039

Note. Negative values indicate weaker complement ranking.

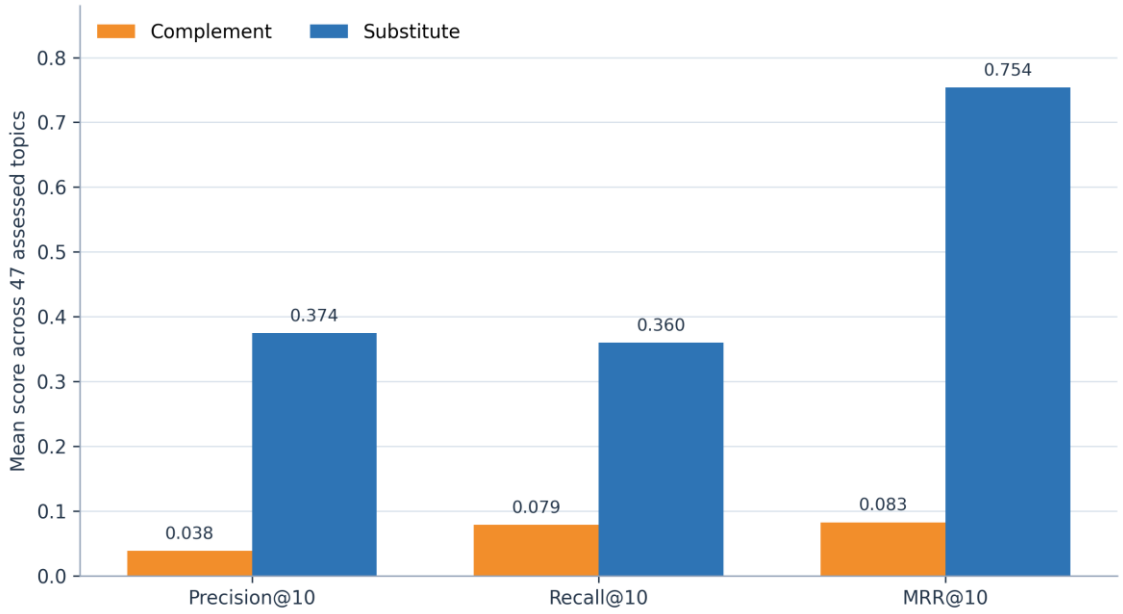

Figure 5. Relation-aware lexical precision, recall, and reciprocal rank at 10.

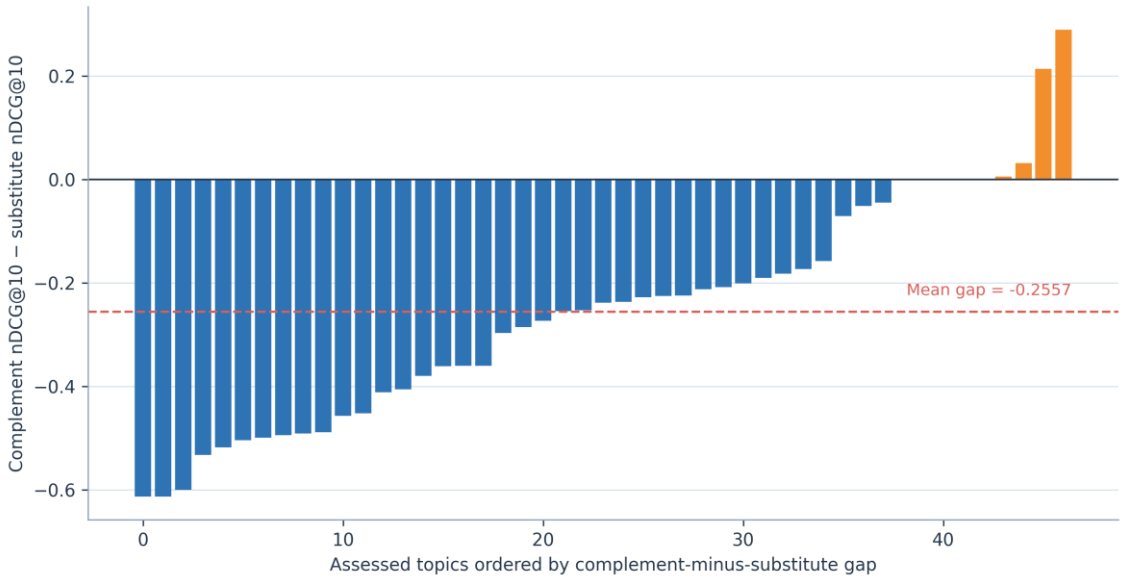

Figure 6. Topic-level complement-minus-substitute nDCG@10 gaps for the relation-aware lexical run.

Table 8. Official scoring and reference-item sensitivity.

Relation	Official nDCG@10	No-self IDCG diagnostic	Qrels match@10	Effective assessed@10
Complement	0.0562	0.0562	57.87%	57.87%
Substitute	0.3119	0.4196	79.36%	79.36%
Average	0.1840	0.2379	68.62%	68.62%

Note. The official IDCG discounts all positive qrels and retains the reference product's S2 judgment even though that product is excluded from every submitted ranking.

Scoring Sensitivity and Assessment Coverage

Table 8 separates official scoring from the diagnostic that removes the seed from the substitute ideal ranking. The relation-aware substitute score rises from 0.3119 to 0.4196 under the diagnostic, while complement is unchanged because the seed carries a substitute label. The

official values remain the primary results. The table also shows that the relation-aware lists match a qrels identifier at 57.87% of complement ranks and 79.36% of substitute ranks, underscoring the effect of pooled judgments.

Table 9. Confirmed supplementary image-URL coverage.

Scope	Unique products	Confirmed URL IDs	Confirmed rate
Corpus	48,616	1,047	2.15%
Query seeds	50	1	2.00%
All qrels candidates	7,103	171	2.41%
All positive relation candidates (including seeds)	1,304	20	1.53%
All positive relation candidates (self-excluded)	1,258	19	1.51%
Positive complement candidates	425	6	1.41%
Positive substitute candidates	841	13	1.55%

Note. These rates are lower bounds based on exact matches to the supplementary URL file. Absence from the file is not evidence that a product lacks an image.

Image Availability

Confirmed supplementary image URLs are sparse, as shown in Table 9. The join matches 1,047 of 48,616 corpus products (2.15%), 171 of 7,103 assessed candidates (2.41%), six of 425 unique positive complement candidates (1.41%), and 13 of 841 unique positive substitute candidates (1.55%). These are confirmed-match rates, not estimates of the true prevalence of product imagery. Figure 7 places the coverage audit beside the image-constrained stress test. The unconstrained relation-aware lists have a mean confirmed-URL rate at 10 of 0.96%, whereas the constrained lists reach 100%. That display guarantee reduces average nDCG@10 from 0.1840 to 0.0138. A hard image filter is therefore not an acceptable proxy for relevance when URL coverage is this limited; a fallback layout preserves stronger candidates.

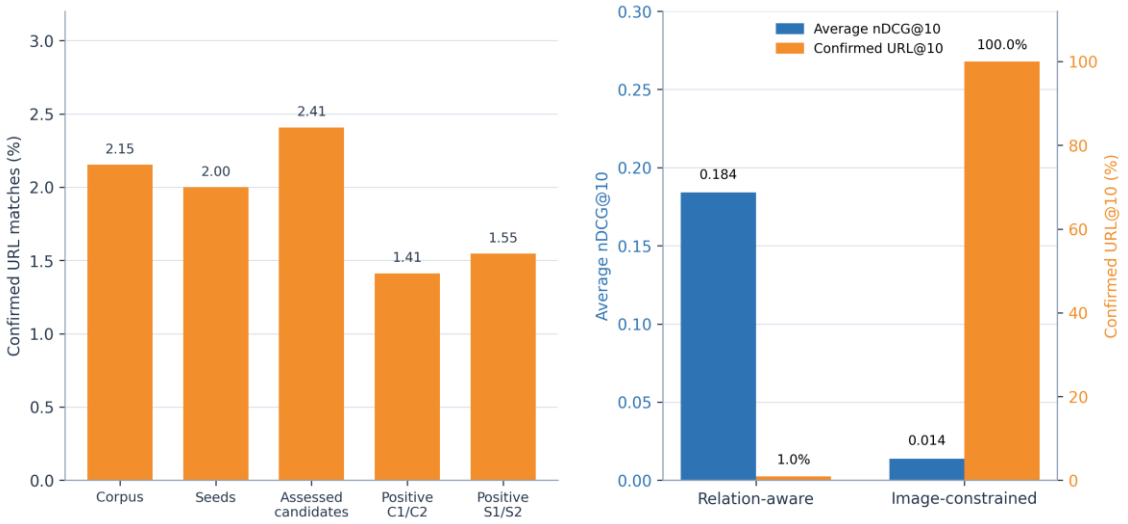

Figure 7. Confirmed URL coverage and the relevance cost of a hard image constraint.

DISCUSSION

Relation-Specific Interpretation

The relation-specific gap is the central finding. Substitute products often share terms that identify function, category, model family, or brand. Complements may instead be linked through use, attachment, power, fit, or compatibility, and those relations are not reliably recoverable from sparse product text. The small differences among the lexical methods do not change that conclusion. The cards can organize and qualify complementary suggestions, but the observed complement scores do not support automatic basket completion or compatibility assurance. Complement cards should consequently use language such as "may pair with" or "related accessory" when the evidence is broad, reserve "compatible with" for a verified compatibility field, and retain visible alternatives. Substitute cards can use more direct language such as "alternative" when title and full-text evidence agree. This asymmetry is not merely stylistic: it translates the measured difference in retrieval reliability into a difference in interface certainty (Jin, 2025a).

Table 10. Evidence-to-card field specification.

Card field	Evidence source	Display rule
Image region	Confirmed URL state	Use an image when confirmed; otherwise, keep a stable text-led region.
Relation label	Complement or substitute list	Use "May pair with" for complements and "Alternative" for substitutes.
Evidence phrase	Title, description, brand, or bullet points	Show one concise phrase traceable to a populated source field.
Qualification	Relation reliability and compatibility evidence	Avoid compatibility assurance unless a dedicated field verifies it.
Alternative control	Ranked candidates	Keep alternatives visible for complementary suggestions.
Primary action	Item destination	Use a neutral action such as "View item"; do not create urgency.
Renderer boundary	Fixed schema values	Templates or constrained generators may vary wording but not facts.

Note. The schema specifies evidence and hierarchy. It does not imply a measured change in shopper comprehension or behavior.

Image State and Card Hierarchy

Image presence should also remain separate from relation evidence. The SQID supplement verifies a URL for only a small portion of this corpus and cannot establish that the unmatched products lack images. The large effectiveness loss under the hard image constraint shows why image readiness belongs in the rendering layer (Wang et al., 2025). A card with a confirmed image URL can use the image as its primary visual anchor; an unmatched item needs a stable text-led fallback with the same relation, evidence, and action structure. Table 10 translates these results into the display contract, and Figure 8 shows the corresponding anatomy. The order is deliberately

stable: image state first, relation second, one evidence phrase third, and a neutral action last. Limiting the number of competing fields follows visibility, signifier, information-density, and cognitive-load principles (Nielsen, 1994; Norman, 2013; Sweller, 1988; Tufte, 2001). It remains a design implication rather than a measured user preference.

Figure 8. Evidence-grounded recommendation-card anatomy.

Renderer Boundary and Limitations

The schema separates evidence selection from verbalization (Xin, 2026a). A deterministic template can fill every field. A constrained language model could vary surface wording, but its quality, factuality, and comparative value are not evaluated by the present ranking experiment (Su, Chen, & Qian, 2026; Zhong et al., 2026). The empirical contribution lies in the relation scores, the image-state audit, and the field restrictions, not in a claim that one text-generation method is superior.

Several limitations guide future work. The qrels arise from pooled submissions, so an unjudged retrieval receives zero gain even when it may be useful. The corpus contains uneven descriptions and no verified compatibility graph. The relation-aware weights are fixed rather than learned from independent training data. The image analysis confirms URL membership only and does not inspect pixels, image quality, or live availability. Finally, offline ranking and schema conformance do not establish shopper comprehension, confidence, click-through, or purchase behavior (Bai & Wu, 2026). Those outcomes require a controlled card study or online experiment (Bai et al., 2026; Lu et al., 2025; Tintarev & Masthoff, 2015; Y. Zhang & Chen, 2020).

CONCLUSION

This study presents an evidence-grounded visual schema for complementary and substitute recommendation cards. Evaluation on the final TREC 2025 topics and NIST judgments shows that substitute recommendation is considerably stronger than complementary recommendation, while small aggregate differences among the unconstrained lexical methods are not statistically reliable. The practical contribution is therefore a cautious display contract rather than a claim of automatic basket completion: use relation-specific labels, tie each reason to available product fields, preserve alternatives for complements, and avoid compatibility language that the evidence cannot support. The supplementary SQID analysis also shows that a hard image requirement can sacrifice substantial relevance when confirmed URL coverage is sparse. Image presence is better handled as a layout state with a text-led fallback. The same field contract can be verbalized by templates or a constrained generator, but comparative text quality and shopper response remain separate empirical questions. Together, these boundaries produce a recommendation card whose visual certainty matches the strength of its underlying evidence.

REFERENCES

Al Ghossein, M., Chen, C.-W., & Tang, J. (2024). Shopping Queries Image Dataset (SQID): An image-enriched ESCI dataset for exploring multimodal learning in product search [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2405.15190

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

Bai, J., & Wu, Q. (2026). Privacy-safe marketing mix modeling and budget optimization under identifier loss: A controlled simulation study. International Journal of Electronics and Communications Systems, 6(1), 83–95. https://doi.org/10.24042/ijecs.v6i1.30533

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction, 12(4), 331–370. https://doi.org/10.1023/A:1021240730564

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

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

Crossing Minds. (2026). Shopping Queries Image Dataset (SQID) [Data set]. Hugging Face. Retrieved July 23, 2026, from https://huggingface.co/datasets/crossingminds/shopping-queries-image-dataset

Go, E. M., & Mothelsang, K. (2024). Trends in Modern Typography Design: Visual Preferences on E-Commerce Platforms. International Journal of Graphic Design, 2(2), 248–263. https://doi.org/10.51903/IJGD.V2I2.2147

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

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

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

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

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

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

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in neural information processing systems (Vol. 30, pp. 4765–4774). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

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

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

National Institute of Standards and Technology. (2026, March 17). 2025 Product Search Track. https://trec.nist.gov/data/product2025.html

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94) (pp. 152–158). Association for Computing Machinery. https://doi.org/10.1145/191666.191729

Norman, D. A. (2013). The design of everyday things (Rev. & expanded ed.). Basic Books.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(85), 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

Reddy, C. K., Màrquez, L., Valero, F., Rao, N., Zaragoza, H., Bandyopadhyay, S., Biswas, A., Xing, A., & Subbian, K. (2022). Shopping Queries Dataset: A large-scale ESCI benchmark for improving product search [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2206.06588

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

Ricci, F., Rokach, L., & Shapira, B. (Eds.). (2015). Recommender systems handbook (2nd ed.). Springer. https://doi.org/10.1007/978-1-4899-7637-6

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

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

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

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

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Tintarev, N., & Masthoff, J. (2015). Explaining recommendations: Design and evaluation. In F. Ricci, L. Rokach, & B. Shapira (Eds.), Recommender systems handbook (2nd ed., pp. 353–382). Springer. https://doi.org/10.1007/978-1-4899-7637-6_10

TREC Product Search Track. (2025). Product recommendation 2025 [Data set]. Hugging Face. https://huggingface.co/datasets/trec-product-search/product-recommendation-2025

TREC Product Search Track Organizers. (2025). Recommendation task instructions. https://trec-product-search.github.io/recommendations

Tufte, E. R. (2001). The visual display of quantitative information (2nd ed.). Graphics Press.

Wang, H., Ren, Y., & Chang, X. (2025). Layout-aware progressive PDF rendering: AI prioritization of PDF slices to reduce time-to-functional-first-frame on FUNSD. Journal of Technology Informatics and Engineering, 4(2), 425–446. https://doi.org/10.51903/jtie.v4i2.523

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

Xin, Q. (2026b). Self-supervised customer representation learning for segmentation and next-purchase prediction on UCI Online Retail. Journal of Information and Technology, 14(1), 20–37. https://doi.org/10.32664/j-intech.v14i01.2229

Xu, H., Chen, Y., & Med, A. (2025). Automatic detection and explanation of dark patterns from interface microcopy: Empirical comparison of BERT-style encoders, RoBERTa-style encoders, and LLM-style decoders on the ec-darkpattern dataset. Journal of Technology Informatics and Engineering, 4(3), 590–612. https://doi.org/10.51903/jtie.v4i3.491

Xu, K., Zhou, H., Zheng, H., Zhu, M., & Xin, Q. (2024). Intelligent classification and personalized recommendation of e-commerce products based on machine learning. Applied and Computational Engineering, 64(1), 147–153. https://doi.org/10.54254/2755-2721/64/20241365

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

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

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

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

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

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

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

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