Do LLM-Built Product Graphs Improve Ecommerce Search?

A Controlled Study of Recall, Ranking, and Candidate Shortlists

Vin Lim
Independent engineering research

30 July 2026

Abstract

New products have titles, descriptions, and attributes before they have clicks, purchases, or relevance labels. We test whether an LLM can turn those catalog fields into a product relationship graph that is useful at this cold-start stage. On a fixed corpus of 58,138 ESCI products, the catalog-text graph connects 53,082 products. This includes 16,095 of the 17,876 products with no allowable training annotation, or 90.04%.

Coverage is not the same as quality, so we also measure retrieval. A GRIT-weighted annotation graph raises recall@100 from 0.7810 to 0.7883 with a validation-selected fixed-size replacement rule. The catalog-text graph reaches 0.7978 and remains higher at 500 and 1,000 candidates. Under a shared expansion and RRF operator, it also gives the highest recall, but nDCG@10 and MRR fall. The annotation graph suffers a much larger ranking loss. Graph expansion is therefore better treated as candidate generation than final ranking.

Finally, we test five ways to choose the ten products presented to the LLM. Held-out validation selects BM25 plus dense RRF. In an 8,000-anchor matched pilot, 28.33% of its proposed pairs differ from the original BM25-first shortlist, yet recall@100 improves by only 0.000981, below the predeclared 0.005 practical threshold. The graph can seed alternatives and other related-product modules before behavior exists, but this study does not measure recommendation precision, click-through rate, conversion, or typed-edge accuracy.

Keywords: product search; cold start; product knowledge graph; large language models; candidate generation; first-stage retrieval; ESCI; reciprocal rank fusion.

1 Introduction

Product search usually runs in two stages. A first-stage retriever selects a set of candidate products from the catalog, and a ranker puts them in order. A relevant product that the first stage misses is never ranked, so first-stage recall limits what the ranker can show. Lexical matching alone misses products that describe the same need in different words, so product search engines add learned semantic matching [18].

A product graph is another source of candidates. It links each product to related products, such as substitutes and complements. When a query retrieves one product, the retriever can add that product's neighbors to the candidate set. GRIT [1] shows that product-product graphs can improve first-stage retrieval.

The edges in such graphs come from what shoppers and annotators have already done. Some are computed from purchase histories [14] or learned from product reviews [15]. GRIT's edges require clicks or human annotation, and ESCI [2] supplies such annotation as Exact, Substitute, Complement, and Irrelevant labels.

All of these sources need the product to have a history. A product listed today has a title, a description, and attributes, but no purchases, reviews, clicks, or labels. This is the item cold-start problem [17], and

© 2026 Vin Lim. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

it recurs whenever a catalog adds a product. In the ESCI corpus used here, 17,876 of 58,138 products (30.75%) have no allowable annotation evidence, so an annotation-derived graph cannot connect them.

Catalog text is available on the first day. LLMs are already used to build e-commerce knowledge graphs. FolkScope [21] and COSMO [22] use them to explain observed shopping behavior, which a new product does not have. Wang et al. [12] use an LLM-built product knowledge graph for explainable recommendation, and Yang et al. [13] construct a graph from raw metadata for cold-start recommendation. To our knowledge, no prior study compares a graph built from catalog text alone with a strong annotation-derived graph under the same retrieval operator and the same queries. Section 10 reviews this work in more detail.

We ask what a catalog-text graph contributes before interaction evidence exists, how its retrieval behavior compares with a strong annotation-derived graph, and what happens as annotation evidence accumulates.

We study four conditions: a strong BM25 and dense hybrid baseline; a sparse co-label graph retained as a low-strength ablation; a full-data GRIT-weighted annotation graph; and a full-catalog graph extracted from titles, descriptions, brands, colors, and bullet points. A separate 8,000-anchor pilot is designed to change the shortlist offered to the LLM. Provider route and time remain disclosed confounds.

The paper makes four contributions.

- We compare three sources of graph edges under two fixed retrieval operators on ESCI: a catalog-text graph, a GRIT-weighted annotation graph, and a sparse co-label graph. All three start from the same BM25 and dense hybrid baseline (Sections 3 and 4).
- We evaluate the annotation graph at seven fractions of annotation evidence across five nested samples, and observe no recall crossover with the catalog-text graph on that grid (Section 5).
- We measure cold-start reach. The catalog-text graph connects 16,095 of the 17,876 products that have no annotation evidence (Section 6).
- We test whether the shortlist of candidates shown to the LLM limits the graph, using a shortlist ablation and a matched 8,000-anchor pilot (Section 7).

An evidence kit maps every central quantitative claim to a versioned artifact.

Main result. The catalog-text graph reaches most of the annotation-cold catalog and produces higher recall than both the hybrid baseline and the full-data annotation graph in this ESCI experiment. It does not improve top-ranking quality under simple RRF, and the mixed BM25 plus dense shortlist does not produce a practically meaningful gain in the matched pilot.

2 Experimental design

The fixed universe contains 58,138 ESCI products. Allowable US train evidence contributes 31,437 complete query groups and 114,451 judgments over 40,262 products. During parameter selection, graph-build and validation groups are disjoint. Once the settings are fixed, the annotation graph is rebuilt on all 31,437 train groups. The reporting set contains 1,500 queries and 30,875 qrels and overlaps neither selection split.

Graph sources are swapped inside a fixed retrieval operator (Figure 1). The candidate-generation arm follows GRIT's top-seed and bottom-replacement design. The shared arm gives the annotation graph, catalog-text graph, and sparse co-label graph exactly the same one-hop expansion and RRF treatment.

2.1 Baseline retrieval

The hybrid baseline combines lexical BM25 [4], implemented with BM25S [8], with BAAI/bge-base-en-v1.5 embeddings [7] searched exactly in FAISS [9]. Reciprocal rank fusion [5] merges the two lists. Evaluation uses cumulated-gain metrics [6] through ranx [10]. Every graph condition starts from the same

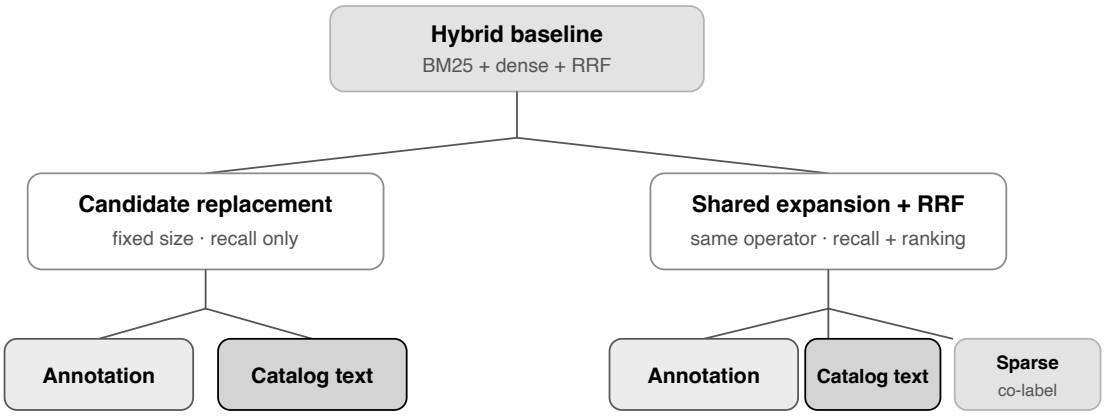


Figure 1: Two controlled integrations. Edge sources are interchangeable within each branch. The sparse co-label graph is retained only as a low-strength ablation.

output. The baseline is strong and reproducible [11]; it does not stand in for every modern embedding model.

2.2 Annotation-graph construction

For every training query, we exclude Irrelevant labels and add each pair of Exact, Substitute, and Complement products. Occurrences accumulate fixed weights: E-E=3, E-S=2, E-C=1, S-S=2, S-C=1, and C-C=1. We preserve raw support and accumulated weight instead of normalizing each sampled graph. Validation selects minimum support one, leaving 329,051 undirected edges and 35,279 covered products.

2.3 Catalog-text graph construction

DeepSeek V4 Flash builds the catalog-text graph from catalog fields only, with no reviews, behavior, or relevance labels. For each source product, BM25 and exact dense search propose candidates; the original full-catalog run keeps the first ten after BM25-first merging. The model classifies them as alternatives, accessories, compatibility, goes-with, or none at temperature zero. The graph contains 301,735 typed rows representing 222,408 undirected pairs and covers 53,082 products.

2.4 Selection and statistics

The GRIT grid has 48 cells; validation selects top 4% seeds and bottom 20% replacement. A deterministic positive control uses the published 2% and 30% settings. It checks direction against the paper, but it is not a bit-for-bit run of the official repository. The shared grid has nine source-neutral cells and selects 50 seeds, fanout 25, and one hop. Effects use query-level 5,000-sample bootstrap intervals, 10,000 paired randomizations, and Holm correction over 18 candidate and 16 shared comparisons.

The 1,500-query set serves only for reporting. Its judgments were available before this analysis, so all graph and retrieval selection is confined to disjoint train groups.

2.5 Candidate shortlist ablation

The LLM can only judge pairs that the blocker puts in front of it. We test five shortlist rules on 6,287 held-out train query groups: BM25-first, BM25-only, dense-only, alternating BM25 and dense results, and RRF. BM25-first fills the cap from its lexical list and admits unseen dense products only if fewer than ten BM25 candidates are available. Of the held-out groups, 3,172 contain at least two positive products and enter the query-macro metric. The metric asks whether a ten-product shortlist recovers products

Table 1: Validation-selected fixed-size candidate replacement on the frozen ESCI test.

Condition	recall@100	recall@500	recall@1000
Hybrid baseline	0.7810	0.8936	0.9167
Annotation graph	0.7883	0.9008	0.9238
Catalog-text graph	0.7978	0.9067	0.9287

Table 2: Identical one-hop expansion and RRF on ESCI. Values are not directly comparable to Table 1 because the integration is different.

Condition	nDCG@10	recall@100	recall@1000	MRR
Hybrid baseline	**0.5771**	0.7810	0.9167	**0.8063**
Sparse co-label graph	0.3359	0.7451	0.9163	0.4472
Annotation graph	0.4013	0.7740	0.9241	0.5755
Catalog-text graph	0.5676	0.8100	0.9290	0.7635

judged Exact, Substitute, or Complement for the same query. It is a proxy for the blocker and does not measure typed relationship quality directly.

At a cap of ten, the alternating rule behaves like five BM25 and five dense results when the lists do not overlap. Shared products count once and the next items fill the open slots, so it is not a hard five-per-channel quota. RRF scores slightly higher. We therefore fuse the top 50 BM25 hits and top 50 exact dense neighbors with RRF at $k = 60$ and keep the ten highest-scoring unique products for the end-to-end test.

The mixed-shortlist pilot selects 8,000 source products by a label-independent hash. It keeps the same named model, prompt, schema, catalog fields, and evidence rules, but uses OpenRouter only; the full graph was assembled from OpenRouter and DeepSeek-native shards. The matched comparison includes 7,998 successful anchors; two anchors ended in provider errors. Before reserve labels are opened, the graph, retrieval runs, manifests, code inventory, and evaluation plan are hashed and locked. Evaluation uses a separate reserve of 8,738 query groups; 7,822 have at least one positive in-corpus judgment and enter the metric denominator. The 1,500-query reporting set is not opened for this pilot.

3 Candidate recall

Table 1 gives the candidate-replacement results. The annotation graph improves over the hybrid baseline by +0.00722 at recall@100, 95% CI [0.00403, 0.01056]; +0.00723 at recall@500 [0.00459, 0.01019]; and +0.00714 at recall@1000 [0.00462, 0.00985]. These prespecified intervals exclude zero.

The catalog-text graph is higher than the annotation graph by +0.00949 at recall@100 [0.00537, 0.01339], +0.00587 at recall@500 [0.00254, 0.00939], and +0.00484 at recall@1000 [0.00160, 0.00804] with the validation-selected replacement rule. With the published 2% seed and 30% replacement settings, the catalog-text advantage remains clear at recall@100, while its intervals against the annotation graph cross zero at 500 and 1,000. The size of the advantage is therefore parameter-sensitive at deeper cutoffs.

4 Shared RRF: deep recall versus ranking

Table 2 gives the shared-operator results. The annotation graph raises recall@1000 by +0.00743 [0.00492, 0.01014] over the hybrid baseline, but lowers recall@100 by −0.00691 [−0.01182, −0.00165], nDCG@10 by −0.17581, and MRR by −0.23088. The catalog-text graph raises recall@100 by +0.02903 [0.02409, 0.03395] and recall@1000 by +0.01234 [0.00976, 0.01523], while lowering nDCG@10 by 0.00954 and MRR by 0.04286.

Both full graphs activate on all 1,500 reporting queries. The annotation graph contributes 1,232 relevant novel top-100 candidates across 471 queries; the catalog-text graph contributes 2,092 across

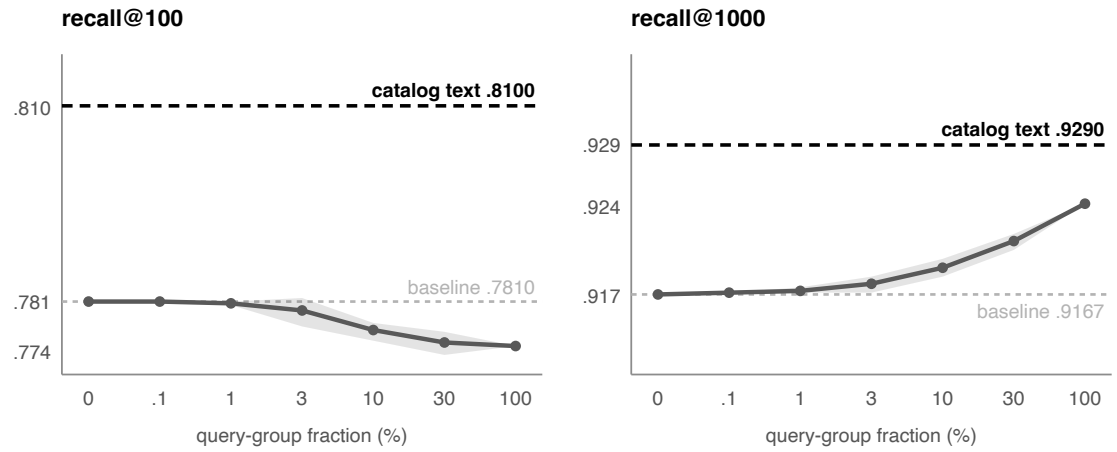


Figure 2: No observed recall crossover on the sampled annotation-evidence grid. The solid line with markers is the annotation-graph mean; the dotted line is the hybrid baseline and the dashed line the catalog-text graph. Shading is the 95% t-interval across five nested query-group samples. Annotation-graph deep recall rises from 0.9167 to 0.9241, but the catalog-text graph remains at 0.9290; top-100 recall moves in the opposite direction.

715. The annotation graph adds more candidates overall, which is consistent with RRF placing too much of that expansion near the top. The loss comes from the integration; it does not show that annotation relationships are poor.

5 Annotation-evidence crossover

We sample complete query groups in five nested chains, keep absolute edge weights and admission rules fixed, and evaluate 0%, 0.1%, 1%, 3%, 10%, 30%, and 100% of annotation evidence. Every 0% ranking equals the hybrid baseline, and every 100% ranking equals the full annotation-graph run through rank 1,000. Figure 2 plots the two recall cutoffs.

By 30% evidence, annotation-graph recall@1000 reaches 0.92105 [0.92045, 0.92165], while recall@100 falls to 0.77456 [0.77301, 0.77611]. Full evidence activates every reporting query and covers 35,279 products, yet the catalog-text graph remains higher on both recall cutoffs. The curve depends on this integration. It does not represent production traffic and should not be extrapolated beyond the sampled grid.

6 Cold-start reach and relevance labels

The structural result does not depend on a claim that graph fusion must improve ranking. Of 58,138 products, 17,876 have no allowable annotation evidence. The catalog-text graph connects 16,095 of them. An annotation-derived graph cannot connect those products until evidence arrives.

6.1 Label strata

The catalog-text graph's largest reliable recall@100 effect is on Substitutes: +0.04050 [0.03252, 0.04916] over the hybrid baseline. Exact also improves. The Complement interval crosses zero, so we do not claim a reliable complement lift. The annotation graph improves deep recall in every label stratum, but not top-100 recall (Figure 3).

6.2 Different edges and the annotation-cold catalog

The annotation graph contains 329,051 undirected pairs and the catalog-text graph contains 222,408. They share 77,242 pairs: Jaccard 0.1629. Only 23.47% of annotation pairs appear in the catalog-text graph, and

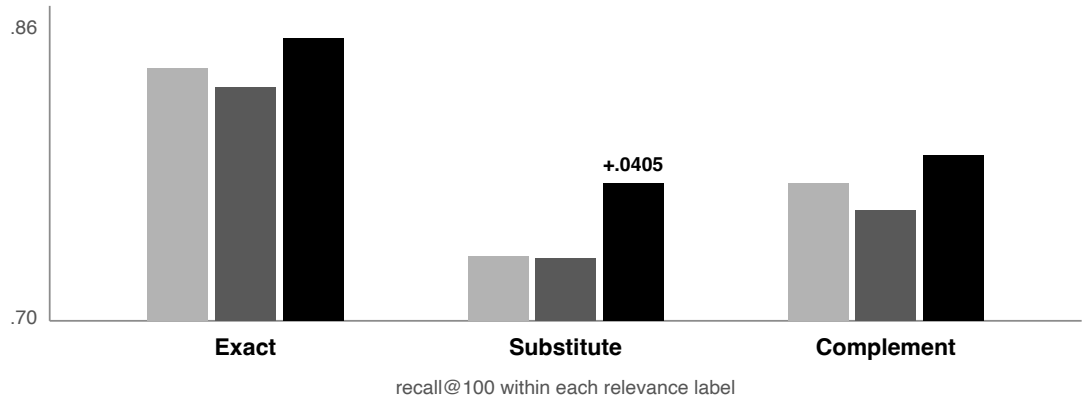


Figure 3: Shared-operator recall@100 by ESCI label. Bars within each group, from left: hybrid baseline (light grey), annotation graph (dark grey), catalog-text graph (black). The catalog-text graph's largest reliable gain is on Substitute. The annotation graph improves the deeper cutoff, not the recall@100 shown here.

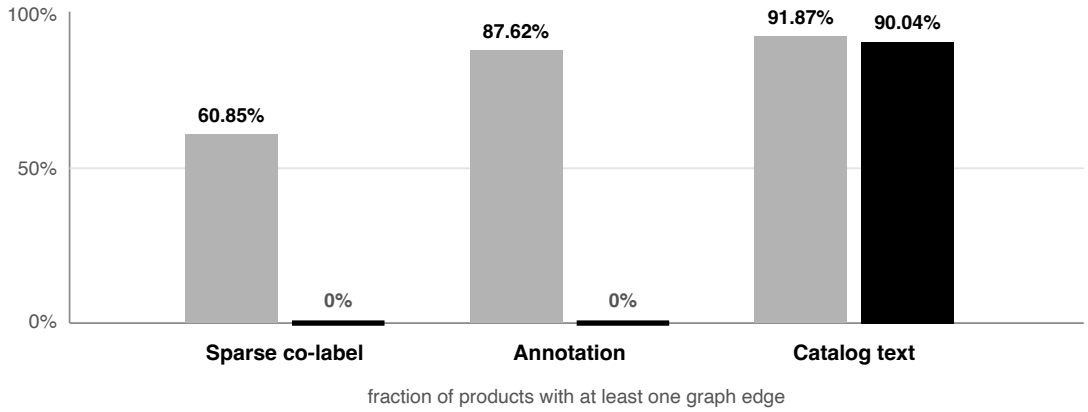


Figure 4: Warm and cold graph coverage under full allowable ESCI supervision. Grey bars are warm coverage and black bars cold coverage. The catalog-text graph reaches 16,095 of 17,876 cold products; annotation-derived graphs reach none until those products receive evidence.

34.73% of catalog-text pairs appear in the annotation graph. Most catalog-text pairs have no counterpart in the annotation graph.

Under the full-data definition, 40,262 products are warm and 17,876 (30.75%) remain cold. The annotation graph covers 87.62% of warm products and no cold products. The catalog-text graph covers 91.87% of warm products and 90.04% of cold products (Figure 4). Coverage shows which products the graph can reach; it does not show that every edge improves retrieval or represents a correct typed relationship.

6.3 Reuse for related-product modules

The graph contains 301,735 directed typed rows: 284,183 alternative-to, 2,210 accessory-of, 3,511 compatible-with, and 11,831 goes-with. Those rows can provide day-one candidates for alternative and related-product modules, including discontinued-item replacements and product-detail-page recommendations. This reuse of the same offline catalog asset was not measured here.

This study does not evaluate typed-edge precision, recommendation relevance, diversity, click-through rate, or conversion. The heavy concentration in alternative-to edges (94.18%) also means the evidence is much stronger for alternative discovery than for accessories, compatibility, or complements. Any recommendation launch needs a separate human-judged evaluation and an online experiment.

Table 3: Held-out candidate-blocker ablation. Higher co-relevance recall@10 is better.

Shortlist rule	Co-relevance recall@10
BM25-first	0.3898
BM25-only	0.3896
Dense-only	0.4146
Alternating BM25 and dense	0.4136
BM25+dense RRF	**0.4181**

Table 4: Mixed shortlist versus the original BM25-first shortlist on partial matched graphs and the sealed secondary reserve. Intervals are paired 95% bootstrap intervals.

Operator and metric	Original shortlist	Mixed shortlist	Paired effect
Replacement recall@100	0.755918	0.756899	**+0.000981** [+0.000035, +0.001960]
Replacement recall@500	0.882407	0.882549	+0.000142 [−0.000462, +0.000776]
Replacement recall@1000	0.914259	0.914591	+0.000331 [−0.000226, +0.000932]
Shared nDCG@10	0.216197	0.218554	+0.002357 [−0.001023, +0.005658]
Shared recall@100	0.750383	0.752402	+0.002019 [−0.000161, +0.004212]
Shared recall@1000	0.914472	0.914946	+0.000474 [−0.000128, +0.001117]
Shared MRR	0.244653	0.245015	+0.000361 [−0.003803, +0.004572]

7 Mixed-shortlist pilot

The blocker ablation asks whether representing both BM25 and dense candidates improves the LLM graph downstream. Dense-only and alternating shortlists beat BM25-first on held-out blocker recall. RRF scores highest (Table 3).

RRF is therefore the mixed method evaluated end to end. The mixed-shortlist graph and original-shortlist matched graph use the same 7,998 source anchors and a separate query reserve. Their absolute scores are not comparable with Tables 1 and 2.

Table 4 gives the matched pilot. The primary recall@100 effect is small and positive: +0.000981, or 0.10 percentage points, with all four fold estimates above zero. Its 95% interval is [0.000035, 0.001960]. Even the top of that interval is less than half the predeclared +0.005 practical threshold. Recall@500 and recall@1000 intervals cross zero. Shared-RRF nDCG passes the one-sided noninferiority check, but its two-sided interval crosses zero, as do the other shared metrics.

The two shortlists differ. Of 79,980 proposed pairs, 22,659 (28.33%) appear only in the mixed shortlist, and 10,002 of those become graph edges. The mixed-shortlist graph has 45,917 edges versus 41,626 for the matched original-shortlist graph; their edge Jaccard is 0.5398. More varied candidates and more edges still do not produce a practically meaningful recall gain.

Recorded spending was $49.05 for the WANDS full-catalog extraction and $52.04 for ESCI. The later pilot campaign cost $10.878571 including the interrupted attempt, for $111.97 across the two full extractions and pilot campaign. Preliminary bake-offs and calibration bring the listed total to approximately $112.10, but those earlier full-run figures are recorded only to cents. Within the completed pilot attempt, the reconciled account delta was $8.923993. Completing the remaining ESCI catalog was projected at $56.02, or $64.43 with the planned buffer; the measured gain did not support the additional $62.20 credit requirement.

What the blocker test shows. At this scale, BM25 crowding is a less convincing explanation for the full-data result. The mixed shortlist produces a material change in graph inputs, yet downstream recall changes by less than the practical threshold. It is robustness evidence and does not replace a second full-data comparison of the annotation and catalog-text graphs.

7.1 Scaling the candidate step

The LLM does not evaluate every possible product pair. Retrieval limits each source product to ten proposals, reducing 1,689,984,453 possible undirected pairs in this corpus to at most 581,380 directed judgments. With a fixed shortlist size $k = 10$, LLM pair evaluation is $O(Nk)$ rather than $O(N^2)$. It still requires one offline prompt per source product, and the shortlist retrieval itself must scale.

The research runs use exact dense search to remove approximation noise. In a separate label-free benchmark over 58,138 normalized 768-dimensional product vectors and 1,000 fixed anchors, FAISS HNSW with *efSearch*=64 reaches 0.9992 neighbor recall@10, 0.99805 at 20, and 0.99156 at 50. Single-anchor median latency is 0.198 ms versus 2.447 ms for exact search. For a batch of 1,000 anchors, exact search is faster per anchor: 0.0593 ms versus 0.1833 ms. HNSW takes 28.44 seconds to build and uses 194.4 MB, compared with 178.6 MB for the exact index.

These measurements show a viable approximate-neighbor option at this catalog size. They are not evidence for million-product production scale. Larger deployments still need sharding, attribute or category filtering, incremental updates, capacity tests, and production latency measurements.

8 Interpretation

The full-data annotation graph improves candidate recall at every measured depth. The catalog-text result is therefore compared with a positive supervised control, not the sparse graph alone.

The catalog-text graph produces higher recall under both controlled operators in this ESCI experiment, with its largest reliable label-level effect on Substitute products. Its more distinctive contribution is structural: it reaches 90.04% of products that have no annotation evidence and can provide cold-start candidates before behavior accumulates.

The shortlist experiment tests whether BM25-first was suppressing useful dense proposals. Dense-only and alternating BM25+dense shortlists beat BM25-first on the held-out blocker metric, and RRF performs best. Once that RRF shortlist passes through LLM judgment, graph construction, and retrieval, however, the gain is only +0.000981 at recall@100, below the +0.005 practical threshold.

Recall and ranking remain different jobs. Both full graphs can bring relevant products into a deeper candidate pool, but shared RRF does not reliably put them in the best order. The catalog-text graph has a modest ranking loss and the annotation graph a large one. A production system should treat graph expansion as candidate generation and learn or otherwise validate the final ranking separately.

9 Limitations

- **Annotations stand in for behavior.** ESCI has no clicks or purchases. A production interaction graph may have different density, coverage, and bias.
- **The GRIT implementation is deterministic, not official.** It follows the published weighting and retrieval design but is not a bit-for-bit run of the authors' repository. The 1,500-query reporting set had been inspected before this analysis; all method selection was therefore kept on disjoint train and validation groups and runs were locked before evaluation.
- **The main comparison is ESCI-only.** WANDS covers the sparse co-label and catalog-text graphs as secondary robustness evidence. The full-data annotation graph was not run there.
- **The crossover has a boundary.** It covers seven annotation fractions, five nested samples, this corpus, and this operator. It does not predict every future evidence level.
- **The mixed-blocker result is partial.** Eight thousand anchors and a separate reserve cannot establish full-catalog recall, ranking quality, or cold coverage. The reserve was opened once after run lock; 916 of its 8,738 groups lacked a positive in-corpus judgment and were excluded from metrics.

- **Provider route and drift may affect the matched pilot.** The original full graph was merged from OpenRouter and DeepSeek-native shards. The mixed-shortlist pilot used OpenRouter only. Although the named model and prompt were held fixed, provider behavior or model drift may contribute to the difference.
- **Coverage is not recommendation quality.** A connected cold product has at least one retained edge; this does not establish that its edge type is correct or that shoppers will find the recommendation useful. Most typed rows are alternatives, and complement evidence is limited.
- **The dense baseline is fixed, not timeless.** BAAI/bge-base-en-v1.5 supports reproducibility, but newer embedding models may change both the hybrid baseline and the candidates offered to the LLM.
- **The scaling benchmark is bounded.** Exact and HNSW search were measured on 58,138 vectors on one machine. The result does not establish latency, cost, or quality at millions of products.

10 Related work and robustness boundary

GRIT [1] is the closest supervised graph-retrieval design. It takes its edges from clicks or human annotation. It expands the top results of a first-stage list through the graph and puts the neighbors in place of the bottom results. Our annotation graph uses its weighting, and our candidate-replacement operator follows its design. ESCI [2] provides explicit Exact, Substitute, Complement, and Irrelevant labels.

Relations derived from behavior have a longer history. Item-to-item collaborative filtering [14] relates products through the purchase histories of the customers who bought them. McAuley et al. [15] predict which products are substitutes or complements, learning mainly from the text of product reviews. P-Companion [16] shows that products bought together are not always complements, and builds cleaner training labels for complement recommendation. All of these methods need interactions, reviews, or labels, which a new product does not yet have [17].

Product knowledge graphs describe products and shopping needs. AutoKnow [19] collects product types and attribute values, and AliCoCo [20] represents user needs as concept nodes. FolkScope [21] and COSMO [22] use LLMs to generate the intentions behind shopping behavior, and both filter the output with human-annotated samples. Wang et al. [12] use an LLM-built product knowledge graph for explainable recommendation and report an online experiment. Yang et al. [13] dynamically construct a graph from raw metadata for cold-start recommendation. None of these papers compares a catalog-text graph with an annotation-derived graph inside one retrieval operator.

WANDS [3] is a secondary robustness check for the sparse co-label and catalog-text graphs. The catalog-text graph is not distinguishable from its strong hybrid on the tested primary metrics, recall@100 and nDCG@10. The sparse co-label graph significantly lowers recall@100. WANDS has no explicit Complement label and provides no evidence about the full-data annotation graph outside ESCI.

11 Conclusion

The catalog-text graph is useful before interaction evidence exists. It connects 16,095 of 17,876 annotation-cold products and can supply candidates for search, alternatives, and related-product modules from catalog text alone. That is a structural capability. Recommendation quality and business impact still need separate evaluation.

In search, the catalog-text graph also produces higher recall than both the hybrid baseline and the full-data annotation graph under the validation-selected replacement rule. The advantage at deeper cutoffs is sensitive to replacement parameters, and simple RRF lowers top-ranking quality. The evidence supports graph-assisted candidate generation, not a claim that the graph should rank the final page.

We tested five proposal rules, including a nominal five-from-each shortlist. RRF scored highest on held-out blocker recall and was used in the matched end-to-end pilot. In that shortlist, 28.33% of pairs are absent from the matched BM25-first shortlist, yet recall@100 improved by only +0.000981, below the

+0.005 practical threshold. BM25 crowding is therefore a less convincing explanation for the full-data result, though the partial pilot is not a second full-catalog comparison.

The practical next step is to keep offline relationship extraction bounded by retrieval, use the graph as a reusable candidate source, and evaluate final ranking and recommendation surfaces independently.

12 Data, materials, and declarations

Data availability. ESCI and WANDS are third-party public datasets governed by their own licenses. Raw data, reporting queries, relevance judgments, API responses, credentials, and account ledgers are not redistributed in the evidence kit. The kit includes dataset provenance, locked configuration records, aggregate reports, graph outputs, hashes for the sealed reporting artifacts, and an inventory of excluded material.

Artifact availability. Evidence kit version 1.0 accompanies this paper and is published with its HTML version at https://astralab.co/insights/llm-product-graphs-ecommerce-search. It contains the first published version of the paper in HTML and PDF, claims-to-evidence mappings, aggregate machine-readable results, graph files, manifests, the prompt and model configuration, selected study code, environment locks, cost records, and archive-wide SHA-256 checksums. Historical provenance files are preserved unchanged, including their internal source paths.

Ethics and data use. The study uses public benchmark catalog data and relevance annotations. It recruited no participants, ran no live user experiment, and intentionally collected no personal data.

Funding and competing interests. The study was conducted as independent engineering research without external grant funding. API costs are disclosed in Section 7. The research predates any product branding. The author may have a commercial interest in future products informed by this work; no commercial product was evaluated.

Author contributions. Vin Lim conceived the study, directed the experiments, reviewed the evidence, and takes responsibility for the paper and release.

Tool assistance. Claude (Anthropic) assisted with software development and with drafting and revising the text. The author reviewed the methods, reruns, artifacts, displayed claims, and prose; tool output is not treated as evidence.

Evidence note. Full-data values come from the locked ESCI final report and five-seed annotation-evidence crossover. The shortlist and pilot values come from the locked selection record, evaluation report, graph statistics, and cost settlement. Figures 1 to 4 use full-data values; Table 4 is the matched partial-catalog pilot. WANDS is secondary robustness evidence because the full-data annotation graph was not run there. Every central quantitative claim is mapped to a versioned artifact in the evidence kit. The primary corpus and relevance labels are the Shopping Queries Dataset (ESCI), https://github.com/amazon-science/esci-data; the secondary robustness corpus is WANDS, https://github.com/wayfair/WANDS.

References

[1] Kulkarni, Kallumadi, MacAvaney, Goharian, and Frieder. GRIT: Graph-based Recall Improvement for Task-oriented E-commerce Queries. Companion Proceedings of the ACM Web Conference, 2025. https://doi.org/10.1145/3701716.3717859

[2] Reddy et al. Shopping Queries Dataset: A Large-Scale ESCI Benchmark for Improving Product Search. arXiv:2206.06588, 2022. https://arxiv.org/abs/2206.06588

[3] Chen, Liu, Liu, Sun, Baltrunas, and Schroeder. WANDS: Dataset for Product Search Relevance Assessment. ECIR, 2022, pp. 128-141. https://doi.org/10.1007/978-3-030-99736-6_9

[4] Robertson and Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval, 2009. https://doi.org/10.1561/1500000019

[5] Cormack, Clarke, and Buettcher. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. SIGIR, 2009. https://doi.org/10.1145/1571941.1572114

[6] Järvelin and Kekäläinen. Cumulated Gain-Based Evaluation of IR Techniques. ACM Transactions on Information Systems, 2002. https://doi.org/10.1145/582415.582418

[7] Xiao et al. C-Pack: Packed Resources for General Chinese Embeddings. arXiv:2309.07597, 2023. https://arxiv.org/abs/2309.07597

[8] Lu. BM25S: Orders of Magnitude Faster Lexical Search via Eager Sparse Scoring. arXiv:2407.03618, 2024. https://arxiv.org/abs/2407.03618

[9] Douze et al. The Faiss Library. arXiv:2401.08281, 2024. https://arxiv.org/abs/2401.08281

[10] Bassani. ranx: A Blazing-Fast Python Library for Ranking Evaluation and Comparison. ECIR, 2022. https://doi.org/10.1007/978-3-030-99739-7_30

[11] Lin. The Neural Hype and Comparisons Against Weak Baselines. ACM SIGIR Forum, 2019. https://doi.org/10.1145/3373767.3373770

[12] Wang et al. Enabling Explainable Recommendation in E-commerce with LLM-powered Product Knowledge Graph. OpenKG at IJCAI, 2024. https://arxiv.org/abs/2412.01837

[13] Yang et al. Adaptive Candidate Retrieval with Dynamic Knowledge Graph Construction for Cold-Start Recommendation. arXiv:2505.20773v3, 2025. https://arxiv.org/abs/2505.20773

[14] Linden, Smith, and York. Amazon.com Recommendations: Item-to-Item Collaborative Filtering. IEEE Internet Computing, 2003. https://doi.org/10.1109/MIC.2003.1167344

[15] McAuley, Pandey, and Leskovec. Inferring Networks of Substitutable and Complementary Products. KDD, 2015. https://doi.org/10.1145/2783258.2783381

[16] Hao et al. P-Companion: A Principled Framework for Diversified Complementary Product Recommendation. CIKM, 2020. https://doi.org/10.1145/3340531.3412732

[17] Schein, Popescul, Ungar, and Pennock. Methods and Metrics for Cold-Start Recommendations. SIGIR, 2002. https://doi.org/10.1145/564376.564421

[18] Nigam et al. Semantic Product Search. KDD, 2019. https://doi.org/10.1145/3292500.3330759

[19] Dong et al. AutoKnow: Self-Driving Knowledge Collection for Products of Thousands of Types. KDD, 2020. https://doi.org/10.1145/3394486.3403323

[20] Luo et al. AliCoCo: Alibaba E-commerce Cognitive Concept Net. SIGMOD, 2020. https://doi.org/10.1145/3318464.3386132

[21] Yu et al. FolkScope: Intention Knowledge Graph Construction for E-commerce Commonsense Discovery. Findings of ACL, 2023. https://doi.org/10.18653/v1/2023.findings-acl.76

[22] Yu et al. COSMO: A Large-Scale E-commerce Common Sense Knowledge Generation and Serving System at Amazon. Companion of SIGMOD, 2024. https://doi.org/10.1145/3626246.3653398