How Query Violations Relate to Relevance in Product Search

Felipe S. F. Paula[1], **Daniel Matheus Kuhn**[2], **Viviane P. Moreira**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

[2]VTEX, Brazil

{fsfpaula,viviane}@inf.ufrgs.br, daniel.kuhn@vtex.com

Abstract. *The concept of relevance has underpinned the research in Information Retrieval since its early days. Specifically, in eCommerce search, where queries typically involve attributes that express constraints, relevance is affected by how well those constraints are satisfied. However, the relationship between relevance and constraint violations was overlooked in existing research. In this work, we analyze 5,000 queries and 50,000 query-product pairs to investigate whether query-intent violation constitutes a ranking signal related to, but distinct from, graded relevance. Our results show that violation is related to, but not reducible to, graded relevance. We further show that simple violation-aware features substantially reduce query violations and remain competitive with, or even slightly improve upon, strong baselines on retrieval quality metrics.*

1. Introduction

Relevance is a fundamental concept in Information Retrieval (IR) that concerns how well search results match user queries. Despite being at the core of IR, it is far from being clearly defined. Early IR experimental methodology [Cleverdon 1960, Voorhees 2007] treated relevance as a binary and universal relationship between a query and a document. The idea of *graded relevance* was introduced in the late 1990s [Kando 1999] to better reflect the notion of partial relevance. Subsequently, concepts such as relatedness, usefulness, and user-dependence were also considered [Schamber et al. 1990]. Relevance is now seen as a multidimensional concept influenced by several factors, including reliability, freshness, novelty, and objectivity [Peikos and Pasi 2024].

In eCommerce search, factors such as brand, seasonality, and price also affect the notion of relevance [Tsagkias et al. 2021, Yang et al. 2021]. A key difference between product search and general web search is that many queries include attributes. For example, in the query "brown female sweater", we can clearly identify that the requested product type is a sweater, the color must be brown, and the clothing gender must be female. If the search engine returns a list full of men's clothes, it could lead to a poor user experience and hinder potential purchases. On the other hand, if the store is out of brown female sweaters, it could show other sweater colors on the search results page, since the shopper could also be interested in those colors. There is a non-trivial relationship between query intent violations, such as showing men's clothes for a query that requested women's, and product relevance, such as semantically similar products that a shopper could also be interested in. To the best of our knowledge, this relationship has not been explored yet.

Large Language Models (LLMs) have become increasingly prominent in IR research due to their strong performance in complex semantic judgment tasks, including synthetic relevance assessment [Faggioli et al. 2023, Soviero et al. 2024]. In particular, their ability to produce consistent, instruction-guided annotations makes them a practical alternative to fully manual labeling pipelines in settings where large-scale human annotation would be prohibitively expensive. This is especially important for studies that require fine-grained judgments over both queries and query-document pairs. In our work, we relied extensively on LLM-based annotation to construct the analysis dataset, covering approximately 5,000 queries and 50,000 query-product pairs. Producing annotations at this scale through manual expert effort alone would likely require thousands of hours of labor, making the study substantially more difficult to execute in practice. The use of LLMs, therefore, played an enabling methodological role, allowing us to investigate query violations and relevance at a scale that would otherwise be difficult to attain.

This paper investigates whether query-intent violation constitutes a ranking signal that is related to, but distinct from, graded relevance in product search. We further examine whether violation-aware structured features, derived from product-type matches and attribute-level constraint satisfaction, can improve ranking effectiveness relative to lexical and semantic similarity baselines, and whether these gains remain feasible under a limited attribute-extraction regime. To guide our inquiry, we define three research questions:

- **RQ1:** To what extent is query-intent violation distinct from graded relevance in product search?
- **RQ2:** Do violation-aware features based on product-type match and attribute-level constraint satisfaction improve ranking effectiveness relative to lexical and semantic similarity baselines?
- **RQ3:** Can a limited attribute extraction scheme retain enough violation information to provide useful ranking gains?

Our contributions are fourfold. First, we provide empirical evidence that query-intent violation is related to, but distinct from, graded relevance in product search, with substitute items playing a particularly important role in distinguishing between the two. Second, we show that simple violation-aware features based on product-type match and attribute-level constraint satisfaction substantially improve relevance label prediction over a majority baseline. Third, we demonstrate that these structured signals can be used for ranking, yielding large reductions in Violation@10 and, with full attribute coverage, competitive or slightly improved NDCG@10 relative to strong lexical and semantic baselines. Finally, through restricted-attribute experiments and error analysis, we show that limited attribute extraction can still provide useful gains, while also clarifying that ambiguity, label inconsistency, and incomplete product evidence remain important factors shaping model performance.

2. Background and Related Work

Human relevance judgments have long underpinned the construction of IR test collections, from early Cranfield-style evaluation to large-scale shared benchmarks such as TREC and CLEF [Cleverdon 1960, Voorhees 2007, Di Nunzio et al. 2006]. In product search, however, users may accept substitutes, complementary items, or only exact matches depending on the shopping intent and context [Su et al. 2018, Yang et al. 2021].

This has motivated the creation of eCommerce benchmarks with *graded* human judgments [Reddy et al. 2022, Chen et al. 2022]. Yet, to the best of our knowledge, public product-search benchmarks provide graded relevance labels without explicitly identifying which query constraints or structured attributes were satisfied or violated. Our work builds on this line of research by asking whether such structured mismatch signals can be recovered from graded relevance judgments and then exploited to improve ranking for constraint-sensitive queries.

Query understanding is a central component of eCommerce search, where queries are typically short, under-specified, and rich in product attributes [Sondhi et al. 2018, Luo et al. 2024]. Early work in this area focused on structuring shopping queries by identifying salient entities and latent structured intents, showing that shopping queries often require more than generic web-style semantic matching [Wu et al. 2017]. More recently, [Luo et al. 2024] explicitly aligned query-side attributes with catalog-side attributes and showed that such query-understanding features improve ranking in a production product-search system. LLMs have also entered this space: they have been used for eCommerce attribute-value extraction on the product side [Fang et al. 2024] and for decomposing product queries into structured attribute-value hints that guide retrieval and reranking [Zhu et al. 2025]. Our work builds on this literature but differs in focus: rather than proposing a new query parser or attribute extractor, we use query-product match signals as an intermediate structure to identify mismatches from graded relevance labels.

Large language models have recently become a major focus in relevance labeling and evaluation research, with early position papers framing them both as a promising opportunity for scaling relevance assessment and as a source of new methodological risks [Faggioli et al. 2023]. Subsequent empirical work has shown that LLMs can often approximate human relevance judgments surprisingly well [Thomas et al. 2024, Soviero et al. 2024]. Product-search-specific work has pushed this direction further, with [Sachdev et al. 2025] using prompting-based LLM pipelines to automate query-product relevance labeling for eCommerce ranking. Our work is adjacent to this literature but uses LLMs in a narrower, more controlled role: rather than replacing human relevance labels, we treat human-graded judgments as the supervision anchor and use LLMs only as auxiliary annotators for latent structure, such as query constraints, product-type compatibility, and attribute-level satisfaction.

3. Core definitions

Relevance. In product search, *relevance* is the degree to which a product is an appropriate result for a query in light of the user's likely shopping goal. Relevance is broader than exact specification matching: a product may still be relevant at some level even when it does not satisfy all requested characteristics, for example, when it is a plausible substitute. In this sense, relevance captures overall usefulness with respect to the query, rather than strict compliance with every expressed requirement.

Attribute. An *attribute* is a product property that can be used to characterize either the user's request or the candidate item. Typical examples include brand, color, size, flavor, material, gender, and capacity. Attributes provide the semantic dimensions along which query requirements and product descriptions can be aligned, compared, and evaluated for satisfaction or mismatch.

Constraint. A *constraint* is an attribute-level condition imposed by the query on the desired product. For example, in a query such as "red nike running shoes," the query specifies constraints over product type (*shoes*), brand (*Nike*), and color (*red*). Constraints are the operational units used to assess whether a candidate product satisfies or violates the query intent. In our setting, violations are computed by checking whether the product satisfies, violates, or leaves unresolved the constraints identified in the query.

Violation. A *violation* occurs when a product fails to satisfy a query requirement that is interpreted as part of the user's search intent. Violations are defined with respect to constraints expressed or strongly implied by the query, such as product type, brand, color, size, or material. A product may therefore be relevant yet still contain one or more violations, which makes violation a concept related to, but not reducible to, graded relevance.

4. Data and Annotation Pipeline

The Amazon Shopping Queries Dataset [Reddy et al. 2022] is a large-scale benchmark that supports product search tasks. This dataset features human relevance labels assigned to query-product pairs. The relevance labels, known as ESCI, are:

- **E - Exact**: When the product is an exact match for the query and satisfies all query specifications.
- **S - Substitute**: When the product is still relevant to the query but differs in some of the specifications. An example would be the user asking for "Nike running shoes," and the retrieved product is "Adidas running shoes".
- **C - Complement**: When the product is not an exact match nor a substitute, but could still be interesting to a shopper. For example, a query could be "Notebooks," and a complementary product could be "Notebook backpack."
- **I - Irrelevant**: The retrieved item is completely irrelevant to the query.

The dataset is in English, Spanish, and Japanese, but we focused only on the English portion. We worked on a random sample of 5,000 queries from the train split of the original dataset and 1,000 from the test split. To create a development set that was not available in the original data, we sampled 20% of the attributed-extracted queries from the sampled training set and completed the set with non-attributed-extracted queries until we reached 1000 queries. With this data selection, we ensured we had sufficient attributed-extracted data for our experiments. Table 1 shows the statistics of the number of queries and query-product pairs in each split of our dataset.

To answer our research questions, we need a large set of query-product pairs with extracted and annotated attributes and violations. Our annotation pipeline has three steps and was performed by an LLM following previous work that confirmed LLM and human annotations are highly correlated [Faggioli et al. 2023, Bencke et al. 2024, Soviero et al. 2024], and that attribute extraction can be reliably performed [Zhang et al. 2021, Loughnane et al. 2024, Luo et al. 2023]. The chosen LLM was GPT-5-mini due to its cost-benefit ratio. Figure 1 shows examples of what is produced in each annotation step.

Step 1 – Query Constraint Annotation. This stage is concerned with annotating whether a query is suitable for attribute extraction. The *constraint profile* is annotated, including whether the query has an extractable product type and attributes. For the *constraint pro-*

Worked examples of the annotation pipeline

Step 1 - Query Constraint Annotation

Query	Constraint profile	Query category
`red running shoes men`	`hard_actionable`	`shoes`
`cheap black t shirt`	`mixed_actionable`	`apparel`
`comfortable office chair`	`soft_preference`	`furniture`
`laptop for programming`	`implicit_or_inf.`	`personal_computers`

Step 2 - Query Attribute Extraction

Query: `huggies diapers size 4` **Extracted product type:** `diapers` **Extracted attributes:** `brand = huggies; size = 4`

Step 3 - Query-Product Attribute Violations

Query: `black leather office chair`
Product title: `Black fabric office chair with padded arms`
Product-type judgment: `exact_match`
Constraint judgments: `color = satisfied; material = violated`

Figure 1. Examples of the three annotation stages used in the pipeline

file, we asked in the prompt that the query be classified into: *hard actionable* (the query contains hard explicit constraints that if violated would hurt the ranking, such as size, gender, and brand), *mixed actionable* (the query combines hard constraints with less stable and subjective filters, such as "cheap nike shoes"), *soft preference* (preferential or desirable qualities, such as "cute shoes" or "on sale laptop"), *implicit or inferential* (queries that require implicit or inferential reasoning for extracting attributes, such as "gifts for grandma"), and *non constraint sensitive* (usually broad queries the constraint satisfaction is not central to relevance, such as "laptop"). At this stage, we also classify whether *extractable fields* are present. The allowed labels for this classification are *type only*, when the query mentions only a product type or a product type and a brand, and *type plus attributes*, when the query mentions a product type and attributes. The query constraint annotation was made in the same prompt, in groups of 10 queries. We also identified and removed queries containing negations since those would reverse the constraints expressed by the attributes. Additionally, we included the "uncertain" classes in each classification, and graded confidence levels from *low* to *high*.

Step 2 – Query Attribute Extraction This stage aims to extract attributes and solicited product types from queries. Since the definition of an attribute can be elusive and subjective, we had to rely on examples and directives to guide the LLM toward the attribute extraction we wanted. For some queries, it can be truly subjective whether attributes exist. For example, the query "wood crates" could ask for "crates" with the material attribute "wood," or we could consider "wood crates" a whole product type since it is a very lexicalized and consolidated item. The same case applies to other compound queries such as "electric toothbrush" and "electric guitar." This issue, if left unchecked, could create hard to manage attribute space. We addressed this problem by asking the LLM to be conservative and trying to match the extracted attributes into one of 10 classes (brand, color, size,

Table 1. Dataset statistics by split.

Split	Queries	CB queries	QP pairs	E (%)	S (%)	C (%)	I (%)
Train	4000	1685	74 744	69.46	19.64	2.15	8.75
Dev	1000	682	18 949	69.42	20.57	1.99	8.02
Test	1000	497	19 182	66.95	21.57	2.01	9.47

CB = constraint-bound. QP = query-product. E = Exact, S = Substitute, C = Complement, I = Irrelevant.

material, flavor, dosage, compatibility, gender, pack_size, and other_explicit_attribute). The last class was used when the LLM could not match the attribute to any of the other classes. This way, we can discover new attributes beyond the examples and constrain what the LLM considers an attribute. For product type extraction, we simply asked the LLM to output the product the query requested. If a product was present in the query, we asked the LLM to extract it; if not, return an empty object.

Step 3 – Query-Product Attribute Violations The goal of this stage is to find which query attributes were violated or satisfied by the retrieved product. We send the LLM the query, the extracted attributes, the extracted product type (if any), and the product title. For each attribute, now called a *constraint*, the LLM should determine whether it was *satisfied*, *violated*, or *cannot determine*. We also asked the LLM to return the product type compatibility. The possible compatibilities are: *exact match*, the retrieved product is exactly what the query requires, *compatible match*, the product is a sub-type/super-type, or an obvious near-equivalent to the requested product, *clear mismatch* when the product clearly is a different kind than what was requested, and, finally, the *cannot determine* label is reserved for when the requested type is missing, or the product title is too vague.

5. Relevance vs Violations

In this section, we address [RQ1] *To what extent is query-intent violation distinct from graded relevance in product search?* In order to answer that, we looked at the proportion of annotations provided by Step-3 Query-Product Attribute Violations in each ESCI label. Complement items were excluded from this analysis because they represented only 2% of the dataset and this was not considered enough to allow for reliable conclusions. Table 2 shows that product-type alignment and attribute-level violation are strongly associated with graded relevance, but they do not collapse into the same notion. Exact items are dominated by exact product-type matches (78.9%) and have by far the lowest overall violation rate (10.1%), which is consistent with the intuition that exact results should satisfy the core product request and most attribute-level constraints. At the other extreme, Irrelevant items show a majority of clear product-type mismatches (56.3%) and a much higher violation rate (40.8%), indicating that both off-target product type and constraint failure are characteristic of low-relevance results.

Substitute items occupy an intermediate position and are particularly informative for distinguishing relevance from violation. The majority of Substitute pairs still exhibit an exact product-type match (59.0%), indicating they remain on target for the requested item category. However, their violation rate rises sharply to 44.7%, substantially above

Table 2. Product-type match distribution by ESCI label.

Product-type / violation summary	Exact	Substitute	Irrelevant
Exact product-type match	**78.9**%	**59.0**%	23.7%
Compatible product-type match	10.9%	15.4%	9.2%
Clear product-type mismatch	6.3%	16.8%	**56.3**%
Product-type undetermined	3.9%	8.9%	10.8%
Pairs with any violation	10.1%	44.7%	40.8%

Percentages may not sum to exactly 100% because of rounding. The last row is computed over the subset of constraint-bound query-product pairs.

that of Exact items and even slightly above that of Irrelevant items. This pattern suggests that many substitutes remain relevant at the product-type level while failing one or more requested attributes. In other words, relevance judgments can still treat a product as useful even when it violates explicit query constraints. Taken together, these results support the claim that query-intent violation is related to graded relevance, but is not fully captured by it. Product-type mismatch is a strong marker of irrelevance, whereas attribute-level violations are especially important for explaining why substitute items can remain relevant despite not fully satisfying the query intent.

These results suggest that query-intent violation should be treated as related to, but distinct from, graded relevance in product search. Product-type mismatch behaves largely as expected, acting as a strong marker of irrelevance, whereas attribute-level violations reveal a more nuanced pattern: many Substitute items remain relevant enough to receive a positive ESCI label despite failing explicit query constraints. This is the key evidence that relevance alone does not fully represent intent satisfaction. In practical terms, the findings motivate ranking approaches that model violations explicitly rather than assuming that graded relevance judgments are sufficient to capture all aspects of query compliance.

6. Ranking experiments

This section addresses RQ2: *Do violation-aware features based on product-type match and attribute-level constraint satisfaction improve ranking effectiveness relative to lexical and semantic similarity baselines?* and RQ3: *Can a limited attribute extraction scheme retain enough violation information to provide useful ranking gains?*

6.1. Label prediction ranker

To answer RQ2, we assessed whether the product type matches and attribute-violation data provide valuable information for ranking products. We devised a simple classifiers that predict ESCI labels to rank products with respect to queries. The Random Forest classification algorithm was used because of its simplicity and its ability to model non-linear dependencies among features. We defined three models, or feature sets, with varying degrees of complexity in attribute extraction. All models make use of the product type match data. This data is encoded in three binary features indicating whether the product type was an exact match, a clear mismatch, or could not be determined. The use of attribute data can be separated into:

- **All-Atbs:** The full range of attributes is used to compute the number of satisfied constraints, the number of violated constraints, and the total number of constraints.
- **Top10-Atbs:** Only the top 10 attributes from the training data are used to compute the features of the number of satisfied constraints, the number of violated constraints, and the total number of constraints. This feature set simulates the data from a more limited attribute extraction approach.

To rank the data with these models, we predicted the product labels with the multiclass classifier and ranked the products with the precedence $E > S > C > I$. Ties are broken using the class probability.

Ranking Baselines We implemented two baselines: lexical and semantic search. For the lexical search, we used the BM25 score between the product and the query. The term-frequency and inverse document frequency statistics used by BM25 were derived from processing the entire ESCI product catalog. For the semantic search baseline, we used the BGE-M3[1] embedding model. The product score is the cosine similarity between the query and the product vector embeddings.

Evaluation Metrics We evaluated ranked results using both a graded relevance metric and a constraint-adherence metric. For graded relevance, we used the Normalized Discounted Cumulative Gain (NDCG) [Järvelin and Kekäläinen 2002] at rank k (NDCG@k), which rewards placing highly relevant products near the top of the ranking. Let $y_r \in \{E, S, C, I\}$ denote the relevance label of the item at rank r, and let $g(\cdot)$ be a gain function mapping labels to nonnegative utilities (e.g., $g(E) = 7$, $g(S) = 3$, $g(C) = 1$, $g(I) = 0$). Then, $DCG@k = \sum_{r=1}^{k} \frac{g(y_r)}{\log_2(r+1)}$ and $NDCG@k = \frac{\text{DCG@}k}{\text{IDCG@}k}$, where IDCG@$k$ is the maximum possible DCG@k for the query under the same gain function. To measure how often top-ranked products violate the query's intended attributes, we use Violation@k. Let $v_r \in \{0, 1\}$ indicate whether the product at rank r violates the query intent, where $v_r = 1$ means a violation is present. Then $Violation@k = \frac{1}{k} \sum_{r=1}^{k} v_r$. Lower values are better; 0 indicates that no item among the top-k results violates the query intent. Although related, these metrics capture different ranking properties. NDCG@k is a position-sensitive, graded relevance metric: it distinguishes exact, substitute, and complementary matches and rewards placing better matches earlier in the ranking. In contrast, Violation@k is a direct constraint-adherence metric: it only measures the proportion of returned items that violate query intent, regardless of whether a non-violating item is exact, substitute, or complementary. As a result, a ranking may achieve strong NDCG@k while still exposing intent violations, or reduce violations while changing its graded relevance profile.

6.2. Results

Table 3 shows that the label prediction models substantially outperform the majority-class baseline, indicating that product-type and attribute-based violation features carry useful signal for ESCI relevance prediction. The baseline attains a low F1-macro of 0.2664, with all of its performance concentrated on the dominant Exact class ($F1\text{-}E = 0.7993$) and no ability to recover the Substitute or Irrelevant classes. In contrast, both violation-aware models improve performance across all reported classes. TOP10-ATBS raises F1-macro to 0.4692, already showing that a restricted attribute inventory is sufficient to recover

[1] https://huggingface.co/BAAI/bge-m3

Table 3. Classification performance of the label prediction models.

Feature set	F1-macro	F1-E	F1-S	F1-I
Baseline (Majority class)	0.2664	0.7993	0.0000	0.0000
Top10-Atbs	0.4692	0.8077	0.3658	0.2340
All-Atbs	**0.5173**	**0.8170**	**0.4223**	**0.3126**

Table 4. Reranked performance by run

Model	NDCG@10 (↑)	Violation@10 (↓)
(baseline) BGE-M3	0.9206	0.1746
(baseline) BM25	0.9038	0.1746
Top10-Atbs	0.9096	0.1261
All-Atbs	**0.9217**	**0.1130**

part of the relevance structure. The strongest results are obtained by ALL-ATBS, which achieves the best F1-macro (0.5173) and the highest class-wise F1 scores for Exact, Substitute, and Irrelevant. The gain is especially notable for the non-dominant classes, suggesting that broader attribute coverage helps the model better distinguish between exact matches, partial mismatches, and clearly irrelevant products. Overall, these results indicate that richer violation-aware representations improve label prediction quality and provide a stronger basis for downstream ranking.

In Table 4, the violation-aware rankers outperform the lexical baseline and, in the case of ALL-ATBS, also slightly surpass the semantic baseline in ranking quality while substantially reducing intent violations. The strongest overall result is obtained by ALL-ATBS, which achieves the best NDCG@10 (**0.9217**) and the lowest Violation@10 (**0.1130**), improving over the BGE-M3 baseline both in effectiveness (0.9217 vs. 0.9206) and, more markedly, in violation reduction (0.1130 vs. 0.1746). This corresponds to an absolute reduction of 0.0616 in Violation@10 relative to both baselines, indicating that product-type match and attribute-level constraint signals provide information not captured by similarity alone. The TOP10-ATBS variant also yields a substantial decrease in violations (0.1261), although with a modest drop in NDCG@10 (0.9096), suggesting that a limited attribute extraction scheme still preserves useful violation information, but that broader attribute coverage leads to the best trade-off. Overall, these results support RQ2 by showing that violation-aware features improve ranking beyond lexical and semantic baselines, and support RQ3 by indicating that even a restricted attribute set remains beneficial, though clearly less effective than the full attribute representation.

6.3. Error Analysis

Since the NDCG@10 scores of the best model and the best baseline are very close, we conducted a comprehensive error analysis to understand why these models yield similar results. In this analysis, we focused on the All-Atbs model as our target model and BGE-M3 as our baseline. As the first step, we obtained the per-query NDCG@10 values for the target and baseline. Then, we computed the ΔNDCG@10 between the target and baseline as NDCG@$10_{All-Atbs}$ – NDCG@10_{BGE-M3}. Next, we selected the queries in the test set with $\Delta\text{NDCG@10} < -0.1$. This threshold selects 49 queries into a group we call "low

Table 5. Error-analysis for low-Δ queries and other queries.

Feature	Low-Δ Queries	Other Queries
number of gold E items in the candidate set	7.6122	13.0692
number of gold S items in the candidate set	7.5714	4.3080
% of top-10 products with predicted violations	0.3240	0.1501
% of top-10 products with product-type mismatch	0.1796	0.0785

delta", and places the rest in a group called "other".

The query-level feature summary in Table 5 suggests that the largest losses of the target model are concentrated in a qualitatively different ranking regime from the rest of the dataset. In the low-Δ subset, the judged candidate sets contain substantially fewer gold Exact items on average (7.61 vs. 13.07) and substantially more gold Substitute items (7.57 vs. 4.31), indicating that these queries are less dominated by clearly exact matches and instead present denser near-match competition. At the same time, the top-10 results for these queries exhibit markedly higher rates of predicted violations (0.324 vs. 0.150) and product-type mismatch (0.180 vs. 0.079). Taken together, these patterns suggest that the target model underperforms particularly on ambiguity-heavy queries in which Exact and Substitute items coexist more densely and in which violation-related signals are more prevalent, raising the possibility that structured mismatch cues are being weighted too strongly relative to the semantic evidence captured by the embedding-based baseline.

To better understand whether the observed disagreements between the reference ranker and the target model reflect genuine model failures or limitations of the data and observable evidence, we conducted a manual audit of three error slices derived from the pair-level comparison. The *exact demotion* slice contains cases in which gold Exact items were demoted by the target model (*i.e.,* ranked lower), typically to Substitute; the *promotion* slice contains cases in which lower-relevance items, especially Substitute or Irrelevant items, were promoted by the target model (*i.e.,* ranked higher); and the *boundary* slice contains disagreements near label boundaries, particularly cases suggestive of Exact/Substitute ambiguity. This audit is important because aggregate ranking metrics alone cannot distinguish between errors caused by model miscalibration and disagreements driven by noisy labels, ambiguous queries, or insufficient evidence in the product title. The annotation taxonomy used in the audit includes the following classes:

- **Query ambiguity**: the query is underspecified or compatible with multiple plausible user intents.
- **Gold-label inconsistency**: the gold ESCI label appears implausible given the visible query-product evidence.
- **Product-type resolution failure**: the main disagreement concerns the intended product type or whether the candidate is on-target.
- **Product-text insufficiency**: the product title does not provide enough information for a reliable judgment.
- **ESCI boundary ambiguity**: the pair lies near a plausible relevance boundary, especially between Exact and Substitute.
- **Evidence conflict / partial evidence**: the visible evidence is mixed, incomplete, or points in conflicting directions.

Table 6. Manual audits by error slice. Rows are sorted by overall frequency.

Annotation	Exact demotion	Boundary	Promotion	Total
Query ambiguity	8	4	4	16
Gold-label inconsistency	5	5	4	14
Product-type resolution failure	0	4	4	8
Product-text insufficiency	0	3	3	6
ESCI boundary ambiguity	3	1	1	5
Evidence conflict / partial evidence	0	1	4	5
Constraint overweighting	3	1	0	4
Query-text insufficiency	1	1	0	2
Total	**20**	**20**	**20**	**60**

- **Constraint overweighting**: the target model appears to penalize an apparent mismatch too strongly relative to ESCI relevance.
- **Query-text insufficiency**: the query text omits key information needed for a stable structured judgment.

The audit results shown in Table 6 indicate that the disagreement region is dominated less by clean target-model failures than by ambiguity and limitations of the benchmark evidence. Across all slices, the most frequent annotations were *query ambiguity* (16/60) and *gold-label inconsistency* (14/60), followed by *product-type resolution failure* (8/60) and *product-text insufficiency* (6/60). In the exact demotion slice, only a minority of cases were attributed to *constraint overweighting* (3/20), whereas most were explained by query ambiguity, gold-label inconsistency, or genuine ESCI boundary uncertainty. The boundary slice was likewise not dominated by true boundary ambiguity; instead, it was more often associated with gold-label inconsistency, query ambiguity, and product-type problems. The promotion slice showed the most heterogeneous profile, with substantial contributions from query ambiguity, gold-label inconsistency, product-type resolution failure, and evidence conflict. Taken together, these findings suggest that a substantial portion of the observed errors arise from unstable labels, ambiguous intent, or insufficient observable evidence rather than from unambiguous failures of the target model itself.

6.4. Discussion

Taken together, the results provide a positive answer to RQ2 and a qualified positive answer to RQ3. First, the ranking results show that violation-aware features based on product-type match and attribute-level constraint satisfaction are useful beyond lexical and semantic similarity alone. In particular, ALL-ATBS achieves the best overall trade-off, slightly improving over the strong BGE-M3 baseline in NDCG@10 while producing a much larger reduction in Violation@10. This pattern is important because it shows that the proposed features are not merely recovering topical similarity; rather, they capture a complementary notion of query compliance that standard relevance-oriented baselines do not fully represent. The classification results are consistent with this interpretation: both feature sets strongly outperform the majority baseline, and the gains are especially visible on the non-dominant classes, suggesting that structured mismatch information helps the model separate exact matches from acceptable substitutes and clearly irrelevant results. At the same time, the comparison between TOP10-ATBS and ALL-ATBS indicates

that broader attribute coverage remains valuable. A restricted attribute inventory already preserves enough signal to substantially reduce violations, but it does not match the full feature set in either label prediction quality or the final ranking trade-off, suggesting that long-tail or less frequent attributes still carry meaningful information for ranking.

The error analysis also helps qualify these findings. Although ALL-ATBS improves violation control and matches the best baseline on NDCG@10, its gains are not uniform across queries. The largest losses are concentrated in a harder subset with fewer gold Exact items, more gold Substitute items, and higher rates of predicted violations and product-type mismatch, suggesting that the model becomes less reliable when the candidate set is dominated by near-match competition rather than clear exact matches. However, the manual audit shows that many of these disagreements cannot be attributed to straightforward model failure. Instead, they are frequently associated with query ambiguity, gold-label inconsistency, product-type resolution difficulty, and insufficient textual evidence in the product title. This weakens a purely negative interpretation of the low-Δ cases and suggests that the observed ceiling against a strong semantic baseline may be partly imposed by the dataset itself. Overall, the results support the view that violation-aware signals are useful and practically relevant, but they also indicate that their benefits are mediated by annotation quality, product-text observability, and the intrinsic ambiguity of ESCI-style relevance boundaries.

7. Conclusion

This work addressed three research questions about the role of query-intent violations in product search. For RQ1, which addressed the extent to which query-intent violation is distinct from graded relevance in product search, the results show that query-intent violation is related to graded relevance, but is not reducible to it: product-type mismatch is a strong marker of irrelevance, while many Substitute items remain relevant despite violating explicit query constraints. For RQ2, which assessed whether violation-aware features based on product-type match and attribute-level constraint satisfaction improve ranking effectiveness, the answer is affirmative: violation-aware features derived from product-type match and attribute-level constraint satisfaction improve ranking effectiveness relative to lexical and semantic baselines, particularly by substantially reducing Violation@10 while remaining competitive in NDCG@10. For RQ3, which investigated whether a limited attribute extraction scheme could retain sufficient violation information to yield useful ranking gains, the answer is also positive but qualified: a restricted attribute extraction scheme retains enough information to yield useful ranking gains, although broader attribute coverage produces the strongest overall results. Taken together, these findings support the view that query violations constitute a meaningful and practically useful ranking signal in product search, while also highlighting that ambiguity, label inconsistency, and incomplete product evidence remain important performance limits.

Future work includes using smaller, alternative language models to extract attributes and study interactions between attributes and other variables, such as query categories. We also plan to incorporate other datasets and languages in our analyses.

Acknowledgments: This work has been financed in part by VTEX Labs, CAPES Finance Code 001, and CNPq. Use of AI: GPT-5-mini was used for data annotation as described in Section 4, and Grammarly was used for text correction.

References

Bencke, L., Paula, F. S., dos Santos, B. G., and Moreira, V. P. (2024). Can we trust llms as relevance judges? In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 600–612. SBC.

Chen, Y., Liu, S., Liu, Z., Sun, W., Baltrunas, L., and Schroeder, B. (2022). Wands: Dataset for product search relevance assessment. In *European Conference on IR Research*, page 128–141.

Cleverdon, C. (1960). The aslib cranfield research project on the comparative efficiency of indexing systems. *Aslib Proceedings*, 12(12):421–431.

Di Nunzio, G. M., Ferro, N., Jones, G. J. F., and Peters, C. (2006). Clef 2005: Ad hoc track overview. In *Accessing Multilingual Information Repositories*, pages 11–36.

Faggioli, G., Dietz, L., Clarke, C. L. A., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., and Wachsmuth, H. (2023). Perspectives on large language models for relevance judgment. In *Conference on Theory of Information Retrieval*, ICTIR '23, page 39–50.

Fang, C., Li, X., Fan, Z., Xu, J., Nag, K., Korpeoglu, E., Kumar, S., and Achan, K. (2024). Llm-ensemble: Optimal large language model ensemble method for e-commerce product attribute value extraction. In *Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2910–2914.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Kando, N. (1999). Overview of ir tasks at the first ntcir workshop. In *NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, Tokyo, Japan, 1999*.

Loughnane, R., Liu, J., Chen, Z., Wang, Z., Giroux, J., Du, T., Schroeder, B., and Sun, W. (2024). Explicit attribute extraction in e-commerce search. In *Workshop on e-Commerce and NLP @ LREC-COLING 2024*, pages 125–135.

Luo, C., Goutam, R., Zhang, H., Zhang, C., Song, Y., and Yin, B. (2023). Implicit query parsing at amazon product search. In *Conference on Research and Development in Information Retrieval*, SIGIR '23, page 3380–3384.

Luo, C., Tang, X., Lu, H., Xie, Y., Liu, H., Dai, Z., Cui, L., Joshi, A., Nag, S., Li, Y., Li, Z., Goutam, R., Tang, J., Zhang, H., and He, Q. (2024). Exploring query understanding for amazon product search.

Peikos, G. and Pasi, G. (2024). A systematic review of multidimensional relevance estimation in information retrieval. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14(5):e1541.

Reddy, C. K., Màrquez, L., Valero, F., Rao, N., Zaragoza, H., Bandyopadhyay, S., Biswas, A., Xing, A., and Subbian, K. (2022). Shopping queries dataset: A large-scale esci benchmark for improving product search.

Sachdev, J., D Rosario, S., Phatak, A., Wen, H., Kirti, S., and Tripathy, C. (2025). Automated query-product relevance labeling using large language models for e-commerce search. In *International Conference on Natural Language Processing and Information Retrieval*, NLPIR '24, page 32–40.

Schamber, L., Eisenberg, M. B., and Nilan, M. S. (1990). A re-examination of relevance: toward a dynamic, situational definition. *Information processing & management*, 26(6):755–776.

Sondhi, P., Sharma, M., Kolari, P., and Zhai, C. (2018). A taxonomy of queries for e-commerce search. In *Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1245–1248.

Soviero, B., Kuhn, D., Salle, A., and Moreira, V. P. (2024). Chatgpt goes shopping: Llms can predict relevance in ecommerce search. In *Advances in Information Retrieval*, pages 3–11.

Su, N., He, J., Liu, Y., Zhang, M., and Ma, S. (2018). User intent, behaviour, and perceived satisfaction in product search. In *ACM International Conference on Web Search and Data Mining*, WSDM '18, page 547–555.

Thomas, P., Spielman, S., Craswell, N., and Mitra, B. (2024). Large language models can accurately predict searcher preferences. In *Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1930–1940.

Tsagkias, M., King, T. H., Kallumadi, S., Murdock, V., and De Rijke, M. (2021). Challenges and research opportunities in ecommerce search and recommendations. In *ACM Sigir Forum*, volume 54, pages 1–23.

Voorhees, E. (2007). Overview of trec 2006.

Wu, C.-Y., Ahmed, A., Kumar, G. R., and Datta, R. (2017). Predicting latent structured intents from shopping queries. In *International Conference on World Wide Web*, WWW '17, page 1133–1141.

Yang, H., Gupta, P., Fernández Galán, R., Bu, D., and Jia, D. (2021). Seasonal relevance in e-commerce search. In *ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4293–4301.

Zhang, D., Li, Z., Cao, T., Luo, C., Wu, T., Lu, H., Song, Y., Yin, B., Zhao, T., and Yang, Q. (2021). Queaco: Borrowing treasures from weakly-labeled behavior data for query attribute value extraction. In *ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4362–4372.

Zhu, Y., Vedula, N., and Malmasi, S. (2025). Hint-augmented re-ranking: Efficient product search using LLM-based query decomposition. In *Joint Conference on Natural Language Processing and Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 200–216.