Improving E-commerce Search with Category-Aligned Retrieval

Rauf Aliev
Independent Researcher
r.aliev@gmail.com

August 26, 2025

Abstract

Traditional e-commerce search systems often struggle with the semantic gap between user queries and product catalogs. In this paper, we propose a Category-Aligned Retrieval System that improves search relevance by first predicting the product category from a user's query and then boosting products within that category. We introduce a novel method for creating "Trainable Category Prototypes" from query embeddings. We evaluate this method with two models: a lightweight `all-MiniLM-L6-v2` and OpenAI's `text-embedding-ada-002`. Our offline evaluation shows this method is highly effective, with the OpenAI model increasing Top-3 category prediction accuracy from a zero-shot baseline of 43.8% to **83.2%** after training. The end-to-end simulation, however, highlights the limitations of blindly applying category boosts in a complex retrieval pipeline: while accuracy is high, naive integration can negatively affect search relevance metrics such as nDCG@10. We argue that this is partly due to dataset-specific ambiguities (e.g., polysemous queries in the Amazon ESCI corpus) and partly due to the sensitivity of retrieval systems to over-constraining filters. Crucially, these results do not diminish the value of the approach; rather, they emphasize the need for confidence-aware and adaptive integration strategies.

1 Introduction

In the competitive landscape of modern e-commerce, the efficiency and relevance of the search function are paramount. A common strategy to enhance search relevance is to first predict the most likely product category from a user's query and then use this prediction to filter or boost the search space. The core problem lies in the nature of user queries, which are often short, colloquial, and do not directly map to formal category names. Simple keyword-based systems fail to capture this nuance, while traditional machine learning models require vast, manually-labeled training sets.

To address this challenge, this paper proposes a hybrid approach that combines zero-shot semantic search with a targeted training mechanism. We represent both queries and categories in a shared vector space using pre-trained language models. The key innovation is the creation of **Trainable Category Prototypes**—new vector representations for categories derived from the weighted average of embeddings of real user queries. This effectively "tunes" the semantic location of each category to better align with user intent.

We present a two-stage evaluation. First, we measure the offline accuracy of the category prediction model itself. Second, we conduct an online simulation to measure the end-to-end impact of these predictions on the final relevance of search results from a real document collection, providing a holistic view of the method's practical utility.

2 Related Work

Research in e-commerce search increasingly integrates semantic embeddings to predict user intent, such as product categories or types, from short and ambiguous queries [3]. A key challenge is translating high-accuracy offline models into practical online gains without degrading metrics like nDCG due to over-constraining filters [4]. Studies emphasize multi-locale and multimodal extensions, addressing cultural and visual aspects that our method could build upon. Evidence leans toward hybrid approaches combining behavioral data with embeddings for robustness, though controversies arise around dataset-specific biases, similar to the ESCI ambiguities we noted (e.g., polysemous terms like "mandoline") [5].

The work by Tigunova et al. (2024) on query-to-product type prediction directly parallels our category prediction task [3]. They propose transfer learning from high-resource to low-resource locales to achieve performance parity, highlighting the need for models that can handle linguistic and cultural diversity. Their findings support our call for adaptive integration, as low-resource scenarios amplify the negative impact of model errors, similar to our online simulation results.

Our approach differs in its core mechanism for representing categories. Rather than training a classifier on top of embeddings, we construct the category representations themselves from user queries. This method of creating "Trainable Category Prototypes" directly embeds user intent into the category vector. The final prototype, a weighted combination of this query-derived vector and the embedding of the category's name, is a novel hybrid representation designed to be robust yet semantically grounded. This technique offers a lightweight and direct way to align the search space with user language.

3 Methodology

In this work, our primary focus is on developing and evaluating a training procedure for constructing weighted query-based category prototypes. To this end, we experiment with two embedding models: the `all-MiniLM-L6-v2` Sentence-Transformer (lightweight) and OpenAI's `text-embedding-ada-002` (large-scale SaaS). Category prototypes are learned through our proposed weighting mechanism and indexed using Elasticsearch for k-NN retrieval. To assess the effectiveness of the resulting models, we conduct end-to-end experiments on the Amazon ESCI dataset, where the final search evaluation is performed against a Solr index.

3.1 Dataset Characteristics

The experiments are based on the Amazon ESCI dataset [1], a large-scale collection of shopping queries.

- **Document Collection:** The full dataset contains 1.7 million products. Our end-to-end search evaluation was performed against the English-language subset, comprising 642,389 documents indexed in Solr.
- **Query Set:** We utilized a ground truth file derived from the ESCI corpus, sampling a total of 10,000 queries which were split into a training set of 8,000 and a test set of 2,000. On average, each query in the ground truth is associated with 12.94 relevant products.
- **Category Taxonomy:** The product data, including hierarchical category information, was sourced from an expanded version of the ESCI dataset [2]. For our experiments, we truncated these paths to the first level, resulting in 112 unique root categories.

3.2 Category Prototype Generation

The core of our method is to create robust vector representations for categories. The process is as follows:

1. **Category Path Truncation:** We truncate all category paths to Level 1 to create broad, useful targets for search filtering (e.g., 'Electronics >Headphones' becomes 'Electronics').
2. **Ground Truth Aggregation:** Using the training set, we create a map of $\{query \to \{category : probability, ...\}\}$.
3. **Prototype Computation:** The prototype vector for a category C, $\vec{v}_{\text{proto}}(C)$, is the weighted average of the embeddings of its associated training queries.

   $$\vec{v}_{\text{query}}(C) = \frac{\sum_{i \in Q_C} p(C|q_i) \cdot \vec{e}_{q_i}}{\sum_{i \in Q_C} p(C|q_i)} \tag{1}$$

   where Q_C is the set of training queries for category C, and $\vec{e}_{q_i}$ is the embedding of query q_i.

   Next, we compute the final hybrid prototype, $\vec{v}_{\text{hybrid}}(C)$, by interpolating between the query prototype and the embedding of the category's name, $\vec{e}_{\text{name}}(C)$:

   $$\vec{v}_{\text{hybrid}}(C) = \alpha \cdot \vec{v}_{\text{query}}(C) + (1 - \alpha) \cdot \vec{e}_{\text{name}}(C) \tag{2}$$

   Here, α is a weighting factor (set to 0.85 in our experiments) that balances the influence of user intent from queries against the formal semantics of the category name. For categories with no training data, the prototype is simply $\vec{e}_{\text{name}}(C)$. These final prototypes are indexed into Elasticsearch.

3.3 End-to-End Search Relevance Evaluation

To measure the real-world impact, we ran the test set of 2,000 queries against the Solr collection. We compared two search configurations:

1. **Baseline (‘title boost x2‘):** A standard keyword search with a boost on the title field:
   `description_t:(<query>) OR title_t:(<query>)^2`.

2. **CARS (OpenAI, K=3, K=5):** The baseline search query augmented with a powerful boost for documents belonging to the Top3/Top5 predicted Level 1 categories, with predictions generated by the trained `text-embedding-ada-002` model.

4 Results

Our evaluation is twofold: we first assess the accuracy of the category prediction model in isolation (offline), and then measure the performance of the end-to-end search system (online).

4.1 Category Prediction Accuracy (Offline Evaluation)

We evaluated the category prediction accuracy on a test set of 2,000 queries for both the `all-MiniLM-L6-v2` and `text-embedding-ada-002` (OpenAI) models. Table 1 compares their Top-3 accuracy in a zero-shot scenario and after full training on 8,000 queries.

Table 1: Comparison of Top-3 Prediction Accuracy for L1 Categories

Model	**Zero-Shot Accuracy**	**Full Training Accuracy**
`all-MiniLM-L6-v2`	28.9%	71.5%
`text-embedding-ada-002`	43.8%	**83.2%**

Table 2: Comparison of Top-5 Prediction Accuracy for L1 Categories

Model	**Zero-Shot Accuracy**	**Full Training Accuracy**
`text-embedding-ada-002`	57.8%	**89.6%**

Both models benefit significantly from the prototype training process. The larger OpenAI model demonstrates superior semantic understanding, outperforming the smaller model in both scenarios and ultimately reaching a very high **83.2%** Top-3 accuracy and **89.6%** Top-5 accuracy.

5 Online Evaluation Simulation

Given the high accuracy of the OpenAI model, we used it to power the Category-Aligned Retrieval System (CARS) in our online simulation. For each query, the Top 5 predicted categories were used to apply a strong boost. The results were compared against the keyword baseline over 2,000 test queries.

5.1 Results

Surprisingly, despite the high offline accuracy, the Category-Aligned Retrieval System underperformed the baseline, as shown in Table 3.

Table 3: Comparison of Search Relevance Metrics. CARS is powered by the OpenAI model with K=5 boosting.

Metric	Statistic	Baseline Search	CARS
2*nDCG@10	Mean	**0.273**	0.255
	Median	**0.174**	0.148
2*Reciprocal Rank	Mean	**0.433**	0.412
	Median	0.250	0.250

The mean scores for the baseline were higher across both nDCG@10 and Reciprocal Rank. This suggests that, on average, the negative impact of incorrect category predictions outweighed the benefits of correct ones.

5.2 Statistical Significance

To determine if these differences were meaningful, we performed a Wilcoxon signed-rank test. The results confirmed that the degradation in performance was statistically significant.

- For nDCG@10, the p-value was **0.00071**.
- For Reciprocal Rank, the p-value was **0.0162**.

This provides strong evidence that the CARS strategy, even when powered by a highly accurate model, was detrimental to overall search relevance in this configuration.

6 Discussion

The disconnect between high offline accuracy and degraded online performance is the most critical finding of this work. It shows that even with a powerful classification model, blindly applying its predictions as a hard filter or strong boost can be counterproductive. This is not necessarily a failure of the model, but rather a failure of a naive integration strategy that does not account for the inherent ambiguity of user queries. The fact that a large portion of queries performed better with CARS while the overall average decreased indicates that the negative impact of a few catastrophic errors outweighs the moderate gains on many other queries.

6.1 Qualitative Analysis of Problematic Queries

A manual analysis of queries where the model failed reveals that many are inherently unclassifiable, even for a human without additional context. These fall into several archetypes, as shown in Table 4.

Table 4: Examples of Inherently Ambiguous or Problematic Queries

Query Example	Amazon's Top Category
Category 1: Too Broad or Ambiguous	
`best offers`	Clothing, Shoes & Jewelry
`rings`	Clothing, Shoes & Jewelry
`planer`	Tools & Home Improvement
`simple human`	Home & Kitchen
Category 2: Non-Transactional Phrases	
`just gonna send it`	Automotive
`embrace the suck`	Clothing, Shoes & Jewelry
`decorating pumpkins without carving`	Toys & Games
`spongebob memes`	Cell Phones & Accessories
Category 3: Unclear or Misspelled	
`real sords`	Sports & Outdoors
`killz`	Tools & Home Improvement
`dapper dan`	Beauty & Personal Care
`turnatable`	Electronics

These examples demonstrate that for a significant subset of queries, reliably determining a single correct category is impossible. Boosting on a wrong prediction for these queries is what drives the average relevance metrics down.

6.2 Illustrative Case Studies

Case Study 1: The Win (”raleigh bicycle”). A query for ”raleigh bicycle” in the baseline system returns a ”Raleigh Crossbody Bag” as a top result due to a keyword match on ”Raleigh”. CARS, however, correctly predicts the category as ‘Sports Outdoors‘. The resulting boost elevates actual Raleigh bicycles to the top of the results, showcasing the system's potential when the prediction is correct and semantically useful.

Case Study 2: The Loss (”mandoline slicer spiralizer”). For this query, CARS incorrectly predicts ‘Musical Instruments‘. The polysemy of ”mandoline” (a cooking utensil vs. a musical instrument) confuses the model. Consequently, CARS boosts completely irrelevant products. This demonstrates that the end-to-end system is critically vulnerable to the model's errors, and the negative impact of such an error can be far greater than the positive impact of a correct prediction.

Case Study 3: The Unseen (”little trees fresh shave”). This query did not appear in the training set. The user's intent is to find the ”Little Trees” brand car air freshener, which belongs to the ‘Automotive‘ category. However, based purely on the semantics of the query string, CARS predicts the top categories as (1) ‘Grocery Gourmet Food‘, (2) ‘Beauty Personal Care‘, (3) ‘Arts, Crafts Sewing‘, (4) ‘Patio, Lawn Garden‘, and (5) ‘Health Household‘. None of these is correct. By applying a strong boost to these irrelevant categories, the actual ‘Automotive‘ product is pushed far down the rankings, effectively becoming undiscoverable. This highlights the model's brittleness when encountering novel or brand-specific queries whose semantics diverge from their product category.

7 Directions for Future Research

The gap between offline accuracy and online performance motivates several avenues for future work aimed at more intelligent integration strategies.

- **Confidence-Based Adaptive Boosting:** Instead of applying a uniform boost, the system should modulate the boost strength based on the model's confidence. Confidence could be measured by the cosine similarity of the query embedding to the top predicted category prototype. For high-confidence predictions (e.g., similarity > 0.9), a strong boost is applied. For moderate confidence, a weaker boost is used. For low-confidence predictions, the system should gracefully fall back to the baseline keyword search, thus avoiding catastrophic failures on ambiguous queries.

- **Explicit Ambiguity Detection:** A separate classification model could be trained to identify inherently problematic queries (e.g., non-transactional, overly broad, polysemous). Queries flagged as ambiguous would bypass the category boosting mechanism entirely. This acts as a protective layer for the retrieval system.

- **Hybrid Retrieval and Re-ranking:** Rather than replacing the baseline, a hybrid approach could be more robust. The final results page could be a combination of the top-k results from the baseline search and the top-k results from CARS. A subsequent learning-to-rank (LTR) model could then re-rank this combined set, using the predicted category and its confidence score as features.

- **Leveraging Category Hierarchy:** Our current method truncates categories to Level 1. Future iterations could predict deeper into the category taxonomy. This would allow for more precise filtering. Additionally, the hierarchy itself can be used to regularize predictions. For instance, if the model predicts 'Electronics ¿ Headphones' and 'Electronics ¿ Speakers' with high confidence, the confidence for the parent category 'Electronics' should be further increased.

8 Conclusion

In this work, we developed a method for query-to-category prediction using "Category Prototypes," achieving a high Top-3 accuracy of **83.2%** with the OpenAI `text-embedding-ada-002` model. Our primary contribution, however, is the clear demonstration of the gap between offline classification performance and end-to-end search relevance.

Our simulated online evaluation showed that a naive boosting strategy based on these highly accurate predictions resulted in a **statistically significant degradation** in search performance. This nuanced result underscores a critical lesson: in complex systems like search, the negative impact of model errors on a subset of queries can easily offset, and even outweigh, the benefits on others, especially when the dataset contains a long tail of ambiguous or non-transactional queries.

Future work should focus not on improving the classifier's accuracy in isolation, but on developing an adaptive mechanism to decide *when* to trust the category prediction. A confidence-based model that applies the category boost only for high-certainty predictions, while defaulting to the baseline for ambiguous queries, could bridge this gap and translate offline accuracy into a tangible online win.

References

[1] Reddy, C., Nangi, M., et al. (2020). *A Shopping Queries Dataset for E-commerce.* Proceedings of The Web Conference 2020.

[2] Shuttie. (2022). *esci-s - A parallel corpus of shopping queries and annotated results for 3 languages*. GitHub repository. `https://github.com/shuttie/esci-s`.

[3] Tigunova, A., Duboue, P., Ganesan, K., & Cubuk, E. D. (2024). *Transfer Learning for E-commerce Query Product Type Prediction.* arXiv preprint arXiv:2410.07121.

[4] Vasilev, F., Antufiev, S., D'yakonov, A., Gusev, G., Tokarev, M., Vasiliev, A., Zha, L., Drach, K., Chernousov, G., & Egorov, E. (2024). *Mind the Gap: From Offline Evaluation to Online Gains for Query-to-Product-Type Prediction in E-Commerce.* Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track.

[5] Rakesh, V., Wang, Y., Malladi, S., Zhao, T., Jain, V., Singh, G., Vu, Q., Chen, H.-S., Hong, L., & Chi, E. H. (2023). *Query Attribute Recommendation at Amazon Search.* arXiv preprint arXiv:2308.03869.