Locale-Aware Product Type Prediction for E-commerce Search Queries

Anna Tigunova
Amazon
Berlin, Germany
tigunova@amazon.de

Thomas Ricatte
Amazon
Luxembourg, Luxembourg
tricatte@amazon.lu

Ghadir Eraisha
Amazon
Luxembourg, Luxembourg
eraishag@amazon.lu

Abstract

Search query understanding (QU) is an important building block of the modern e-commerce search engines. QU extracts multiple intents from customer queries, including intended color, brand, etc. One of the most important tasks in QU is predicting which product category the user is interested in.

In our work we are tapping into query product type classification (Q2PT) task. Compared to classification of full-fledged texts, Q2PT is more complicated because of the ambiguity of short search queries, which is aggravated by language and cultural differences in worldwide online stores. Moreover, the span and variety of product categories in modern marketplaces pose a significant challenge.

We focus on Q2PT inference in the global multi-locale e-commerce markets, which need to deliver high quality user experience in both large and small local stores alike. The common approach of training Q2PT models for each locale separately shows significant performance drops in low-resource stores and prevents from easily expanding to a new country, where the Q2PT model has to be created from scratch.

We use *transfer learning* to address this challenge, augmenting low-resource locales through the vast knowledge of the high-resource ones. We introduce a *unified, locale-aware* Q2PT model, sharing training data and model structure across worldwide stores.

We show that the proposed unified locale-aware Q2PT model has superior performance over the alternatives by conducting extensive quantitative and qualitative analysis on the large-scale multilingual e-commerce dataset across 20 worldwide locales. Our online A/B tests have shown that using locale-aware model improves over the previous user experience, increasing customer satisfaction.

CCS Concepts

• **Information systems → Query intent**.

Keywords

Query Understanding; Product Search

ACM Reference Format:
Anna Tigunova, Thomas Ricatte, and Ghadir Eraisha. 2025. Locale-Aware Product Type Prediction for E-commerce Search Queries. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3746252.3761568

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761568

1 Introduction

Problem statement. Query understanding in e-commerce extracts customer shopping intent from their search queries, by classifying the query as having a target brand, color, size, etc. The extracted attributes are extremely important for search results ranking, query augmentation, recommendations and many other usecases [2]. One of the most critical components in query understanding is a Query-to-Product Type (Q2PT) classifier, which associates search query with a product type (PT) the customer intended to buy.

Q2PT signal has a direct impact on the customer experience, influencing the item selection shown to the user. Moreover, Q2PT signal improves search latency by narrowing the retrieval to a specific product category shard [6].

Query category prediction has received significant attention in related works [5, 15, 16], which address various challenges associated with Q2PT prediction, such as short queries [5, 7], long-tail queries [14, 15], product type hierarchy [16], etc. However, to the best of our knowledge, there has not yet been studies, which target the issues of Q2PT classification in the multi-locale setting.

In international marketplaces, query understanding models are often trained on a per-locale basis, using the data from a single store [1, 8]. That ensures that the local peculiarities are captured: for instance, query 'pants' in UK would mean that the user is searching for *underwear*, while in the US it means the intent to buy *trousers*.

Most studies develop and test query understanding models for high-resource locales, such as United States, achieving remarkable results. However, many emerging stores suffer from data shortage [4], which is a major blocker for applying per-locale Q2PT models. The issue is further aggravated by the long-tail nature of the product category distribution, particularly skewed in emerging stores.

Proposed approach and contributions. To overcome the mentioned issues with sparse and unbalanced data in low-resource locales, we propose to leverage rich data in high-resource locales via *knowledge transfer* for Q2PT prediction.

We experiment with a unified multilingual multi-locale Q2PT model, which shares model parameters and training data across all stores. This solution allows small locales to benefit from the knowledge of the established stores. This solution reduces memory requirements and model training time and complexity.

A drawback of this approach is that the unified *locale-agnostic* model can transfer biases from the large stores to the low-resource ones [1]. To alleviate this, we propose a *unified locale-aware* model variant, by conditioning the prediction on the locale-id. Our experiments show that locale-aware model performs better than the agnostic one by preserving locale-specific traits.

To support our findings we conduct large-scale analysis with an e-commerce data set sampled from real data, including 20 locales and 1414 product types. Additionally, we run an online study to assess the impact of locale-aware model in real settings, which have shown positive impact in customer engagement and purchase behavior.

2 Related Work

Search query classification. Query understanding (QU) signals are essential in e-commerce and thus have received significant attention [5, 9, 14–16].

Long-tail queries with scarce training data is a common challenge in QU, usually solved by transferring knowledge from the similar frequent queries [14, 15]. Multiple studies propose solutions for short and ambiguous search queries Jiang et al. [5], Liu et al. [7]. Another issue is the shortage of QU training data, which is tackled by generating synthetic queries using catalogue items [9, 12].

Handling data from multiple locales. Cross-market recommendations and search is an important problem for international e-commerce stores: the developed solutions need to meet the performance bar globally and be able to generalize to new locales. However, there has been only a few studies addressing this [1, 4, 10].

Roitero et al. [10] compare the performance of global and locale-specific models in cross-market music recommendation, showing that market knowledge plays an important role. Bonab et al. [1] adapt the globally pretrained recommender to the individual stores with additional fine-tuning, complicates model training and serving.

Close to our work is [8], proposing a BERT-based query classification model with a separate product type classification layer for each locale. The unified backbone reduces storage requirements, however, training the classification layer for each locale separately affects the model's ability to transfer knowledge. We use this architecture as a baseline in our experiments.

3 Methodology

Q2PT prediction a multi-label classification task: for an input query in each locale, the Q2PT model needs to return all product types associated with the query. In our study we propose an alternative to an existing method of training Q2PT models for each locale separately [8], replacing it with a unified model, which shares model parameters and training data across all stores.

We use BERT-based encoder [3] to create the representations of the input queries, following current research in query classification [8, 9, 15, 16]. The encoder is followed by a fully-connected layer with sigmoid activation, applied to the [CLS] token.

Building on this base architecture, we compare three variants:

- **non-unified (*NU*)** [8] - which has a common DistilBert backbone but separate classifiers for all locales. Each locale-specific classifier is trained only on the data from that store.
- **unified locale-agnostic ($\mathbf{U}_{ag}$)** - in this model both DistilBert encoder and classifier are shared across all locales and trained on the mix of global data.
- **unified locale-aware ($\mathbf{U}_{aw}$)** - this model is similar to U_{ag}, however, it conditions the product type prediction on the locale-id, by prepending the locale-id token to the input keywords, separated with a special [SEP] token.

U_{ag} solves the problem of data shortage in low-resource stores, but leads to biases transferred from the bigger ones. U_{aw} overcomes this problem by encoding the input locale information alongside with the query.On the other hand, U_{ag} model is more practical for the cases of cold-start launches of new stores, where U_{aw} needs to be retained to learn new locale-id.

Additionally we experimented with prepending the English name of the locale instead of its id, to add semantic information and use the model's general knowledge of the country. However, We did not observe any performance difference of this strategy over prepending locale id.

4 Experimental Setup

4.1 Datasets

We use an experimental data set sampled from real data, covering 1414 product types and 20 locales with 12 distinct languages. Our dataset has immense discrepancy in sample distribution among the locales with large stores taking over 90% share of the dataset.

We split the locales into 2 buckets: *High-Resource* (Hi-Re), including (**US** - United States, **DE** - Germany, **UK** - United Kingdom, **JP** - Japan, **IN** - India, **IT** - Italy, **CA** - Canada, **FR** - France, **ES** - Spain, and *Low-Resource* (Lo-Re) locales, including **MX** - Mexico, **BR** - Brazil, **AE** - United Arab Emirates, **AU** - Australia, **SA** - Saudi Arabia, **EG** - Egypt, **NL** - Netherlands, **TR** - Turkey, **SE** - Sweden, **SG** - Singapore, **PL** - Poland based on the number of training samples.

The data for training Q2PT models is created by aggregating fully anonymized customer click-through behavior, following previous research [6, 14]. The product type for the user search query is derived as the majority product type of items, that the user clicked following that query.

We use 100s of millions samples for model training and 10s of millions for validation and hyperparameter selection.

4.2 Evaluation Data

As the customer click-through data is noisy and prone to trends and seasonality, we cannot rely on it for model performance evaluation. Instead, we create two evaluation datasets: human-annotated and automatically weakly labeled.

Human-labeled data. We recruited professional annotators, who have undergone specific training, to label search queries with applicable product types, resulting in around 1k human-labeled queries. Each query was labeled by two annotators and an arbitrator, who resolved the disagreements.

The human-annotated dataset has high-quality, but mostly consists of queries associated with popular PTs. Thus, this dataset has a very low coverage of the product type label space: on average around 600 product types (42% of the whole list) are included for each locale.

Automatically labeled data. This dataset was collected, following [13]. To create this dataset we leverage relevance labels for the <*query, item*> pairs, obtained from a pre-trained classifier. For each query we collect the categories of all items that are predicted relevant to it. The query label is then selected as a majority category from relevant items.

The resulting dataset consists of 2.7M labels in total, with an average of 120k labels per locale. To check the correctness of this labelling, we manually inspected 200 queries from different locales, and verified that this labeling method achieves almost 90% accuracy. Importantly, this method allowed us to collect a sufficient number of evaluation samples for all 1414 product types.

We fine-tuned a multilingual DistilBert [11] checkpoint on the Q2PT task with Adam optimizer with 8e-5 learning rate, 0.001 dropout and 2^{11} batch size until convergence, using binary cross-entropy loss. The hyperparameters were chosen through grid search on the validation split. We report recall at 0.8 precision, as the fixed high-precision setting is a standard in evaluating customer-facing applications, such as e-commerce sites

5 Experiment Results

Results on the human-labeled data. In Table 1 we report recall at 0.8 precision for all 20 locales. Additionally, we aggregate the results separately for established High-Resource (Hi-Re) and emerging Low-Resource (Lo-Re) stores, and worldwide (WW).

We observe that both variants of the unified model outperform the non-unified one for all locales, with the total increase of 2% recall worldwide. Notably, the most benefiting locales are small ones, e.g. PL (+6%) and SE (+5%). It shows that the unified models can efficiently transfer knowledge from high-resource to the low-resource locales. Importantly, we notice that even on the Hi-Re locales a generalist U_{aw} model performs slightly better than a specialist U_{aw} model.

Between the two variants of a unified model, U_{aw} has slightly more pronounced gains over U_{ag}. This illustrates that the task of product type prediction is not locale-invariant: conditioning on the locale information helps the model to distinguish locale-specific peculiarities (we discuss it further in Section 6.1).

To further assess the models on low- and high-resource locales we plot precision-recall curves for one of the high-resource (US) and one of the small-resource (PL) locales in Figure 1. For US, U_{aw} slightly dominates for the high precision values, thanks to the both more diverse training data from the other locales and preserved information about the current locale. In PL, there is a significant gap between *NU* and consolidated models, because the latter were trained on 1000 times more data.

Additionally, we experimented with a completely disjoint model architecture: for each locale all model parts are shared among the locales, and they are trained and stored separately. We found that this variation performs on par with U_{aw} and has 2% performance increase compared no NU. Given that this model takes 7 times longer to train (sequentially on all locales) and 20 times greater memory requirements, without significant performance improvements, we did not consider this model for our analyses.

Results on the automatically-labeled data. We compute recall at 0.8 precision on the automatically-labeled evaluation set, and present the aggregated results in Table 3. On this dataset the gap between *NU* and unified models considerably increases, which can be attributed to the automated data having more long-tail product types, which are challenging to predict, and which have extremely scarce training data in low-resource locales. At the same time, human data is largely composed of the most popular product types, on which the performance of all models is on par.

In this dataset, however, the trend between both unified models changes, with U_{ag} having slightly better performance. After examining the performance of the models per locale, we found that U_{ag} has superior results in all but two of the biggest locales: US and UK (in these locales U_{ag} performance is 1% worse than U_{aw}). It shows that conditioning on locale-id helps the model to develop some locale-specific knowledge.

Qualitative analysis. We also note that the knowledge transfer can have negative effect on the Q2PT predictions in some cases. To validate it, we checked the evaluation queries for which the *NU* model made a correct prediction, while a unified model (we use U_{aw} in this study) made an error. For instance, the query '*roma*' should yield different product types depending on the store: it should return *laundry_detergent* for MX (local brand), *personal_fragrance* for DE and other European stores (name of perfume popular in Europe), and for the remaining locales the correct prediction will be a *book*. We notice that in SG U_{aw} model, biased by the data from the other locales, predicts *laundry_detergent* for the query '*roma*', while *NU* correctly predicts *book*.

Another interesting dimension of analysis is the differences between the errors of U_{ag} and U_{aw} on the automatically-labeled dataset. Given that US and UK are the only locales where U_{aw} outperforms the locale-agnostic model, we investigated the US queries where U_{ag} fails, while U_{aw} is correct. One example query is '*boys drawers*': while 'drawers' is a common name for underwear in Asia, in the US this term is not used, and the locale-aware model predicts the correct product type as *dresser*. At the same time, U_{ag} model, which does not have access to the locale information, erroneously predicts PT *underpants*.

5.1 Analysis on the Product Type Level

In this analysis we aim to evaluate the models on different product type groups, based on their frequency in the training data. The distribution of product types is very skewed, and thus it is desirable to assess the models' capabilities to correctly infer long-tail PTs.

For this analysis we split the list of PTs into 3 parts: *head* (very frequent PTs, e.g. *book*), *torso* (average frequency, e.g. *wall ornament*) and *tail* (niche PTs, e.g. *mounted storage system kit*), so that each part has 1/3 of query mass of the whole dataset. As a result we got 48 head, 198 torso and 1168 tail product types.

For this experiment we use automatically-labeled data, because it evenly covers all product types. We compare the performance of non-unified model against unified locale-aware model, as it has shown superior performance to the U_{ag} on the human-labeled benchmark. We compute per-PT accuracy, which is defined as the number of correctly predicted occurrences of a product type, over all occurrences of this product type in the evaluation set. Additionally, we compute Pearson correlation of the PT frequency and PT accuracy, to see how much the models overfit on the high-frequency product types.

The results of this analysis are shown in Table 3, aggregated for low- and high-resource locales and worldwide.

	MX	BR	AE	AU	SA	EG	NL	TR	SE	SG	PL	Lo-Re
NU	0.91	0.81	0.9	0.84	0.83	0.72	0.66	0.87	0.53	0.91	0.54	0.77
$\mathbf{U}_{ag}$	0.92	0.78	0.9	0.85	0.87	0.77	0.69	0.9	0.55	0.91	0.59	0.79 (+2%)
$\mathbf{U}_{aw}$	0.92	0.82	0.91	0.87	0.87	0.76	0.70	0.90	0.58	0.92	0.60	0.80 (+3%)

	US	DE	UK	JP	IN	IT	CA	FR	ES	Hi-Re	WW
NU	0.79	0.9	0.88	0.76	0.85	0.93	0.92	0.89	0.92	0.87	0.82
$\mathbf{U}_{ag}$	0.81	0.91	0.9	0.79	0.83	0.95	0.93	0.91	0.93	0.89 (+2%)	0.83 (+1%)
$\mathbf{U}_{aw}$	0.80	0.91	0.90	0.80	0.86	0.95	0.92	0.91	0.93	0.89 (+2%)	0.84 (+2%)

Table 1: Results for recall at 0.8 precision, comparing the baseline (*NU*) and the unified methods ($\mathbf{U}_{ag}$, $\mathbf{U}_{aw}$), in Low-Resource (top) and High-Resource (bottom) locales on the human-labeled evaluation set.

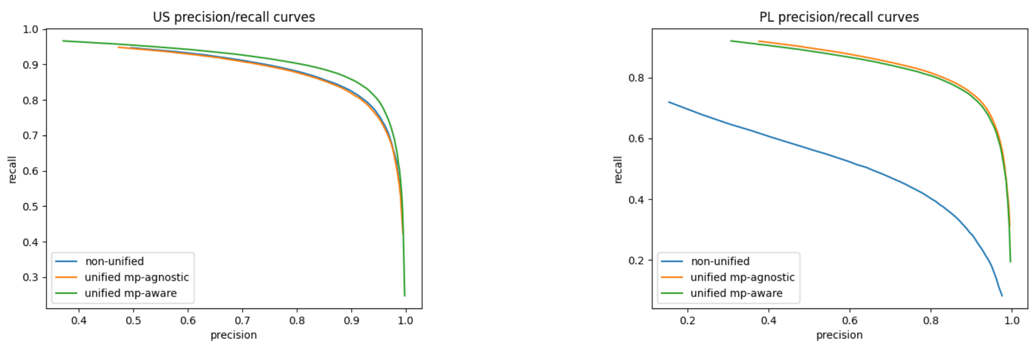


Figure 1: Precision-recall curves in US (left) and PL (right) locales.

	Lo-Re	Hi-Re	WW
NU	0.64	0.80	0.71
$\mathbf{U}_{ag}$	0.83 (+17%)	0.87 (+7%)	0.85 (+14%)
$\mathbf{U}_{aw}$	0.81 (+15%)	0.86 (+6%)	0.83 (+12%)

Table 2: Results for recall at 0.8 precision, comparing the baseline (*NU*) and the unified methods ($\mathbf{U}_{ag}$, $\mathbf{U}_{aw}$), on the automatically labeled evaluation set.

	WW		Lo-Re		Hi-Re	
	NU	$\mathbf{U}_{aw}$	*NU*	$\mathbf{U}_{aw}$	*NU*	$\mathbf{U}_{aw}$
correlation	0.2	0.15	0.22	0.15	0.16	0.14
head	0.85	0.9	0.83	0.89	0.87	0.91
torso	0.79	0.85	0.77	0.84	0.82	0.87
tail	0.7	0.8	0.65	0.78	0.75	0.83

Table 3: Pearson correlation of the number of samples and accuracy per-PT (top line), accuracy per product type on the head/torso/tail PT splits.

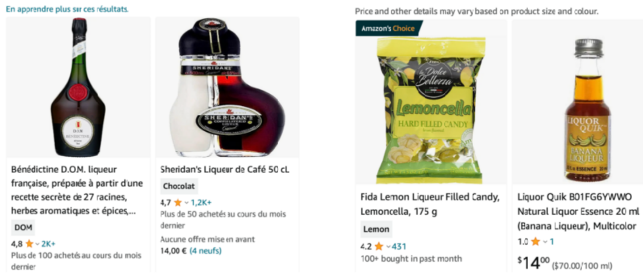
Unsurprisingly, both models perform better on the more frequent groups of product types, however, the accuracy drop from head to tail PT group is more prominent for the non-unified model. Additionally, we noted that the unified model dramatically reduces the number of product types, which do not meet a high accuracy bar of at least 0.5. Non-unified model has 105 such low-accuracy PTs, compared to 27 for a unified model. Most of those 27 PTs are related to digital content, which are easy to confuse to each other (e.g., *music track* and *music album*).

Consistently with the previous results, the correlation of accuracy and PT frequency is greater in NU model. One interesting observation is that the correlation difference is small on Hi-Re locales and is more pronounced on the Lo-Re stores. Additionally, we compared correlations in the biggest locale (US) and the smallest (PL). The correlation in US was similarly low for both *NU* and U_{aw} (0.12 and 0.9 resp.), however, in PL the correlation of *NU* accuracy is extremely high (0.23) and it is getting smoothed with U_{aw} (0.11).

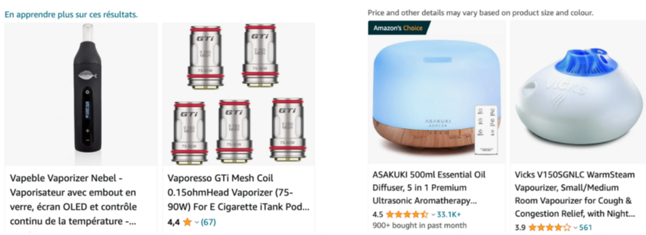
6 Online User Study

We conducted an A/B experiment in all 20 locales to evaluate the impact of our proposed unified model for Q2PT prediction, compared to the non-unified model. The unified model demonstrated significant improvements on human-evaluated search relevance metrics, positively impacting both customer engagement and purchase behavior. Customers were more likely to make a purchase when presented with the unified model's predictions (+0.01% increase in revenue worldwide and number of items purchased). Additionally, the model simplification would potentially lead to a significant reduction in latency and infrastructure costs.

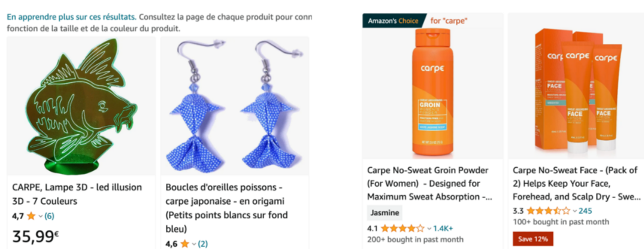
Additionally, the unified model improved the purchases of newly published products (+0.06% increase in revenue worldwide coming from this category of product; +0.02% in terms of number of items purchased). This can be explained by the fact that in low-resource



(a) **liquere** products



(b) **vaporizer** products



(c) **carpe** products

Figure 2: Examples of product type discrepancies between FR (left 2 products) and CA (right 2 products) stores.

locales new items might be coming from the product categories that are currently emerging on those markets, and thus are underrepresented in the training data. In this case the non-unified model, not having seen enough training samples of those product types, will refrain from prediction or make an incorrect one.

6.1 Categorization of Local Differences

During analysis of query differences, we found multiple cases where the product type distributions significantly vary across locales. We attribute those discrepancies to one of the following cases:

1. Dialectal differences: despite the same language, the query has a different meaning to the users in different stores. One particular example is the word '*liqueur*', which means an alcoholic drink in France, but a non-alcoholic drink or syrup in Canada (see Figure 2a). Another example is '*vaporizer*', which means a smoking gadget in France and an air humidifier appliance in Canada (depicted in Figure 2b). These discrepancies can be explained by cultural and historical factors, e.g. language drift from the neighbouring countries. Although such cases are pretty rare in the data, they can have a profound impact on the user experience, especially in the cases of products under legal regulations.

2. Selection differences: the query has different meaning with respect to the product type, that is caused by a mismatch in the selection of products offered in corresponding markets. For example, a query "carpe" that means "a carp" in French leads to fishing accessories in FR store. However, in Canada it is mapped to a popular cosmetic brand marketed under the same name (see Figure 2c). PTs with a limited offer in particular store also belong to this category.

3. Noisy differences: finally, the discrepancies can be attributed to the noise in the user behaviour, which can be amplified for the locales with scarce data. As a toy example, a PT with 5 clicks and 10 impressions should be treated differently than PT with 10K clicks and 20K impressions, even though the probability p will be equal to 0.5 in both cases.

7 Conclusion

In this study we investigate the task of e-commerce query product type prediction in a multi-locale setup and propose a transfer learning solution to augment Q2PT predictions in low-resource stores to achieve global parity.

Our experiments show that the unified Q2PT model, sharing the parameters and training data across all locales, outperforms the non-unified model, on both low- and high-resource locales, additionally decreasing infrastructure requirements. We compare locale-agnostic and locale-aware variants of the unified model, showing that it is important to capture store-specific characteristics. The findings from our work will be useful for practitioners developing multi-locale query classification models.

8 GenAI Usage Disclosure

No GenAI tools have been used to prepare this manuscript or any parts of the experimental code/analysis.

References

[1] Hamed Bonab, Mohammad Aliannejadi, Ali Vardasbi, Evangelos Kanoulas, and James Allan. 2021. Cross-market product recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 110–119.

[2] Yi Chang and Hongbo Deng. 2020. *Query understanding for search engines*. Springer.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[4] Bruce Ferwerda, Andreu Vall, Marko Tkalcic, and Markus Schedl. 2016. Exploring music diversity needs across countries. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. 287–288.

[5] Haoming Jiang, Tianyu Cao, Zheng Li, Chen Luo, Xianfeng Tang, Qingyu Yin, Danqing Zhang, Rahul Goutam, and Bing Yin. 2022. Short text pre-training with extended token classification for e-commerce query understanding. *arXiv preprint arXiv:2210.03915* (2022).

[6] Heran Lin, Pengcheng Xiong, Danni Danqing Zhang, Fan Yang, Ryoichi Kato, Mukul Kumar, William Headden, and Bing Yin. 2020. Light feed-forward networks for shard selection in large-scale product search. (2020).

[7] Xianjing Liu, Hejia Zhang, Mingkuan Liu, and Alan Lu. 2019. System Design of Extreme Multi-label Query Classification using a Hybrid Model.. In *eCOM@SIGIR*.

[8] Chen Luo, William Headden, Neela Avudaiappan, Haoming Jiang, Tianyu Cao, Qingyu Yin, Yifan Gao, Zheng Li, Rahul Goutam, Haiyang Zhang, et al. 2022. Query attribute recommendation at amazon search. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 506–508.

[9] Yiming Qiu, Chenyu Zhao, Han Zhang, Jingwei Zhuo, Tianhao Li, Xiaowei Zhang, Songlin Wang, Sulong Xu, Bo Long, and Wen-Yun Yang. 2022. Pre-training tasks for user intent detection and embedding retrieval in e-commerce search. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4424–4428.

[10] Kevin Roitero, Ben Carterette, Rishabh Mehrotra, and Mounia Lalmas. 2020. Leveraging behavioral heterogeneity across markets for cross-market training of recommender systems. In *Companion Proceedings of the Web Conference 2020*. 694–702.

[11] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[12] Michael Skinner and Surya Kallumadi. 2019. E-commerce Query Classification Using Product Taxonomy Mapping: A Transfer Learning Approach.. In *eCOM at SIGIR*.

[13] Anna Tigunova, Ghadir Eraisha, and Ezgi Akcora. 2025. sQuIrRel: Large-scale evaluation of e-commerce query classification models. In *The Web Conference*.

[14] Junhao Zhang, Weidi Xu, Jianhui Ji, Xi Chen, Hongbo Deng, and Keping Yang. 2021. Modeling Across-Context Attention For Long-Tail Query Classification in E-commerce. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 58–66.

[15] Lvxing Zhu, Hao Chen, Chao Wei, and Weiru Zhang. 2022. Enhanced representation with contrastive loss for long-tail query classification in e-commerce. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*. 141–150.

[16] Lvxing Zhu, Kexin Zhang, Hao Chen, Chao Wei, Weiru Zhang, Haihong Tang, and Xiu Li. 2023. HCL4QC: Incorporating Hierarchical Category Structures Into Contrastive Learning for E-commerce Query Classification. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 3647–3656.