Technical Report Draft
E-commerce Product Search Competition

DcuRAGONS

September 11, 2025

1 Introduction

Over the past two years, the rapid advancement of large language models (LLMs) has fundamentally reshaped the landscape of information retrieval (IR). By leveraging their superior natural language understanding and encapsulated global knowledge, LLMs are increasingly integrated into search pipelines, enhancing components such as query intent understanding, query rewriting, retrieval and ranking, and the automation of manual annotation workflows. Industry adoption of LLM-powered search solutions has accelerated, with tangible improvements observed in both relevance metrics and user engagement indicators.

This competition is designed to accelerate research and practical applications of LLMs in multilingual search technologies, with a particular focus on e-commerce domains. It addresses two core tasks: (1) the *Multilingual Query-Category Relevance Task*, which measures the semantic alignment between a user's search query and a candidate product category, and (2) the *Multilingual Query-Item Relevance Task*, which evaluates whether a multilingual query is relevant to a specific product listing. Both tasks are formulated to reflect real-world challenges in e-commerce environments, where queries are often noisy, ambiguous, or code-mixed, and where users expect accurate and efficient retrieval across diverse languages.

The problem setting is motivated by Alibaba AIDC's global e-commerce platforms, including AliExpress, Lazada, Daraz, Trendyol, and Mirivia. Collectively, these platforms serve millions of users in more than 100 countries, spanning over 20 languages. In such settings, the e-commerce search engine is a critical component of the user experience, enabling effective navigation of vast product catalogs. To support this competition, datasets have been curated from real-world multilingual search logs, with annotations provided

by domain experts. However, due to the wide linguistic diversity, some languages lack sufficient expert annotators, making it essential to develop models with strong generalization capabilities that can perform well on unseen languages without dedicated training data.

A unique focus of this competition is the Query-Category Relevance Task, which requires mapping user queries to hierarchical product category paths (e.g., "Electronics > Audio Devices > Headphones"). This process is crucial in the early stages of retrieval, filtering irrelevant categories and improving downstream relevance. For example, a query such as "wireless noise-canceling earbuds" should correctly map to the "Headphones" category. Achieving high accuracy in this task is non-trivial due to several challenges: (i) queries may come from multiple languages, including low-resource ones unseen during training; (ii) real-world queries often contain noise such as misspellings, abbreviations, or mixed-language tokens; and (iii) the hierarchical nature of product categories requires models to capture both fine-grained semantics and structural consistency.

In summary, this competition provides a challenging yet practical benchmark for advancing multilingual LLM-based search solutions in e-commerce. By tackling noisy, multilingual queries and requiring models to generalize beyond seen languages, the competition aims to push forward the development of robust and deployable multilingual search technologies.

2 Our methodology

2.1 Data Preparation

Data splitting A key aspect of our methodology lies in the data cross-validation splitting strategy. We split data into three folds, using stratified splitting on the main category path. To ensure a reliable validation setup and avoid information leakage, we enforced that identical queries do not appear in different folds. For QC data specifically, we balanced the distribution of category paths across folds while preventing highly similar paths from being split into different folds.

Data augmentation During preprocessing, we noticed highly noisy texts in the data, thus applied simple text cleaning process. Steps include: stripping redundant spaces, lowering cases, removing empty queries.

We also observed that concatenating query text in two languages enable better comprehension of the LLM, especially due to the category paths and item labels are in English. In particular, we augmented the original queries

with their English translation versions, then use hypens to concatenate them into a single query. We observed consistent outperformance of it over single-language representations, highlighting the benefit of multilingual context.

Exploratory Data Analysis To support the final decision of model selection, we did the brief survey about the given data in Appendix A. Based on the variety of languages, we focused on the families of two multilingual LLMs as Qwen (Qwen2.5 and Qwen3) and Gemma (Gemma2 and Gemma3), and a Transformer-based multilingual model XLM-RoBERTa.

2.2 Model Selection and Setup

Model Choice We experimented with a range of models of different scales to identify an effective balance between performance, multilingual coverage, and the competition's constraint of using models with at most 15 billion parameters. Initial experiments began with smaller models such as `XLM-RoBERTa` and Qwen2.5 with less than 7B params (Appendix B), which allowed us to quickly validate the implementation pipeline and test the impact of basic feature engineering techniques. Building on this foundation, we scaled up to larger multilingual LLMs, including Qwen2 and Qwen3 series, and the Gemma family of models. In practice, the performance trend observed was roughly ordered as: Qwen2 < Qwen3 < Gemma-2 9B < Gemma-3 12B.

For our best submissions, we focused on **Gemma-2-9B**, Qwen2-14B, and Qwen3-14B, as these models provided strong multilingual capabilities (supporting up to 29 languages) while satisfying the parameter limit. These choices allowed us to leverage both model scale and multilingual robustness, while ensuring compliance with competition requirements. A summary of tested models, including parameter size, multilingual support, and whether they were used in our final submission, is provided in Table **??**.

Model Name	# Parameters	Multilingual Support	Used in Best Submission
XLM-Roberta	550M	✗	✗
Qwen2	7B	✓(29)	✗
Qwen3	7B	✓(29)	✗
Gemma-2-9B	9B	✗	✓
Gemma-3-12B	12B	✓(140)	✓
Qwen2-14B	14B	✓(29)	✗
Qwen3-14B	14B	✓(29)	✗

Table 1: Summary of tested models

2.3 Training & Inference Methodology

Training Environments The model training experiments were conducted on a cluster equipped with four NVIDIA A100 GPUs (80GB each). We adopted the HuggingFace Transformers [1] library with DeepSpeed [2] integration to enable efficient distributed training, mixed-precision computation, and memory optimization. Models were fine-tuned for five epochs using cross-entropy loss, with parameter-efficient fine-tuning applied via Low-Rank Adaptation (LoRA).

For optimization, we employed the AdamW optimizer with a learning rate of $1e^{-5}$, batch size of 8, and linear learning rate scheduling. Gradient accumulation was enabled to effectively increase the batch size under GPU memory constraints. Hyperparameters were tuned empirically based on validation performance across folds. External datasets were not used in this iteration of our experiments, ensuring compliance with competition rules.

Inference Settings During inference, we adopted an ensemble strategy across cross-validation folds. Specifically, we trained models on three folds and performed inference on the test set using each of the trained models. Final predictions were obtained by averaging the softmax probabilities across models, which consistently yielded more stable and accurate results compared to single-model predictions.

2.4 Evaluation Results

Evaluation Metrics Both competition tasks are formulated as binary classification problems, with labels indicating whether a query is relevant or not relevant. Model performance on the test set is measured using the F_1 score of the positive class (relevant). The final evaluation metric, F_1^{avg}, is defined as the average of the F_1 scores across the two tasks: Query-Category Relevance and Query-Item Relevance. Formally,

$$F_1^{avg} = 0.5 \times F_1^{\mathrm{QC}} + 0.5 \times F_1^{\mathrm{QI}},$$

where F_1^{QC} and F_1^{QI} represent the positive-class F_1 scores on the Query-Category and Query-Item tasks, respectively. For a given task, the F_1 score is computed as

$$F_1 = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}},$$

[1] https://huggingface.co/

[2] https://www.deepspeed.ai/

with Precision and Recall defined with respect to the positive (label = 1) class. This metric emphasizes the model's ability to balance precision and recall on relevant predictions, which is critical for practical deployment in multilingual e-commerce search.

Architecture	**CA**	**CV**	**2lang**	**QC F_1 %**	**QI F_1 %**
XLM-Roberta	✗	✗	✗	$82_{\pm 0.801}$	$80.52_{\pm 1.287}$
Qwen3-0.6B	✗	✓	✗	-	78.50
Qwen3-14B	✗	✓	✗	88.38	-
Gemma-2-9B	✓	✓	✓	$88.83_{\pm 0.099}$	$88.20_{\pm 0.21}$
Gemma-3-12B	✓	✓	✓	$89.14_{\pm 0.152}$	-
Best	✓	✓	✓	**89.36**	**88.45**

Table 2: Comparison of methods across different architectures and training strategies. Results are calculated on the public testset, report average and deviation across cross-validation folds. **CA**: category path augmentation; **CV**: stratified + category-aware split; **2lang** = English translation + concatenation of queries

Ablation Results Our experiments, as shown in Table 2, highlight several key insights regarding model performance and design choices. First, smaller architectures such as XLM-Roberta (QC: 82.0%, QI: 80.5%) and sub-billion parameter Qwen variants (e.g., Qwen3-0.6B with QI: 78.5%) provided useful baselines but consistently underperformed compared to larger multilingual LLMs. This confirms that lightweight models, while attractive for efficiency, lack the representational capacity required to capture the complexity of multilingual, noisy queries in this competition setting.

Second, multilingual pretraining proved critical. Models such as Qwen3-14B (QC: 88.4%) and Gemma series demonstrated clear gains over monolingual or limited-language models. This finding aligns with the multilingual diversity of the dataset and supports our augmentation strategy of translating queries into English and concatenating them, which consistently improved performance across both QC and QI tasks.

Thirdly, model size was positively correlated with accuracy, though with diminishing returns. Scaling from XLM-Roberta to Gemma-2-9B improved QC performance by more than six points (82.0% → 88.8%), but further scaling to Gemma-3-12B yielded only a marginal increase (88.8% → 89.1%). This trend suggests that while scale contributes significantly to handling complex multilingual semantics, further improvements may require architectural innovations or task-specific pretraining rather than brute-force scaling.

Finally, ensembling across folds consistently delivered the best results. Our final submission achieved QC: 89.36% and QI: 88.37% by averaging predictions across multiple models, surpassing any individual run. The averaging of softmax probabilities reduced variance and mitigated overfitting to specific folds, reinforcing ensemble strategies as a robust choice for multilingual search tasks.

3 Discussion & Analysis

The source code for our project is published at `https://github.com/nhtlongcs/e-commerce-product-search`

Limitations Despite these positive results, several limitations remain. First, we did not fully explore unified modeling approaches where a single model jointly learns Query-Category and Query-Item tasks. Such multitask training could potentially improve generalization and efficiency but was constrained by resource limitations. Second, while our augmentation strategy of translation and concatenation proved effective, it may introduce noise in low-quality translations, and we did not systematically evaluate the impact of translation errors. Third, our models were not explicitly optimized to exploit the hierarchical structure of category paths, which could provide an additional inductive bias to improve QC predictions. Finally, computational constraints limited the extent of our hyperparameter sweeps, leaving open the possibility that further tuning could yield additional gains.

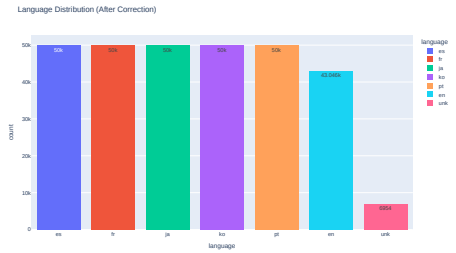
Future Works Building on these insights, we identify several promising directions for future research. A natural extension is to develop a unified framework for joint modeling of QC and QI tasks, allowing for shared representations and potential cross-task transfer benefits. In addition, integrating the structure of the category taxonomy into the modeling process could restrict invalid predictions and improve interpretability. We also envision exploring more sophisticated data augmentation techniques, such as back-translation or paraphrasing, to further address multilingual imbalance. Finally, investigating task transferability—for example, training on QC and testing on QI—could shed light on shared semantic features across tasks and inform more efficient training paradigms.

4 Conclusion

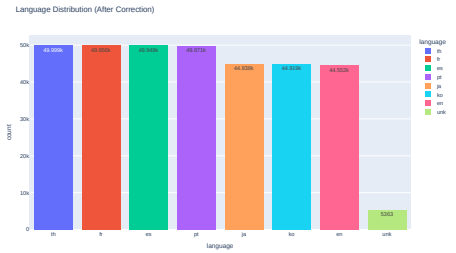
Our experiments demonstrate that careful data handling, translation-based augmentation, and parameter-efficient fine-tuning on strong multilingual LLMs are keys to achieve robust performance in noisy, multilingual e-commerce search tasks. Larger models such as Gemma-2-9B and Gemma-3-12B proved most effective, and ensemble inference further stabilized results across folds. While promising, our approach leaves open directions for improvement, including unified multitask modeling, leveraging category taxonomies, and exploring richer augmentation strategies to better handle unseen languages. These steps offer potential to build even more resilient and efficient search systems for real-world deployment.

A Exploratory Data Analysis

The datasets cover multiple languages, with the distribution shown in Figure 1. This highlights linguistic diversity and multilingual imbalance, motivating our use of multilingual models such as Qwen.



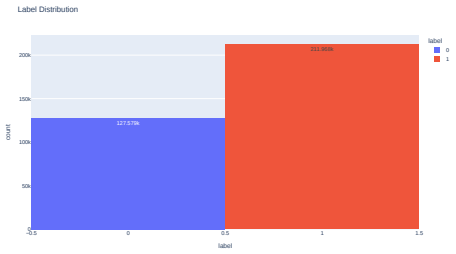
(a) QC language distribution after correction.



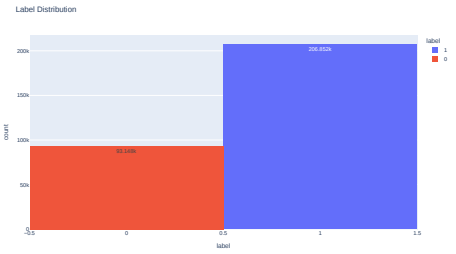
(b) QI language distribution after correction.

Figure 1: Language distribution across datasets.

The label distributions for both QI and QC are shown in Figure 2. We observe a significant imbalance, with relevant samples (label 1) underrepresented. This motivates stratified splitting to ensure fair training and evaluation.



(a) Label distribution in QI.

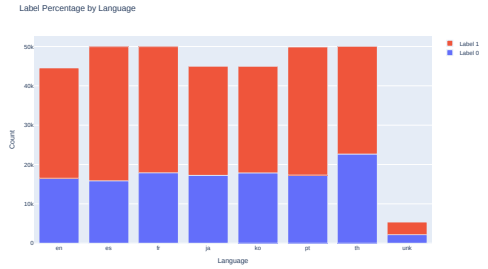


(b) Label distribution in QC.

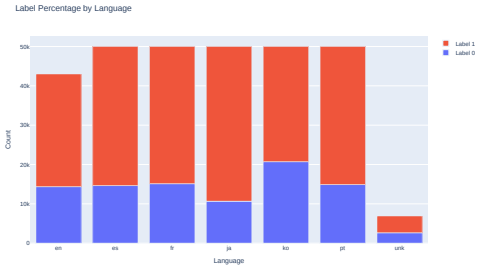
Figure 2: Label distributions across datasets.

Figure 3 illustrates label distribution across languages, highlighting the need for translation and concatenation augmentation, and justifying the use of multilingual models.

We observe in Figure 4 that relevant queries may have different length distributions, suggesting that query concatenation can provide more context, especially for short queries.

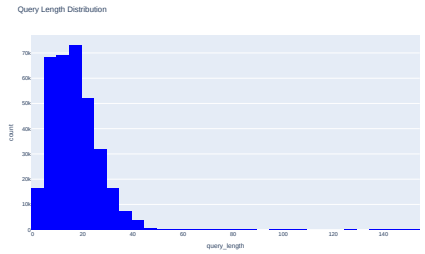


(a) QI: label distribution per language.

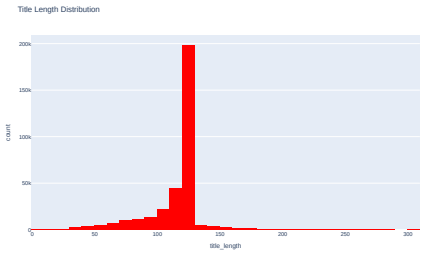


(b) QC: label distribution per language.

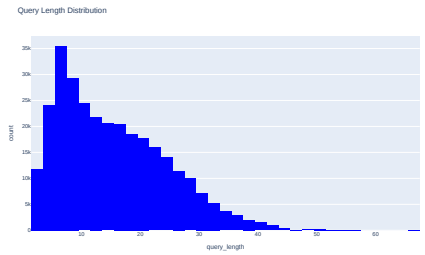
Figure 3: Label distribution across languages.



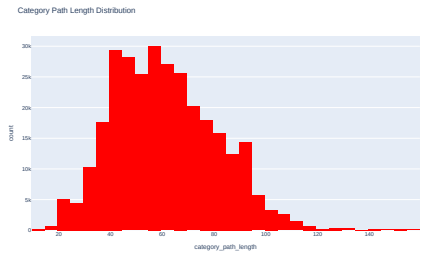
(a) QI: query length distribution.



(b) QI: item title length distribution.



(c) QC: query length distribution.

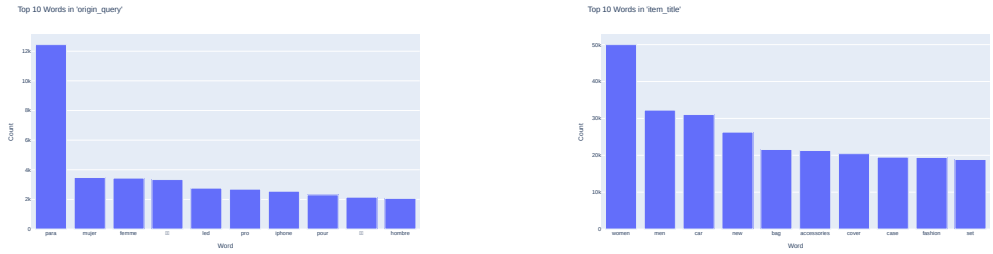


(d) QC: category path length distribution.

Figure 4: Length distributions across datasets.

Comparing the overlap between top query words and titles/categories motivates the use of lexical similarity features such as Jaccard similarity, and highlights the need for semantic models.

Overall, these analyses motivate our use of multilingual models, translation-based augmentation, and semantic similarity features in our approach.



(a) Top 10 words in QI queries.

(b) Top 10 words in QI item titles.

Figure 5: Top lexical features in QI.

B Small Model Performance

Before the final selection of $\geq$ 7B-parameter models, we benchmarked the available ones having less than 7B parameters to optimise the hyperparameter values. We selected two small versions of Qwen2.5 with 0.5B and 1.5B, and changed the setting in the given baseline code. When the model is applied a function to detect all the linear modules instead of only applying LoRA on the o-q-k-v modules, we observed that it gains the better results in Table 3. We hypothesised that applying LoRA only on o-q-k-v adapts *attention patterns* (where to look), while applying it on all linear modules also adapts *feature transformations* (how to think), giving stronger overall results.

Model	**LoRA**	**QC F_1 %**	**QI F_1 %**
Qwen2.5-0.5B	o-k-q-v	82.41	81.47
	Linear	83.34	82.33
Qwen2.5-1.5B	o-q-k-v	83.19	82.33
	Linear	83.32	82.81

Table 3: Results when applying LoRA on Linear modules and o-k-q-v modules (baseline). Blue results indicate the best ones in this comparison setting.