A Data-Centric Approach to Multilingual E-Commerce Product Search: Case Study on Query-Category and Query-Item Relevance

Yabo Yin
yinyabo22@outlook.com
No Affiliation
China

Yang Xi
23140666@bjtu.edu.cn
Beijing Jiaotong University
China

Jialong Wang
wjl_906@xauat.edu.cn
Xi'an University of Architecture and Technology
China

Shanqi Wang
19917805190@163.com
Harbin University of Science and Technology
China

Jiateng Hu
1178183680@qq.com
Cangzhou Jiaotong College
China

Alibaba International Tech

Reported for CIKM AnalytiCup 2025 Competition Workshop
Reported by Yabo Yin(几小只)
Nov 14, 2025

Introduction

The Multilingual Search Challenge in Global E-Commerce

Platforms like AliExpress serve millions of users across languages — yet achieving high-quality, cross-lingual search relevance remains a major challenge.

Core Tasks:

Query-Category (QC) & Query-Item (QI) Relevance.

Key Challenges:

Language Imbalance & Cold Start

In new markets or for uncommon languages, we often start with very limited training data.

Data Quality & Model Bias

Noisy labels and class imbalance are common in real-world datasets.

LLMs Are Not Magic

Marginal gains from large-scale LLMs are offset by their high deployment costs and latency.

Motivation and Insight

Leverage, Don’t Modify, the Base LLM

Powerful & Ready-to-Use

Modern foundational models are already powerful,general-purpose engines out-of-the-box.

High Cost of Modification

Architectural changes increase risk, reduce portability, and harm maintainability.

Higher ROI Strategy

Focus on Data & Fine-Tuning
Proven by success stories like Qwen3-Embedding/Reranker[1].

Key Techniques for Robust Data & Fine-Tuning

Translation-based Augmentation: Mitigates language imbalance and cold-start issues.

Self-validation Filtering & Semantic Hard Negative sampling: Improves data quality and reduces model bias.

[1] Zhang Y, Li M, Long D, et al. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models[J]. arXiv preprint arXiv:2506.05176, 2025.

Method

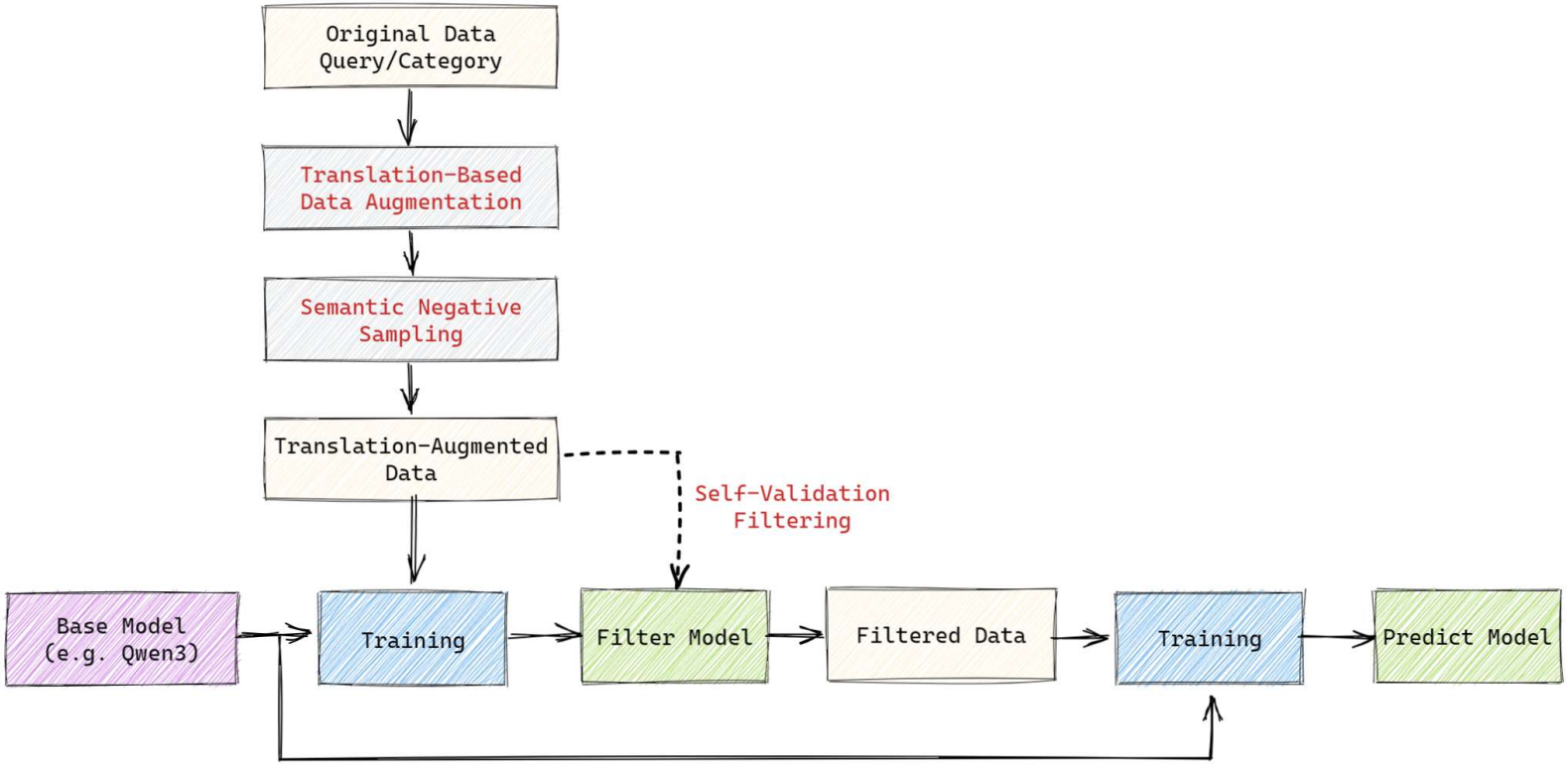
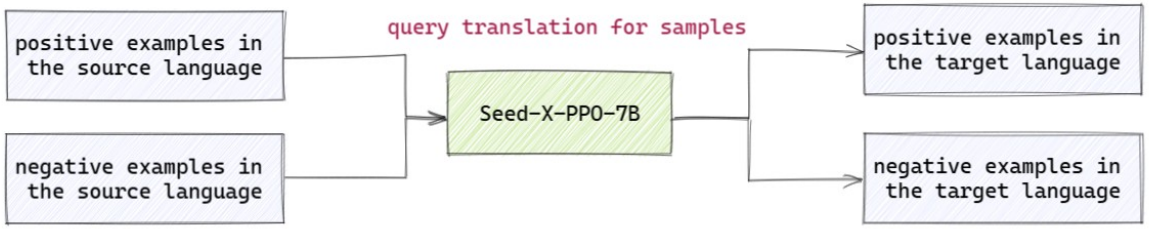
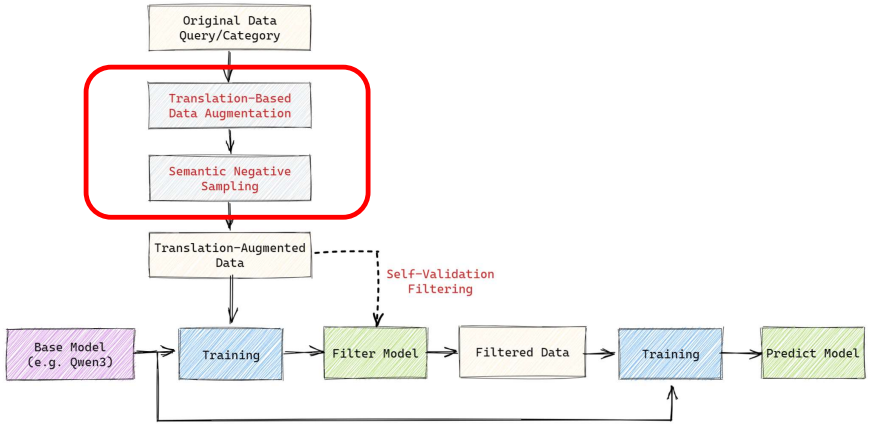


Figure 1: Overview of the proposed data-centric framework for the Query-Category (QC) relevance task.

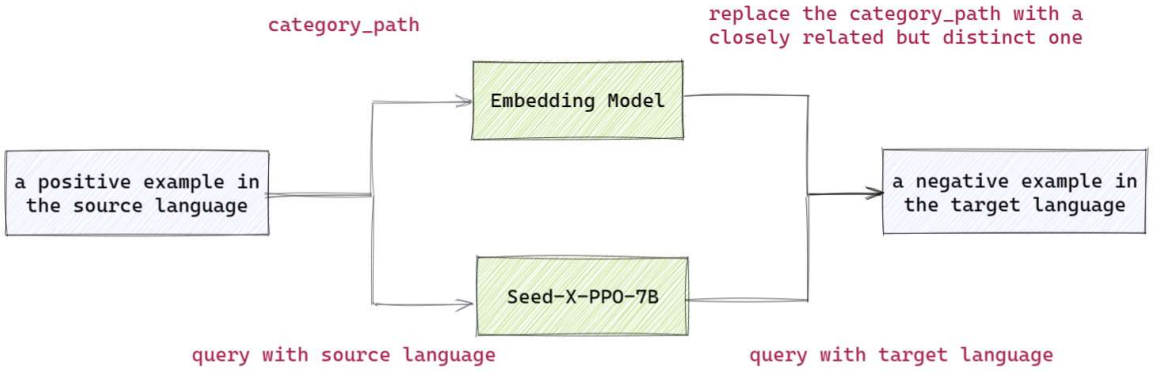
Method

Translation-based Augmentation



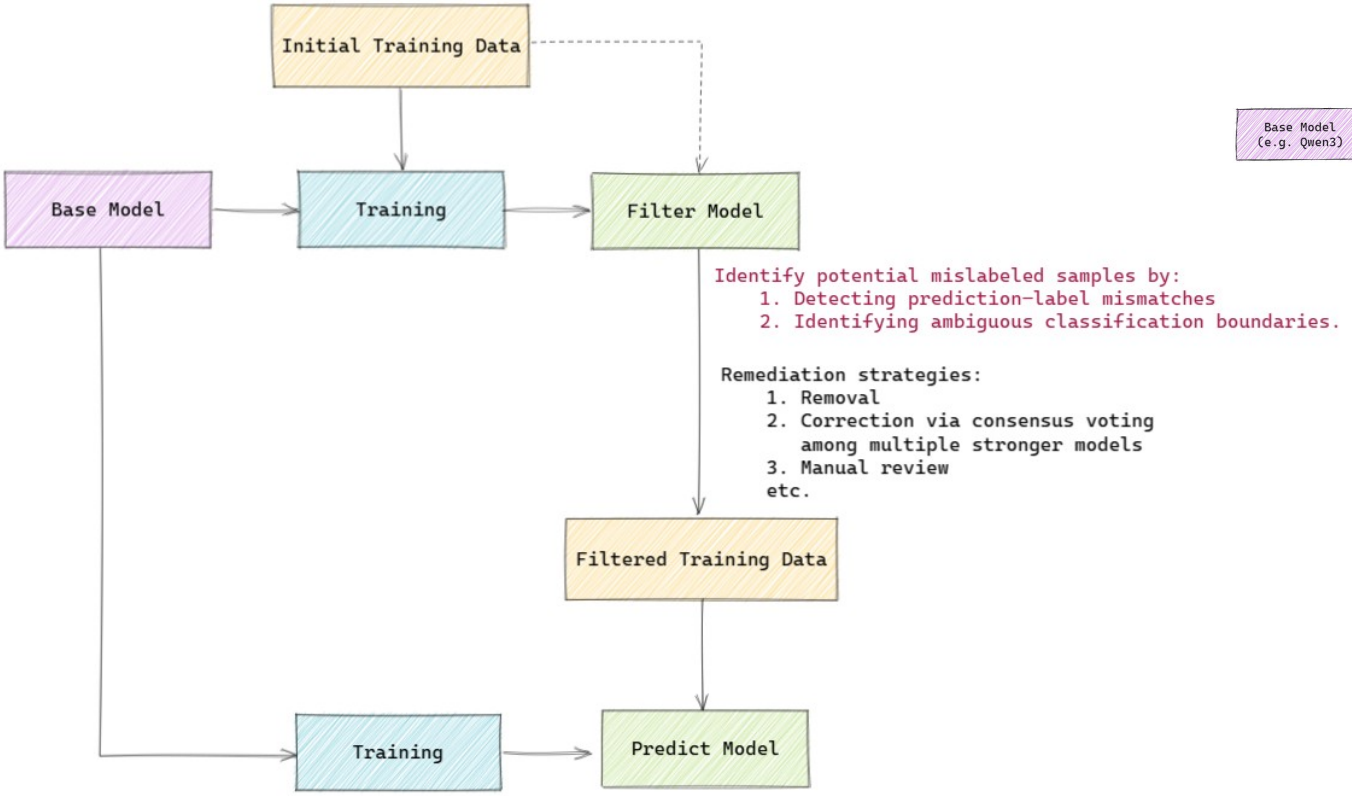


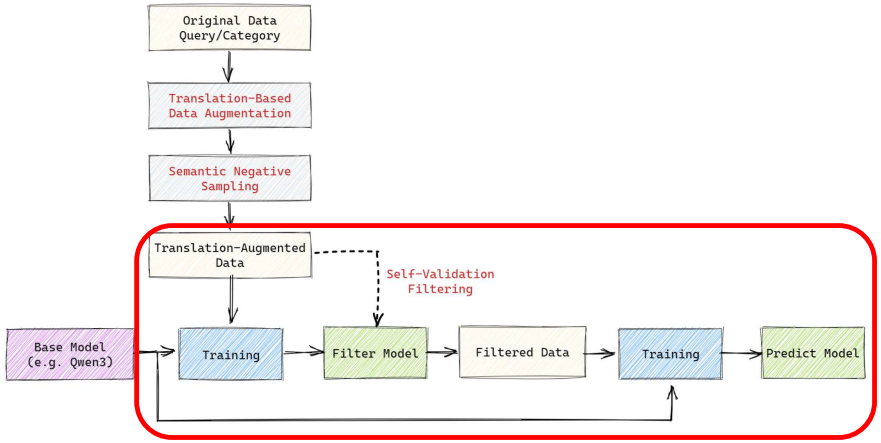
Semantic Hard Negative Sampling



Method

Self-validation Filtering

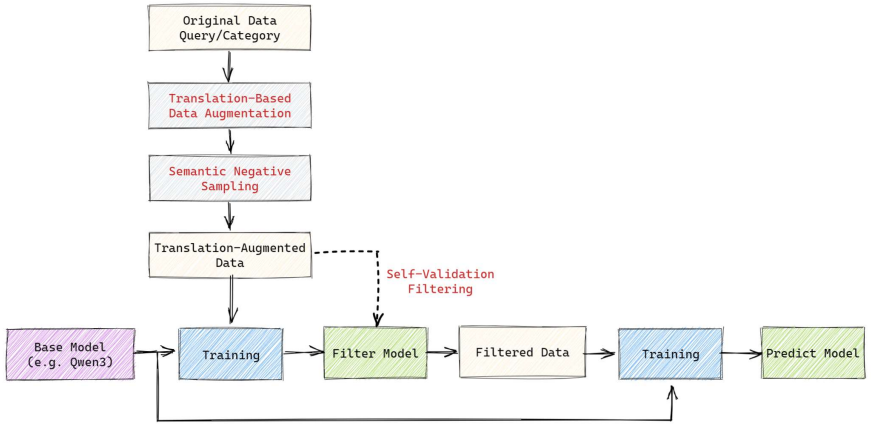




Method

Model Training

$$\mathcal{L} = -\log p(l \mid P(q, c)), \tag{1}$$



Model Inference

$$p(\text{yes}) = \frac{e^{p(\text{yes})}}{e^{p(\text{yes})} + e^{p(\text{no})}} \tag{2}$$

A task-specific threshold is then applied to this probability to obtain the final binary prediction:

$$\text{pred} = \begin{cases} \text{yes}, & \text{if } p(\text{ yes }) \geq \text{ threshold} \\ \text{no}, & \text{if } p(\text{ yes }) < \text{ threshold} \end{cases} \tag{3}$$

Experiments

Table 1: Implementation Details

Setting	Value
Base Model	Qwen/Qwen3-14B-Base
LoRA Rank	32
LoRA Dropout	0.1
Infer Framework	vLLM
Threshold	QC = 0.4, QI = 0.2

Table 2: Dataset Statistics for QC and QI Tasks

Task		English	French	Spanish	Korean	Portuguese	Japanese	German	Italian	Polish	Arabic	Thai	Vietnamese	Indonesian
QC	Train	50k	50k	50k	50k	50k	50k	–	–	–	–	–	–	–
	Dev	10k	10k	10k	10k	10k	10k	10k	10k	10k	10k	–	–	–
	Test	10k	10k	10k	10k	10k	10k	10k	10k	10k	10k	–	–	–
QI	Train	40k	40k	40k	45k	40k	45k	-	-	-	-	40k	-	-
	Dev	10k	5k	5k	5k	5k	5k	5k	5k	5k	5k	5k	5k	5k
	Test	10k	10k	10k	10k	10k	10k	10k	10k	10k	10k	10k	10k	10k

Experiments

Table 3: Performance comparison of different data-centric strategies on the dev set (preliminary competition phase). Results are reported in F1 score for Query-Category (QC) and Query-Item (QI) relevance tasks. Best results are in bold.

Base Model	Training Data Type	Method	TASK	
			QC	QI
Qwen3-8B	Original	SFT	0.8733	0.8537
Qwen3-8B	Original	DPO	0.8765	-
Qwen3-14B	Original	SFT	0.8760	0.8662
Qwen3-14B	Augmented	SFT + **Translation Augmentation**	0.8834	0.8738
Qwen3-14B	Augmented	SFT + **Translation Augmentation + Semantic Negative Sampling**	0.8885	-
Qwen3-14B	Augmented	SFT + **Translation Augmentation + Semantic Negative Sampling + Self-Validation Filtering**	0.8873	-
Qwen3-14B	Augmented	SFT + **Translation Augmentation + Semantic Negative Sampling + Task-Specific Threshold**	**0.8919**	-
Qwen3-14B	Augmented	SFT + **Translation Augmentation + Self-Validation Filtering**	-	0.8799
Qwen3-14B	Augmented	SFT + **Translation Augmentation + Self-Validation Filtering + Task-Specific Threshold**	-	**0.8839**

Experiments

Table 4: Inference Configuration

Component	Specification
PyTorch	2.8.0
Transformers	4.55.0
vLLM	0.11.0
GPU	RTX 4090 * 1

Table 5: Inference speed on the two tasks: QC and QI.

Base Model	Task	Infer Speed
Qwen3-14B-base	QC	35.9 it/s
	QI	35.6 it/s

Experiments

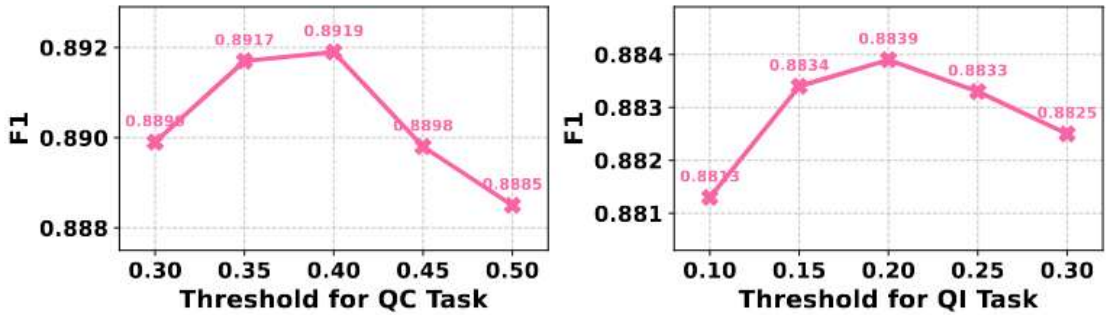


Figure 2: Performance comparison of QC and QI tasks across different prediction thresholds.

Table 6: Final Online Competition Results (Top 3 Teams)

Rank	Team	Overall	QC	QI
1	DcuRAGONS	**0.8931**	**0.8965**	**0.8897**
2	Industry_AI	0.8889	0.8928	0.8851
3	**Ours**	0.8865	0.8896	0.8833

Future Work

Spelling-Robust Data Augmentation

We noticed that many real-world queries contain typos or slang. Using training data with realistic typos and slang to build more robust models.

Human-in-the-Loop Data Refinement

Building more refined data correction pipelines is promising. Mislabeled samples are often high-value.It is inefficient to simply delete mislabeled samples, as it overlooks their significant value.

Task-Specific Optimization Methods

Exploring training objectives and loss functions that are specifically designed for QC/QI task characteristics.

Thanks

Contact: yinyabo22@outlook.com

Alibaba International Tech