LLM-Augmented Machine Translation for Scalable, Context-Aware Cross-Lingual E-Commerce Search

Nicole McNabb[1,*], Dayron Rizo-Rodriguez[1], Jesus Perez-Martin[1], Yuanliang Qu[1], Clement Ruin[1], Alina Sotolongo[1], Pankaj Adsul[1] and Leonardo Lezcano[1]

[1]*Walmart Global Tech, Sunnyvale, CA, USA*

Abstract

E-commerce search in the US and Canada presents a unique opportunity for Cross-Lingual Information Retrieval (CLIR), allowing non-English-speaking customers to benefit from English-language search systems. Machine Translation (MT) enhances search performance by translating customer queries into English before processing them. However, traditional MT systems face challenges in this domain, including polysemy, high latency, limited contextual information in queries, and the presence of non-translatable entities such as brand names, making generic MT approaches suboptimal. We present a scalable three-step MT system for CLIR that delivers precise, context-aware translations for multilingual users across markets. First, we construct an LLM-powered Translation Memory that leverages product and search session data to generate accurate translations for context-scarce queries and those with non-translatable entities such as brand names. Second, we show the effectiveness of customizing translatability by language and locale. Third, we introduce an 8-bit quantized Neural Machine Translation (NMT) model enhanced with an LLM-driven contextual rule engine, achieving 3x higher throughput, 40%+ lower latency, and 58% lower inference cost than previous NMT approaches without compromising translation quality. Deployed to *www.walmart.ca/fr*, our system shows a statistically significant increase in customer conversion rate, +8.2% weighted nDCG, and +3.3% precision in search results compared to monolingual search.

Keywords

Cross-lingual Search, Large Language Models, Entity-Aware Translation, Cross-lingual Ambiguity, Translatability, Neural Machine Translation (NMT), Integer Quantization

1. Introduction

There is a growing demand for B2C e-commerce search engines to address language barriers and cultural differences [1]. To improve query understanding, as well as search precision and recall [2, 3, 4, 5], recent approaches [6, 7] have explored automatic query translation as an early step in the search process. This strategy is especially important for platforms serving a global audience, where the ability to process a wide range of languages and cultural contexts is essential. This is particularly relevant for online stores and marketplaces in the US, where 13% of the population speaks Spanish as a first language [8], and Canada, where 22% of the population speaks French as a first language (primarily in Quebec) [9].

While physical stores allow customers to visually find products, navigating an e-commerce site often requires proficiency in the store's native language. For example, 38% of Québécois citizens speak only French [9] and may struggle to shop on English-only e-commerce sites. Additionally, 40% of customers avoid purchasing from websites not in their native language [1]. These findings highlight the importance of multilingual support in modern e-commerce search engines to broaden market reach.

Cross-Lingual Information Retrieval (CLIR) systems for e-commerce search often leverage Machine Translation (MT) to convert user queries into the search engine's language. However, traditional MT

ECOM'25: SIGIR Workshop on eCommerce, Jul 17, 2025, Padua, Italy

*Corresponding author.

nicole.mcnabb@walmart.com (N. McNabb); dayron.rizo.rodriguez@walmart.com (D. Rizo-Rodriguez); jesus.perez-martin@walmart.com (J. Perez-Martin); yuanliang.qu0@walmart.com (Y. Qu); clement.ruin@walmart.com (C. Ruin); alina.sotolongo@walmart.com (A. Sotolongo); pankaj.adsul@walmart.com (P. Adsul); leonardo.lezcano@walmart.com (L. Lezcano)

0009-0006-7951-2720 (N. McNabb); 0009-0004-9607-350X (D. Rizo-Rodriguez); 0000-0002-5719-5043 (J. Perez-Martin); 0009-0007-3516-1399 (Y. Qu); 0009-0003-5224-4648 (C. Ruin); 0000-0002-0522-6127 (A. Sotolongo); 0009-0002-6495-3970 (P. Adsul); 0009-0001-2664-2867 (L. Lezcano)

systems face high latency and domain-specific translation challenges such as cross-lingual ambiguity, regional and dialect variations, non-translatable entities like brands, and limited query context. These issues must typically be addressed differently for each language and locale.

We introduce an efficient, scalable MT system designed for cross-lingual e-commerce search that addresses these challenges across languages and markets. The key contributions of the system are:

1. An **LLM-powered Translation Memory**, the first offline use of large language models (LLMs) to resolve cross-lingual ambiguity and perform entity-aware translation for e-commerce search. This system is *context-aware*, integrating product catalog and user behavior data to improve translation quality.
2. **Language-Tuned Translatability** logic to effectively manage code switched queries (eg. Spanglish) and regional and dialect variations (eg. Québécois, Puerto Rican Spanish) across markets, with the flexibility to extend to new locales.
3. A **quantized Neural Machine Translation (NMT) model** for CPU-based inference that delivers cost-effective, scalable performance at sub-10ms latency without sacrificing translation quality.

We extend the system from Spanish search in the US to French search in Canada, validating our approach through end-to-end search improvements on *www.walmart.ca/fr*.

2. Related Work

Prior work in CLIR for e-commerce has leveraged MT to convert user queries into the search system's primary language [6, 7, 10]. To deliver translations at scale with low latency, Yao et al. [7] introduced an asynchronous strategy combining the speed of Statistical Machine Translation (SMT) online with the accuracy of NMT offline. Recently, several fast NMT frameworks have been developed [11, 12, 13]. Perez-Martin et al. [10] adapted the highly-optimized *Marian-NMT* framework [13] for synchronous e-commerce search in Spanish. We extend this work by quantizing the *Marian-NMT* model to enable faster, more cost-effective inference in production across markets. [1]

To improve contextual translations in e-commerce, Laenen and Moens [14] leveraged visual context from product images to improve quality of product description translations. Gao et al. [15] adapted LLMs using domain-specific tokenizer optimization and fine-tuning on product title corpora. However, as noted in Section 1, translating user queries poses additional challenges, such as cross-lingual ambiguity and identification of non-translatable entities, that neither work on product title or description translation addresses. These issues remain unexplored in LLM-augmented CLIR. We address them using user engagement signals and product catalog data to improve translation quality and to generate rules for handling non-translatable entities in NMT.

Prior work on adapting MT systems to multiple languages has focused on machine translation rather than translatability. Gupta et al. [16] proposed a cross-lingual decoder for low-resource language adaptation using incremental training. However, fine-tuning a language-specific NMT model remains more effective for medium- and high-resource languages. Moslem et al. [17] addressed stylistic variation in languages using LLMs with in-context learning, though this approach is unsuitable for low-latency applications. Perez-Martin et al. [10] built a dialect-sensitive lexicon from Wiktionary for efficient Spanish query detection, but its application to other languages remains unexplored. We experiment with this approach for Canadian French, confirming the importance of locale-specific lexicons and showing that detection logic must be tailored to each language due to varying degrees of English overlap and usage across locales.

3. System Architecture

We present an efficient multi-language, multi-locale query translation system that detects the source language of the query and returns its English translation for retrieval by the underlying search engine.

[1] Marian-NMT website: marian-nmt.github.io

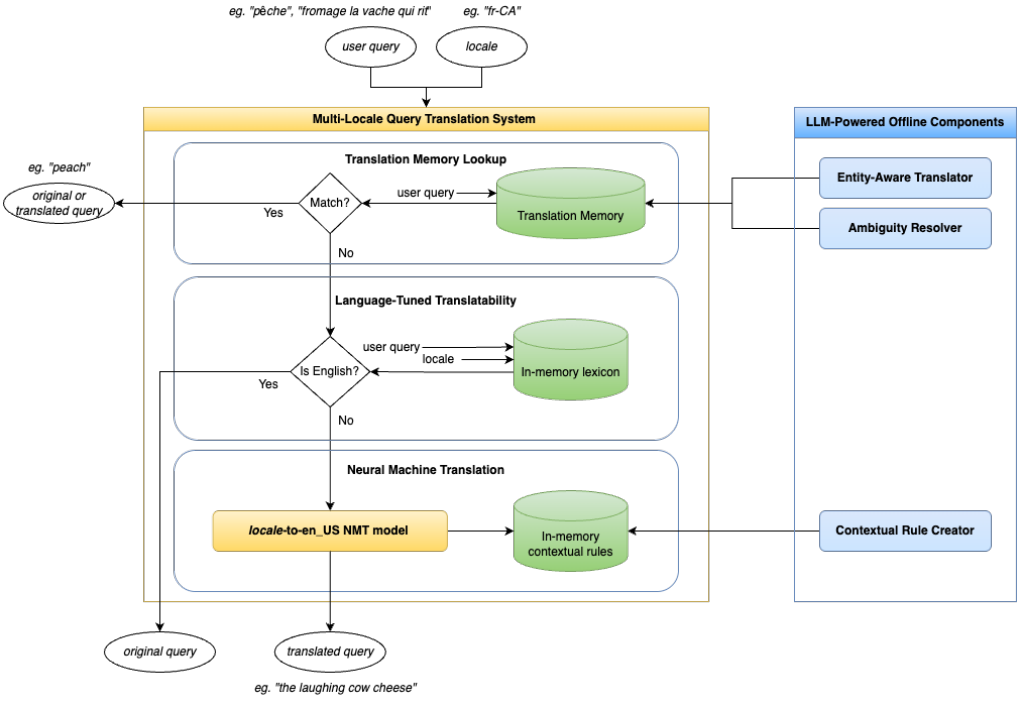

Figure 1: Proposed query translation system, including runtime flow (yellow) and offline components (blue).

Figure 1 shows the query translation system architecture.

The first backend module, as shown in Figure 1, is the Translation Memory, which caches translations of high-frequency queries. This enables low-latency lookups for repeated queries. Due to the heavy skew in search traffic, the Translation Memory can serve a large majority of requests. For reference, Yao et al. [7] report a 90% cache-hit rate in their MT system, and Perez-Martin et al. [10] report 80% for Spanish queries. In our case, French Canadian queries are further skewed, allowing a 100 MB translation memory to cover over 95% of search traffic.

We enhance the Translation Memory with LLM-generated translations, detailed in Section 4. While a domain-adapted prompt enables high translation accuracy overall, LLMs struggle with two query types: those containing rare or new *non-translatable entities* (e.g., the yogurt brand Liberté, which also means "freedom" in French), and ambiguous terms with multiple meanings (e.g., pêche, meaning either "peach" or "fishing"). To achieve high translation quality for these query types, we introduce two specialized LLM components: the Entity-Aware Translator (Section 4.1) and Ambiguity Resolver (Section 4.2).

If a query is not found in the Translation Memory, the Translatability module is triggered. It detects language using token-based analysis with a lexicon covering regional and dialect variations. Importantly, language detection logic is tailored to each language and locale. Section 5 details the specific modifications made for Canadian French and their impact.

The query is then passed to the Neural Machine Translation layer, which encompasses a fine-tuned *Marian-NMT* model supported by a fast contextual rule engine that handles non-translatable entities not yet learned by the NMT model. Traditionally, rule creation and validation is a time-consuming, manual process. To streamline this, we introduce the Contextual Rule Creator (Section 6.1), an LLM-based module that automates rule generation. Finally, Section 6.2 explains how we fine-tuned and quantized *Marian-NMT* for efficient CPU execution, ensuring fast, scalable translations without sacrificing quality.

4. Translation Memory

The Translation Memory can be populated using a domain-specific MT model or an LLM guided by a domain-specific prompt and optional in-context examples. We use GPT-4o (version 2024-05-13) with an

Table 1
Example entities, queries, and translations before and after using the Entity-Aware Translator.

Entity Name	Product Categories	Query	Translation Before	Translation After
Liberté	Yogurts	yogourt liberte lego statue de la liberte	freedom yogurt lego statue of liberty	liberte yogurt lego statue of liberty
Royale	Toilet Paper, Facial Tissue, Paper Towels	kleenex 3 epaisseur royale laine bernat bleu royale	kleenex 3-ply royal bernat royal blue yarn	royale 3-ply kleenex bernat royal blue yarn

e-commerce search-specific prompt including the following specifications:

1. Do not translate brands, product lines, model names, media titles, or other named entities.
2. The translation must be in English.
3. The most accurate translation is the most concise translation that completely preserves the query's original intent.
4. The translation must refer to a product. If the query is already in English or the translation does not refer to a product, return the original query as the translation.

We chose GPT-4o for its strong multilingual support and cost-effectiveness compared to competitors such as Claude Opus 3. As a result, the incremental cost of the LLM-based translations is less than 5% of the total system cost.

Evaluated on a sample of 5,500 popular, unique French queries with human-curated reference translations, the LLM-based approach achieves a significant BLEU [18] score improvement, from 69.8 using domain-adapted *Marian-NMT* to 97.7. This gain is primarily due to broader knowledge of entities like brands and books often seen by NMT models, and the ability to detect and implicitly correct grammar and spelling errors. Despite these strengths, GPT-4o still struggles with ambiguous queries and those involving unknown non-translatable entities. We introduce two LLM modules that address these shortcomings: the Entity-Aware Translator and the Ambiguity Resolver.

4.1. Entity-Aware Translator

The Entity-Aware Translator begins with a data pipeline that extracts structured entity data including brands, product lines, franchises, characters, sports teams, and media titles from the product catalog along with their associated product categories. This information is then matched with historical queries from the past year where the entity appears as a proper sub-string. The module supplies the LLM with three key pieces of information: the query, the identified non-translatable entity, and the product categories associated with the entity. The LLM is prompted to translate the query according to the earlier translation guidelines, with an added instruction: the entity must not be translated within the context of the given product categories. Because the module relies solely on product catalog data to contextualize queries, it can handle novel and long-tail entities mentioned in queries as long as they are present in the catalog. Table 1 illustrates examples of entity-aware translations.

Compared to the generic translation prompt (Section 4), this approach improved exact match translation accuracy by 2.6% and increased the BLEU score by 1.4 on a representative random sample of 3,000 unique queries containing non-translatable entities, as validated by professional linguists. As such queries comprise about 15% of French search traffic, this boosts overall translation accuracy by 0.4%.

4.2. Ambiguity Resolver

The LLM-based Ambiguity Resolver translates common but ambiguous queries like "gomme", "ballon", "trésor", or "pêche", where the intent cannot be determined from the query alone and there is insufficient context for the Entity-Aware Translator to apply. These queries may have multiple translations or refer to non-translatable entities. To resolve these ambiguities, the module combines LLM-based translation with product catalog data and session-level user behavior signals to infer the most likely intent.

Table 2
Example queries, session data including add-to-cart (ATC) actions, and translations with the Ambiguity Resolver.

Query	Candidate Translation	Session Reformulations	ATC Category	% of ATCs	Translation
gomme	eraser gum	gomme a effacer, gomme crayola gomme à mâcher excel, gomme sans sucre	Art Pencils Chewing Gum	40% 60%	gum
pêche	peach fishing	peche fruits, jus peche, peche en de peche sport, peche mouche	Juices Fishing Tackle Boxes	92% 8%	peach

The module first uses the product catalog to extract all queries linked to non-translatable entities (eg. product line ”Trésor”). For these and other single-token queries, it analyzes past session data including query refinements to determine the distribution of add-to-cart (ATC) actions across distinct product categories. The module then prompts the LLM to generate the most accurate translation of each query using three key pieces of context: the product categories customers engaged with, the percentage of ATC events linked to each category, and examples of query refinements. While this approach chooses the single most relevant translation for simplicity, in cases where ATC distribution is nearly uniform across categories, it may be preferable to blend search results from multiple plausible translations. Table 2 presents examples illustrating this workflow.

We find that 26% of French Canadian search traffic consists of potentially ambiguous single-token queries, thousands of those overlapping with known non-translatable entities. By leveraging session behavior, this approach selects translations that are more likely to drive customer engagement.

5. Language-Tuned Translatability

As shown in Figure 1, the Translatability component is customized for each language and locale. Perez-Martin et al. [10] showed that for Spanish, building a lexicon of language-specific terms, including regional and dialectal variants, from *wiktionary.org* and using lexicon lookups at runtime outperforms pre-trained language classifiers such as those proposed by Joulin et al. [19].

We find similar results for French in Canada. Lexicons derived from external sources are essential for capturing Québécois terms and translations. For example, “cartable” means “school bag” in France but “binder” in Canada; “espadrille” refers to a light shoe in France but a sport shoe in Canada; “bleuet” means “blueberry” in Canada but “cornflower” in France. Terms like “tuque” (winter hat) and “duo tang” (folder) are uniquely Canadian. We use *wiktionary.org* to develop a lexicon of 172k unique French Canadian terms for language detection.

In the US, queries from Hispanic users often include Spanish-English code switching (*e.g.*, ”cake de fresa”) [10]. To handle this, the language detection logic must tolerate partial Spanish queries. We achieve optimal translation performance by requiring approximately 30% of query tokens to appear in the Spanish lexicon. However, French Canadian queries are more linguistically consistent. Most containing a French token are either entirely French, include a non-translatable entity, or contain a word identical in both French and English. Extending the 30% threshold from Spanish misclassified 40.8% of French queries in a 10,000 GPT-4o-labeled query sample, hurting relevance. Instead, we found that classifying any query with at least one French token as French raised recall to 100% on the sample and improved BLEU from 80.5 to 82.5.

We also evaluated removing language detection entirely, relying on the fine-tuned *Marian-NMT* model to preserve the 18% of queries that are English or non-translatable entities. However, the model correctly preserves English queries only 80.6% of the time, introducing errors like altered numerical values (eg. ”058336173” to ”58336000”), verb tense shifts (eg. ”food weighing scale” to ”food weight scale”), and truncation of long queries (eg. ”bissel powergroom swivel rewind pets” to ”bissel powergroom swivel rewind p”). Thus, we conclude that robust language detection remains essential for maintaining translation quality using *Marian-NMT*.

Table 3
Corpus details. The average length, vocabulary size, and data split.

Source	Size	Queries (French)		Translations (English)	
		Avg. len.	Vocab.	Avg. len.	Vocab
In-Domain	3.6M	3.87	388,102	3.50	241,997
Out-of-Domain	1.2M	8.96	302,744	9.81	320,122
$\mathscr{D}$	**4.8M**	**4.53**	**284,584**	**3.79**	**241,617**
-train (70%)	3.4M	4.53	246,915	3.79	210,873
-validation (20%)	958K	4.53	144,982	3.79	125,766
-test (10%)	479K	4.53	105,212	3.79	91,181

6. Neural Machine Translation

To enable real-time query translation at scale, we require a lightweight and efficient NMT model. We use the Tiny.Untied [20] model architecture, which at only 16.9M parameters is well-suited for ultra-low inference latency. We fine-tune Fr-En Tiny.Untied on a bilingual parallel corpus combining in-domain French queries with LLM-generated English translations (Section 4) and out-of-domain data from the OPUS-MT benchmark [21] (Table 3). This joint training strategy helps the model learn both general-domain content and e-commerce-specific patterns, including terminology, entities, and code switched queries [22, 23, 24]. We evaluate model performance on BLEU and CHRF [25] computed on our held-out test set of 479k queries (Table 3), achieving a BLEU of 49.77 and CHRF of 72.37 (Table 6).

6.1. Contextual Rule Creator

While the LLM-powered Entity-Aware Translator (Section 4.1) enables high-quality offline translation of queries containing non-translatable entities by leveraging product catalog context, our lightweight production NMT model does not fully capture this entity-specific knowledge, especially for new and rare entities. To bridge this gap, we introduce the Contextual Rule Creator, a module that *distills learned LLM behaviors* into explicit, token-based rules applied during real-time inference.

Rather than relying solely on heuristic rules, the Contextual Rule Creator extracts patterns from LLM translations and codifies them into a structured decision framework. For each detected non-translatable entity, the system performs:

1. **Entity Context Extraction**: Collects historical queries containing the entity along with its associated product categories.
2. **Candidate Rule Generation**: Prompts the LLM to infer translation behaviors (translate vs. not-translate) conditioned on co-occurring tokens, capturing domain-specific nuances.
3. **Candidate Rule Validation**: Proposes structured rules, which are then validated using a lightweight multi-stage evaluation process (Section 6.1.1).

This process encodes the LLM's implicit entity knowledge in a form that the NMT model can apply during inference with minimal latency and cost.

While the contextual rule engine provides a necessary bridge today, it is a *transitional mechanism*. Our end goal is to retrain the NMT model directly on LLM-augmented datasets, progressively reducing the need for explicit rules. Meanwhile, this distillation approach allows the NMT system to immediately benefit from the improvements of the Entity-Aware Translator without costly daily retraining or added instability. Table 4 illustrates examples of LLM-proposed contextual rules.

6.1.1. Rule Validation and Deployment

The primary goal of the Contextual Rule Creator is to distill LLM translation behavior into lightweight rules for NMT, requiring a validation process that is fast, scalable, and minimally disruptive. Our approach balances the need for linguistic precision with the recognition that the LLM already captures the correct behavior in most cases. The rule validation and deployment process follows four stages:

Table 4
Example entities, queries, and contextual rules generated by the Rule Creator.

Entity	Categories	Queries	T Tokens	NT Tokens	Default
brut	Deodorants	cuir brut, miel brut, bois brut, brut deodorant	cuir, miel, bois	deodorant	T
liberte	Yogurts	yogourt liberte, liberte grec, kefir liberte, oeuf en liberte	oeuf	yogourt, grec, kefir	NT

Table 5
Latency of our domain-adapted NMT models at 50 QPS.

Inference engine	Latency (ms)		
	Average	p95	p99
1 T4 GPU	17.23	21.98	36.28
2 Intel Ice Lake CPUs	**9.9**	**12.63**	**15.74**

Step 1. Impact Simulation on Large Query Sample. Each rule is evaluated offline on a representative sample of 10 million pre-translated queries. The system computes the number and percentage of queries impacted by the rule, returns the list of impacted queries and outputs after applying the rule, and automatically promotes impacting rules.

Step 2. Alignment Check with LLM Behavior. Promoted rules are evaluated by re-translating impacted queries using the LLM without the rule applied. This verifies that no rule introduces new behavior inconsistent with the LLM translation, and that a high percentage of LLM translations already conform to the rule's action (preserve or translate the entity). We find that well-formed rules align with the LLM output in over 90% of impacted queries.

Step 3: Targeted Linguist Review Trained linguists review each rule using a lightweight UI that displays the rule logic, example queries with and without rule application, and the LLM translations. Each rule undergoes a 5–10 minute review to ensure it improves NMT inference without causing semantic drift. Rules are either approved, lightly adjusted, or rejected.

Step 4: Long-Term Rule Quality Monitoring. Deployed rules are continuously monitored for residual query impact after LLM translation and for cases of minimal ongoing impact, identifying candidates for retraining. Rules with persistent high residual impact or low relevance are re-evaluated. The ultimate goal is to transition knowledge distilled into rules back into model retraining, eliminating the need for manual intervention over time.

6.2. Scalable and Cost-Efficient NMT Deployment

Expanding to new regions requires replicating NMT instances, but GPU-based inference is expensive and sensitive to traffic surges. To address scalability and cost constraints, we leverage 8-bit quantization via the Marian-NMT [26] toolkit, deploying the fine-tuned, quantized model to CPU.

We test this configuration on a dataset of 1M queries with token length ≤ 7, a constraint sufficient to cover 99.98% of French Canadian search queries. As shown in Table 5, the int8-quantized model deployed on an instance with two Intel Ice Lake CPUs exhibits substantial performance improvements over the GPU-based setup. Additional outcomes include:

1. Throughput of up to 150 QPS per instance, a 3x increase compared to the GPU configuration
2. p99 latency of 27 ms under maximum load
3. 58% reduction in monthly NMT inference costs

Table 6 shows translation quality metrics across configurations. Importantly, quantization introduces negligible degradation in translation quality, while delivering substantial improvements in latency, scalability, and cost-efficiency.

Table 6
Evaluation of domain-adapted Tiny.Untied for French-English translation on our test set.

Model	BLEU	CHRF
Pre-trained	37	65
Fine-tuned	**49.77**	**72.37**
Fine-tuned + int8 quantization	**49.62**	**72.25**

7. Impact on Key Business Metrics

We deployed our CLIR system in Canada by integrating our MT system into the existing English search engine. To measure its impact, we benchmarked the CLIR system against a production baseline that retrieves search results directly from French-language product content without translation. We first evaluated translation quality and search relevance impact offline, then ran a two-week AB test to validate system performance on key business metrics.

For the offline evaluation, we extracted a representative random sample of 2,000 unique queries weighted by page impressions from Canadian French search traffic over three months post-model training. This sampling strategy increases the likelihood of including queries with unseen non-translatable entities, helping evaluate the ability of the system to handle cold-start scenarios. Our MT system achieved **90%** exact match accuracy in translation and a BLEU score of **82** on this sample, based on comparisons with reference translations provided by professional linguists.

To measure search relevance, for both control and treatment, bilingual human judges manually graded the relevance of the top 5 search results for each query on a 4-point scale depicted in Table 7. The evaluation showed **+8.2%** weighted nDCG and a **3.3%** increase in Relevant results under the CLIR system, both achieving statistical significance (p-value < 0.05).

Table 7
Relevance Evaluation Scoring

Score	Label	Description	Example: "black nike shoes"
1	Relevant	Fully matches query intent	A black Nike shoe
-1	Partially Relevant	Valid substitute but partial intent mismatch	A white Nike shoe
-2	Irrelevant	Unrelated to the query	A pair of Adidas socks
-3	Embarrassing	Clearly inappropriate result	A swimming pool

After validating the improvement in search relevance, we ran a two-week AB test. Users that issued at least one search on *www.walmart.ca/fr* qualified for the test. Half were randomly assigned to the baseline search experience, while the other half received the CLIR experience. The test revealed a statistically significant lift in conversion rate. We also observed significant reductions in zero-result pages and search abandonment rate, showing that the CLIR system improves both customer satisfaction and engagement for non-English-speaking customers.

8. Conclusion

In this paper, we presented a machine translation system for cross-lingual e-commerce search that delivers high translation quality, performance, and scalability across languages and markets. Our key contributions include the first offline use large language models to resolve cross-lingual ambiguity and perform entity-aware translation in e-commerce search, language-tuned translatability logic to handle code switched and dialectal queries across regions, and a quantized neural machine translation model for low-latency, CPU-based inference that maintains translation quality while reducing cost. This work opens the door for future research into deeper LLM integration with low-latency, multilingual, and localized search applications.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (GPT-4) in order to: Grammar and spelling check, paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] K. Vashee, The impact of MT on the Global Ecommerce Opportunity, 2022. URL: https://blog.modernmt.com/the-impact-of-mt-on-the-global-ecommerce-opportunity/.

[2] A. Ahuja, N. Rao, S. Katariya, K. Subbian, C. K. Reddy, Language-agnostic representation learning for product search on e-commerce platforms, in: WSDM 2020 - Proceedings of the 13th International Conference on Web Search and Data Mining, Association for Computing Machinery, Inc, 2020, pp. 7–15. URL: https://doi.org/10.1145/3336191.3371852. doi:10.1145/3336191.3371852.

[3] H. Lu, Y. Hu, T. Zhao, T. Wu, Y. Song, B. Yin, Graph-based Multilingual Product Retrieval in E-Commerce Search, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers, Association for Computational Linguistics, Stroudsburg, PA, USA, 2021, pp. 146–153. URL: https://www.aclweb.org/anthology/2021.naacl-industry.19. doi:10.18653/v1/2021.naacl-industry.19.

[4] S. Mangrulkar, A. Bengaluru, I. M. Ankith S, I. Vivek Sembium, A. M. S, Multilingual Semantic Sourcing using Product Images for Cross-lingual Alignment, in: Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), volume 1, ACM, 2022, p. 11. URL: https://doi.org/10.1145/3487553.3524204. doi:10.1145/3487553.3524204.

[5] X. Zhang, K. Ogueji, X. Ma, J. Lin, D. R. Cheriton, Towards Best Practices for Training Multilingual Dense Retrieval Models (2022). URL: https://arxiv.org/abs/2204.02363v1. doi:10.48550/arxiv.2204.02363.

[6] Q. Hu, H.-F. Yu, V. Narayanan, I. Davchev, R. Bhagat, I. S. Dhillon, Query transformation for multilingual product search, in: SIGIR 2020 Workshop on eCommerce, 2020. URL: https://sigir-ecom.github.io/ecom2020/ecom20Papers/paper6.pdf.

[7] L. Yao, B. Yang, H. Zhang, W. Luo, B. Chen, Exploiting Neural Query Translation into Cross Lingual Information Retrieval, in: SIGIR eCom 2020, 2020. URL: https://arxiv.org/abs/2010.13659v1. doi:10.48550/arxiv.2010.13659.

[8] A. Flores, 2015, Hispanic population in the United States statistical portrait, 2020. URL: https://www.pewresearch.org/hispanic/2017/09/18/2015-statistical-information-on-hispanics-in-united-states/.

[9] C. Heritage, Some facts on the Canadian Francophonie, 2024. URL: https://www.canada.ca/en/canadian-heritage/services/official-languages-bilingualism/publications/facts-canadian-francophonie.html.

[10] J. Perez-Martin, J. Gomez-Robles, A. Gutiérrez-Fandiño, P. Adsul, S. Rajanala, L. Lezcano, Cross-lingual search for e-commerce based on query translatability and mixed-domain fine-tuning, in: Companion Proceedings of the ACM Web Conference 2023, 2023, pp. 892–898. URL: https://doi.org/10.1145/3543873.3587660. doi:10.1145/3543873.3587660.

[11] G. Klein, Y. Kim, Y. Deng, J. Senellart, A. Rush, OpenNMT: Open-Source Toolkit for Neural Machine Translation, in: Proceedings of ACL 2017, System Demonstrations, Vancouver, Canada, 2017, pp. 67–72. URL: https://aclanthology.org/P17-4012/.

[12] O. Kuchaiev, B. Ginsburg, I. Gitman, V. Lavrukhin, C. Case, P. Micikevicius, OpenSeq2Seq: Extensible Toolkit for Distributed and Mixed Precision Training of Sequence-to-Sequence Models, in: Proceedings of Workshop for NLP Open Source Software (NLP-OSS), Association for Computational Linguistics (ACL), 2018, pp. 41–46. URL: https://aclanthology.org/W18-2507. doi:10.18653/V1/W18-2507.

[13] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. H. K. Heafield, T. Neckermann, F. Seide,

U. Germann, A. F. Aji, N. Bogoychev, A. F. Martins, A. Birch, Marian: Fast Neural Machine Translation in C++, in: ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations, Association for Computational Linguistics (ACL), 2018, pp. 116–121. URL: https://aclanthology.org/P18-4020. doi:10.18653/V1/P18-4020.

[14] K. Laenen, M.-F. Moens, Multimodal neural machine translation of fashion e-commerce descriptions, in: N. Kalbaska, T. Sadaba, F. Cominell, L. Cantoni (Eds.), Fashion Communication in the Digital Age. FACTUM 2019, Springer, 2019, pp. 46–57. URL: https://doi.org/10.1007/978-3-030-15436-3_4. doi:10.1007/978-3-030-15436-3_4.

[15] D. Gao, K. Chen, B. Chen, H. Dai, L. Jin, W. Jiang, W. Ning, S. Yu, Q. Xuan, X. Cai, L. Yang, Z. Wang, LLMs-based machine translation for e-commerce, Expert Systems with Applications 258 (2024) 125087. URL: https://doi.org/10.1016/j.eswa.2024.125087.

[16] K. K. Gupta, S. Chennabasavraj, N. Garera, A. Ekbal, Pre-training synthetic cross-lingual decoder for multilingual samples adaptation in E-commerce neural machine translation, in: H. Moniz, L. Macken, A. Rufener, L. Barrault, M. R. Costa-jussà, C. Declercq, M. Koponen, E. Kemp, S. Pilos, M. L. Forcada, C. Scarton, J. Van den Bogaert, J. Daems, A. Tezcan, B. Vanroy, M. Fonteyne (Eds.), Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, Ghent, Belgium, 2022, pp. 241–248. URL: https://aclanthology.org/2022.eamt-1.27/.

[17] Y. Moslem, R. Haque, J. D. Kelleher, A. Way, Adaptive machine translation with large language models, 2023. URL: https://arxiv.org/abs/2301.13294. arXiv:2301.13294.

[18] K. Papineni, S. Roukos, T. Ward, W. Zhu, BLEU: a method for automatic evaluation of machine translation, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (2002) 311–318. URL: http://dl.acm.org/citation.cfm?id=1073135. doi:10.3115/1073083.1073135.

[19] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of Tricks for Efficient Text Classification - ACL Anthology, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Association for Computational Linguistics, Valencia, Spain, 2017. URL: https://aclanthology.org/E17-2068/.

[20] N. Bogoychev, R. Grundkiewicz, A. F. Aji, M. Behnke, K. Heafield, S. Kashyap, E.-I. Farsarakis, M. Chudyk, Edinburgh's Submissions to the 2020 Machine Translation Efficiency Task, in: Proceedings of the Fourth Workshop on Neural Generation and Translation, Association for Computational Linguistics, Online, 2020, pp. 218–224. URL: https://aclanthology.org/2020.ngt-1.26. doi:10.18653/v1/2020.ngt-1.26.

[21] B. Zhang, P. Williams, I. Titov, R. Sennrich, Improving massively multilingual neural machine translation and zero-shot translation, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1628–1639. URL: https://aclanthology.org/2020.acl-main.148. doi:10.18653/v1/2020.acl-main.148.

[22] M. Dhar, V. Kumar, M. Shrivastava, Enabling Code-Mixed Translation: Parallel Corpus Creation and MT Augmentation Approach, in: Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing, 2018, pp. 131–140. URL: https://aclanthology.org/W18-3817/.

[23] D. Gautam, P. Kodali, K. Gupta, A. Goel, M. Shrivastava, P. Kumaraguru, CoMeT: Towards Code-Mixed Translation Using Parallel Monolingual Sentences, in: Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching, 2021. URL: https://aclanthology.org/2021.calcs-1.7/. doi:10.18653/v1/2021.calcs-1.7.

[24] A. Pratapa, M. Choudhury, S. Sitaram, Word Embeddings for Code-Mixed Language Processing, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 3067–3072. URL: https://aclanthology.org/D18-1344/. doi:10.18653/v1/D18-1344.

[25] M. Popović, chrF: character n-gram F-score for automatic MT evaluation, in: 10th Workshop on Statistical Machine Translation, WMT 2015 at the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015 - Proceedings, Association for Computational Linguistics (ACL), 2015, pp. 392–395. URL: https://aclanthology.org/W15-3049. doi:10.18653/V1/W15-3049.

[26] N. Bogoychev, R. Grundkiewicz, A. F. Aji, M. Behnke, K. Heafield, S. Kashyap, E.-I. Farsarakis,

M. Chudyk, Edinburgh's submissions to the 2020 machine translation efficiency task, in: A. Birch, A. Finch, H. Hayashi, K. Heafield, M. Junczys-Dowmunt, I. Konstas, X. Li, G. Neubig, Y. Oda (Eds.), Proceedings of the Fourth Workshop on Neural Generation and Translation, Association for Computational Linguistics, Online, 2020, pp. 218–224. URL: https://aclanthology.org/2020.ngt-1.26/. doi:10.18653/v1/2020.ngt-1.26.