Cost-Effective E-Commerce Catalog Translation at Scale Ensuring Named Entity Protection

Asier Gutiérrez-Fandiño[1], Jorge Yero Salazar[1], Clement Ruin[1], Alejandro Quintero-Roba[1], Shang Ravichandran[1], Jesus Perez-Martin[1], Pankaj Adsul[1], Suruchi Garg[1], Leonardo Lezcano[1]

[1]Walmart Translation Platform, Walmart Global Tech

Correspondence: leonardo.lezcano@walmart.com

Abstract

We present an enterprise-grade translation platform for global e-commerce that combines daily batch and real-time API pipelines with optimized T5-based models and a Reference Generator to enforce >99% non-translatable entity preservation. A linguist-driven rule engine and explainable evaluation framework (BLEU, COMET, and a custom e-commerce metric) enable continuous quality improvements. Deployed on GPU-accelerated inference servers and CPU-based processing nodes, our system processes millions of listings per day with sub-second latency and achieves 10×–100× cost savings over general-purpose LLMs for English→Spanish and English→French translation, all while version-tracking every update for robust enterprise rollouts.

1 Introduction

Global e-commerce platforms must deliver high-quality, contextually accurate multilingual content to drive market expansion and engagement: 7 out of 10 users always choose their native language (Nimdzi Insights LLC, 2023), 57% will abandon sites lacking local-language support, 71% distrust poor translations, and 69% find them hard to navigate (Nimdzi Insights LLC (2024, 2025)); automated translation is used by 75–82% of shoppers, though only 49% rate its quality as "good." Spanish, spoken by 65.2M Hispanics (>19% of the U.S.) of whom 75% report proficiency, represents a $3.78T purchasing-power market (Mark Hugo Lopez and Passel; Latino Donor Collaborative, 2024). In Canada, 22% of citizens identify French as their mother tongue and 29.1% as proficient in it, anchored by Quebec's C$ 552.7B GDP.[1][2]

Our platform processes over 500M requests per month, ranks among the top 70 most-visited sites worldwide, and sits in the top five global online retailers.[3] Our U.S. and Canadian catalogs each contain hundreds of millions of items with tens of thousands of daily updates. While large language models achieve state-of-the-art translation quality (Kudugunta et al. (2023); Cui et al. (2025); Team et al. (2022)), their inference latency and computational cost are prohibitive at this scale, and they often alter critical non-translatable entities. In this paper, we introduce a scalable, cost-efficient translation platform that matches or exceeds state-of-the-art quality and throughput, and reduces translation costs.

2 Related Work

Machine translation and multilinguality in e-commerce have spurred extensive research: language-agnostic embedding methods align queries and product descriptions into a shared vector space, yielding up to 23% relative F1 improvements (Ahuja et al., 2020), and combining token-, subword-, and character-level representations with re-ranking significantly boosts retrieval for German, French, and Japanese (Zhang and Tan, 2021), though such approaches often falter on specialized terminology and out-of-vocabulary issues. Translating queries into the store's primary language, exemplified by Walmart's Spanish→English NMT fine-tuned with engagement data, back-translation, and a translatability classifier, provided a +70 % nDCG lift and GMV gains (Perez-Martin et al., 2023), while empirical analysis of BLEU and chrF improvements highlighted diminishing returns for retrieval precision (Zhang and Misra, 2023). Unsupervised domain adaptation using only monolingual query data achieved 20 BLEU-point gains for Hindi–English (Kulkarni and Garera, 2022). On the modeling side, eBay's

[1]French Language in Canada: https://en.wikipedia.org/wiki/French_language_in_Canada

[2]Languages of Canada: https://en.wikipedia.org/wiki/Languages_of_Canada

[3]Semrush April 2025: https://www.semrush.com/website

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1148–1157
November 4-9, 2025

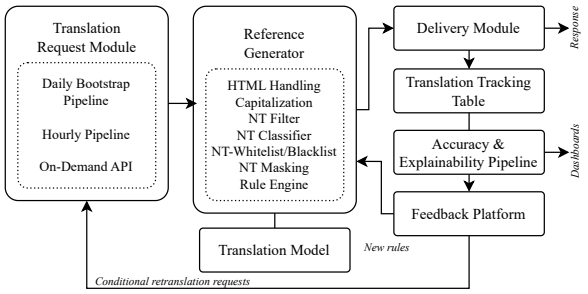

Figure 1: Translation Platform diagram from the translation flow perspective.

in-house LiLiuM family (1B/7B/13B) trained on a 3T-token corpus with custom tokenization and RedPajama-V2 data matched LLaMA-2 on English NLU and outperformed it on non-English and MT tasks (Herold et al., 2024; Weber et al., 2024; Touvron et al., 2023)), and their earlier cross-lingual search pipeline integrated user behavior and conversion metrics alongside traditional quality measures (Guha and Heger, 2014). Two-stage fine-tuning regimens such as G2ST and LEMT, leveraging self-contrastive semantic enhancement, annotated parallel corpora, lexicons, and domain-aligned bilingual resources, consistently outperform general-purpose NMT systems including GPT-4 on e-commerce benchmarks (Chen et al., 2024; Gao et al., 2024). More recently, hybrid translation–summarization and retrieval-augmented generation (RAG) techniques address cross-lingual product title brevity and specificity: length-aware filtering and open-source LLM evaluators (Zhang et al., 2024a), and few-shot inventory retrieval prompts have yielded up to 15.3 % chrF improvements on low-resource pairs (Zhang et al., 2024b; Popović, 2015).

3 Translation Platform

The Translation Platform combines large-scale batch workflows with a low-latency API to cover every use case and needs in our catalog translation. The Translation Platform diagram is shown in Figure 1.

3.1 Architecture

Our translation system comprises two modes of operation; batch processing and on-demand (API-based) translation, of which batch processing handles over 99% of all translations. Within the batch mode, we employ two distinct pipelines:

- **Bootstrap Pipeline**: Designed to populate the initial catalog, this pipeline runs once per day. Each execution selects the top N million most frequently accessed items that remain untranslated, processes them through our translation engine, and completes well before the next scheduled run.

- **Hourly Pipeline**: To keep our catalog up to date, this pipeline executes every hour, targeting listings that have been modified since the previous run. The hourly pipeline treats each product as an atomic unit, leveraging associated metadata and fields to produce higher-fidelity translations.

The on-demand translation API is invoked during the setup of new items. Whenever an item is submitted in English, the API generates translations in real time.

Translated items are first routed through our catalog processing pipeline and then recorded in a Translation Tracking Table (TTT), which stores key metadata such as the translation model's name and version.

In terms of scalability, our batch processing infrastructure, comprising both CPU-based and NVIDIA T4 cost-effective GPU-based nodes, scales with the current depth of the task queue. Conversely, scales in response to increases in HTTP request traffic or GPU utilization.[4]

3.2 Translation Models

Before selecting the models, we defined the translation task by identifying four distinct text categories:

- **Product title:** Up to 200 characters long, contain few verbs and often include non-translatable entities (e.g., brand, product line, model), American-style title capitalization, and a sequence of features without comma separation.[5]

- **Long text:** Multi-paragraph content that may include HTML markup, non-translatable entities, emojis, and special characters (e.g., ©, ♦, ◄, ●, ♡, ★, ∞) which must be preserved and handled correctly.

[4]We took some ideas from: https://www.trainy.ai/blog/gpu-utilization-misleading

[5]For example: *Samsung Galaxy Buds 3 Bluetooth 5.2 Wireless Noise-Canceling – 40 mm Hi-Res Drivers 20 Hour Battery Life Foldable Over-Ear Design Built-in Mic Touch Controls Fast Charge EQ Presets Compatible with iOS Android.*

- **Specification:** Texts of variable length, ranging from a single character to several paragraphs, which can also contain HTML, emojis, and special symbols. The translation of specifications relies on the attribute name for a guided translation.[6]

- **Review:** User-generated free-text, often containing typos, abbreviations, and other informal language phenomena; these reviews may likewise include emojis and special characters.

Our initial translation framework relied on models pre-trained for individual language pairs, since many-to-many systems, such as No Language Left Behind (Team et al., 2022) or M2M100 (Fan et al., 2020), 1. produced inconsistent outputs, 2. contained a large number of parameters that led to slow, resource-intensive execution, and 3. did not achieve the accuracy of pair-specific models.

After generating our training data, we fine-tuned Marian-based models (Tiedemann and Thottingal, 2020; Tiedemann, 2020) for targeted translation tasks. The resulting models either surpassed or closely matched the performance of OpenAI instruction-based and Google Translate systems on texts of increased complexity. However, because they relied on an outdated tokenizer without byte-level encoding, they were unable to generate certain tokens (e.g., uppercase "Ú," emojis, and other special characters). This situation led us to pre-training our own T5 models (Raffel et al., 2023).

We first trained our own SentencePiece[7] tokenizer on an internal dataset of various e-commerce text types: product titles, long text descriptions, specifications, and reviews. This new tokenizer uses 32% less tokens than the Marian tokenizer and allows us to customize the vocabulary. We then pre-train a sequence-to-sequence translation model using curated and deduplicated versions of the OPUS[8] corpora, for English→Spanish (22B tokens) and English→French (17B tokens) language pairs. We used a more modern language modeling architecture with T5-base and set the maximum context length to 128 tokens, a reasonable limit given the data we translate. The combination of the custom tokenizer with the more efficient architecture led to important increases in BLEU and COMET scores once the pre-trained model is fine-tuned on specific e-commerce tasks (see Table 1).

Eval. Dataset	Base model	Fine-tuned	En→Es BLEU/COMET	En→Fr BLEU/COMET
tatoeba-test v2021-08-07	opus-mt	–	57.26/92.36	53.13/90.39
	Our T5	–	55.77/92.12	53.09/90.87
Titles (Proprietary)	opus-mt	–	30.16/72.20	28.99/65.10
	Our T5	–	30.85/72.58	28.75/65.57
	opus-mt	title	67.00/84.64	63.38/80.02
	Our T5	multitask	69.12/84.93	62.78/79.95
	Our T5	title	**69.40/85.07**	**64.93/80.59**
Reviews (Proprietary)	opus-mt	–	43.67/79.50	47.03/80.87
	Our T5	–	46.56/80.80	52.42/82.29
	opus-mt	review	67.65/89.56	67.07/88.70
	Our T5	multitask	71.18/90.09	71.59/90.13
	Our T5	review	**71.96/90.23**	**71.91/90.42**

Table 1: BLEU/COMET scores for English→Spanish (En→Es) and English→French (En→Fr) on the generic Tatoeba test set and proprietary e-commerce datasets. `opus-mt` stands for opus-mt-tc-big-**.

We store translations in our tracking table using a semantic versioning scheme of the form $X.Y.Z$, where X denotes the base pretrained model version, Y indicates the fine-tuning iteration, and Z corresponds to the code-base revision. This structured versioning enables us to identify which translations may be affected when a defect in the translation pipeline or a particular model version is discovered to produce incorrect outputs under specific conditions. We also continuously refresh legacy translations, and in particular those for high-frequency ("head") products.

3.3 Translation Logic

Although state-of-the-art translation models are trained or prompted to preserve non-translatable segments, our empirical observations indicate that augmenting these models with auxiliary safeguards further reduces mistranslation rates. We have therefore implemented a dedicated translation framework, hereafter referred to as the *Reference Generator*, which enforces non-translatability constraints uniformly across all text categories described above.

For inputs containing HTML markup, we first parse and sanitize the HTML to extract the raw textual content, thereby yielding standardized, markup-free source segments suitable for processing by downstream translation modules.

To address the non-translatable-entity problem, a central challenge in contemporary Machine Trans-

[6]For instance, in our initial analysis we found that "N", referring to "No" was translated to "North" in Spanish. Adding the context ensures correct translation.

[7]SentencePiece library: https://github.com/google/sentencepiece

[8]OPUS webpage: https://opus.nlpl.eu/

lation,[9] we extract metadata from each item, including brand, product line, model, sports team, and sports league. Because merchant-provided metadata fields are often noisy, irrelevant or overly general (e.g., a brand field containing an entire product title), we clean those fields and apply fuzzy-matching techniques to align these metadata entries with substrings in the source text. Then, to automate the classification of potential non-translatable entities from the matched metadata, we use a quantized RoBERTa-based classifier, offering a lightweight, fast (latency under 5ms on CPU) and language-agnostic solution at inference time. Despite a large amount of noisy metadata fields, representing 73% of all entries, the classifier solution preserves more than 99% of non-translatable entities while removing 96% of noisy metadata fields. Additionally, when metadata alone is insufficient to capture all non-translatable instances, we reference three curated entity repositories: a *Whitelist* comprising context-specific non-translatables (e.g., proper brand names such as "Way To Celebrate"), a *Blacklist* containing common terms that should be translated despite superficial resemblance to named entities, and a *Potential Whitelist* for ambiguous tokens that may require context-dependent treatment.

As a preprocessing step for product title inputs, all alphabetic tokens that do not correspond to identified non-translatable entities are converted to lowercase. This normalization is motivated by empirical observations that neural translation engines (e.g., Google Translate and Marian models) exhibit degraded performance when confronted with text in full or title-style capitalization.

Following normalization, the processed strings are sent to our GPU-accelerated translation models. Upon receipt of the model outputs, each translation is programmatically inspected to verify the integrity of the non-translatable entities. Although this procedure may appear simplistic, it serves a dual purpose: first, allowing the translation model to leverage contextual cues from surrounding non-translatable tokens (which, when correctly preserved, can improve overall translation fidelity), and second, identifying instances in which the model erroneously alters or omits these critical segments.

When a non-translatable entity is altered, the text is routed through a Product-Category-Aware Entity Masking module: the original token is swapped for a category-specific, untranslatable placeholder (e.g., "My Jeans" → "Nike" in Clothing), ensuring the translation engine leaves it intact. After translating the masked text reliably, the placeholder is replaced with the true entity in postprocessing.

Our target locales include U.S. Spanish, a hybrid dialect blending Latin American regionalisms and English[10] and Canadian French, using models trained on localized data. After translation, a rule engine refines translations for accuracy, consistency, and tone, including softening certain offensive terms.

In cases where the source field contains HTML markup, translated segments are reintegrated into the original HTML structure. For product title inputs, we restore title-style capitalization and then reorder tokens within the translated string so that any non-translatable entities originally occupying the lead positions continue to appear first. Such non-translatable refactoring constitutes a nontrivial string transformation, since the translation model may insert intervening function words (e.g., the Spanish preposition "de") that must be accounted for in postprocessing.

The *Reference Generator* leverages an LLM to produce high-quality synthetic parallel corpora, enabling the training of compact models optimized for e-commerce translation.

3.4 Accuracy & Explainability Pipeline

From a business perspective, BLEU and COMET metrics often lack explainability. Their scientific foundations make them difficult to interpret for stakeholders, and they may fail to flag critical translation errors in e-commerce contexts, such as incorrect brand translations or the omission of meaningful tokens.

For any of the models presented in Section 3.2, the following definitions apply. Fix a model X for a given translation task—e.g., translating *product titles*. Let $\mathsf{T} = \{T_i\}_{i=0}^{n-1}$ be a sample of n item titles, stratified by user impressions and product types to ensure a representative and homogeneous product sample. For the selected task, define a fixed set $\mathsf{S} = \{S_j\}_{j=0}^{k-1}$ of *severe issues* (e.g., mistranslated brands or missing tokens). Let $f : (\mathsf{T}, \mathsf{S}) \mapsto [0, 1]$ be defined as

[9] SemEval 2025 task for Entity-Aware Machine Translation: https://sapienzanlp.github.io/ea-mt/

[10] A phenomenon also identified by linguists: https://theconversation.com/linguists-have-identified-a-new-english-dialect-thats-emerging-in-south-florida-205620

$$f(T_i, S_j) := \begin{cases} 0 & \text{if } T_i \text{ contains issue } S_j, \\ 1 & \text{otherwise.} \end{cases}$$

Then, e-commerce quality metric is defined as:

$$m(\mathrm{T}) := \frac{1}{n} \sum_{i=0}^{n-1} \left(\prod_{j=0}^{k-1} f(T_i, S_j) \right)$$

We refer to the functions $f(\bullet, S_j)$ as scoring signals, as they determine whether a given text contains a specific translation issue. Some of these signals rely on a *LLM-powered translation*[11] as a ground-truth reference.

Scoring signals include:

- **Semantic similarity**: Measures cosine similarity between embeddings generated by a pre-trained sentence transformer model.

Text to translate		Our translation
Rubber Chrome Kettlebell	↗	Campanilla Cromada de Goma
	↘	**LLM translation**: Pesa Rusa de Cromo y Goma

- **Non-translatable protection**: Verifies that non-translatable entities from the item metadata are properly preserved, according to predefined *blacklist/whitelist* rules.

Text to translate		Our translation
Well Woven Medusa Nord Nordic Lattice Pattern Outdoor...	⟶	Alfombra de 9'3 "x 12'6" Gris Claro Al Aire Libre con Patrón...

- **New entity detection**: Uses a context-aware LLM to identify non-translatables not present in the item metadata or whitelist, which should have been protected in the translation.

Text to translate		Our translation
Straight Talk Motorola Razr 2024, (...) [Locked to Straight Talk]	⟶	Hablar Claro Motorola Razr 2024 (...) [Bloqueado para Hablar Claro]

- **Uncommon token analysis**: Compares the unique token sets (excluding stop words) of our model's translation and the LLM reference, identifying meaningful tokens missing from our model's output.

Our translation		LLM translation
Calvin Klein Alfombra de Vapor CK970	⟶	Calvin Klein Alfombra de Área Vapor CK970

Each week, we run our evaluation pipeline on a freshly drawn stratified random sample, using an LLM-generated translation as the reference so that, alongside our custom metric m, we can compute BLEU scores; although neither BLEU nor m is strictly normally distributed, the *Central Limit Theorem* ensures their sample means are approximately normal (a fact supported by Shapiro–Wilk $p > 0.05$), allowing us to assess their relationship via Pearson, Spearman, Kendall, and Normalized Mutual Information (NMI). Results for two fine-tuned models, one tailored to product titles, the other to longer texts, from Table 2 show that despite m's design for enhanced explainability and domain-specific error alignment, it remains strongly correlated with BLEU across both short- and long-form translation tasks.

Statistical test	Title	Long text
Pearson	0.837	0.982
Spearman	0.824	0.957
Kendall	0.667	0.867
NMI	0.769	0.835

Table 2: Correlation between m and BLEU for our models fine-tuned on product titles and long-form text.

3.5 Annotation & Feedback Platform

A team of e-commerce–specialized linguists conduct a systematic verification of translated content, drawing both from the live website and from the evaluation framework pipeline from section 3.4. To prevent ad-hoc modifications or stylistic adjustments, thereby preserving consistency and comparability, annotators are not permitted to alter translations directly. Instead, they author constraint rules via a purpose-built interface tailored to the translation task. Each rule encodes a simple, human-interpretable conditions of the following types: • The source text must [not] contain a specified substring. • The translated text must [not] contain a specified substring. Annotators can specify the exact substrings to be replaced and define the corresponding replacement strings.

Given the high volume of rules generated by annotators on a daily basis, manual review by the engineering team is infeasible. To ensure that each rule is grounded in empirical evidence, we implemented an automated validation loop that integrates rule evaluation with real-world translation data. The annotation workflow depicted in Figure 2 proceeds as follows:

1. A linguist identifies a linguistic issue and formulates a constraint.

2. The system retrieves a small but representative sample of items to show the linguist which items would be impacted by the new rule. A second linguist examines the proposed rule

[11] This translation uses the same pre- and post-processing as our model, but employs a high-quality large language model as the translation engine.

against the sample and approves it if it correctly targets the issue.

3. Once validated, an automated process (a) extracts all affected items from the production database and (b) generates translation outputs for each item using the OpenAI model.

4. For each impacted item, the system presents four versions to the linguists, original translation with/without the rule applied and OpenAI translation with/without the rule applied.

5. Linguists review these four renderings side by side and assign one of the following outcomes to the rule: • Archive: The rule is unnecessary or counterproductive. • P0, critical: The rule addresses a serious error that must be enforced immediately. • P1, language flavor: The rule refines stylistic or regional preferences without correcting a critical error. • P2, minor style: The rule addresses superficial issues (e.g., capitalization, punctuation) that do not impact core meaning.

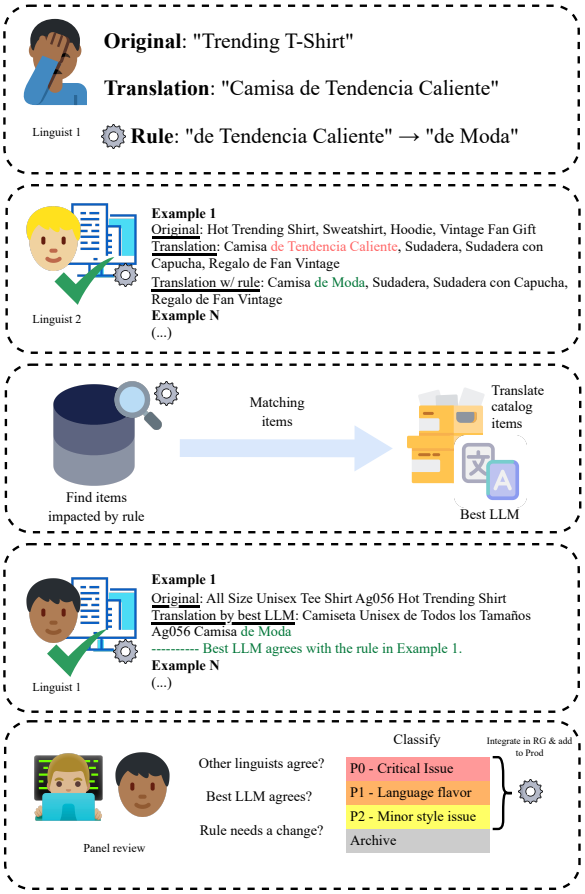

Figure 2: Translation Platform diagram from the translation flow perspective.

Model	Title ($)	Review ($)
Google Translate API	35.30	4.34
gpt-4o-2024-05-13	18.06	3.59
ours	0.17	0.10

Table 3: Translation average cost comparison by element type and model. The cost reflects the price for translating a batch of 1,000 elements.

In the final validation stage, linguists examine paired translation instances in which the OpenAI model either satisfies the newly introduced rule (rendering the rule redundant) or violates it (indicating that the rule remains necessary). Empirical analysis demonstrates that the OpenAI model conforms to the vast majority of rules classified as P0, whereas it frequently diverges from rules classified as P1 or P2. Notably, despite providing prompts intended to elicit regionally or stylistically "flavored" translations, the OpenAI model rarely produces such variants; as a result, linguists often formulate P1 rules to enforce flavor-specific output, but the OpenAI translations fail to satisfy most of these constraints.

4 Conclusions and Future Work

Our purpose-built translation platform combines daily "bootstrap" and hourly batch pipelines with a low-latency API, GPU-accelerated T5-based English→U.S. Spanish and English→Canadian French models, and a metadata-driven Reference Generator (fuzzy matching, RoBERTa classifier and curated white/blacklists for non-translatables) to enforce over 99% non-translatable entity preservation. A linguist-managed rule engine and an explainable evaluation framework with a custom e-commerce quality metric drive continuous improvements. In production, this system handles millions of listings per day with sub-second API latency, matches or exceeds state-of-the-art MT models in both translation quality and end-to-end throughput, and reduces costs by 10×–100× compared to general-purpose LLM deployments—all under semantic versioning and a Translation Tracking Table for enterprise-grade traceability and rollback.

Future work will explore more compact (distilled and quantized) model variants, broaden language support (e.g., Mandarin, Hindi), and enhance our rule-engine analytics and dynamic supervision features to further automate rule creation and refine localization quality.

Limitations

Despite being an e-commerce platform with hundreds of millions of listings, most of our attention is focused on items with high impressions. Consequently, our training and evaluation datasets, as well as our non-translatable whitelists and blacklists, and the data samples used for linguist feedback, are biased towards popular listings. While parts of our evaluation metrics and training data generation rely on state-of-the-art LLMs, we recognize that these systems are not perfect translators and may introduce their own biases (e.g., a tendency to shorten or summarize content). As these systems continue to improve, we are well positioned to quickly adapt and match their performance for our specific tasks, while maintaining high efficiency and low cost. Finally, our feedback and annotation platform has automated a significant portion of data sampling, rule impact evaluation and rule deployment. However, this process remains highly labor-intensive and exception-prone, requiring the involvement of multiple linguists over extended periods. In future work, we plan to explore more agentic capabilities, as well as more natural ways to express and apply constraint rules, to further increase the level of automation in rule creation.

Acknowledgements

This work wouldn't have been possible without the collective efforts of the broader team, including: Nicole McNabb, Sushant Chaudhari, Dayron Rizo-Rodriguez, Vikas Murthy, Jorge Ortiz-Fuentes, Soumyakant Mishra, Alina Sotolongo, Yuanliang Qu, Pavan Malyala, Bruno Gutierrez, John Chau, Henry Rosales, Chellappan Lakshmanan, Rajat Sharma, Ketki Kulkarni, Brian Linneman, Dilpreet Singh, Niranjan Joshi, Arpitha Shetty, Anna Lavinia Dambrosio, Julio Madrid, Diana Liceaga, Elda Benet, Giselle Ojeda, Carlos Ramos, Nicole Aguilera, Paresh Mondkar, Gauthier Dumoulin, Akhil Vishwanatham, Tracy Poulliot, Mai Le.

References

Aman Ahuja, Nikhil Rao, Sumeet Katariya, Karthik Subbian, and Chandan K. Reddy. 2020. Language-agnostic representation learning for product search on e-commerce platforms. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 7–15, New York, NY, USA. Association for Computing Machinery.

Kaidi Chen, Ben Chen, Dehong Gao, Huangyu Dai, Wen Jiang, Wei Ning, Shanqing Yu, Libin Yang, and Xiaoyan Cai. 2024. General2specialized llms translation for e-commerce. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 670–673, New York, NY, USA. Association for Computing Machinery.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. *Preprint*, arXiv:2502.02481.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *Preprint*, arXiv:2010.11125.

Dehong Gao, Kaidi Chen, Ben Chen, Huangyu Dai, Linbo Jin, Wen Jiang, Wei Ning, Shanqing Yu, Qi Xuan, Xiaoyan Cai, Libin Yang, and Zhen Wang. 2024. Llms-based machine translation for e-commerce. *Expert Systems with Applications*, 258:125087.

Jyoti Guha and Carmen Heger. 2014. Machine translation for global e-commerce on eBay. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Users Track*, pages 31–37, Vancouver, Canada. Association for Machine Translation in the Americas.

Christian Herold, Michael Kozielski, Leonid Ekimov, Pavel Petrushkov, Pierre-Yves Vandenbussche, and Shahram Khadivi. 2024. Lilium: ebay's large language models for e-commerce. *Preprint*, arXiv:2406.12023.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Preprint*, arXiv:2309.04662.

Mandar Kulkarni and Nikesh Garera. 2022. Vernacular search query translation with unsupervised domain adaptation. *Preprint*, arXiv:2208.03711.

Latino Donor Collaborative. 2024. 2024 u.s. latino gdp grows by 13% to $3.6 trillion; the world's fifth-largest economy is now projected to surpass japan and germany by 2027. Press release.

Jens Manuel Krogstad Mark Hugo Lopez and Jeffrey S. Passel. Who is Hispanic? — pewresearch.org. https://www.pewresearch.org/short-reads/2024/09/12/who-is-hispanic. [Accessed 30-05-2025].

Nimdzi Insights LLC. 2023. The 2023 NIMDZI 100: The ranking of the largest language service providers in the world – Including the top 100 largest language service providers. Technical report, Nimdzi Insights LLC, Sandpoint, ID. Copyright © 2023 by MultiLingual Media LLC; published and distributed by MultiLingual Media LLC.

Nimdzi Insights LLC. 2024. The 2024 NIMDZI 100: The ranking of the largest language service providers in the world – Including the top 100 largest language service providers. Technical report, Nimdzi Insights LLC, Sandpoint, ID. Copyright © 2023 by MultiLingual Media LLC; published and distributed by MultiLingual Media LLC.

Nimdzi Insights LLC. 2025. The size and state of the language services industry: The 2025 NIMDZI 100: Including the ranking of the top 100 largest language service providers. Technical report, Nimdzi Insights LLC, Mercer Island, WA. Distributed by MultiLingual Media LLC.

Jesus Perez-Martin, Jorge Gomez-Robles, Asier Gutiérrez-Fandiño, Pankaj Adsul, Sravanthi Rajanala, and Leonardo Lezcano. 2023. Cross-lingual search for e-commerce based on query translatability and mixed-domain fine-tuning. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 892–898, New York, NY, USA. Association for Computing Machinery.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. Redpajama: an open dataset for training large language models. *Preprint*, arXiv:2411.12372.

Bryan Zhang and Amita Misra. 2023. Machine translation impact in e-commerce multilingual search. *Preprint*, arXiv:2302.00119.

Bryan Zhang, Taichi Nakatani, Daniel Vidal Hussey, Stephan Walter, and Liling Tan. 2024a. Don't just translate, summarize too: Cross-lingual product title generation in e-commerce.

Bryan Zhang, Taichi Nakatani, and Stephan Walter. 2024b. Enhancing e-commerce product title translation with retrieval-augmented generation and large language models. *Preprint*, arXiv:2409.12880.

Hang Zhang and Liling Tan. 2021. Textual representations for crosslingual information retrieval. In *Proceedings of the 4th Workshop on e-Commerce and NLP*, pages 116–122, Online. Association for Computational Linguistics.

A Model Serving Benchmark

Our internally hosted translation pipeline incurs costs primarily based on computational resources used, linking expenses directly to throughput and efficiency. Effective resource management not only increases performance, but substantially reduces operational costs.

A benchmark study (Figure 3 and Figure 4) comparing TorchServe[12], OpenNMT[13], and VertexAI[14] reveals the superior performance of OpenNMT, with consistently lower inference durations and higher translations per second at larger batch sizes. This efficiency directly translates into cost savings, especially evident when accumulating translations before processing. The cost calculations presented in Table 3 utilize this optimized OpenNMT framework.

[12]TorchServe: https://docs.pytorch.org/serve/

[13]OpenNMT: https://opennmt.net/

[14]VertexAI: https://cloud.google.com/vertex-ai

We recommend ongoing monitoring and optimization of batch sizes and resources to maintain scalability and efficiency. Future evaluations will include frameworks such as vLLM[15] and Text Generation Inference[16], and further model optimizations like quantization and distillation.

B Evaluation Data Availability

To promote transparency and open science, we provide a small evaluation sample of the task-specific data used during model development. This dataset includes titles, long text, specifications, and reviews for both En→Es (U.S.) and En→Fr (Canada) language pairs. The sampling process ensured an equal mix of varied popular items from top product categories, as well as a selection of random catalog items. The *text_translation* data was generated using an OpenAI GPT-4o model in combination with our Reference Generator, and serves as our gold standard data. We also provide the *non_translatables*, which are entities supplied by the seller that appear in the English *text_original* and must be preserved during translation. We hope the community can leverage this dataset to drive innovation in the protection of non-translatable named entities.

[15] vLLM: https://docs.vllm.ai/

[16] Text Generation Inference: https://huggingface.co/docs/text-generation-inference

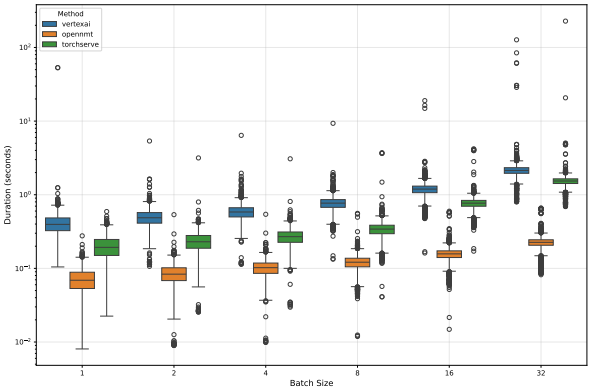

Figure 3: Inference duration vs batch size for different serving frameworks.

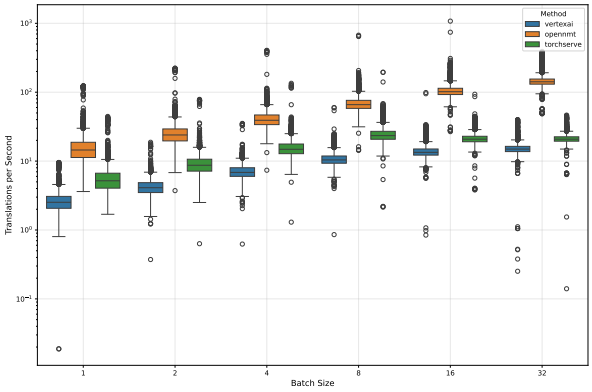

Figure 4: Translations per second vs batch size for different serving frameworks.