

---

# The Cross-Border E-Commerce Selection Benchmarking Model: A Multi-Modal Framework for Improved Zero-Shot Cross-Lingual Product Classification

---

Kriti Mahajan<sup>1</sup>

## Abstract

Within-country selection benchmarking models fall short for cross-border selection benchmarking because cross-border competitor data is typically a) multilingual (thus requiring cross-lingual comparisons) and b) sparse and noisy, thus eliminating supervised fine-tuning options. To address these challenges, this paper presents the Cross-Border E-commerce Selection Benchmarking model - a novel multi-modal, zero-shot, cross-lingual product search and comparison engine. The model combines pre-trained large language models, vision transformers, and generative models to determine if a given pair of products - drawn from a reference platform catalog and a foreign market competitor catalog - are exact matches without needing big data or fine-tuning. Evaluated across 10 cross-border language arcs and 17 product categories, the model achieves an F1 score of 79%, substantially outperforming embedding based classification (F1 score: 27%) and prompt-based (zero-shot/ToT/CoT/Reasoning) classification (F1 score : 65%) - across both high and low resource languages, providing greater performance uplift for low-resource languages. Online deployment delivers gains of 7.8% in product discovery and 20% in revenue. We also demonstrate that generated image synthesis can effectively mitigate image sparsity in product catalogs and improve downstream product pair classification accuracy.

## 1. Introduction

For multinational e-commerce platforms, selection benchmarking - systematically identifying overlaps and gaps between a platform's own product catalog and competitor prod-

uct catalogs in destination marketplaces - is a foundational capability. Accurate selection benchmarking enables detection of missing products, assessment of assortment breadth, and downstream analyses such as price competitiveness. Cross-border competitor product comparison fundamentally differs from within-market matching due to two compounding factors. First, linguistic heterogeneity: source and destination catalogs are typically in different languages, and comparison becomes harder when low-resource languages are involved, where translation quality and large language model (LLM) performance are weaker. Second, data sparsity and noise: in our empirical analysis of cross-border competitor catalogs, 40% of products lack images and 90% lack universal product identifiers (GTIN/EAN/UPC) (see Appendix Table 1). Noise is further driven by poor catalog quality management (ex: 70% of the samples reviewed on competitor websites had poor product placement, lighting, irrelevant products in the frame and blurry images) and catalog localization artifacts such as mistranslations and transliterations. This asymmetry between well-structured internal catalogs and sparse external catalogs introduces additional complexity for cross-catalog product comparison at scale.

This combination of cross-lingual text comparisons, sparse and noisy competitor data and limited availability of labeled training examples has three consequences for selection benchmarking: (1) traditional keyword/text-similarity methods are ineffective; (2) the sparse and noisy data eliminates the possibility of fine-tuning LLMs and ViTs<sup>1</sup>; and (3) only a limited set of features - product title, brand, and a high-level product category - are reliably available as inputs. Thus, motivated by two key problems - comparing localized-cross-lingual text and working in a noisy, sparse data environment - we present the Cross-Border E-commerce Selection Benchmarking Model : a multi-modal, cross-lingual, zero-shot product comparison engine that uses an ensemble of pre-trained LLMs, vision transformers (ViTs), and Stable Diffusion. The model operates in three stages: 1) candidate generation via cross-lingual semantic search, 2)

<sup>1</sup>Cross Border Tech & Science, Amazon. Correspondence to: Kriti Mahajan <kritimhj@amazon.com>.

<sup>1</sup>Experiments with fine-tuning using masked language modeling resulted in catastrophic forgetting, which led to that experimental approach being abandoned in the early days of model building

candidate ranking using semantic and sub-string similarity, and 3) multi-modal classification combining text reasoning, visual similarity, and identifier checks (see Figure 1). In this paper, we will **focus on the motivation, experiments and results that allowed us to arrive at a multi-modal classifier** for determining if any given product pair is an exact match or not, with an F1 score of 79% across all arcs and all GLs (see Appendix A for a brief discussion regarding the remaining model architecture experiments). **Our primary contributions are:** (1) a novel framework for cross-lingual zero-shot selection benchmarking; (2) a zero-shot multi-modal classification architecture using pre-trained LLMs, ViTs, and image generators; (3) comprehensive evaluation across 10 language pairs and 17 product categories demonstrating generalization across high and low resource languages; and (4) online experiments corroborating offline findings. **Related Work:** The literature on multi-modal cross-lingual e-commerce product comparison is limited; supported by the meta-review in [4], our work appears to be the first to combine text and image data for this task. Our approach is philosophically aligned with MURAL [1], which demonstrates that visual contexts enhance cross-lingual understanding.

## 2. Model Overview

Let  $C_r = \{p_r^1, \dots, p_r^N\}$  denote the reference platform’s internal product catalog and let  $C_c = \{p_c^1, \dots, p_c^N\}$  denote a competitor catalog in a foreign marketplace. Each product is represented by a tuple:  $p = (t, b, g, I)$  where  $t$  is the product title (unstructured text),  $b$  is the brand,  $g$  is a product category and  $I$  is a set of product images (if available). Given a competitor product  $p_c \in C_c$ , the task of the selection benchmarking model is to determine if there exists an exact match  $p_r \in C_r$  i.e. determine if the competitor product has an exact match in the reference catalog or not. An exact match requires agreement on key attributes (brand, product type, model number, quantity), allowing missing attributes in  $p_c$  if remaining information is consistent; a false positive is a non-match classified as a match, and a false negative is a true match classified as a non-match. **We seek a decision function:**  $f(p_c, C_r) \rightarrow 0, 1$  where  $f = 1$  if an exact match exists and  $f = 0$  otherwise.

**2.1 Cross-catalog Brand Mapping:** For each  $p_c$  with brand  $b_c$ , we first map  $b_c$  to a reference catalog brand  $b_r$  using cosine similarity over all-MiniLM-L6-v2 embeddings, enabling alignment of semantically similar or transliterated brand names across languages and catalogs. If  $\cos(b_c, b_r) < 0.7$ ,  $p_c$  is assumed to have no exact match in  $C_r$  and is excluded from further processing.

**2.2 Candidate Generation:** We restrict the reference catalog to the subset  $C_r^{b_r, g}$  matching the mapped brand  $b_r$  and product group  $g$ . Competitor titles are normalized

and translated into the reference catalog language using M2M100<sup>2</sup>. Let  $\theta_t(\cdot)$  denote the LaBSE embedding function. Candidate reference products are retrieved via cosine similarity search over text embeddings:  $K(p_c) = \text{Top-}k \arg \max_{p_r \in C_r^{b_r, g}} \cos(\theta_t(t_c), \theta_t(t_r))$  where  $t_c$  is the foreign competitor title and  $t_r$  is the reference catalog title.

**2.3 Candidate Ranking:** The candidate set  $K(p_c)$  is re-ranked using a composite text similarity score: a sub-string match score followed by a LaBSE embedding based cosine similarity as a tie breaker -  $s(p_c, p_r) = \alpha \cdot \cos(\theta_t(t_c), \theta_t(t_r)) + \beta \cdot \text{SubstringSim}(p_c, p_r)$  where  $\text{SubstringSim}(\cdot)$  captures sub-string and token overlap features. The top-ranked reference candidate  $\hat{p}_r$  is passed to the classification stage.

**2.4 Multi-modal Classifier Architecture:** (Figure 2) The top-ranked candidate  $\hat{p}_r$  is evaluated using a multi-modal classifier that combines text-based reasoning, visual similarity, and unique identifier checks (when available) to determine whether the pair constitutes an exact match. The classifier component integrates three complementary signals in a sequential manner: **2.4.1 Identifier consistency:** If universal product identifiers (ex: UPC, EAN, GTIN) are available for both products and match exactly, the pair is classified as an exact match and model inference ends; else we move to **2.4.2 Text Semantic Similarity:** We define a binary, chunked text cosine similarity decision function (section 3.1.1.a):  $f_{text}(p_c, \hat{p}_r) \in \{0, 1\}$  using LaBSE embeddings. The model assesses whether the two product titles refer to the same product under translation and cross-lingual variation; if the pair is classified as an exact match and model inference ends; else we move to **2.4.3 Visual Similarity:** When competitor images are missing, we generate a synthetic image  $\hat{i}_c$  using Stable Diffusion conditioned on  $(t_c, b_c, g_c)$  through a tuned prompt. Let  $\theta_v(\cdot)$  denote a ViT embedding function. Visual similarity is computed as:  $s_v(p_c, \hat{p}_r) = \cos(\theta_v(\hat{i}_c), \theta_v(i_r))$ . A visual match is said to be found if:  $s_v(p_c, \hat{p}_r) \geq \tau_v$  where  $\tau_v$  is empirically calibrated. **2.4.4 Final Decision Rule:** The final classification decision is as follows:

$$f(p_c, \hat{p}_r) = \begin{cases} 1 & \text{if } f_{id} = 1 \vee f_{text} = 1 \vee f_{vis} = 1 \\ 0 & \text{otherwise} \end{cases}$$

If  $f(p_c, \hat{p}_r) = 0$ , the competitor product  $p_c$  is flagged as a selection gap. This 3-step design allows the model to scale to large catalogs while reserving the most computationally expensive inference for a small subset of candidate pairs.

<sup>2</sup>Because translated text with LaBSE achieved significantly higher hit@k (50%, 70%, 100% for  $k = 1, 50, 100$ ) than cross-lingual xlm-r-multilingual-v1 (5%, 11%, 21%), we adopt this approach and omit further discussion given this paper’s focus on the classification architecture.

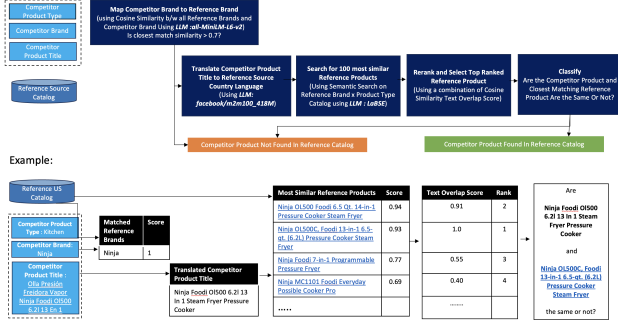


Figure 1. Cross-Border E-commerce Selection Benchmarking Model

### 3. Experimental Methodology

**3.1 Baselines:** We evaluate three classifier classes<sup>3</sup> -

**3.1.1. Text Classifiers:** This category includes three classifier types. The baseline, **a. Cosine Similarity Threshold**, is a rule-based binary classifier that uses a 0.7 cosine similarity threshold between LaBSE-embedded representations of reference and competitor product titles. Scores  $> 0.7$  are classified as matches (1), otherwise non-matches (0). However, this classifier exhibits a high false negative rate (84%; Appendix Table 5). We therefore evaluate additional approaches: (b) **Chunk cosine similarity** ( $> 0.7$ ): product titles are split into 3-word chunks and similarity is computed across chunks: (b.1) **Mean chunk similarity**: classify as match if average similarity  $> 0.7$ ; (b.2) **Max chunk similarity**: classify as match if maximum chunk similarity  $> 0.7$ . (c) **Generative prompt-based classifiers**: Claude models (v2, Sonnet, Haiku) classify pairs as matches if the response is “SAME” (1, else 0). Four prompting strategies are tested: zero-shot, reasoning, chain-of-thought (CoT), and tree-of-thought (ToT).

**3.1.2. Image Classifiers:** This family uses Stable Diffusion to generate images for foreign competitor products to address sparse/noisy inputs and mimic the reference platform’s product image display style. We evaluate: (a) **Image cosine similarity** ( $> 0.15$ ): using a pre-trained ViT, embeddings of generated and reference images are compared; pairs are matches if similarity exceeds 0.15. (b) **SSIM** ( $> 0.75$ ): structural similarity is computed between generated and reference images; pairs are matches if SSIM exceeds 0.75.

Given the sensitivity of generative models to input text, we test six preprocessing strategies (Appendix Table 3): (1) original text; (2) cleaned original text (remove special characters, symbols, digits); (3) destination text translated to source language; (4) cleaned version of (3); (5) both texts

<sup>3</sup>All thresholds in this section are selected based on highest F1 in human-in-the-loop evaluation compared to alternative thresholds.

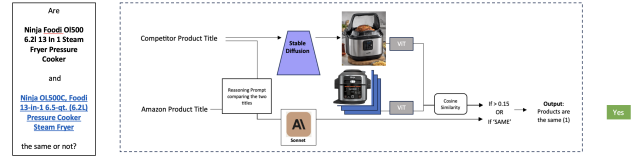


Figure 2. Classification Architecture. To reduce cost, during deployment the reasoning prompt is replaced with max chunked semantic similarity (section 3.1.1.a) without loss of accuracy.

translated to English; (6) cleaned version of (5). This yields 90 text classifiers and 12 image classifiers.

**3.1.3 Multi-modal Classifiers:** We combine all text and image classifiers using: (1) **AND** (both must predict match) and (2) **OR** (either predicts match), resulting in 2160 multimodal classifiers.

**3.2 Evaluation:** We construct a manually labeled dataset of  $\sim 1000$  foreign competitor–reference product pairs spanning 10 arcs and 17 product categories identified by the benchmarking model as highest-ranked matches. The sample is stratified by arc. Human auditors label pairs as exact match (1) or not (0). The dataset includes arcs with high-resource languages (e.g., English in US) and low-resource languages (e.g., Japanese in JP). Source marketplaces: US, DE, JP. Destination marketplaces: Mexico (MX), Singapore (SG), Germany (DE), Turkey (TR) (catalogs in Spanish (ES), English (EN), Germany (DE), Turkish (TR)), yielding 10 language pairs. Performance is measured using F1 score due to class imbalance.

### 4. Results

**4.1 Offline Evaluation :** For each classifier category (Text, Image and Multi-modal) the best in class classifier is defined as the classifier with the highest F1 score within that category. Figure 3.A shows that the aggregate best in class multi-modal classifier (see Figure 2 for architecture) outperforms all classifiers across all categories, demonstrating that extracting information from both text and images using an ensemble of large models provides the greatest uplift - 53% - in classification performance. The baseline cosine similarity classifier achieves  $F1 = 0.27$ , compared to 0.80 for the best multi-modal model. Across all arcs and product types (Figs. 3.B-C), 3 trends emerge: 1. for every source–destination pair and GL, the best multi-modal classifier outperforms the baseline and the top text and image models; 2. The best-in-class classifier in each category significantly surpasses the baseline, with image classifiers showing an average 43% F1 uplift and text classifiers showing a 40% F1 up-

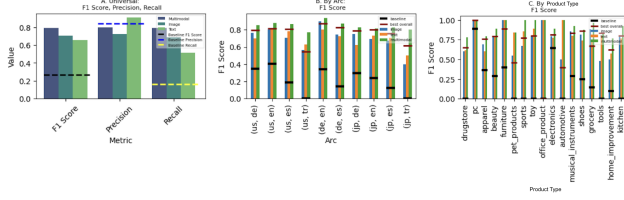


Figure 3. Best In Class Classifier Comparison

lift; 3. Although the aggregate best-in-class multi-modal classifier does not outperform its counterparts at the source-destination country pair or Product Type level, it generally surpasses the image and text classifiers across most arcs and product types, with exceptions in the arcs US>TR, DE>EN and the product types for toys and office products. Multi-modal classification generalizes well across all 10 language pairs, with larger gains in low-resource arcs ( 67%) than high-resource arcs ( 59%) (Appendix Table 8). **4.1.1 Product Type Analysis:** The consistent uplift provided by the best in class multi-modal classifier in each Product Type shows that multi-modal classification generalizes across all product domains. The baseline cosine similarity classifier has an F1 score greater than 0.5 for only 2 product types - PC and Electronics - implying that it worked well only for head selection hardline product types. In comparison, the best in class multi-modal classifier at the Product Type level results in comparably high F1 scores across all product types. The best in class multi-modal classifier at the aggregate level also results in comparably high F1 score, with only 3 product types having an F1 score less than 0.5. This evidence supports the efficacy of multi-modal classification in delivering consistent and reliable performance across both softlines and hardlines. **4.1.2 Arc Level Analysis:** (to avoid stratification bias) shows multi-modal models are optimal in every arc (Appendix Fig. 4). In each arc, every leading multi-modal classifier uses cosine similarity (ViT embeddings) over SSIM for image distance measurements, likely because image differences are not purely noise-based. Text cleaning is critical across all languages, especially in low-resource settings. Maximum chunk similarity generally outperforms mean chunk similarity and Claude in high-resource arcs due to improved embeddings and translations, while Claude requires more complex prompting (e.g., reasoning, ToT) to outperform cosine-based methods.

**4.2 Online Experimentation:** Across 6 arcs (3 source, 6 destination countries) and 6 languages, deployment of the model yields a 7.8% increase in correct cross-lingual exact match discovery and a 20% revenue uplift across 660k products, improving product availability and outperforming the prior version of the model that used a unimodal text cosine similarity classifier.

**4.3 Error Analysis:** Compared to the baseline (see Figure 3.A), the best-in-class multi-modal classifier achieves

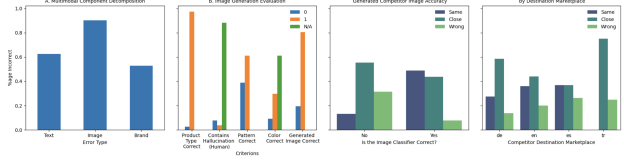


Figure 4. Error Analysis

a 63% recall uplift but with a 5% precision drop, yielding a 14% false positive rate and 21% false negative rate (see Appendix Table 5), indicating a greater tendency toward false negatives (i.e., missed matches). Qualitatively, it handles fuzzy model numbers, incomplete attributes, transliterations, and mistranslations well, but struggles with product type synonyms (causing false negatives) and complementary products (causing false positives). We decompose each multi-modal classification into its component parts - the image classifier and the text classifier and examine the contribution of each type of classifier to the error. We find that a higher percentage of errors can be attributed to the image classifier (see Figure 4.A). Evaluation of 100 Stable Diffusion-generated images (split evenly between correct and incorrect classifications) using six quality criteria shows high fidelity overall (see Appendix Table 9): 97.40% of the generated images had correct product type, 88.31% were without hallucinations (hallucinations mainly due to unnecessary human figures), correct color where present in the product title (though 61.04% lacked color), 100% accurate logos when generated, 61.04% correct patterns, and 80.52% broadly matching real products (with slight degradation in TR). Multi-modal classification errors increase when generated images deviate from real products indicating that image faithfulness is critical.

## 5. Conclusion

We present the Cross Border E-commerce Selection Benchmarking Model, a zero-shot multimodal framework for cross-lingual product search and benchmarking. By integrating text similarity, visual similarity, and synthetic data augmentation, it addresses sparse, noisy, and multilingual competitor data, achieving 79% F1 across 10 language arcs and 17 categories, with online gains of 7.8% in product discovery and 20% in revenue. It consistently outperforms unimodal baselines and generalizes across languages, categories, and marketplaces, supporting large-scale deployment. Limitations include inherited biases from pre-trained models, potential noise from Stable Diffusion-generated images, empirically tuned similarity thresholds that may not generalize, and performance variation due to uneven language coverage. Future work will explore adaptive thresholds and domain-specific fine-tuning to address the same.

## References

- [1] A. Jain et al., "Mural: Multimodal, multi-task retrieval across languages," Google Research, 2021. [Online]. Available: <https://research.google/blog/mural-multimodal-multi-task-retrieval-across-languages/>. [Accessed: 08 June 2024].
- [2] M. S. U. Miah et al., "A multimodal approach to cross-lingual sentiment analysis with ensemble of Transformer and LLM," *Nature News*, 2024. [Online]. Available: <https://www.nature.com/articles/s41598-024-60210-7>. [Accessed: 08 June 2024].
- [3] T. Tayir and L. Li, "Unsupervised multimodal machine translation for low-resource distant language pairs," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 4, Art. no. 55, pp. 1–22, 2024. doi: 10.1145/3652161.
- [4] S. Yin et al., "A Survey on Multimodal Large Language Models," n.d. [Online]. Available: <https://arxiv.org/html/2306.13549v2>. [Accessed: 08 June 2024].

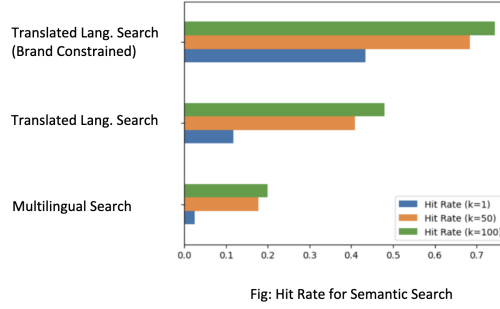


Figure 5. Ablation Analysis for Search

## A. Appendix

### Ablation Analysis for Search and Re-ranking:

Building on prior work leveraging paraphrase-xlm-r-multilingual-v1 for product substitution modeling, we implemented a baseline cross-lingual semantic similarity approach using cosine distance to identify selection gaps between a source e-commerce catalog and competing regional platforms. Product matching was performed at the product-category  $\times$  market-pair granularity.

**Baseline (Multilingual Search):** The baseline paraphrase-xlm-r-multilingual-v1 model yielded poor results: a 2% hit rate at rank 1 (i.e., for only 2% of competitor products, the top-ranked source catalog result was a correct match) and 20% hit rate at rank 100 (i.e., even among the top 100 nearest neighbors, only 20% of competitor products had any correct match in the source catalog).

**Translated-Language Search:** Through continued literature review of multilingual embedding strategies, we explored the LaBSE model—a language-agnostic BERT sentence embedding model fine-tuned for cross-lingual retrieval across 109 languages. Rather than relying on cross-lingual matching in native text, we translated competitor product titles into the source catalog language, trading cross-lingual alignment errors for potentially lower translation errors. Experiments the open-source M2M-100 translation model produced comparable accuracy with lower latency and reduced infrastructure costs as compared to Google Translate. Under this configuration, hit rate at rank 1 improved to 11% and hit rate at rank 100 improved to 50%.

**Brand Constrained Translated-Language Search:** The above approach was computationally prohibitive due to exhaustive  $n \times m$  pairwise comparisons across all source and competitor products within each category–market combination. We hypothesized that constraining the candidate search space by brand would simultaneously improve precision and reduce runtime. We implemented fuzzy cross-catalog brand matching using all-MiniLM-L6-v2 embeddings, achieving approximately 90% brand-matching accuracy (with failures attributable primarily to localization-induced brand name variations). This reduced runtime by approximately 90% while improving downstream matching quality. Brand- and category-constrained semantic search using LaBSE embeddings yielded a hit rate at rank 1 of 45% and hit rate at rank 100 of 78%.

**Ablation Analysis and Ensemble Reranking:** Ablation analysis on the improved pipeline revealed that while LaBSE embeddings performed well for candidate retrieval, ranking quality remained poor—correct matches frequently appeared within the top 100 candidates but not at rank 1, particularly for products with distinguishing numeric attributes (e.g., size, quantity) or noisy competitor titles. To address this reranking failure mode, we implemented an ensemble reranker combining substring-level token matching with cosine similarity as a tie-breaking signal, yielding significant improvements in top-1 precision, increasing to 60%.

Table 1. Destination Country Sample Size and Key Attribute Null Rate

| Destination Country | N<br>Prod-<br>ucts | Title | Brand<br>Name | Product<br>Type-<br>Prod-<br>uct<br>Group | Category<br>Desc | GTIN | EAN  | UPC  | Bread<br>Crumb1 |
|---------------------|--------------------|-------|---------------|-------------------------------------------|------------------|------|------|------|-----------------|
| Singapore           | 34,119             | 0.00  | 0.07          | 0.00                                      | 1.00             | 1.00 | 1.00 | 1.00 | 0.00            |
| Mexico              | 209,576            | 0.03  | 0.18          | 0.00                                      | 0.70             | 0.85 | 0.85 | 0.91 | 0.00            |
| Germany             | 290,740            | 0.00  | 0.50          | 0.00                                      | 0.86             | 0.00 | 0.00 | 0.91 | 0.00            |
| Turkey              | 764,575            | 0.00  | 0.04          | 0.00                                      | 0.80             | 0.72 | 0.73 | 0.94 | 0.01            |

Table 2. Number of Classifiers by Category

| Classifier Type | # Classifiers |
|-----------------|---------------|
| Text            | 90            |
| Image           | 12            |
| Multimodal      | 2160          |

Table 3. Summary of Text Preprocessing for Product Titles

| Text Preprocessing Type                                           | Source Marketplace reference plat-<br>form’s Product Title                        | Destination Marketplace Competi-<br>tor Product Title |
|-------------------------------------------------------------------|-----------------------------------------------------------------------------------|-------------------------------------------------------|
| Original Text                                                     | [Tamaris] 1-1-27227-28 001 38 /// 5 UK                                            | Tamaris 2536441 001 black Stiefelette                 |
| Cleaned Original Text                                             | Tamaris UK                                                                        | Tamaris black Stiefelette                             |
| Destination Text Normalized to<br>Source Country Language         | [Tamaris] 1-1-27227-28 001 38 /// 5 UK                                            | 2536441 001                                           |
| Cleaned Destination Text Normalized<br>to Source Country Language | Tamaris UK                                                                        |                                                       |
| Text Normalized to English                                        | [Tamaris] Heel Sandals Tamaris<br>Damen 1-1-27227-28 001 38<br>Black///White 5 UK | Tamaris 2536441 001 black ankle<br>boots              |
| Cleaned Text Normalized to English                                | Tamaris Heel Sandals Tamaris Damen<br>38 Black White 5 UK                         | Tamaris black ankle boots                             |

Table 4. Human Reviewed Sample Construction

| Source Marketplace ID | Destination Language | Num. Samples |
|-----------------------|----------------------|--------------|
| US                    | German (de)          | 140          |
| US                    | English (en)         | 78           |
| US                    | Spanish (es)         | 71           |
| US                    | Turkish (tr)         | 71           |
| DE                    | English (en)         | 47           |
| DE                    | Spanish (es)         | 75           |
| JP                    | German (de)          | 78           |
| JP                    | English (en)         | 112          |
| JP                    | Spanish (es)         | 93           |
| JP                    | Turkish (tr)         | 197          |

Table 5. Performance Metrics of Best in Class Classifiers (Aggregate)

| Classifier Name                                                | AUC  | Precision | Recall | F1   | TPR  | TNR  | FPR  | FNR  |
|----------------------------------------------------------------|------|-----------|--------|------|------|------|------|------|
| Baseline Cosine Classifier                                     | 0.57 | 0.84      | 0.16   | 0.27 | 0.16 | 0.98 | 0.02 | 0.84 |
| Multimodal (Sonnet Reason (OT) x Image Cosine Similarity (OT)) | 0.82 | 0.80      | 0.79   | 0.79 | 0.79 | 0.86 | 0.14 | 0.21 |
| Image Cosine Similarity (OT) Classifier                        | 0.75 | 0.73      | 0.68   | 0.70 | 0.68 | 0.82 | 0.18 | 0.32 |
| Text Sonnet Reason (OT) Classifier                             | 0.74 | 0.91      | 0.52   | 0.66 | 0.52 | 0.96 | 0.04 | 0.48 |

Table 6. Best In Class Multi-Modal Classifier Performance by Arcs

| source-destination country pair | Resource Class | AUC  | Precision | Recall | F1   | TPR  | TNR  | FPR  | FNR  |
|---------------------------------|----------------|------|-----------|--------|------|------|------|------|------|
| (US > SG)                       | High           | 0.80 | 0.81      | 0.96   | 0.88 | 0.96 | 0.63 | 0.37 | 0.04 |
| (US > MX)                       | High           | 0.84 | 0.80      | 0.95   | 0.86 | 0.95 | 0.74 | 0.26 | 0.05 |
| (US > TR)                       | Low            | 0.88 | 0.71      | 0.83   | 0.77 | 0.83 | 0.93 | 0.07 | 0.17 |
| (DE > SG)                       | High           | 0.76 | 0.90      | 0.97   | 0.94 | 0.97 | 0.56 | 0.44 | 0.03 |
| (DE > MX)                       | High           | 0.87 | 0.87      | 0.87   | 0.87 | 0.87 | 0.86 | 0.14 | 0.13 |
| (JP > DE)                       | Low            | 0.82 | 0.81      | 0.85   | 0.83 | 0.85 | 0.79 | 0.21 | 0.15 |
| (JP > SG)                       | Low            | 0.80 | 0.81      | 0.81   | 0.81 | 0.81 | 0.79 | 0.21 | 0.19 |
| (JP > MX)                       | Low            | 0.84 | 0.84      | 0.82   | 0.83 | 0.82 | 0.86 | 0.14 | 0.18 |
| (JP > TR)                       | Low            | 0.83 | 1.00      | 0.67   | 0.80 | 0.67 | 1.00 | 0.00 | 0.33 |

Table 7. Best In Class Multi-Modal Classifier Performance by product types

| Product Category    | AUC  | Precision | Recall | F1   | TPR  | TNR  | FPR  | FNR  |
|---------------------|------|-----------|--------|------|------|------|------|------|
| Drugstore           | 0.90 | 0.70      | 1.00   | 0.80 | 1.00 | 0.80 | 0.20 | 0.10 |
| PC                  | 1.00 | 1.00      | 1.00   | 1.00 | 1.00 | 1.00 | 0.00 | 0.00 |
| Apparel             | 0.90 | 0.70      | 0.90   | 0.80 | 0.90 | 0.90 | 0.10 | 0.10 |
| Beauty              | 0.80 | 0.80      | 1.00   | 0.90 | 1.00 | 0.60 | 0.40 | 0.00 |
| Furniture           | 1.00 | 1.00      | 1.00   | 1.00 | 1.00 | 1.00 | 0.00 | 0.00 |
| Pet Products        | 0.90 | 0.80      | 0.90   | 0.80 | 0.90 | 0.90 | 0.10 | 0.10 |
| Sports              | 1.00 | 1.00      | 1.00   | 1.00 | 1.00 | 1.00 | 0.00 | 0.00 |
| Toys                | 1.00 | 1.00      | 1.00   | 1.00 | 1.00 | 1.00 | 0.00 | 0.00 |
| Office Products     | 1.00 | 1.00      | 1.00   | 1.00 | 1.00 | 1.00 | 0.00 | 0.00 |
| Electronics         | 0.90 | 0.90      | 0.90   | 0.90 | 0.90 | 0.90 | 0.10 | 0.10 |
| Automotive          | 1.00 | 1.00      | 1.00   | 1.00 | 1.00 | 1.00 | 0.00 | 0.00 |
| Musical Instruments | 0.80 | 0.90      | 1.00   | 0.90 | 1.00 | 0.50 | 0.50 | 0.00 |
| Shoes               | 0.80 | 0.90      | 0.90   | 0.90 | 0.90 | 0.70 | 0.30 | 0.10 |
| Grocery             | 1.00 | 0.90      | 1.00   | 1.00 | 1.00 | 0.90 | 0.10 | 0.00 |
| Tools               | 0.90 | 0.80      | 0.80   | 0.80 | 0.80 | 0.90 | 0.10 | 0.20 |
| Home Improvement    | 0.90 | 0.70      | 0.80   | 0.80 | 0.80 | 0.90 | 0.10 | 0.20 |
| Kitchen             | 1.00 | 0.90      | 1.00   | 0.90 | 1.00 | 1.00 | 0.00 | 0.00 |

Table 8. Multi-Modal Classifier Performance Across High and Low Resource Arcs

| Resource Class | Baseline | Multimodal | Text  | Image |
|----------------|----------|------------|-------|-------|
| High Resource  | 0.286    | 0.880      | 0.763 | 0.782 |
| Low Resource   | 0.132    | 0.808      | 0.638 | 0.617 |

Table 9. Evaluation of Generated Image Characteristics

| Classification Criteria        | Incorrect (%) | Correct (%) | N/A (%) |
|--------------------------------|---------------|-------------|---------|
| Product Type Correct           | 2.60          | 97.40       | -       |
| Contains Hallucination (Human) | 7.79          | 3.90        | 88.31   |
| Pattern Correct                | 38.96         | 61.04       | -       |
| Color Correct                  | 9.09          | 29.87       | 61.04   |
| Generated Image Correct        | 19.48         | 80.52       | -       |

Table 10. Baseline Classifier Performance by Arcs

| source-destination country pair | Resource Class | AUC  | Precision | Recall | F1 Score | TPR  | TNR  | FPR  | FNR  |
|---------------------------------|----------------|------|-----------|--------|----------|------|------|------|------|
| (US → DE)                       | High           | 0.58 | 0.81      | 0.22   | 0.35     | 0.22 | 0.94 | 0.06 | 0.78 |
| (US → SG)                       | High           | 0.59 | 0.81      | 0.27   | 0.41     | 0.27 | 0.90 | 0.10 | 0.73 |
| (US → MX)                       | High           | 0.54 | 0.80      | 0.11   | 0.19     | 0.11 | 0.97 | 0.03 | 0.89 |
| (US → TR)                       | Low            | 0.50 | 0.00      | 0.00   | 0.00     | 0.00 | 1.00 | 0.00 | 1.00 |
| (DE → SG)                       | High           | 0.55 | 0.89      | 0.21   | 0.34     | 0.21 | 0.89 | 0.11 | 0.79 |
| (DE → MX)                       | High           | 0.53 | 0.75      | 0.08   | 0.14     | 0.08 | 0.97 | 0.03 | 0.92 |
| (JP → DE)                       | Low            | 0.59 | 1.00      | 0.18   | 0.30     | 0.18 | 1.00 | 0.00 | 0.83 |
| (JP → SG)                       | Low            | 0.57 | 1.00      | 0.14   | 0.24     | 0.14 | 1.00 | 0.00 | 0.86 |
| (JP → MX)                       | Low            | 0.52 | 0.75      | 0.07   | 0.13     | 0.07 | 0.98 | 0.02 | 0.93 |
| (JP → TR)                       | Low            | 0.50 | 0.00      | 0.00   | 0.00     | 0.00 | 0.99 | 0.01 | 1.00 |

Table 11. Baseline Classifier Performance by GL

| Product Category    | AUC  | Precision | Recall | F1 Score | TPR  | TNR  | FPR  | FNR  |
|---------------------|------|-----------|--------|----------|------|------|------|------|
| Drugstore           | 0.50 | 0.00      | 0.00   | 0.00     | 0.00 | 1.00 | 0.00 | 1.00 |
| PC                  | 0.90 | 1.00      | 0.80   | 0.89     | 0.80 | 1.00 | 0.00 | 0.20 |
| Apparel             | 0.60 | 0.67      | 0.25   | 0.36     | 0.25 | 0.96 | 0.05 | 0.75 |
| Beauty              | 0.55 | 0.80      | 0.17   | 0.29     | 0.17 | 0.93 | 0.07 | 0.83 |
| Furniture           | 0.63 | 1.00      | 0.25   | 0.40     | 0.25 | 1.00 | 0.00 | 0.75 |
| Pet Products        | 0.50 | 0.00      | 0.00   | 0.00     | 0.00 | 1.00 | 0.00 | 1.00 |
| Sports              | 0.47 | 0.00      | 0.00   | 0.00     | 0.00 | 0.94 | 0.06 | 1.00 |
| Toys                | 0.50 | 0.00      | 0.00   | 0.00     | 0.00 | 1.00 | 0.00 | 1.00 |
| Office Products     | 0.50 | 0.00      | 0.00   | 0.00     | 0.00 | 1.00 | 0.00 | 1.00 |
| Electronics         | 0.73 | 0.90      | 0.50   | 0.64     | 0.50 | 0.97 | 0.03 | 0.50 |
| Automotive          | 0.50 | 0.00      | 0.00   | 0.00     | 0.00 | 1.00 | 0.00 | 1.00 |
| Musical Instruments | 0.58 | 1.00      | 0.17   | 0.29     | 0.17 | 1.00 | 0.00 | 0.83 |
| Shoes               | 0.57 | 1.00      | 0.14   | 0.25     | 0.14 | 1.00 | 0.00 | 0.86 |
| Grocery             | 0.54 | 1.00      | 0.08   | 0.14     | 0.08 | 1.00 | 0.00 | 0.92 |
| Tools               | 0.50 | 0.00      | 0.00   | 0.00     | 0.00 | 1.00 | 0.00 | 1.00 |
| Home Improvement    | 0.53 | 1.00      | 0.05   | 0.10     | 0.05 | 1.00 | 0.00 | 0.95 |
| Kitchen             | 0.50 | 0.00      | 0.00   | 0.00     | 0.00 | 1.00 | 0.00 | 1.00 |

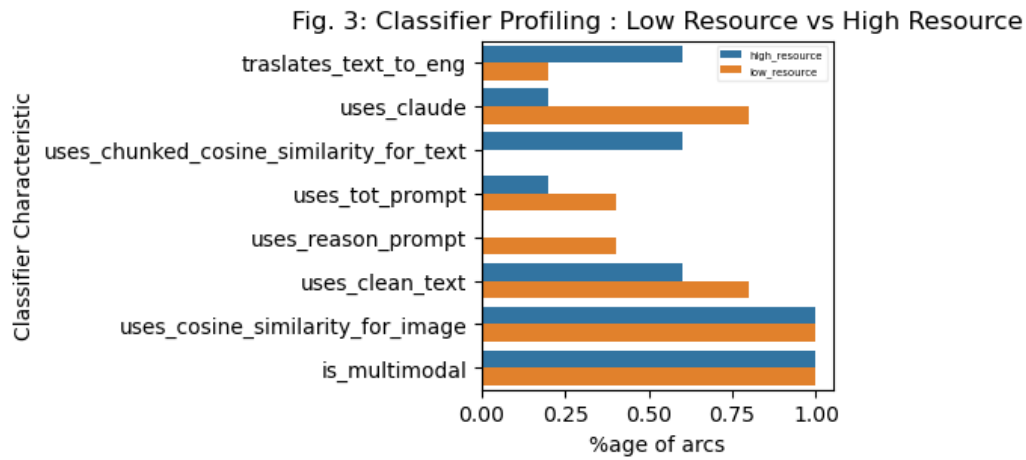


Figure 6. Classifier Profiling: Low vs High Resource

1. Competitor Product Title - Fender 2311000000 Frontman 10 Amplificador Guitarra | Same (Correct)



Generated By Stable Diffusion



Actual Product Image from the Internet

2. Competitor Product Title - tenis skechers dama mujer blanco memory foam original | Same (Correct)



Generated By Stable Diffusion



Actual Product Image from the Internet

3. Competitor Product Title - Marc O'Polo Rucksack Mod. Alexa | Close (Correct)



Generated By Stable Diffusion



Actual Product Image from the Internet

4. Competitor Product Title- Fisiogel Cuidado Hidratante Diario Crema Líquida Para Piel | Far (Incorrect)



Generated By Stable Diffusion



Actual Product Image from the Internet

Figure 7. Examples Of Competitor Product Images Generated By Stable Diffusion and Their Accuracy Classification