LocRegen: Cost-Efficient Redundancy Removal in Multilingual E-commerce Titles with Small Language Models

Bryan Zhang
Amazon.com
bryzhang@amazon.com

Stephan Walter
Amazon.de
sstwa@amazon.de

Luca Lomanto
Amazon.de
lucalo@amazon.de

Merve Arinik
Amazon.nl
merveari@amazon.nl

Abstract

E-commerce product titles often include redundant information that negatively impacts the user experience. Removing repeated words through restructuring and paraphrasing can make titles more concise and improve readability. While large language models can optimize titles, their computational cost makes them impractical for large-scale applications. In this paper, we first analyze the sources of repetition in multilingual product titles, then present LocRegen, a system that uses smaller language models to efficiently remove redundancies while preserving essential product attributes. Our experiments across five languages show that LocRegen with a 7B model substantially outperforms a 47B mixture-of-experts model: LocRegen achieves a 2.4% redundant title rate compared to 3.5% for the 47B model, and maintains a 3.8% overall error rate across all error categories including key product attribute omission compared to 8.4% for the 47B model. These results demonstrate that LocRegen delivers superior performance on cost-effective hardware with acceptable latency, making it practical for large-scale deployment where much larger models would be computationally prohibitive.

1 Introduction

With e-commerce shopping websites worldwide, products are accessible in different languages through global marketplaces. However, e-commerce catalogs (e.g., Amazon, Walmart) often contain products with excessively long titles that are difficult to read or exceed screen size limits (Rozen et al., 2021; Zhang et al., 2021). This leads to poor readability and customer experience, particularly when titles are used in other contexts such as being read aloud by voice assistants. Studies show that 65% of product titles contain 15 or more words (Rozen et al., 2021), often intentionally lengthened by sellers who include redundant keywords and additional product attributes for search engine optimization (SEO) (Xiao and Munro, 2019).

The challenge is further complicated by modern e-commerce stores that enable multilingual product discovery (Rücklé et al., 2019; Nie, 2010; Saleh and Pecina, 2020; Bi et al., 2020; Jiang et al., 2020; Lowndes and Vasudevan, 2021) and localize product information using machine translation systems (Way, 2013; Guha and Heger, 2014; Zhou et al., 2018; Wang et al., 2021). Title length often naturally increases during translation depending on the language pair (Zhang et al., 2024), necessitating additional optimization to improve customer experience by adhering to the *Gricean maxim of quantity*-being informative as required, no more and no less.

In industry settings, e-commerce product titles must also conform to *title templates*: structured formats that specify the key attributes relevant to a given product type. When optimizing titles, this attribute-related information in the original title must be preserved. Removing repeated words offers a safe and effective approach to such optimization for readability, and we focus on content words excluding function words. We define a repeated word as a content word

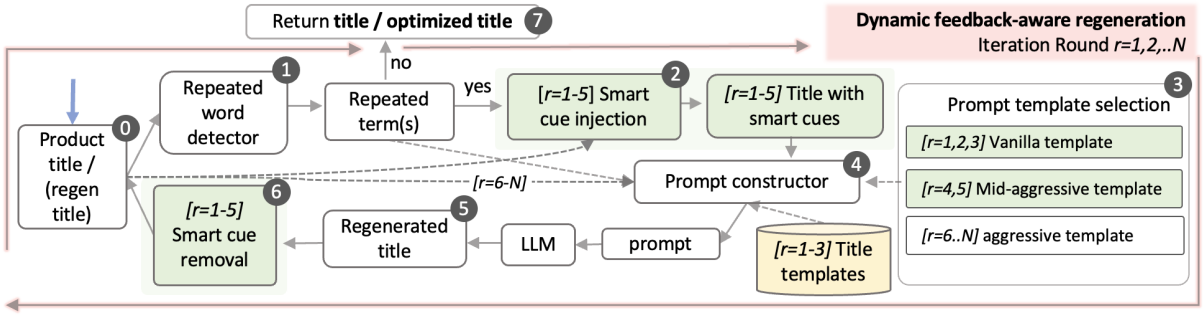

Figure 1: `LocRegen` system that uses smaller language models to efficiently remove repeated words while preserving essential product attributes.

having more than two occurrences in a title, and a redundant title as containing one or more repeated content words. The Repeated Word Removal (RWR) task aims to (1) restructure and paraphrase titles to reduce content word repetition (limiting occurrences to 2 or less) while (2) preserving all key attributes in the title template, as shown below:

Original title: [MASKED BRAND] Large Desk Mat, Office Desk Pad, Computer Desk Mat, Laptop Mat for Desk, Desk Protector Mat, Desktop Mat, Desk Writing Pad, Desk Blotter Pad, Desk Cover Mat (80x40cm, Green)

Optimized title: [MASKED BRAND] Large Desk Mat for Office Computers, Laptops, and Desktops (80x40cm, Green)

The original title contains repeated words *Desk, Mat* and *Pad*, which are reduced to 2 occurrences or fewer in the optimized version. The optimized title is more concise while retaining all key attributes for the *desk mat* product type including "brand", "product type", "color", "size".

While multilingual large language models (LLMs) have shown promising results for summarization tasks and could potentially perform RWR through title regeneration, there are still significant challenges remaining in the e-commerce context:

Cost and Scalability: Regenerating a large volume of multilingual titles using LLMs requires costly and powerful hardware, posing significant challenges when processing titles at industrial scale. While smaller LLMs require significantly less powerful hardware — enabling deployment on cost-effective configurations such as a single mid-range GPU, they can have inferior performance in complex tasks such as removing repeated words while preserving key product attributes.

Dynamic Business Requirements: Title templates undergo frequent changes to accommodate business needs, either through refinements to existing product types or additions for new categories. This requires models to adapt quickly.

Therefore, in this paper we first analyze redundancy in multilingual e-commerce titles, to understand the need for repeated word removal (RWR). Second, we propose `LocRegen`, a cost-efficient system that leverages a smaller language model for RWR through title regeneration. Our system consists of three key components: (1) *Dynamic Feedback-aware Regeneration (DFR) framework*, (2) *smart cue augmentation*, and (3) a *Repeated Word Detection (RWD)*. These components enhance the small model's performance through iterative title regeneration and provide effective guidance for preserving key attributes while removing repeated words. Additionally, the observation-driven smart cue injection for small LLMs can generalize to other generation tasks.

The manual audit results of the experiments show that `LocRegen`, using a 7B model, reduces the redundant title rate to 2.4% on average of five languages, compared to 3.5% with the substantially larger 47B model without `LocRegen`. Additionally, `LocRegen` maintains a 3.8% error rate across all error categories and languages including key product attribute omission, substantially outperforming 47B model's 8.4%. These results demonstrate that our system can enable smaller models to outperform much larger alternatives, offering a practical and cost-efficient solution for industrial-scale applications.

2 Redundancy in Multilingual E-commerce Titles

While sellers can intentionally include repeated content words for marketing and search optimization, the localization process itself can also introduce repeated content words, which are natural translation outcome and not localization errors. Our goal is to optimize such redundant titles for readability through RWR task.

language pairs	redund. rate	language pairs	redund. rate
German-Spanish	50%	Italian-French	29%
German-English	46%	Spanish-Italian	27%
Spanish-French	41%	Spanish-English	27%
French-German	35%	English-German	27%
Italian-German	35%	Italian-Spanish	25%
Spanish-German	34%	English-Italian	23%
French-Italian	32%	English-Spanish	22%
Italian-English	32%	English-French	19%

Table 1: The percentage of the e-commerce non-redundant source titles that become redundant titles in the target language after localization. We use random experimental data sample which has over 10K titles per language pair.

A product title containing one or more repeated content words is considered redundant. While source titles with redundancies typically remain redundant after localization, a significant portion of non-redundant source titles become redundant during the process. We observe redundancy rates range from 19% to 50% depending on the language pair as shown in Table 1. Therefore, a large volume of "native" titles and localized titles need RWR task in e-commerce.

2.1 Localization-Induced Repeated words

Through an extensive analysis of multilingual product titles, we have identified three primary mechanisms through which localization naturally creates repeated words. These repeated words are a normal outcome of translation and should not be considered errors in the localization process.

Compound Word Decomposition: Languages such as German and Swedish naturally use compound words such as the examples below: the German word *Flasche (bottle)* in compound words *Trinkflasche*, *Sportflasche* and *Wasserflasche*, the Swedish word *spö (rod)* in compound words *fiskespö*, *Spinnspö* and *saltvattensfiskespö*. When translated into languages that do not use such compounds, like English, these words are typically split into their components, resulting in repeated occurrences of words like *Bottle*.

Example 1

de: [MASKED BRAND] Trinkflasche, 1l, 700ml, 500ml Trinkflasche Kinder Auslaufsicher, Wasserflasche mit Motivierender Zeitmarkierung, BPA Frei Tritan Sportflasche für Schule, Sport, Fahrrad, Arbeit, Fitness

en: [MASKED BRAND] Drinking Bottle, 1 Litre, 700 ml, 500 ml Children's Water Bottle, Leak-Proof, with Motivational Time Marker, BPA-Free Tritan Sports Bottle for School, Sports, Cycling, Work, Fitness

Example 2

sv: [MASKED BRAND] Spinnspö lättvikts 24T kolfiberrämne slitstarkt fiskespö, premium korkhandtag, mångsidigt sötvattens-och saltvattensfiskespö för gädda, abborre och gös-tillgängliga i storlekar

en: [MASKED BRAND] Spinning Rod Lightweight 24T Carbon Fiber Subject Heavy Duty Fishing Rod, Premium Cork Handle, Versatile Freshwater and Saltwater Fishing Rod for Pike, Perch and Zander-available in sizes

Vocabulary Asymmetry: Vocabulary asymmetry between languages can cause different source words to map to the same target word. Following Nida and Taber's principle of functional equivalence in translation, this mapping often leads to repetition (Nida and Taber, 1974). As the following example shows, the distinct English words *Kids*, *Boys* and *Girls*, *Children* may all translate to *Kinder* in German, which is correct but creating repetition in the target title.

en: [MASKED BRAND] Kids Animal costumes Boys Girls Pijamas Fancy Dress outfit Cosplay Children (Tiger, XL (For kids 120-140 cm tall))

de: [MASKED BRAND] Tierkostüme für Kinder, Jungen, Mädchen, Pyjama, Kostüm, Cosplay, Kinder (Tiger, XL (für Kinder 120–140 cm groß))

Morphological Richness Differences: Morphological differences between languages can create repetition when a source language's richer forms collapse into fewer target forms. As the example below shows, Italian distinguishes gender in nouns, with *Neonato* (masculine) and *Neonata* (feminine) both translating to *Newborn* in English, where grammatical gender doesn't exist. This morphological simplification in English leads to repetition in the translation.

it: [MASKED BRAND], Body Neonato e Neonata, Senza Manica, con Comoda Apertura a Patello, Designed in Italy, Abbigliamento Neonato e Neonata 0-24 Mesi, Idee Regalo Nascita

en: [MASKED BRAND], Newborn and Newborn Baby Bodysuit, Sleeveless, with Comfortable Snap-Button Opening Opening, Designed in Italy, Newborn and Newborn Clothing 0-24 Months, Birth Gift Ideas

3 LocRegen system

3.1 System Overview

Our approach is based on the observation that while larger LLMs can perform this task with simple instructions in a single prompt, smaller LLMs require a more strategic approach. By analyzing error patterns and general model behavior in smaller LLMs, we can introduce additional cues and hints that reduce task difficulty, enabling smaller models to achieve comparable or even su-

perior performance to much larger LLMs. Moreover, smaller LLMs typically have significantly lower inference costs and latency, which allows for multiple prompts and inference iterations. We propose `LocRegen`, a cost-efficient and effective title regeneration system for the Repeated Word Removal (RWR) task. The system consists of three key components: First, a **multilingual Repeated Word Detection (RWD)** component identifies repeated words and provides a direct "feedback" to the model. Second, the **Dynamic Feedback-aware Regeneration (DFR) framework** leverages this feedback to iteratively regenerate titles until either all repeated words are eliminated (positive feedback) or a predefined iteration threshold is reached. Throughout the regeneration process, different prompt templates are selected based on the iteration round: prompt templates with milder instructions are applied in early iterations, while progressively those with harsher instructions are used in later rounds, thereby increasing the likelihood of successfully removing repeated words. Third, the **Smart Cues component** strategically inserts cues around specific repeated words that appeared earlier in the title, signaling to the LLM in the prompt not to remove them. This mechanism helps the LLM to have better focus and preserve essential attribute-related information defined in the title template.

Figure 1 illustrates the `LocRegen` workflow, where steps 0–6 constitute one regeneration round. Given an input product title, the RWD component first detects repeated words (Step 1). If none are detected, the title is returned immediately. Otherwise, the regeneration process proceeds as follows: For rounds 1–5, smart cues markers are injected around the first and second occurrences of the repeated words (Step 2). Based on the current round number, an appropriate prompt template is selected (Step 3), where each template requires different inputs: rounds 1–3 utilize product-type-specific title templates, rounds 1–5 incorporate titles with smart cues, and rounds beyond 5 use titles directly without cue markers (Step 4). The selected prompt is then used to regenerate the title via a small LLM (Step 5). For rounds 1–5, cues markers are subsequently removed from the regenerated title (Step 6). The process returns to Step 1 for repeated word detection. If repetitions persist, regeneration continues to the next round; otherwise, the optimized title is returned as the final output.

3.2 Observation-based solution: Smart Cues

We observe that smaller language models tend to remove repeated words from the beginning of titles, potentially eliminating crucial brand/product type information. This phenomenon becomes more pronounced when regenerating titles in certain languages such as Italian or Spanish[1]. To address this, we introduce smart cues: we implement smart cue marker insertions around the first one or two occurrences of repeated words from the beginning of a title, and include instructions in the prompt not to remove marked words-*Don't remove those redundant words surrounded by "<MARKER><\MARKER>"*...[2].This enables the LLM to focus on reducing repeated words later in the title while preserving key attribute information. Those markers essentially reduce difficulty of the task by giving the model the guidance. Once the regeneration is complete, we remove all the markers in the regenerated titles.

3.3 Multiple Prompt Templates

Redundancy removal is a delicate task that requires preserving key attributes while minimizing unnecessary paraphrasing. To address this challenge, we introduce a progressive prompt template strategy ranging from gentle to aggressive instructions for repeated word removal. Within our dynamic feedback-aware generation framework, we progressively adjust prompt templates as the regeneration process continues, as illustrated in Figure 1.

We begin with a *vanilla prompt* for the initial rounds (1–3). This template incorporates two key components: (1) the title with smart cue injection, and (2) a product-type-specific title template. For example, the product type *Cabinet* uses the template [*brand, style, room-type, size, material, mounting-type, door-style*]. The prompt includes explicit instructions to preserve attribute information: *"Please identify the key information in the title for each attribute in the list `[TITLE TEMPLATE]`. The new title should retain all the attribute information if they exist in the original title..."*[4]

As the process advances, we transition to a *mid-aggressive prompt* for rounds 4–5. This template maintains smart cue integration, but removes the explicit title template guidance, allowing for more

[1]Refer to Table3 in section 5, [partial]/[full] system comparison for error *Key info Omitted*

[2]See the prompt templates in the section 7

flexible redundancy removal. Finally, for rounds 6 and beyond, we employ an *aggressive prompt* that eliminates both smart cue injection and title template considerations, focusing solely on aggressive redundancy removal.

Table 2 summarizes the key differences among these templates. This progressive strategy aligns prompt intensity with redundancy removal difficulty, maximizing the likelihood of successful title optimization while preserving essential product information. Table 7 in Appendix presents the major prompt content and instructions for each template.

Prompt template	Round	Smart cue	Title template
vanilla	1st-3rd	yes	yes
mid-aggressive	4th-5th	yes	no
aggressive	6th- above	no	no

Table 2: Prompt templates throughout different rounds of title regeneration and associated components

3.4 Repeated Word Detector (RWD)

The multilingual Repeated Word Detector (RWD) serves as the critical feedback generator in the `LocRegen` system. To process multilingual product titles from worldwide stores, it must handle morphologically rich languages where words can appear in various inflected forms. We use `spaCy-v3.6`[3] and its Medium ML models for lemmatization and Part-of-Speech (POS) tagging. Lemmatization produces actual words and requires understanding of word context while maintaining distinctions between different word meanings, making it more appropriate to group the words.

For repeated word detection, we first obtain the base form of each word through lemmatization for occurrences counting. We exclude functional words (such as *"the", "a"* in English, *"el", "la"* in Spanish) using the following POS tags: `ADP, CCONJ, CARDINAL, SCONJ, DET, NUM`. Additionally, we maintain a special expression cache to exclude brand names and other legitimate expressions containing repeated words. A lower-cased lemma appearing more than twice (>2) in the title is considered a repeated word.

4 Experimental Setup

Languages: we use titles in the following five languages for our experiment: English, German, Spanish, Italian, and French.

LLMs and inference: Qwen2.5-7B-Instruct-GPTQ-Int8 (8-bit quantized)[4] is used for the `LocRegen` system, requiring approximately 7 GB of GPU memory. We use `vllm 0.6.3.post1`[5] as the inference framework. We use a 47B model `Mixtral-8x7B-Instruct` (Jiang et al., 2024)[6] for comparison. With 40B more parameters, matching the throughput of our 7B model requires more powerful and costly hardware—a critical consideration when processing titles at industrial scale. Temperature is set to 0.1 for LLM inference.

Manual audit test data: For each language, we have randomly sampled approximately over 1,000 product titles from their respective e-commerce stores (EN, DE, ES, IT, FR). Each language dataset includes both original titles created in that language and titles localized from the other four languages. All sampled titles contain repeated content words and span over 150 product types.

Manual audit: The auditors are native speakers of the target language and are provided with: (a) original and regenerated titles, (b) product type information of the titles, and (c) a list of title templates-essential attributes for each product type, as defined by business teams. Auditors received task-specific training. Ambiguous cases were resolved through collaborative discussion to ensure consistency. We note that the reported error rates are point estimates from manual audits without confidence intervals; given the sample sizes ($\sim$1K titles per language), minor differences between systems should be interpreted with caution.

Title error metrics: manual audit is conducted on the regenerated titles on the following metrics: (1) *Redundancy Present*: auditors need to detect whether any repeated words (more than two occurrences) present. (2) *Key Information Omitted*: the auditors can check whether any key attribute information present in the original title for a given product type is missing in the regenerated titles (3) *Hallucination Present*: containing information about the product in the regenerated title that did not exist in the original title and can materially change the product's offering and mislead customers (4) *Linguistic Errors Present*: containing

[3]https://github.com/explosion/spaCy (MIT License)

[4]https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-GPTQ-Int8 (Apache license 2.0) (Team, 2024)

[5]https://github.com/vllm-project/vllm (Apache-2.0)

[6]https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1 (Apache license 2.0)

	LocRegen												Mixtral-8x7B-Instruct					
	[Partial system]: DFR + RWD						[Full system]: DFR + RWD + smart cues						RWD (Single-pass)					
	DE	**IT**	**ES**	**EN**	**FR**	**ave**	**DE**	**IT**	**ES**	**EN**	**FR**	**ave**	**DE**	**IT**	**ES**	**EN**	**FR**	**ave**
Redundancy	0.0%	4.0%	2.0%	0.0%	1.0%	1.4%	2.0%	1.6%	5.0%	2.0%	1.5%	2.4%	0.0%	7.0%	8.5%	0.5%	1.5%	3.5%
Key Info Omitted	1.0%	34.3%	45.2%	8.8%	17.8%	21.4%	0.5%	18.2%	8.5%	13.0%	17.1%	11.5%	4.5%	25.1%	5.5%	17.0%	15.1%	13.4%
Hallucination	0.5%	5.1%	8.1%	6.2%	0.5%	4.1%	1.0%	1.5%	1.5%	5.5%	0.5%	2.0%	2.5%	14.6%	14.0%	7.0%	2.5%	8.1%
Linguistic Errors	0.0%	2.5%	7.6%	0.0%	1.0%	2.2%	0.0%	3.5%	0.0%	0.0%	0.5%	0.8%	2.5%	19.6%	15.5%	0.0%	5.5%	8.6%
Context Change	7.2%	23.2%	29.4%	8.8%	11.2%	16.0%	1.0%	3.5%	1.0%	4.0%	1.5%	2.2%	6.5%	17.1%	7.5%	7.5%	3.0%	8.3%
average						9.0%						3.8%						8.4%

Table 3: Manual audit error rates for LocRegen (1) [Full system] with all the proposed components, LocRegen (2)[Partial system] with Dynamic Feedback-aware Regeneration (DFR framework) including 3 prompt templates and RWD without smart cues component. (3) Mixtral-8x7B-Instruct using RWD is for comparison.

words that do not exist in the same language, or are written in a nonsensical way that would be unintelligible to a native speaker (5) *Context Change Present*: A contextual change is where the regenerated title no longer means exactly the same thing as the original title as a result of newly arranged words or phrases inadvertently changing the context.

Experiment configurations:

LocRegen **[partial system]**: uses Dynamic Feedback-aware Regeneration (DFR framework) with maximum 6 iterations with 3 prompt templates (*vanilla, mid-aggressive, aggressive*). It does not include the smart cues component, and there are also no instructions in prompts related to the smart cues component.

LocRegen **[full system]**: includes all the proposed components, specifically the *[partial system]* configuration plus the smart cue component.

Mixtral-8x7B-Instruct: regenerate the title once (single pass) using RWD and vanilla prompt template (standard prompt) with no smart cue-related instructions.

5 Results and Analysis

RWD Accuracy: We conducted a separate manual audit on the RWD component. The accuracy exceeds 97% across all five languages, for both identifying redundant titles and detecting repeated words. Additionally, we observe that RWD labels approximately 1% of titles as redundant that auditors deem non-redundant. These false positives are primarily function terms we intend to exclude, with most errors attributable to POS tagging inaccuracies. Given this high accuracy, RWD provides a reliable feedback signal for the DFR framework.

Standalone LLM Performance:

We first assessed the standalone repeated word removal capability of two Qwen2.5 variants—the 8-bit quantized 7B and the 1.5B—by regenerating titles in a single pass using RWD and a vanilla prompt (without smart cue instructions). Using RWD as an automatic evaluation metric on over 10K redundant titles per language, the 7B model achieved a 32% average redundancy rate across five languages (Table 4, r1), while the 1.5B model showed 41% (Table 5, r1). With iterative regeneration, the 7B model reduced redundancy to 0.8%–2.7% within 6 rounds, while the 1.5B model plateaued at approximately 16% with minimal improvement beyond round 6, indicating a capacity limitation. Based on these results, we selected the 7B model and conducted a manual audit of LocRegen [partial system] using our designated test sets. The key question is whether the proposed components of LocRegen can bridge the performance gap between a standalone small LLM and a substantially larger model.

	r1	**r2**	**r3**	**r4**	**r5**	**r6**
DE	28.6%	16.5%	12.7%	11.1%	7.1%	1.8%
FR	32.1%	16.6%	11.4%	9.6%	6.3%	1.6%
IT	32.0%	15.3%	11.0%	9.5%	6.2%	2.2%
ES	29.4%	13.9%	9.3%	4.8%	3.9%	0.8%
EN	41.1%	22.7%	18.2%	16.1%	9.7%	2.7%
Aver	32.7%	17.0%	12.5%	10.2%	6.6%	1.8%

Table 4: LocRegen with base model Qwen2.5-7B-Instruct-GPTQ-Int8. The percentage of redundant title detected by the repeated word detector (RWD) during each round of the Dynamic feedback-aware iterative regeneration, each language has 1K titles

	r1	**r2**	**r3**	**r4**	**r5**	**r6**	**r7**	**r8**	**r9**
DE	36%	23%	19%	19%	18%	18%	18%	17%	17%
FR	43%	27%	21%	19%	18%	17%	17%	17%	17%
IT	39%	24%	20%	18%	17%	17%	17%	16%	16%
ES	40%	23%	19%	17%	16%	15%	15%	15%	15%
EN	47%	25%	19%	18%	17%	17%	17%	17%	17%
Aver	41%	25%	20%	18%	17%	17%	17%	16%	16%

Table 5: LocRegen with base model Qwen2.5-1.5B-Instruct. The percentage of redundant title detected by the repeated word detector (RWD) during each round of the dynamic feedback-aware iterative regeneration, each language has 1K titles

Manual Audit Results: Table 3 shows the results of manual audits comparing regenerated titles from LocRegen [partial system] and [full system]. LocRegen [partial system] reduces redundancy rate from approximately 32% (standalone LLM baseline) to 1.4% across all five languages, demonstrating the effectiveness of the iterative feedback mechanism. However, it performs worse than Mixtral-8x7B-Instruct in key information omission (21.4% vs 13.4%) and shows only modest improvement in hallucinations (4.1% vs 8.1%). These results suggest that while the iterative feedback mechanism alone can remove repeated words, it does not effectively preserve attribute-related information defined in title templates—the core requirement of the task. This motivated us to investigate small LLM behavior and develop the smart cue solution described in Section 3.2.

LocRegen [full system] with smart cues reduces key information omission from 21.4% to 11.5%, validating our observation in Section 3.2 that small LLMs tend to remove repeated words from title beginnings. It also reduces hallucination, linguistic errors, and context change because smart cues allow the model to focus on repeated words appearing later in the title, requiring less paraphrasing and restructuring. Although an average redundancy rate shows a slight increase (1.4% to 2.4%), the overall error rate is more balanced, dropping from 9.0% to 3.8% compared to [partial system]. Moreover, LocRegen [full system] outperforms the substantially larger Mixtral-8x7B-Instruct across all error categories. This better performance is particularly significant given that LocRegen accomplishes this with substantially fewer parameters. While the DFR framework could in principle be applied to Mixtral-8x7B-Instruct, this would further multiply its already substantial computational cost—requiring multi-GPU infrastructure for each iteration—making it impractical for industrial-scale deployment. The purpose of LocRegen is precisely to demonstrate that a smaller model, when enhanced with observation-driven techniques, can surpass the single-pass performance of a much larger model while remaining deployable on cost-effective hardware.

Computation platform and latency The primary cost advantage of LocRegen lies not in reduced computation per se, but in substantially lower hardware requirements: the quantized 7B model can be served on a single commodity GPU, whereas Mixtral requires multi-GPU infrastructure. Batch inference of the test sets was performed using vLLM on a machine equipped with 8 NVIDIA A10G GPUs (178 GiB total GPU memory), 192 vCPUs, and 768 GiB system RAM. LocRegen using Qwen2.5-7B-Instruct-GPTQ-Int8 with up to 6 iterations completed processing in approximately 70% of the time required by Mixtral-8x7B-Instruct in a single pass as most titles converge within 1–3 iterations. On a cost-effective configuration with 1 NVIDIA T4 GPU (16 GiB GPU memory), 4 vCPUs (Intel Xeon P-8259L), and 16 GiB system RAM, LocRegen latency increased by approximately 53% compared to the 8×A10G configuration, while Mixtral-8x7B-Instruct was computationally infeasible on this hardware as its model weights far exceed the available GPU memory.

Title length reduction: As repeated word removal in titles can intuitively optimize title length, we further investigate the title length reduction. Table 6 shows consistent title length reductions across all five languages, with mean reductions ranging from 27.5% to 32.1% and median reductions from 32.1% to 36.5%. Italian and French demonstrate the strongest compression rates (mean reductions of 32.1% and 31.4% respectively), while English shows a more moderate 27.5% reduction. These consistent results across different languages demonstrate LocRegen's effectiveness in reducing title length while preserving essential information. Our analysis shows the length of regenerated titles is reduced approximately 30% on average across 5 languages as Table 6, while preserving essential information.

These results demonstrate that our proposed approaches in LocRegen can effectively overcome the limitations of the small LLM's capacity and offer a more cost-efficient and effective approach in the industrial setting to remove the repeated words in product titles while preserving the key attribute information.

6 Related work

In an industry setting, Repeated Words Removal (RWR) task for product titles is typically conducted through a title length optimization step,

Language	Δ Median (%)	Δ Mean (%)
FR	-34.8%	-31.4%
IT	-36.5%	-32.1%
ES	-32.1%	-29.6%
EN	-32.1%	-27.5%
DE	-32.1%	-29.2%

Table 6: Percentage reduction in title length between original and regenerated titles across languages. Values show both median and mean reductions, measured in characters. Negative percentages indicate shorter regenerated titles.

which employs techniques such as monolingual summarization (Fetahu et al., 2023; Sun et al., 2018), text truncation (Wang et al., 2020; Guan et al., 2022) and manual editing - focusing on length reduction rather than explicitly targeting redundancy. Recent work has explored neural models for product title optimization, including masked text scoring (Samar et al., 2018) and user-sensitive adversarial training (Wang et al., 2020). While these approaches show promise, they typically require significant training data and computational resources. The output length of neural machine translation is studied (Lakew et al., 2019), they focus on general translation instead of titles and their requirements in ecommerce. The multilingual title length problem has also been studied in the context of localization (Zhang et al., 2024) studies encoder-decoder transfer models for cross-lingual title summarization rather than addressing redundancy removal and product attributes preservation from title templates. Other approach utilizes product title templates to structure information (Xiao and Munro, 2019), though primarily for title generation rather than redundancy removal. To our knowledge, our work is the first to specifically address multilingual product title redundancy through a cost-efficient approach that: 1) explicitly preserves template-specified key attributes, 2) easily adapts to changes of product-type-specific title templates in e-commerce, 3) leverages small language models with specialized augmentation techniques, and 4) provides an iterative feedback mechanism for precise redundancy removal across languages.

7 Conclusion

In this paper, we analyze redundancy in multilingual e-commerce titles, demonstrating how repeated words naturally emerge during localization through various linguistic phenomena. We present `LocRegen`, a cost-efficient system that enables a 7B parameter model to effectively remove redundancy while preserving essential product attributes. Our experiments across five languages show that LocRegen with a 7B model substantially outperforms a 47B mixture-of-experts model: `LocRegen` achieves a 2.4% redundant title rate compared to 3.5% for the 47B model, and maintains a 3.8% overall error rate across all error categories including key product attribute omission compared to 8.4% for the 47B model. Combined with consistent title length reductions of 27.5–32.1%, these results demonstrate that complex multilingual tasks can be accomplished efficiently with smaller models when enhanced through observation-driven techniques, offering a practical blueprint for cost-efficient industrial-scale applications.

Sustainability Statement

This work specifically aims to reduce computational costs for multilingual title optimization by using a quantized 7B parameter model deployable on a single commodity GPU, rather than relying on substantially larger models requiring multi-GPU infrastructure. Experiments were conducted on a machine with 8 NVIDIA A10G GPUs, though the proposed system is designed to operate on hardware as minimal as a single NVIDIA T4 GPU (16 GiB). By enabling smaller models to outperform larger alternatives, this work contributes to reducing the environmental impact of large-scale NLP applications in e-commerce.

Appendix

Prompt Templates

Table 7 presents the instruction part of each template. The notation used throughout the prompts is defined as follows: `[LANGUAGE]` denotes the language of the title (e.g., "English", "German"); `[REDUNDANT WORD LIST]` represents the list of repeated words identified by the Redundant Word Detector (RWD); `[TITLE]` refers to either the title with smart cue injection (rounds 1–5) or the original title without smart cues (rounds 6 and beyond); `[SMART CUE]` indicates marking tags such as "<MARKER><\MARKER>" in our experiment; `[WORDS BETWEEN SMART CUES]` denotes repeated words enclosed by smart cues (e.g., <MARKER>bag<\MARKER> for the repeated word *bag*); and `[TITLE TEMPLATE]` specifies a product-type-specific set of relevant attributes that varies across different product categories.

Template Name	Instruction parts of the prompt templates
Vanilla	... reduce the redundant words [REDUNDANT WORD LIST] in the [LANGUAGE] title "[TITLE]", restructure the title and make sure the redundant words [REDUNDANT WORD LIST] appear only once or twice in the new [LANGUAGE] title, you can reduce their occurrences either through combining phrases, rephrasing like a linguist. Don't remove those redundant words surrounded by "[SMART CUE]" e.g. [WORDS BETWEEN SMART CUES]. You should only restructure title when it is necessary to reduce the redundant word occurrences, and such restructure also needs to be minimized. Please also identify the key information in the title for each attribute in the list [TITLE TEMPLATE]. The new title should retain all the attribute information if they exist in the original title, and also retain as many words as possible from the original title, and maintains the original context and meaning...
mid-aggressive	... reduce the redundant words [REDUNDANT WORD LIST] at the end of the [LANGUAGE] title "[TITLE]", restructure the title like a linguist and make sure the redundant words [REDUNDANT WORD LIST] appear only once or twice in the new [LANGUAGE] title, you can reduce their occurrences either through rephrasing like a linguist, or combining phrases or simply remove those redundant words appearing at the end of the title if it doesn't change the meaning of the title. Don't remove the those redundant words surrounded by "[SMART CUE]" e.g. [WORDS BETWEEN SMART CUES]. The new title should maintain the original context and meaning. ...
aggressive	... reduce the redundant words [REDUNDANT WORD LIST] at the end of the [LANGUAGE] title "[TITLE]", restructure the title like a linguist and make sure the redundant words [REDUNDANT WORD LIST] appear only once or twice in the new [LANGUAGE] title, you can reduce their occurrences either through rephrasing like a linguist, or combining phrases or simply remove those redundant words appearing at the end of the title if it doesn't change the meaning of the title. The new title should maintain the original context and meaning...

Table 7: Prompt templates for repeated word removal in product title optimization. Three templates are used in iterative regeneration cycles until no repeated words remain. The templates provide progressively aggressive instructions: the *vanilla* template emphasizes minimal restructuring while preserving essential attributes and original phrasing; the *mid-aggressive* template allows removal of redundant words at title endings when meaning is preserved; and the *aggressive* template prioritizes redundancy elimination over attribute preservation. All templates protect words marked with [SMART CUE] tags from removal.

References

Bi, Tianchi, Liang Yao, Baosong Yang, Haibo Zhang, Weihua Luo, and Boxing Chen. 2020. Constraint translation candidates: A bridge between neural query translation and cross-lingual information retrieval.

Fetahu, Besnik, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023. InstructPTS: Instruction-tuning LLMs for product title summarization. In Wang, Mingxuan and Imed Zitouni, editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 663–674, Singapore, December. Association for Computational Linguistics.

Guan, Xinyi, Shun Long, Weiheng Zhu, Silei Cao, and Fangting Liao. 2022. Mask-based text scoring for product title summarization. In 2022 8th International Conference on Systems and Informatics (ICSAI), pages 1–6.

Guha, Jyoti and Carmen Heger. 2014. Machine translation for global e-commerce on ebay. In Proceedings of the AMTA, volume 2, pages 31–37.

Jiang, Zhuolin, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. 2020. Cross-lingual information retrieval with BERT. In Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020), pages 26–31, Marseille, France, May. European Language Resources Association.

Jiang, Albert Q., Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Lakew, Surafel Melaku, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. In Niehues, Jan, Rolando Cattoni, Sebastian Stüker, Matteo Negri, Marco Turchi, Thanh-Le Ha, Elizabeth Salesky, Ramon Sanabria, Loic Barrault, Lucia Specia, and Marcello Federico, editors, Proceedings of the 16th International Conference on Spoken Language Translation, Hong Kong, November 2-3. Association for Computational Linguistics.

Lowndes, Mike and Aditya Vasudevan. 2021. Market guide for digital commerce search.

Nida, Eugene Albert and Charles Russell Taber. 1974. The theory and practice of translation, volume 8. Brill Archive.

Nie, Jian-Yun. 2010. Cross-language information retrieval. Synthesis Lectures on Human Language Technologies, 3(1):1–125.

Rozen, Ohad, David Carmel, Avihai Mejer, Vitaly Mirkis, and Yftah Ziser. 2021. Answering product-questions by utilizing questions from other contextually similar products. In Toutanova, Kristina, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 242–253, Online, June. Association for Computational Linguistics.

Rücklé, Andreas, Krishnkant Swarnkar, and Iryna Gurevych. 2019. Improved cross-lingual question retrieval for community question answering. In The World Wide Web Conference, WWW '19, page 3179–3186, New York, NY, USA. Association for Computing Machinery.

Saleh, Shadi and Pavel Pecina. 2020. Document translation vs. query translation for cross-lingual information retrieval in the medical domain. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6849–6860, Online, July. Association for Computational Linguistics.

Samar, Thaer, Myriam C. Traub, Jacco Ossenbruggen, Lynda Hardman, and Arjen P. Vries. 2018. Quantifying retrieval bias in web archive search. Int. J. Digit. Libr., 19(1):57–75, mar.

Sun, Fei, Peng Jiang, Hanxiao Sun, Changhua Pei, Wenwu Ou, and Xiaobo Wang. 2018. Multi-source pointer network for product title summarization. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18, page 7–16, New York, NY, USA. Association for Computing Machinery.

Team, Qwen. 2024. Qwen2.5: A party of foundation models, September.

Wang, Manyi, Tao Zhang, Qijin Chen, and Chengfu Huo. 2020. Selling products by machine: a user-sensitive adversarial training method for short title generation in mobile e-commerce.

Wang, Haifeng, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2021. Progress in machine translation. Engineering.

Way, Andy. 2013. Traditional and emerging use-cases for machine translation. Proceedings of Translating and the Computer, 35:12.

Xiao, Joan and Robert Munro. 2019. Text summarization of product titles. In eCOM@SIGIR.

Zhang, Xueying, Yunjiang Jiang, Yue Shang, Zhaomeng Cheng, Chi Zhang, Xiaochuan Fan, Yun Xiao, and Bo Long. 2021. Dsgpt: Domain-specific generative pre-training of transformers for text generation in e-commerce title and review summarization. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, page 2146–2150, New York, NY, USA. Association for Computing Machinery.

Zhang, Bryan, Taichi Nakatani, Daniel Vidal Hussey, Stephan Walter, and Liling Tan. 2024. Don't just translate, summarize too: Cross-lingual product title generation in E-commerce. In Malmasi, Shervin, Besnik Fetahu, Nicola Ueffing, Oleg Rokhlenko, Eugene Agichtein, and Ido Guy, editors, Proceedings of the Seventh Workshop on e-Commerce and NLP @ LREC-COLING 2024, pages 58–64, Torino, Italia, May. ELRA and ICCL.

Zhou, Mingyang, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. CoRR, abs/1808.08266.