Break-Ideate-Generate (BrIdGe): Moving beyond Translations for Localization using LLMs

Swapnil Gupta* International Machine Learning Amazon swapgupt@amazon.com

Lucas Pereira Carlini* Latam Machine Learning Amazon lcarlini@amazon.com

Prateek Sircar International Machine Learning Amazon sircarp@amazon.com

Deepak Gupta International Machine Learning Amazon dgupt@amazon.com

Abstract

Language localization is the adaptation of written content to different linguistic and cultural contexts. Ability to localize written content is crucial for global businesses to provide consistent and reliable customer experience across diverse markets. Traditional methods have approached localization as an application of machine translation (MT), but localization requires more than linguistic conversion – content needs to align with the target audience's cultural norms, linguistic nuances, and technical requirements. This difference is prominent for long-form text, where multiple facts are present in a creative choice of language. We propose a novel prompt approach for Large Languages Models (LLMs), called **Br**eak-**Id**eate-**Ge**nerate (BrIdGe), for language localization. BrIdGe 'breaks' the source content into granular facts, 'ideates' an action plan for content creation in the target language by organizing the granular facts, and finally executes the plan to 'generate' localized content. This approach emulates the cognitive processes humans employ in writing that begin with identifying important points, followed by brainstorming on how to structure and organize the output. We evaluated the BrIdGe methodology from multiple perspectives, including impact of BrIdGe prompt on different LLMs and performance comparisons with traditional MT models and direct translation through LLMs on public benchmark and proprietary e-commerce datasets. Through human and LLM-based automated evaluations across content in multiple languages, we demonstrate effectiveness of BrIdGe in generating fluent localized content while preserving factual consistency between source and target languages.

1 Introduction

With the globalization of businesses and the need to cater to diverse audiences worldwide, content localization has become crucial (Okonkwo et al., 2023). Localization adapts content originally designed for a source region to meet the cultural, linguistic, and technical requirements of different target regions (Paton, 2024). For businesses with diverse customer bases, effective localization is paramount to create accessible experiences for customers, regardless of their location, language, or cultural background. Specifically, for written content, localization goes beyond translation, as the latter only focuses on linguistic conversion keeping same structure and stylistic expressions from source to target language (Sorrentino, 2023). Whereas content localization allows modification in content structure, idiomatic expressions, and information organization to ensure native-like fluency while preserving factual alignment. For instance, the English idiom "boat neck dresses can be dressed up or down easily" imply that the dress can be used for both formal and casual occasions. However, machine translation (MT) tools like AWS Translate[1] and Google Translate[2], translate this idiom to Portuguese as "Este vestido pode ser facilmente vestido para cima ou para baixo" which is an incorrect literal translation meaning boat neck dresses can be worn on top as well as on bottom. Figure 1 shows nuances of localization which are missed by translation.

Large Language Models (LLMs) pre-trained on large text corpus (Anthropic, 2024; Touvron et al., 2023; Rastogi, 2024) have demonstrated exceptional abilities to abstract the factual knowledge in their weights (Petroni et al., 2019), follow instructions and perform Chain-of-Thought (CoT) reasoning (Wei et al., 2023). This has enabled them to break down complex problems into smaller, more manageable steps, mirroring human cognitive processes. LLMs have also showed impressive multilingual capabilities with promising results on

*These authors contributed equally to this work

[1]https://aws.amazon.com/translate/

[2]https://translate.google.com.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 627–637

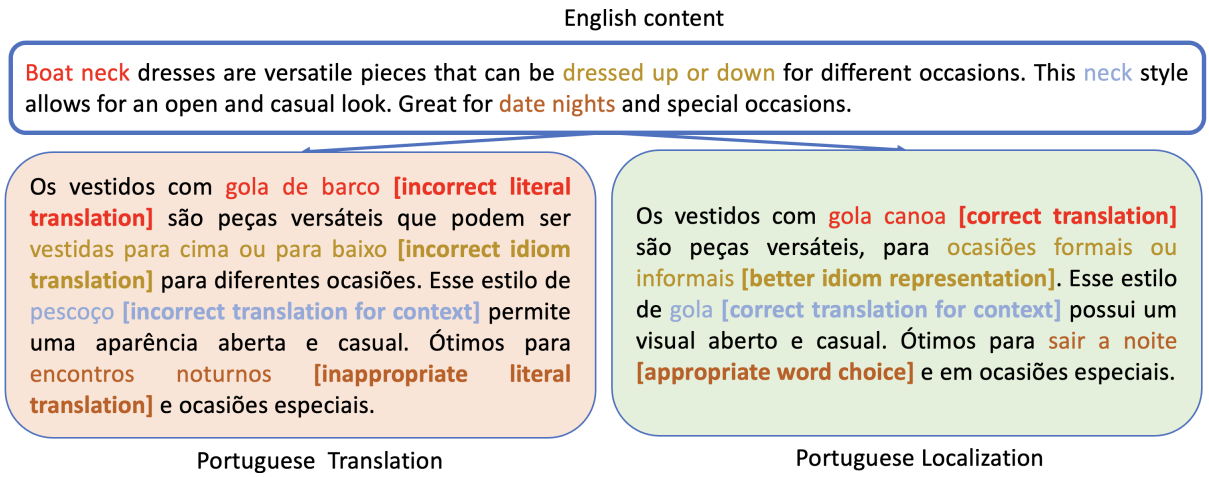

Figure 1: Comparison between Translation and Localization from English→Portuguese. Here, AWS Translate is used to get the Portuguese translation. Localization is a more holistic adaptation of content from source to target language. In the example, Localization makes multiple modifications in choice of words and phrases, which is missing in Translation.

numerous multi-lingual natural language processing (NLP) tasks (Zhu et al., 2024; Aggarwal et al., 2024; Ahuja et al., 2023). In this work, we leverage LLMs to emulate the human writing behavior (Hillocks, 1986; Du et al., 2022), where we first note down our initial and granular thoughts, followed by contextually structuring the information as per the requirements of final use-case. And we demonstrate its efficacy for the task of textual content localization from a source language to a target language. To achieve this, we propose a novel prompting approach called **Br**eak-**Id**eate-**Ge**nerate (**BrIdGe**) for LLMs. Given content in a source language, BrIdGe first 'breaks' it into granular facts, then 'ideates' an execution plan and finally 'generates' content in the target language. We perform extensive experiments on public benchmark datasets for multiple languages pairs and demonstrate superior performance of the BrIdGe prompt in comparison to standard translation prompts for multiple LLMs. We also show effectiveness of BrIdGe in a real-world e-commerce application of localizing educational content. In this application, we generate educational content about product attributes and benefits with the objective of aiding customers in taking informed shopping decisions. For example, given a chair with an attribute "finish type" as "lacquer", we generate content around properties and benefits of chairs with lacquer finish. Here the original content is generated in English language and the task is to localize it to languages of Non-English-speaking marketplaces. Manual audit by language and marketplace experts demonstrates that BrIdGe outperforms state-of-the-art translation strategies on fluency, while maintaining factual consistency between source and target languages.

The major contributions of this paper are:

(1) We identify an important and relatively under-explored problem - content localization. We propose BrIdGe - a novel LLM-based approach for content localization inspired by human writing.

(2) Via extensive experiments on public benchmark datasets comprising several language pairs, we show that BrIdGe outperforms translation-based prompting strategies across LLMs.

(3) We study effectiveness of BrIdGe on a real-world e-commerce application of localizing educational content originally generated in English to Non-English-speaking marketplaces. The study indicated superior performance of BrIdGe in comparison to state-of-the-art baselines.

2 Related Works

With the rise of internet and social media, the need for effective language localization has become increasingly important. Traditionally, human translation was the primary approach for localization, with professional translators adapting content to suit different linguistic and cultural contexts. However, human translation is time-consuming and expensive. With machine learning, statistical (Koehn, 2009) and neural (Koehn, 2020) MTs became dominant approaches. While MT has shown significant improvements in recent years, it still faces challenges in terms of accuracy and fluency (Koehn and Knowles, 2017). Also, its performance in trans-

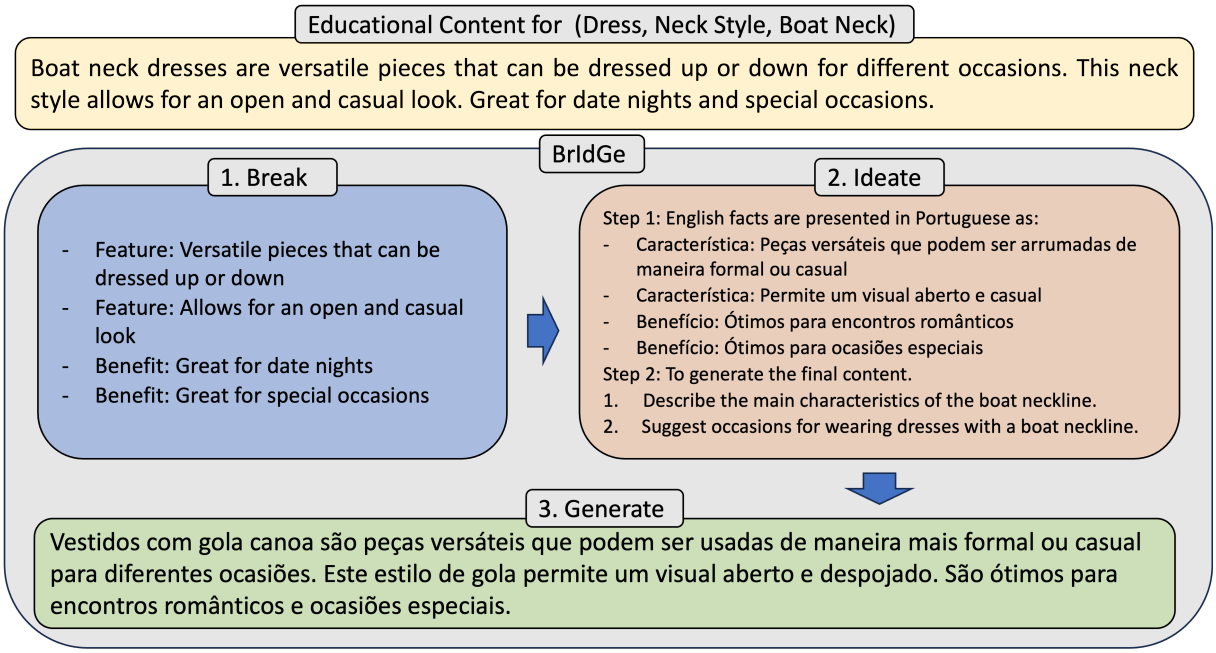

Figure 2: **BrIdGe Workflow:** The figure demonstrates how BrIdGe localize an LLM generated educational content for a quadruple *(product type, attribute name, attribute value)* from English→Portuguese.

lating cultural-specific items remains poor due to the gap between the cultural differences associated with languages (Akinade et al., 2023).

In this paper, we combine various lines of research on multi-lingual LLMs and its reasoning capabilities to localize content. Our approach primarily comprises of the following steps. The first step is named **Break**, which segments the original content into granular facts. This strategy is being widely adopted for hallucination detection and correction (Dhuliawala et al., 2023; Min et al., 2023; Zhao et al., 2023). To the best of our knowledge, this is the first work which adopts the strategy of breaking content into granular facts for Localization. LLMs have demonstrated improved performance in analytical tasks achieved by encouraging them to generate internal thoughts or logical chains before responding (Wei et al., 2023; Wang et al., 2022), and allowing them to update their initial responses through self-critique (Madaan et al., 2023). This strategy is named Chain-of-Thought (CoT). We leverage these techniques in the BrIdGe prompt, to execute all the instructions in the prompt step-by-step generating outputs at intermediate step conditioning the final generated localized content to effectively adhere all the steps.

3 BrIdGe: Break-Ideate-Generate

In this section, we describe our approach to localize content, which assumes access to an LLM that can be prompted, and content generation in both source and target language. Another key assumption of our method is that this language model, when suitably prompted, can both create and execute a plan to generate responses adhering to specific criterion and instructions.

We introduce BrIdGe, a novel method for content localization inspired by human writing behavior. Our approach is illustrated in Figure 2. BrIdGe prompt first instructs LLM to break input content into granular facts (section 3.1), next to ideate content generation plan appropriate for the specified marketplace and use-case (section 3.2), and finally execute the plan to generate the target language content by organizing the granular facts (section 3.2). While there are multiple steps in our workflow, we created a unified prompt, which can perform these steps and generating the localized content in a single LLM call.

3.1 Break

Recent works have noted that textual content, especially long-form, is a combination of several pieces of factual information (Dhuliawala et al., 2023; Min et al., 2023; Zhao et al., 2023). While processing any textual content, humans also inherently recognize all the facts as the first step. This allows humans to develop a comprehensive understanding of the content. To emulate this human behavior, the first instruction in the BrIdGe prompt is to break down the source content into granular facts. For

instance, given a statement “Lace dresses have a delicate and intricate fabric made from interwoven yarn or thread”, it can be separated into two granular facts: 1) "Lace dresses have a delicate and intricate fabric" and 2) "Lace dresses are made from interwoven yarn or thread". To deepen the content understanding, as the next instruction in the prompt is to categorize each fact in domain-specific categories. In the context of educational content of e-commerce product attributes, these categories are "Physical Features", "Benefits", and "Suitable use-cases". Applications where such categories are not pre-determined, LLM is instructed to infer them from the content itself.

3.2 Ideate

After identifying the list of granular facts in the source language, the next set of instructions in BrIdGe prompt are about setting up the additional context about the task and organizing the facts in a logical, coherent way suited to the target language as per the additional context. The LLM is instructed to deliberate over the segmented facts and task requirements before generating the final response. For educational content generation, these instructions include marketplace-related metadata if available like name of marketplace, measurement units, etc. and language-related requirements for the educational content task.

3.3 Generate

Finally, the BrIdGe prompt ends with CoT instructions (Wei et al., 2023) to go over the entire prompt step-by-step, generating in-between thoughts and outputs at each step before generating the final response. The prompt is also augmented with manually crafted in-context learning examples to guide the LLMs CoT reasoning.

4 Experiments

4.1 Datasets

Experiments used two datasets, described below:

1. FLORES-200: The FLORES-200 multilingual MT benchmark (NLLB Team, 2022; Goyal et al., 2021; Guzmán et al., 2019) consists of translations from English into 200 languages. The dataset contains 997 samples for each language, sampled from Wikinews[3], Wikijunior[4], and Wikivoyage[5]. We considered 4 language pairs, with English being the source language in all pairs, and Portuguese, Spanish, Czech and Hindi are the 4 target languages.

2. Educational Content: We considered a real-world e-commerce application of generating educational content for product attribute values. For example, in the product category *"Chair"* for the attribute *"finish type"*, a valid attribute value is *"Lacquer"*. To create this dataset, we selected a list of 10K triplets of the form *(product category, attribute, attribute value)* which spanned across 400 different product categories and finally selected a random sample of 500 triplets for experimentation. For each triplet, educational content containing information about features, benefits and common utility of the attribute value in the product category is generated by prompting Claude-3.5-sonnet (Anthropic, 2024). We present examples of generated educational content in Table 3 in appendix A. The task here is to localize the English language content to different non-english speaking marketplaces. For this work, we considered 4 marketplaces which (along with their primary language) which are Brazil (Portuguese), Mexico (Spanish), Germany (German) and India (Hindi).

4.2 Baselines

On FLORES-200 dataset, our primary objective is to demonstrate that our proposed BrIdGe prompting strategy is more effective for LLM-based content localization as compared to a standard translation prompt. Therefore, on FLORES-200 dataset, we compare BrIdGe with a standard translation prompt instructing the LLM to translate the English content to a target language. For a fair comparison with the BrIdGe prompt, we provided the same task-specific context as well as added standard CoT instructions ("think step-by-step") to the prompt. We call this prompt as **Translation-CoT**.

We compare the two prompting strategies with four instruction-tuned LLMs to ensure generalization of BrIdge: Claude-3.5-sonnet (Anthropic, 2024), Llama3.1-70B (Touvron et al., 2023), Command R+ (Rastogi, 2024), and Mixtral 8x7B (Jiang et al., 2024). We use greedy decoding during text generation for stable outputs.

For the educational content dataset, we take the best performing LLM in the Flores-200 experiments (Claude-3.5-Sonnet) and compare it against

[3]https://en.wikinews.org/wiki/Main_Page
[4]https://en.wikibooks.org/wiki/Wikijunior
[5]https://en.wikivoyage.org/wiki/Main_Page

3 different localization strategies: a) Translation-CoT b) AWS Translate (a powerful commercial translation system) and c) Direct Generation. In direct generation, we prompt the LLM to generate educational content directly in the target language independent of content in source language. We keep the exact prompt used for content generation in English with additional instructions to generate content in the target language, and we also added "in-context learning" examples in target language with the help of human expert. This strategy enables a better comprehension of the model's latent information regarding the task domain in a language.

4.3 Evaluation Metrics

Several works have demonstrated that standard translation metrics like BLEU (Papineni et al., 2002), BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020) correlate poorly with human judgement and has pivoted to LLM-based translation quality metrics (Kocmi and Federmann, 2023; Chen et al., 2024). Here, we adopt an LLM-based evaluation method to assess two aspects: a) fluency, and b) adequacy (factual consistency). For computing LLM-based fluency metric, we follow the direct assessment prompting strategy as proposed in (Kocmi and Federmann, 2023) having the best correlation with human annotations.

For LLM-based adequacy computation, we design a two-step process. First, we extract all the facts in source and target language content, followed by identifying the matching facts in the two lists. Let S be the set of facts in the source content and U be the set of facts identified in the localized target language content and $I = S \cap U$, a set of facts present in both the contents. For each sample, we define precision (P) as $|I|/|U|$ and recall (R) as $|I|/|S|$ and hence F1 score as $2 * P * R/(P + R)$. We define "adequacy score" as the mean F1 score across all the samples in the dataset.

For the educational content dataset, we conducted a thorough assessment by conducting manual audits by language experts from the respective marketplaces. For fluency assessment, we defined four grades A-D, where A is the best and D is worst. We provide a description of the 4 Grades in appendix B. Language expert annotators were asked to provide a fluency grade basis their judgement for each of the generated content. Based on these grades, we define two metrics for fluency comparison: **a) High Quality Fluency**: Defined as the percentage of generated content graded as A or B. **b) Risky Generation**: Defined as percentage of generated content belonging to Grade D. A good content is expected to have high "High Quality Fluency" metric and low "Risky Generation" metric.

Note, for easier comparison and to maintain confidentiality as mandated by company policy, we present results as relative lifts compared to the worst performing baseline as 1.00x.

5 Results

5.1 Quantitative Results

Tables 1 and 2 show our quantitative results.

1. Flores-200 All the four LLMs (Claude 3.5 Sonnet, Llama 3.1-70B, Command R+ and Mixtral 8x7B) showed improvement in fluency when prompted through the proposed BrIdGe prompt as compared to *Translation-CoT* across all languages (Portuguese, Spanish, Czech and Hindi). Specifically, Claude 3.5 Sonnet showed consistent and significant improvements across all languages, ranging from 1.27x (Czech) to 1.68x (Portuguese). Whereas, Mixtral 8x7B showed maximum fluency improvements of 2.06x in Portuguese and 1.90x in Czech. This primarily highlights the importance of "break" step in BrIdGe which allows LLMs structural flexibility in framing target language content.

2. Educational Content In the Flores-200 experiment, we observed that Claude 3.5 Sonnet had the best absolute metrics in terms of adequacy and fluency. Therefore, we leverage Claude 3.5 Sonnet for localization of educational content. Here, we observe that, BrIdGe has better holistic performance compared to AWS Translate / Direct Content Generation / Translation-CoT. Approaches involving direct translation struggle with *high quality fluency* and tend to generate risky outputs more frequently. Whereas, Direct Content Generation suffers from low adequacy but has higher fluency. Meanwhile, BrIdGe achieved balanced values across all metrics. It demonstrated *high quality fluency* increment ranging from 1.29x in Spanish to 2.18x in Hindi when comparable to AWS Translate. Compared to Direct Content Generation, adequacy of BrIdGe is significantly higher for all languages.

Observations on Adequacy Scores In both the dataset, for some languages, we observe a slight decrease in adequacy scores. In Flores-200, BrIdGe adequacy was 0.99x across languages as compared to Translation-CoT and in educational content,

LLM	Prompting Method	Portuguese		Spanish		Czech		Hindi	
		Adequacy	Fluency	Adequacy	Fluency	Adequacy	Fluency	Adequacy	Fluency
Claude 3.5 Sonnet	Translation-CoT	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x
	BrIdGe	0.99x	1.68x	1.00x	1.40x	0.99x	1.27x	1.00x	1.48x
Llama 3.1-70B	Translation-CoT	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x
	BrIdGe	1.00x	1.31x	0.99x	1.11x	0.99x	1.08x	0.99x	1.27x
Command R+	Translation-CoT	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x
	BrIdGe	0.99x	1.58x	0.98x	1.37x	0.99x	1.34x	0.99x	1.36x
Mixtral 8x7B	Translation-CoT	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x
	BrIdGe	1.00x	2.06x	0.99x	1.38x	0.99x	1.90x	0.98x	1.00x

Table 1: Adequacy and Fluency results on Portuguese, Spanish, Czech and Hindi languages on the Flores-200 dataset. In all the cases English is the source language.

Method	Portuguese			Spanish		
	Adequacy	High Quali. Fluency	Risky Gen.	Adequacy	High Quali. Fluency	Risky Gen.
AWS Translate	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x
Direct Content Generation	0.53x	2.18x	0.49x	0.65x	1.10x	0.45x
Translation-CoT	1.00x	1.90x	0.27x	0.94x	1.23x	0.20x
BrIdGe	0.97x	2.12x	0.15x	1.00x	1.29x	0.11x

Method	German			Hindi		
	Adequacy	High Quali. Fluency	Risky Gen.	Adequacy	High Quali. Fluency	Risky Gen.
AWS Translate	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x
Direct Content Generation	-	-	-	0.74x	2.18x	0.12x
Translation-CoT	0.95x	1.23x	0.39x	1.01x	2.09x	0.40x
BrIdGe	1.00x	1.38x	0.18x	0.99x	2.18x	0.20x

Table 2: Adequacy and Fluency results on Portuguese, Spanish, German and Hindi language localization of educational content with English as the source language. Note that for "risky gen.", lower the metric, better it is for content generation.

BrIdGe scores compared to AWS Translate were 0.97x for Portuguese, and 0.99x for Hindi. This can be attributed to the fact that instead of just linguistic conversion, BrIdGe modifies content such that expressions from source language which are not suitable for target language are either replaced with more suitable phrases or removed.

5.2 Qualitative Results

We present a qualitative comparison of localization between AWS translate and BrIdGe approach in Figure 3 in the Appendix. We observed structural nuances of localization that BrIdGe adheres to, which translation itself, by definition, may not necessarily follow. For example, in the localization example for Hindi, the first two sentences were merged to create a more fluent output. Furthermore, the framework has carefully chosen to transliterate words like "support" and"outdoor" instead of translating them, catering to the cultural nuance of code-mixing prevalent in the Indian market. Similarly, in the German example, the final two sentences on "versatile dressing" have been merged, and the idiomatic phrase "dress up and down" has been completely omitted, as it was literally transferred in the German translation. For the Spanish example, the first two sentences have been merged to enhance fluency. Additionally, the subject "solid back" has been replaced with the pronoun "they" in the second sentence to avoid redundancy. The idiomatic expression "fashion statement" has been expressed more appropriately compared to the translation.

6 Conclusions & Future Work

This paper introduced BrIdGe, a novel prompt strategy for performing comprehensive content localization beyond linguistic translation. By emulating human writing through iterative steps of breaking down input, ideating a localization plan, and generating the final output, BrIdGe demonstrates promising localization of content by preserving meaning and achieving fluency. Experiments in four languages for educational content showed the strengths of BrIdGe. It achieved comparative adequacy scores to baselines while outperforming them with fluency. Qualitatively, BrIdGe preserved meaning across long and complex sentences, appropriately handling domain-specific context. Most importantly, we observed the impact of the Break step, going beyond the standard Chain-of-Thought strategy, by segmenting input facts, which allows flexibility to the LLM to organize and reconstruct

the final output generating fluent content.Going forward, we plan to experiment this framework with moderate to small sized LLMs to optimize the cost and latency constraints that come with large LLMs like Claude. We also plan to experiment our approach to more indigenous languages and using low resources languages as the source one.

Limitations

In this section, we enumerate a few limitations of this approach. While the BrIdGe prompting strategy has shown promising results in content localization, the experiments are done on only 6 language pairs, where, except for Hindi, every language follows Roman script. With BrIdGe prompts having almost 4x input tokens and 2x output tokens than Translation-CoT, the user has to trade-off between the cost and latency of such generation and the required localization capabilities. Additionally, given that Localization/Translation is a content generation task, we need to properly assess the method stability by prompting several times with varying hyperparameters, however, such experiment would lead to manual annotation cost increase. Finally, we need an access to large powerful LLMs which can run the whole BrIdGe based localization in one prompt.

References

Divyanshu Aggarwal, Ashutosh Sathe, Ishaan Watts, and Sunayana Sitaram. 2024. Maple: Multilingual evaluation of parameter efficient finetuning of large language models. *Preprint*, arXiv:2401.07598.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. Mega: Multilingual evaluation of generative ai. *Preprint*, arXiv:2303.12528.

Idris Akinade, Jesujoba Alabi, David Adelani, Clement Odoje, and Dietrich Klakow. 2023. Varepsilon kú mask: Integrating Yorùbá cultural greetings into machine translation. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 1–7, Dubrovnik, Croatia. Association for Computational Linguistics.

Anthropic. 2024. Claude 3.5 sonnet.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024. Benchmarking llms for translating classical chinese poetry:evaluating adequacy, fluency, and elegance. *Preprint*, arXiv:2408.09945.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *Preprint*, arXiv:2309.11495.

Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. 2022. Understanding iterative revision from human-written text. *arXiv preprint arXiv:2203.03802*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english.

George Hillocks. 1986. Research on written composition. *Urbana, IL: National Council of Teachers of English*.

Albert Q. Jiang, 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. *Preprint*, arXiv:2401.04088.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Philipp Koehn. 2009. *Statistical Machine Translation*. Cambridge University Press.

Philipp Koehn. 2020. *Neural Machine Translation*. Cambridge University Press.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *Preprint*, arXiv:2303.17651.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

James Cross et al. NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Ifeanyi Okonkwo, John Mujinga, Emmanuel Namkoisse, and Adrien Francisco. 2023. Localization and global marketing: Adapting digital strategies for diverse audiences. *Journal of Digital Marketing and Communication*, 3:66–80.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Brandon Paton. 2024. Content localization: The fundamentals, benefits, significance.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *Preprint*, arXiv:1909.01066.

Ritvik Rastogi. 2024. https://ritvik19.medium.com/papers-explained-166-command-r-models-94ba068ebd2b.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Francesa Sorrentino. 2023. Localization vs translation: The difference explained.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2714–2730, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. *Preprint*, arXiv:2305.03268.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Appendix

A Examples of Educational content

In Table 3, we present some examples for the LLM generated educational content as described in Section 4.1.

B Fluency Grades

For the manual audit of Educational Content data for localization task as described in Section 4.3 we provided the following grade definitions to our auditors.

1. Grade A: The content is aligned with cultural and grammar nuances from target language, all sentences are easy to understand;

2. Grade B: The content may present some minor fluency and writing errors in small parts of the text, like word repetitiveness, or sub-optimal choice of specific words for the context of the PT-attribute-detail;

3. Grade C: The content may present major fluency and writing errors in a larger portion of the text, like complete sentences or multiple distinct phrases. Also, it may present meaningless expressions and attribute details;

4. Grade D: The content presents false, incorrect, offensive, inappropriate, or irrelevant information that can potentially expose Amazon to risks.

C Qualitative Comparison

In Figure 3, we present the qualitative details comparing AWS Translate and BrIdGe.

(PT/AN/AV)	LLM Generated Educational Content
(Paddleboard, Material, PVC)	PVC paddleboards are lightweight yet rigid, making them easy to carry and provide good stability on water. PVC boards are affordable options and are appropriate for beginners and casual paddlers looking for an entry-level board for lakes and calm waters.
(Electric water boiler, Material, Ceramic)	Ceramic electric water boilers have an inner tank made of ceramic material. Ceramic is an insulator that allows water to heat up quickly while retaining heat efficiently. Ceramic boilers are durable, corrosion-resistant and easy to clean. They are widely used for boiling water for tea/coffee and are appropriate for homes and small offices due to fast heating and energy efficiency.

Table 3: Examples of LLM generated educational content. Product Type (PT), Attribute Name (AN) and Attribute value (AV) are given as input to the LLM and is instructed to generate features, benefits and suitable utility for the attribute value.

English Content	Translation (AWS Translate)	Localization (BrIdGe)	Target Lang.
Chairs with a solid back provide good support for your spine and posture. The solid surface evenly distributes weight and pressure. Solid back chairs promote proper alignment of your back, preventing pain and fatigue from sitting for extended periods.	Cadeiras com encosto sólido fornecem um bom suporte para a coluna e a postura. A superfície sólida [word repetitiveness] distribui uniformemente o peso e a pressão. Cadeiras com encosto sólido [word repetitiveness] promovem o alinhamento adequado das costas, evitando dores e fadiga ao ficar sentado por longos períodos.	Cadeiras com encosto sólido proporcionam bom suporte para a coluna e postura, distribuindo o peso e pressão uniformemente [better readability by merging sentences]. [subject hidden to avoid repetitiveness] Promovem o alinhamento adequado das costas, evitando dor e fadiga ao sentar por longos períodos.	Portuguese
Geometric patterned socks have shapes like stripes, dots, zigzags or abstract prints. They add visual interest and personality to an outfit. Geometric socks are recommended as fun statement pieces to liven up a casual or formal look. They can be paired with suits, jeans or shorts. The geometric patterns make them versatile to match with different tops and shoes.	Los calcetines estampados geométricos tienen formas como rayas, puntos, zigzags o estampados abstractos. Añaden interés visual y personalidad a un atuendo. Los calcetines geométricos son recomendables [suboptimal word choice] como piezas divertidas de declaración [unnatural phrasing] para amenizar [suboptimal word choice] un look casual o formal. Se pueden emparejar con trajes, jeans o pantalones cortos. Los patrones geométricos los hacen versátiles para combinar con diferentes tops y zapatos.	Los calcetines con patrón geométrico tienen formas como rayas, puntos, zigzags o estampados abstractos, lo que añade interés visual y personalidad al atuendo. Se recomiendan como piezas llamativas y divertidas [natural phrasing] para animar [appropriate word choice] un look casual o formal, pudiéndose combinar con trajes, jeans o shorts. Los patrones geométricos los hacen versátiles para hacer match [idiomatic expression] con diferentes tops y zapatos.	Spanish
Polka dot dresses have a playful pattern of dots in different sizes scattered all over the fabric. They create a fun, retro, and feminine look. Polka dots are versatile and can be dressed up or down easily. They are ideal for casual daytime wear and recommended for women looking to add a whimsical touch to their wardrobe.	Gepunktete Kleider haben ein verspieltes Muster aus Punkten in verschiedenen Größen, die über den ganzen Stoff verteilt sind. Sie kreieren einen lustigen, retro und femininen Look.[literal translation of idiomatic expression] Polka Dots sind vielseitig und können leicht schick oder lässig angezogen werden. Sie sind ideal für lässige Tageskleidung und werden für Frauen empfohlen, die ihrer Garderobe eine skurrile Note verleihen möchten.	Gepunktete Kleider haben ein verspieltes Punktemuster in verschiedenen Größen, die über den ganzen Stoff verstreut sind. Sie erzeugen damit einen fröhlichen, retroartigen und weiblichen Look. Gepunkte sind vielseitig und ideal für lässige Alltagskleidung. Gepunktete Kleider werden für Frauen empfohlen, die ihrem Kleiderschrank einen launischen Touch hinzufügen möchten. [merging sentence for better readability, managing idiomatic expression better]	German
Metal chairs have a frame made of metal material which is strong, sturdy and durable. They provide good support and stability. Metal chairs are recommended for both indoor and outdoor use as they can withstand weather elements and do not rust or corrode easily.	धातु की कुर्सियों में धातु की सामग्री से बना एक फ्रेम होता है जो मज़बूत, मज़बूत [Word Repetition due to literal translation of strong and sturdy] और टिकाऊ होता है। वे अच्छा समर्थन [wrong choice of word] और स्थिरता प्रदान करते हैं। घर के अंदर और बाहर दोनों जगह उपयोग के लिए धातु की कुर्सियों की सिफारिश की जाती है क्योंकि वे मौसम के तत्वों का सामना कर सकती हैं और आसानी से जंग नहीं लगाती या खराब नहीं होती हैं।	धातु से बनी कुर्सियों का फ्रेम मजबूत, स्थिर और टिकाऊ होता है जो अच्छा सपोर्ट [transliteration instead of translation to suit cultural preferences] देता है। [better readability by merging sentences]. इन्हें इनडोर और आउटडोर दोनों जगहों पर उपयोग के लिए रिकमेंड किया जाता है क्योंकि ये मौसम का असर सहन कर सकती हैं और आसानी से जंग नहीं लगती।	Hindi

Figure 3: Qualitative analysis: Above examples demonstrate that BrIdGe is effective at identifying suitable modifications to the source content both in content structure as well as choosing alternate phrasing based on target language nuances.