Internationalization in Large Language Models

Anonymous ACL submission

Abstract

In this paper, we introduce the concept of internationalization for large language models and its importance for improving global usability. We define internationalization and localization in the context of LLMs, where internationalization refers to the model's ability to adapt to different languages, cultural norms, and regional conventions without requiring changes to its architecture or weights. We present the first dataset specifically designed to evaluate internationalization in LLMs and evaluate GPT-4o, GPT-fo-mini, and LLaMA-3.1-8B-instruct across four locales: Canadian English (en_CA), Iranian Persian (fa_IR), Hijazi Arabic (ar_SU), and Vietnamese (vi_VN). Our evaluation reveals that current LLMs respond incorrectly more than 50% of the time, even when the locale is explicitly specified, demonstrating that multilingual LLMs are not inherently internationalized. This work highlights the gap between multilingualism and true internationalization, offering a new benchmark for advancing the global adaptability of LLMs. Finally, we propose an application that automatically detects user's locale through system settings, web browser APIs, user profiles, or by analysing previous user interactions and passes this information to internationalized LLMs to localize LLMs responses according to user's needs.

1 Introduction

In today's digital age, social media and online platforms can rapidly attract users from across the globe, especially when not constrained by logistical concerns. This phenomenon is even more pronounced for practical and versatile products like large language models (LLMs). For instance, in 2022, ChatGPT reached one million users in just five days (Jarinaa et al., 2023). By comparison, Instagram took 2.5 months to achieve the same milestone after its launch in 2010 (Dhaker and Shekhawat, 2022), and X (formerly, Twitter) took two years following its introduction in 2006.

A significant portion of these users come from linguistic and cultural backgrounds different from those of the developers. As of September 2024, only about 18% of ChatGPT users were from the United States (similarweb). This highlights the importance of adapting digital products to meet the diverse needs of a global audience. Given that developing LLMs is highly expensive, companies must address cultural and linguistic diversity to remain competitive and justify their investment (Keniston, 1997), which is currently a challenge (Maréchal, 2024). One of the most crucial ways to achieve this is through localization.

Localization (l10n) refers to the process of adopting a product for a specific region, language, or culture (Ishida and Miller). Number formatting is a basic example of localization; for instance, the number π is written as 3.14 in the `en_US` locale, $3,14$ in the `de_DE` locale, and ۳.۱۴ in `fa_IR` locale.

Given the vast number of locales in the world, it is not straightforward to localize a product—or, in our case, an LLM—for each locale without a standardized approach. This is where internationalization comes in. *Internationalization* (i18n) refers to the process of designing a product so that it can be easily localized for different locales without requiring significant changes to the product itself (Kersten et al., 2002).

Localization goes beyond simple translation (Keniston, 1997). It requires careful consideration of cultural elements. Aspects such as social norms and idiomatic expressions play a crucial role in effective localization (Pyae, 2018). Other elements, such as calendars, number formatting, and measurement units, might initially seem unrelated to culture, but they vary across cultures and must also be localized. Upon closer examination, these differences are often deeply rooted in the historical

development of each culture.

The main goal of this paper is to argue that multilingual capability alone is not sufficient for LLMs—cultural aspects must also be considered. While previous works have highlighted the importance of addressing cultural norms and mitigating biases (Li et al., 2024a, Kharchenko et al., 2024, Li et al., 2024b), we aim to go a step further by emphasizing the significance of other localization factors, such as calendars, number formatting, and units of measurement. In addition, we introduce the concept of internationalization specifically for LLMs and propose that LLMs, among other aspects, should also be evaluated from this perspective.

We hope this work enhances the cultural diversity of LLMs, ensuring they better reflect the unique characteristics of diverse communities and provide a more inclusive user experience. By addressing the needs of underrepresented and marginalized cultures, we aim to prevent them from being overlooked in global AI systems. This paper marks the beginning of simplifying internationalization in LLMs, ultimately encouraging the preservation and revitalization of endangered languages. In doing so, it promotes a more inclusive and culturally diverse digital environment, where linguistic and cultural diversity is actively supported.

Our contributions are summarized below:

- Introduced the concept of internationalization and localization for LLMs, which go beyond multilinguality to include norms, idiomatic expressions, and regional formats.
- Released the first evaluation dataset focused on LLM internationalization.
- Evaluated the internationalization ability of three LLMs and showed that they are more than half the time unable to produce desired results.
- Proposed the first architecture for an internationalized LLM application LLM.

The rest of this paper is organized as follows. Section 2 we review previous work related to internationalization in large language models. Section 3 introduces the concept of internationalization in LLMs, explaining why multilingual models are not inherently internationalized. In Section 4 , we detail the creation of our i18n dataset and present the results of evaluating several LLMs across different locales. Finally, Section 5 discusses how to develop an internationalized application, including automatic locale detection through system settings, web browser APIs, and user profiles, to ensure that LLMs localize responses according to user-specific needs.

2 Related works

Multilingual large language models have garnered increasing attention, with substantial research efforts aimed at expanding their capabilities beyond English. A foundational work in this area is mT5 (Xue, 2020), which was trained on a dataset spanning 101 languages, setting a precedent for broad language coverage. BLOOM (Le Scao et al., 2023) followed with support for 46 languages, while Aya 23 (Aryabumi et al., 2024) extends multilingual capabilities across 23 languages, building upon its predecessor, Aya (Üstün et al., 2024), which covers 101 languages. The Aya dataset (Singh et al., 2024), with support for 114 languages, further enriches multilingual research resources. Command-R (AI), another notable multilingual LLM, is optimized for retrieval-augmented generation (RAG) (Lewis et al., 2020), primarily supporting 10 languages while extending to 13 more. PersianMind (Rostami et al., 2024) is a LLM based on LLaMA 2 that add supports of Persian language to a foundation model. There are also some works focused on adding support of specific language to existing LLMs.

In addition to the multilingual datasets mentioned earlier, there are several others that contribute significantly to the development of multilingual LLMs. The CulturaX dataset (Nguyen et al., 2023), a large web-crawled corpus support 167 languages, has been widely used in training multilingual models. FLORES (Costa-jussà et al., 2022) dataset is specifically designed to support translation task. BELEBELE (Bandarkar et al., 2023) is designed to evaluate the reading comprehension of language models in 122 languages. Moreover, the World Value Survey (WVS) offers a unique dataset focused on cultural norms, providing insights that could help adapt LLMs to specific cultural contexts, an area that remains under explored in current multilingual LLM research.

3 Internationalization for LLMs

There is currently no widely accepted definition of internationalization for LLMs. Inspired by the concept of internationalization in software, we define

	Calendar	Number Format	Units	Currency
en_US	Gregorian (e.g., 10/25/2024)	Latin (e.g., 123,456.789)	Imperial (e.g., 10 miles)	USD (e.g., $3.75)
en_CA	Gregorian (e.g., 2024-10-25)	Latin (e.g., 123,456.789)	Metric (e.g., 16.09 kilometers)	CAD (e.g., $5.17)
fa_IR	Solar Hijri (e.g.,۱۴۰۳/۸/۴)	Persian (e.g.,۱۲۳،۴۵۶.۷۸۹)	Metric (e.g., ۱۶.۰۹کیلومتر)	IRR (e.g., ۱۵۷،۸۳۷ریال)
ar_SU	Lunar Hijri (e.g.,۱۴۴۶/۴/۲۱)	Arabic (e.g., ١٢٣،٤٥٦.٧٨٩)	Metric (e.g., ١٦.٠٩ كيلومتر)	SAR (e.g., ۱۵۷،۸۳۷ریال)
vi_VN	Gregorian (e.g., 25/10/2024)	Latin (e.g., 123.456,789)	Metric (e.g., 16,09 km)	VND (e.g., 93.618 ₫)

Table 1: Differences among various locales for different calendar systems, number formats, measurement units and currency.

it as the capability of a language model to generate responses suitable for various languages, cultures, and regions without requiring changes to its architecture or additional training. This approach ensures that the model is ready to be localized when necessary while satisfying its functional requirements.

The first step for achieving an internationalized LLM is to provide multilingual support. Fortunately, many earlier work (among academia and industry) have already devoted significant efforts towards this goal. In this paper, we show that multilingual LLMs are not necessarily fully internationalized. An issue stems from LLMs' training data. Many multilingual datasets, particularly those for low-resource languages, are simply translations of English datasets (Chen et al., 2024). Translation models do not necessarily localize the source input into target when translating, which results in a lack of cultural or regional adaptation.

Additionally, past limitations in software i18n, such as in operating systems and databases, have created challenges with older datasets. For instance, it was once difficult or nearly impossible to input text in certain languages in PCs. Unicode Consortium was created and tasked to remove language barriers and support as many of them as possible, including those that were previously extinct.

Beyond data issues, during training, models lack awareness of specific locales associated with the data. While language identification is a relatively easy task, distinguishing locales is far more challenging. Differences between en_US and en_GB locales are shown in Table 1.

Category	Number of Questions
Calendar	50
Length	20
Volume	20
Weight	20
Temperature	20
Currency	50
Total	180

Table 2: Overview of the base dataset

Furthermore, majority of LLM development happen in western countries. Lack of diversity among developers usually leads to neglecting challenges faced by people from other cultures and their specific needs. These reasons highlight why multilingual LLMs are not truly internationalized.

Table 1 compares calendar systems, number formatting, units, and currency across different locales. However, localization extends beyond these aspects. It is equally important to consider norms, biases, and idiomatic expressions when localizing. As there is already extensive research on these elements, our focus will remain on other aspects.

4 Evaluation of i18n in Current LLMs

To the best of our knowledge, no i18n dataset for LLMs currently exists. Hence, we developed one using GPT-4o. The dataset creation process began by generating 180 questions and answers in English to serve as the baseline. These questions covered a variety of topics to ensure diversity in both lan-

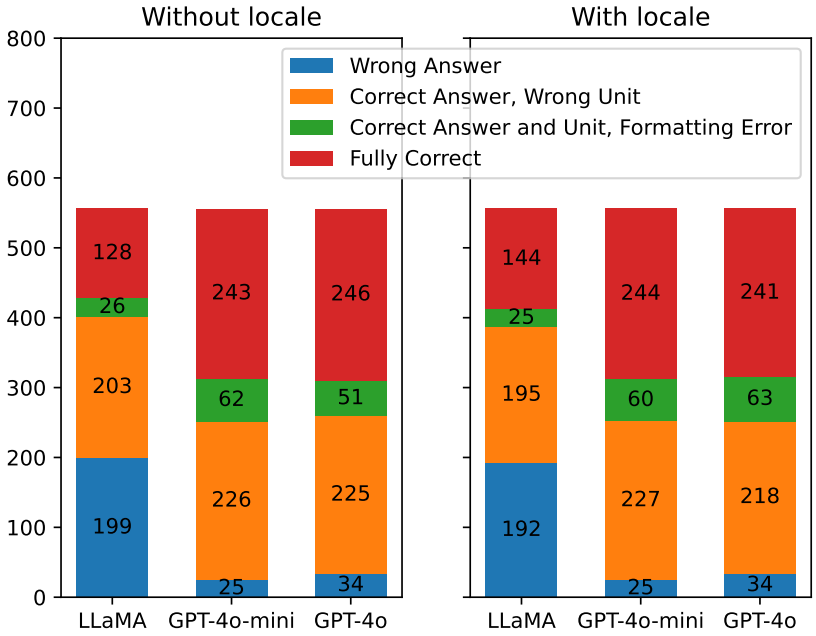

Figure 1: Evaluation of three LLMs for internationalization using our dataset over 5 locales. The left subfigure does not provide the locale explicitly, whereas the right one provides it through a system prompt.

guage and contextual nuances. The overview of base dataset is demonstrated in Table 2.

Next, English questions were translated into multiple target languages using Google Translate to reflect different linguistic and cultural contexts. Table 3 provides an overview of the various locales included in our dataset. The translation process aimed to maintain the integrity of the questions while adapting them to the grammar and syntax of each language. However, translation alone is not sufficient for internationalization, as it does not account for cultural and regional differences. To address this, answers were not just translated but also localized using i18n Python libraries. These libraries handled cultural adaptations, such as modifying dates, times, currencies, units of measurement, and other locale-specific elements to ensure that responses were appropriate for each locale.

The final step involved manually reviewing and validating localized responses to ensure they reflected the nuances of each language and culture while preserving the original meaning. As a result, we successfully created an i18n dataset which can be used to evaluate the internationalization capabilities of large language models. This dataset will be released to the public upon acceptance of the paper.

Language	Country	Locale
English	U.S.	`en_US`
English	Canada	`en_CA`
Vietnamese	Vietnam	`vi_VN`
Persian	Iran	`fa_IR`
Arabic	Saudi Arabia	`ar_SU`

Table 3: Overview of the locales included in our dataset

Several LLMs, including LLaMA-3.1-8B-Instruct, GPT-4o, and GPT-4o-mini, have been evaluated using the newly created i18n dataset. The models were tested in two different modes: one with the locale and another one without the locale. By comparing the models' performance in these two modes, we can assess both their ability to automatically detect the correct locale and their effectiveness in generating appropriately localized responses.

The models' responses, along with the correct answers generated using i18n libraries are manually evaluated. Evaluators were asked to mark responses generated by LLMs as a) fully correct, b) correct answer and unit with formatting error, c) correct answer but wrong unit, and d) wrong answer. For brevity, we summed the number of

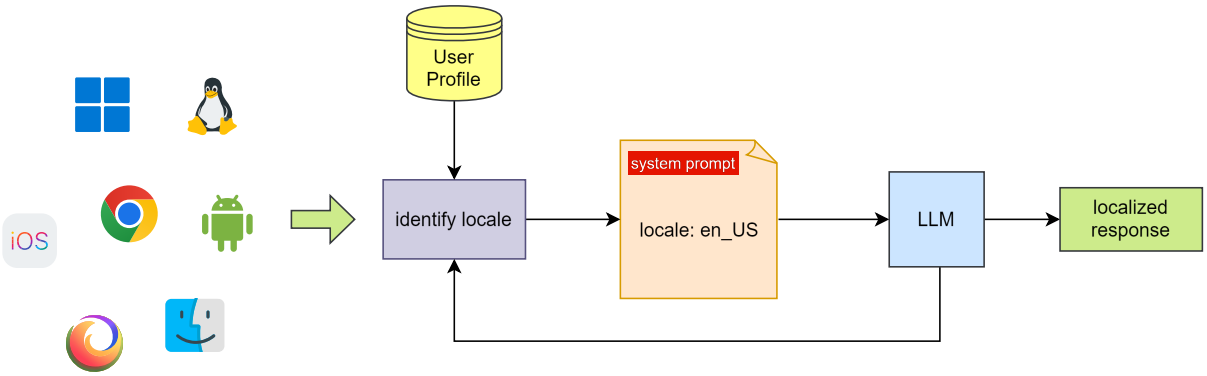

Figure 2: Proposed architecture for an internationalized LLM application. The locale is determined using operating system settings, web browser APIs, user profiles, or by analysing previous user interactions.

each evaluation criteria for each LLM and depicts results in Figure 1. The left figure shows the results where LLMs were not explicitly given the locale, where the right one shows the results where locale was included in LLMs' system prompt. As can be seen, the improvement when LLMs were prompted about the correct locale is minimal. This shows that they have minimal understanding of locales. Even though GPT-4o and GPT-4o-mini provide superior results compared to LLaMA, they still fail to provide expected results in over 50%.

5 Achieving i18n in LLMs

Based on our results, we propose the architecture showed in Figure 2 for building an internationalized application using LLMs. The system determines a user's locale from their operating system settings, web browser API, user profile, or by inferring it from prior user interactions. This information is passed to the LLM using a system prompt. This allows the model to be aware of the user's regional and cultural context, enabling it to generate responses which are both linguistically and culturally appropriate.

By automatically detecting the locale, the system removes the need for users to manually specify their preferences, allowing for seamless interaction across different languages and regions. The localized responses can include correctly formatted dates, times, currencies, and units, ensuring alignment with regional conventions. Furthermore, by embedding cultural and regional awareness into the system, it helps to avoid cultural misunderstandings or biases in the generated responses, enhancing both accuracy and user satisfaction.

6 Conclusion

In this paper, we introduced the concept of internationalization for large language models and emphasised its significance for enhancing global usability. We distinguished true i18n from mere multilingualism, where the former involves adapting to regional and cultural specifics without changing the model's architecture or weights.

We developed the first dataset specifically designed for evaluating i18n in LLMs and evaluated models such as GPT-4o, GPT-fo-mini, and LLaMA-3.1-8B-instruct across multiple locales, including en_CA, fa_IR, ar_SU, and vi_VN. Our results demonstrated that these models often respond incorrectly over 50% of the time, even with explicit locale specifications, highlighting a significant gap between multilingual capabilities and true internationalisation.

To address this issue, we proposed a system for automatically detecting user locales through system settings, web browser APIs, user profiles, or prior interactions. This information can be used by internationalized LLMs to generate culturally and regionally appropriate responses. Our findings provide a foundation for advancing the global adaptability of LLMs and suggest directions for future research.

7 Limitations

Our dataset covers only five locales (including English). It covers only six aspects of internationalization. Future work needs to expand the locale count as well as the aspect count to ensure models are properly evaluated for internationalization. Moreover, we only evaluated three LLMs. To gain a more comprehensive perspective, a broader range of models should be assessed. Finally, a proof of

concept for an application based on the architecture outlined in Section 5 would be valuable.

References

Cohere For AI. c4ai-command-r-v01. https://huggingface.co/CohereForAI/c4ai-command-r-v01. Accessed: 2024-10-16.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Pinzhen Chen, Simon Yu, Zhicheng Guo, and Barry Haddow. 2024. Is it good data for multilingual instruction tuning or just bad multilingual evaluation for large language models? *arXiv preprint arXiv:2406.12822*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Shubhi Dhaker and Rajshree Shekhawat. 2022. Impact of trustworthiness of instagram influencer on purchase intention of customers: An empirical study of udaipur division.

Richard Ishida and Susan K. Miller. Localization vs. Internationalization. https://www.w3.org/International/questions/qa-i18n. Accessed: 2024-10-16.

B Jarinaa, S Aisha Rani, and M Manida. 2023. A study on chat gpt and its user impacts in current scenario. *International Journal of Emerging Knowledge Studies*, 2(4):99–107.

Kenneth Keniston. 1997. *Software localization: Notes on technology and culture*. Program in Science, Technology, and Society, Massachusetts Institute of Technology.

Gregory Kersten, Mik Kersten, and Wojciech Rakowski. 2002. Software and Culture: Beyond the Internationalization of the Interface. *Journal of Global Information Management*, 10:86–101.

Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. *arXiv preprint arXiv:2406.14805*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946*.

Jialin Li, Junli Wang, Junjie Hu, and Ming Jiang. 2024b. How well do llms identify cultural unity in diversity? *arXiv e-prints*, pages arXiv–2408.

Loïc Maréchal. 2024. The flow of investments in the llm space. In *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*, pages 129–135. Springer Nature Switzerland Cham.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

Aung Pyae. 2018. Understanding the role of culture and cultural attributes in digital game localization. *Entertainment computing*, 26:105–116.

Pedram Rostami, Ali Salemi, and Mohammad Javad Dousti. 2024. Persianmind: A cross-lingual persian-english large language model. *arXiv preprint arXiv:2401.06466*.

similarweb. chatgpt.com Traffic Analytics, Ranking & Audience. https://www.similarweb.com/website/chatgpt.com. Accessed: 2024-10-16.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

WVS. WVS Database. https://www.worldvaluessurvey.org/wvs.jsp. Accessed: 2024-10-16.

L Xue. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.