Translating Domain-Specific Terminology in Typologically-Diverse Languages: A Study in Tax and Financial Education

Arturo Oncevay **Elena Kochkina** **Keshav Ramani**
Toyin Aguda[*] **Simerjot Kaur** **Charese H. Smiley**
JPMorgan AI Research
arturo.oncevay@jpmorgan.com,charese.h.smiley@jpmchase.com

Abstract

Domain-specific multilingual terminology is essential for accurate machine translation (MT) and cross-lingual NLP applications. We present a gold-standard terminology resource for the tax and financial education domains, built from curated governmental publications and covering seven typologically diverse languages: English, Spanish, Russian, Vietnamese, Korean, Chinese (traditional and simplified) and Haitian Creole. Using this resource, we assess various MT systems and LLMs on translation quality and term accuracy. We annotate over 3,000 terms for domain-specificity, facilitating a comparison between domain-specific and general term translations, and observe models' challenges with specialized tax terms. We also analyze the case of terminology-aided translation, and the LLMs' performance in extracting the translated term given the context. Our results highlight model limitations and the value of high-quality terminologies for advancing MT research in specialized contexts.[1]

1 Introduction

Accurate translations are critical for regulated or high-stakes areas such as taxation and finance. Errors in translating specialized terms can lead to misunderstanding, legal inconsistencies, and mistrust in multilingual communication (Nunziatini, 2019). While LLMs and large-scale MT systems have made significant advances, they often struggle with low-frequency, domain-specific terminology (Semenov et al., 2023; Oncevay et al., 2025), especially in languages that are typologically distant or low-resource. Existing public terminology datasets for MT research are often limited in language coverage (Jhirad et al., 2023) or domain specificity (Alam et al., 2021), making it difficult to assess or improve MT systems in specialized settings.

In this work, we introduce a gold-standard multilingual terminology for tax and financial education, enabling further research on domain-specific and terminology-aware MT. Terms are extracted from curated government publications, covering seven typologically and script-diverse languages: English (EN), Spanish (ES), Russian (RU), Vietnamese (VI), Korean (KO), Chinese Traditional (ZH(T)) and Simplified (ZH(S)), and Haitian Creole (HT).[2]

Our main contributions are: (i) crafting a multilingual terminology in two domains; (ii) annotating over 3,000 terms for domain-specificity; (iii) extracting a parallel test set containing a subset of the new terminology; (iv) comparing translation performance of domain-specific versus general terms using various MT systems and LLMs. Moreover, we evaluated terminology-aware translation, assessed LLMs' performance in parallel term extraction, and conducted error analyses of model outputs.

2 Related Work

Regarding domain-specific terminologies, Jhirad et al. (2023) developed a financial glossary, and Kang et al. (2021) categorized financial terms by topic, although both limited to English. In legal and financial MT, most studies focus on single language pairs (Ghaddar and Langlais, 2020; Fu et al., 2024; Luo et al., 2018), with few public resources. Exceptions include Volk et al. (2016), who released a four-language parallel corpus, Nakhle et al. (2025), who published a document-level test set in five European language pairs, and Oncevay et al. (2025), who studied the impact of domain-specific terms in translation for European languages. Besides, terminology-aware MT often uses general-purpose dictionaries (Alam et al., 2021; Ghazvininejad et al., 2023) or terms extracted via LLMs (Semenov et al., 2023). In contrast, our terminology is manually curated and covers seven diverse languages.

[*]Contribution done while working at JPMorgan.

[1]Please contact the author(s) if you want to have access to the terminologies and parallel data.

[2]Appendix A includes details about the languages.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 35030–35044
November 4-9, 2025

3 Multilingual Terminology for Tax and Financial Education

We sourced terminology lists from U.S. government publications, specifically the IRS[3] for tax-related terms and the CFPB[4] for financial education terms. The IRS provides comprehensive terminology relevant to taxation (e.g., *tax relief*, *tangible assets*), while CFPB focuses on financial literacy and education (e.g., *loan origination*, *upfront cost*). Both organizations offer curated glossaries for several languages on their websites and in publicly available PDFs, which are accessible for non-commercial use.[5] We parsed the HTML and PDF documents, extracted the entries, and considered many-to-many term translations across all languages. It is worth noting that CFPB includes data only in traditional Chinese, not simplified.

Example The term *excise taxes*[6] is translated in our IRS dataset as follows: ES→*impuestos sobre artículos de uso y consumo*; KO→*특별사용세*; RU→акцизные налоги; VI→*thuế gián thu*; HT→*dwa dasiz*, ZH(S)→工商税; ZH(T)→工商稅.

Domain-Specific Annotation Although the terminology lists are curated by specialized government entities, they encompass a wide range of terms. To enhance their utility, we categorized the terms into three groups: **Domain-Specific** (DS), **Domain-Contextual** (DC), and **General** (G). DS terms are those with primary meanings directly related to taxation or financial education (e.g., *nonresident alien*, *withholding allowance*). DC terms have specific meanings in tax or financial contexts but are also used in other domains (e.g., *exemption*, *deduction*, *filing*). General terms are common words that appear in tax or financial documents but are not specialized (e.g., *housewife*, *conversion*, *university*). The annotation focused on the English entries, and involved a human-in-the-loop approach, using three commercial closed LLMs as automatic annotators (see Appendix F for the prompt). A human expert reviewed all instances of disagreement among the automatic annotations. Appendix C includes the annotation guidelines.

Data Statistics The IRS and CFPB datasets contain 1,359 and 1,992 English terms, respectively, with entries in other languages having at least 83% coverage, and there are 219 terms included in both sets. The category ratios (DS-DC-G) are 34%–22%–44% for IRS and 86%–10%–4% for CFPB.[7]

4 Parallel Corpus for Evaluation

To leverage our terminologies, we extracted a parallel corpus covering our target languages paired with English and including terms from our datasets. We sourced the translated articles from the IRS repository, which cover topics related to tax services, tax forms, procedures and more. Appendix B lists all the articles.[8]

Alignment Using English as a pivot, we aligned paragraphs across articles by leveraging indexes and headers, with additional cleaning steps. For alignment validation, we computed a Mean Similarity Difference score using multilingual embeddings, where we observed a statistically significant difference for all language-pairs. Moreover, we manually verified a sample of the English-Spanish and English-Russian alignments to ensure accuracy. Appendix D expands on the alignment process.

Data Statistics We extracted 8,994 parallel paragraphs (up to 500 English words each) for all language pairs, sampling 6,491 with IRS terms and 4,652 with CFPB terms. These subsets include terms and translations from their respective glossaries, and we refer to them as IRS and CFPB subsets. While there is segment overlap in the sampled subsets, further MT experiments focus more on the specific term entries. Table 2 in the Appendix shows details on the total number of terms and unique term pairs per language pair.

5 MT with Domain-Specific Terminology

We leveraged the terminologies and parallel corpus to study MT performance and model behavior.

5.1 Experimental Setup

Models We evaluated a diverse set of models: First, two multilingual and open-source

[3]Internal Revenue Service (https://www.irs.gov/)

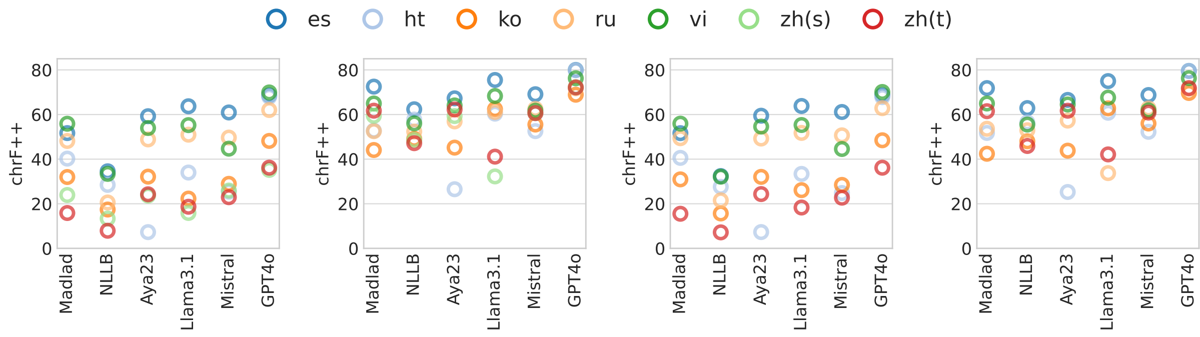
[4]Consumer Financial Protection Bureau (https://www.consumerfinance.gov/)

[5]All content was reviewed for copyright compliance.

[6]We did not inflect or extract the lemmas (e.g., singular form), which can be useful to extend the terminology coverage for inflectional languages. We leave that task for future work.

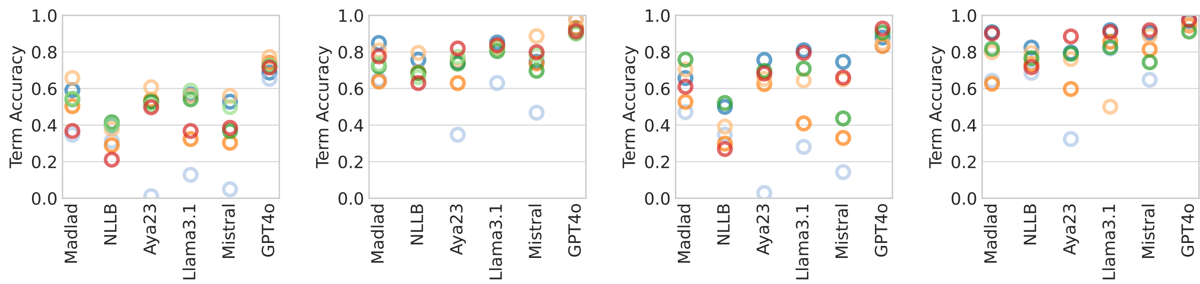
[7]For the MT experiments, these ratios change. The parallel data extracted in §4 includes balanced ratios (DS-DC-G) of 32%–28%–40% for IRS and 49%–33%–18% for CFPB.

[8]Although these articles focus on the tax domain, they also cover some financial education topics (tax procedures, forms). A parallel corpus specifically focused on financial literacy would better enhance the coverage of specialized CFPB terms.



(a) IRS: EN→XX (b) IRS: XX→EN (c) CFPB: EN→XX (d) CFPB: XX→EN

Figure 1: chrF++ scores for IRS and CFPB datasets in both translation directions.



(a) IRS: EN→XX (b) IRS: XX→EN (c) CFPB: EN→XX (d) CFPB: XX→EN

Figure 2: Term Accuracy scores for IRS and CFPB datasets in both translation directions.

MT systems–NLLB (3.3B params.) (NLLB Team et al., 2022) and MADLAD (3B params.) (Kudugunta et al., 2024). Second, light versions of open-weighted LLMs: Aya-23-8B (AYA23; Üstün et al., 2024), Llama-3.1-8B-Instruct (LLAMA3.1; Llama Team, 2024) and Mistral-7B-Instruct-v0.3 (MISTRAL; Jiang et al., 2023). Third, one closed LLM–GPT-4o (OpenAI, 2024).

Inference We used g5.12xlarge AWS instances. For all LLMs, we run a simple zero-shot translation prompt (see Appendix F) with temperature = 0.

Evaluation Metrics For overall translation quality, although COMET (Rei et al., 2020) strongly correlates with human judgment (Kocmi et al., 2021), it cannot reliably handle Haitian Creole. Therefore, for a fair comparison, we used chrF++ (Popović, 2017), a string-based metric.[9] Additionally, we used Term Accuracy (TA) as described by Oncevay et al. (2025), a binary metric that measures whether a term was exactly translated or not. To assess TA significance in all experiments, we used the non-parametric Mann-Whitney U test. The test is suitable for our binary success/failure data as it does not assume normality and compares the full distribution rather than just means.

5.2 Overall Translation and Term Accuracy

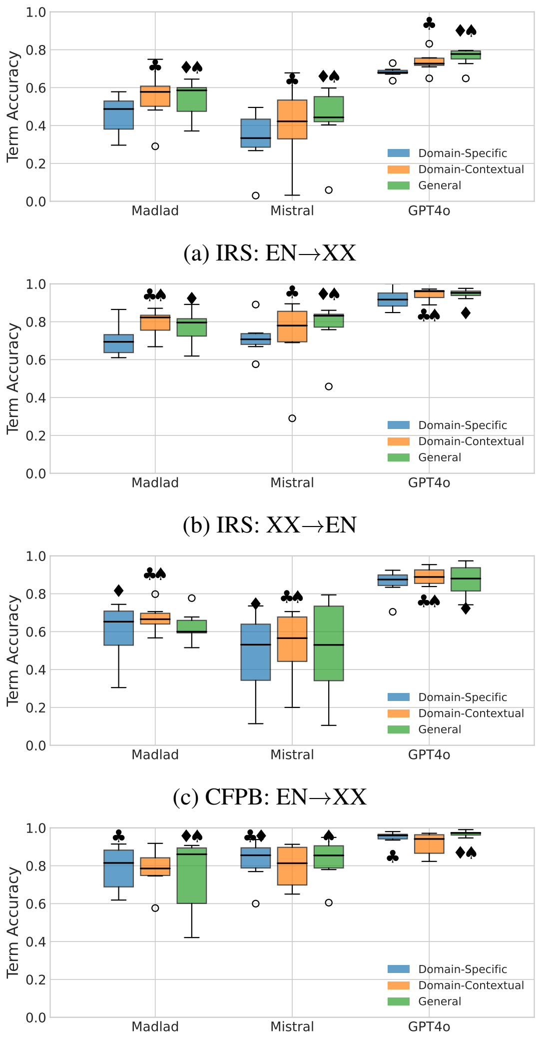
We first analyze the overall translation quality. Figs. 1 and 2 present the chrF++ and TA scores, respectively.[10] As expected, translation performance, both in terms of overall quality and TA, declines when translating out of English across all model types and datasets. Among the light LLMs, performance varies by setting and there is no clear advantage. For EN→XX, ZH(T) consistently emerges as the most challenging language to translate (by chrF++), while HT exhibits the least accurate term translations. Conversely, XX→EN results are more mixed, with ZH(T) and HT frequently appearing as problematic languages in both metrics, and KO occasionally posing challenges. Finally, even GPT-4o, the closed and highest-performing model, shows room for improvement in term accuracy, especially in the EN→XX direction.

5.3 Domain-Specific versus General Terms

We then focus on Term Accuracy across domain specificity (DS, DC, and G). Fig. 3 presents the

[9] As our focus is term-level analysis, chrF++ is used primarily for relative comparison, not to determine the best model.

[10] All scores are detailed in Tables 6 to 13 in the Appendix.



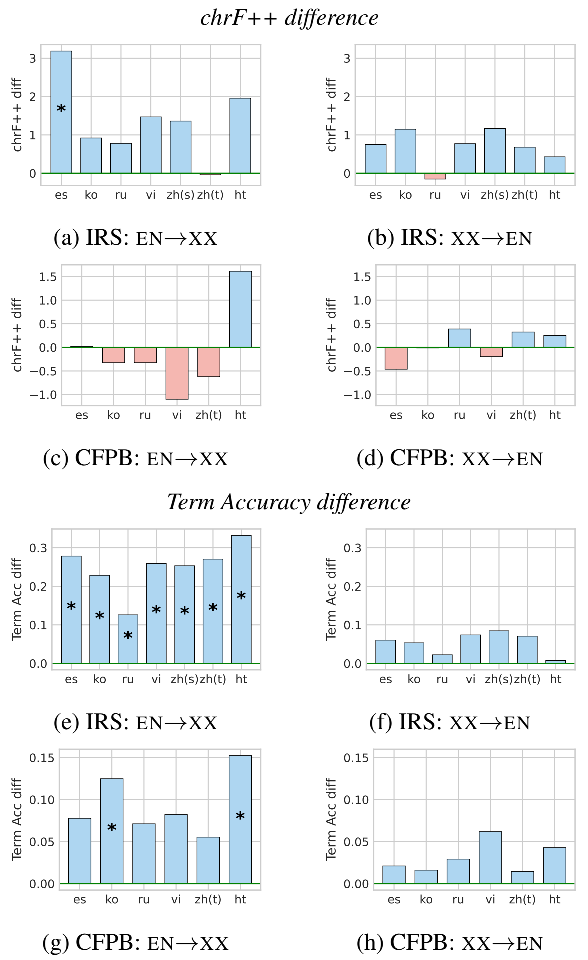
(a) IRS: EN→XX

(b) IRS: XX→EN

(c) CFPB: EN→XX

(d) CFPB: XX→EN

Figure 3: Term Accuracy scores for IRS (a-b) and CFBP (c-d). Statistical significance was determined at $p < 0.05$, and significant differences are marked with symbols above (or below) the better-performing category: ♣ between DS and DC, ♦ between DS and G, and ♠ between DC and G.



(a) IRS: EN→XX

(b) IRS: XX→EN

(c) CFPB: EN→XX

(d) CFPB: XX→EN

(e) IRS: EN→XX

(f) IRS: XX→EN

(g) CFPB: EN→XX

(h) CFPB: XX→EN

Figure 4: Difference in chrF++ (top) and Term Accuracy (bottom) scores using GPT-4o with different prompts. (*) indicate statistically significant differences.

analysis with three different type of models in both translation directions. For the IRS subset (Figs. 3a, 3b), we found that DS terms consistently have lower accuracy scores compared to DC and G terms across all settings, highlighting the challenge they pose in translation. Besides, the mixed results between DC and G categories indicate that models handle these terms with comparable accuracy, likely due to the higher frequency of DC terms in varied contexts within pretraining data.

However, we note that this pattern is not consistent for the CFPB subset (Figs. 3c, 3d). This likely arises because, while CFPB (financial education) terms can include some complex concepts, the distinction between categories is more blurred, leading to biases in differentiating specialized versus general terms, as noted in the imbalanced label ratio for CFPB in §3.

5.4 Terminology-Aided Translation

We measure performance variation using terminology-aware prompts (Ghazvininejad et al., 2023). The new prompt (see Appendix F) incorporates the target terms and their translations. For this experiment, we focus on GPT-4o, the best-performing LLM, to stress-test the contribution of proper terminology in a strong system.

Fig. 4 shows the score differences for chrF++ and Term Accuracy (TA) compared to the baseline translation prompt. As expected, there is no significant variation for XX→EN in either metric or subset, as the baseline scores are already robust. However, for EN→XX in the IRS subset (Fig. 4e), there is a notable improvement in TA across all languages. Results for CFPB are less pronounced, with only 2 out of 6 languages showing significantly better TA for EN→XX (Fig. 4g). This is also expected, as the baseline results for CFPB with GPT-4o were already high (average 0.87; see Fig. 2c), leaving limited room for improvement. Finally, our analysis shows that a metric like chrF++ cannot consistently capture these changes (Figs. 4a, 4c), underscoring the importance of our term-level evaluation.

Domain-specific versus general terms We also examined whether terminology-aided translation prompts have different effects on TA for domain-specific (DS), domain-contextual (DC), and general (G) terms. Figure 5 in the Appendix presents the results. For XX→EN, differences are minimal (below 0.15) across all languages and datasets, likely due to GPT-4o's strong performance in this direction. For EN→XX, IRS shows a greater improvement for DS terms compared to general terms in most languages, except Haitian Creole. In the CFPB subset, the pattern is less consistent, though Haitian Creole remains an exception. These findings suggest that incorporating terminology can be particularly beneficial for translating tax-specific terms.

Prompt size increase We analyzed the prompt size increase when using the terminology-aware prompt compared to the baseline. Table 4 in Appendix E provides the details. The terminology-aided approach increases token consumption by approximately 1.3 times on average, ranging from 1.22 times (Korean or Russian) to 1.4 times (Spanish). The results indicate a trade-off with improved translation quality or terminology accuracy, which should be assessed per language pair.

5.5 Parallel Term Extraction

We further investigate the strength of LLMs in extracting exact term translations using parallel text as context. Appendix F includes the prompt, and Table 1 presents the accuracy of parallel term extraction for EN→XX experiments (English term is given) using Mistral and GPT-4o. As expected, GPT-4o consistently outperforms Mistral, with a smaller performance gap between the best and worst languages (12%) compared to Mistral (44%). Vietnamese poses the greatest challenge for both models, while the top-performing language varies by dataset for each model. Notably, even for low-resource Haitian Creole, high-quality term translations can be extracted, aiding resource creation.

	IRS		CFPB	
Lang.	MISTRAL	GPT-4O	MISTRAL	GPT-4O
ES	0.74	0.94†	0.75	0.95†
HT	0.51	0.94†	0.64	0.98†
KO	0.81	0.91†	0.71	0.86†
RU	0.72	0.88†	0.74	0.91†
VI	0.37	0.84†	0.44	0.91†
ZH(S)	0.76	0.89†	-	-
ZH(T)	0.74	0.88†	0.77	0.95†
Avg.	0.66	0.90†	0.68	0.92†

Table 1: Parallel Term Extraction Accuracy (EN→XX). † indicates stats. significant improvement ($p < 0.05$)

5.6 Error Analysis and Discussion

Finally, we conducted an error analysis on translation and term extraction tasks, with examples and annotations by a Russian-speaking expert provided in Appendix G. In the term extraction task, Mistral often deviates from the specified output format, adding extraneous tokens like synonyms or alternative prompt continuations. A frequent issue is the mismatch in conjugation or plurality between the predicted term and the gold standard, which, in Russian, does not always constitute an incorrect translation. In the translation task (EN→RU), both models demonstrate strong performance. Errors often arise from terminology translation, as well as from translating common terms that lack direct equivalents in the target language, and incorrect grammar. Generally, the models favor a literal translation approach, with Mistral being more literal than GPT-4o, which produces more natural-sounding translations. Notably, Mistral consistently uses descriptive phrases instead of adjectives, such as "income that is subject to tax" instead of "taxable income". We also observed that terminology-aided translation enhances model outputs, improving translation quality.

6 Conclusion

We introduced new typologically and script-diverse language resources for MT research: translated terminologies for tax and financial education domains in seven languages, along with a parallel corpus for the tax domain. We enhanced the terminologies with domain-specificity annotations and observed that different models faced challenges when translating tax-specific terms compared to general terms. Knowing that LLMs can benefit from incorporating translated term pairs in the translation prompts, this study represents a valuable step forward in promoting MT research in specialized domains.

7 Limitations

This study acknowledges several limitations that may affect the generalizability and comprehensiveness of the findings. First, the parallel corpus (§4) is derived from a snapshot of IRS articles published online. Only five out of seven translated articles were published recently (2024), making it impossible to fully mitigate data contamination in large-scale models (MT systems or LLMs, whether open-weighted or closed). Nonetheless, while it is likely that some models have been pretrained with the raw data, it is less likely they have been post-trained specifically for translation tasks using these texts. Moreover, results indicate that all models still struggle, to varying degrees, with accurate term-level translation and overall translation quality.

Second, the study was constrained by access to a Russian-speaking annotator solely for error analysis. While this provided valuable insights, the scope of linguistic expertise was limited, and future research will aim to expand the range of linguistic annotators to enhance the robustness of error analysis across multiple languages.

Third, the selection of MT systems and LLMs was limited, and the chosen models are primarily trained on general-domain data. This may have restricted their ability to handle domain-specific terminology and contexts. Future research should explore fine-tuning these models on tax and finance data, and/or incorporate a broader range of LLMs to capture a wider spectrum of linguistic nuances and improve the depth of analysis.

Lastly, the prompt engineering for each scenario (translation and term extraction) was limited due to the exploratory nature of this study. As the focus was on contributing resources and laying groundwork for terminology discovery, exhaustive prompt engineering was not pursued. Future research will aim to refine and expand prompt engineering techniques to optimize the performance and applicability of LLMs in various linguistic contexts.

Acknowledgments

We are thankful to Samuel Mensah, Joy Sain, Xiaomo Liu and other members of the JPMorgan AI Research team for their insightful feedback since early stages of this work. We also appreciate the feedback of the anonymous reviewers and meta-reviewer.

Disclaimer

References

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663, Online. Association for Computational Linguistics.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.

Yuxin Fu, Shijing Si, Leyi Mai, and Xi-ang Li. 2024. Ffn: a fine-grained chinese-english financial domain parallel corpus. In *2024 International Conference on Asian Language Processing (IALP)*, pages 127–132. IEEE.

Abbas Ghaddar and Phillippe Langlais. 2020. SEDAR: a large scale French-English financial domain parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3595–3602, Marseille, France. European Language Resources Association.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *Preprint*, arXiv:2302.07856.

James Jhirad, Edison Marrese-Taylor, and Yutaka Matsuo. 2023. Evaluating large language models’ understanding of financial terminology via definition modeling. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 93–100, Nusa Dua, Bali. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Juyeon Kang, Ismail El Maarouf, Sandra Bellato, and Mei Gan. 2021. FinSim-3: The 3rd shared task on learning semantic similarities for the financial domain. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, pages 31–35, Online. -.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. Madlad-400: a multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Linkai Luo, Haiqin Yang, Sai Cheong Siu, and Francis Yuk Lun Chin. 2018. Neural machine translation for financial listing documents. In *Neural Information Processing*, pages 232–243, Cham. Springer International Publishing.

Mariam Nakhle, Marco Dinarelli, Raheel Qader, Emmanuelle Esperança-Rodier, and Hervé Blanchon. 2025. DOLFIN - document-level financial test-set for machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5544–5556, Albuquerque, New Mexico. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Mara Nunziatini. 2019. Machine translation in the financial services industry: A case study. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 57–63, Dublin, Ireland. European Association for Machine Translation.

Arturo Oncevay, Charese H. Smiley, and Xiaomo Liu. 2025. The impact of domain-specific terminology on machine translation for finance in european languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, New Mexico, US. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the WMT 2023 shared task on machine translation with terminologies. In *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore. Association for Computational Linguistics.

UncleCode. 2024. Crawl4ai: Open-source llm friendly web crawler & scraper. `https://github.com/unclecode/crawl4ai`.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Martin Volk, Chantal Amrhein, Noëmi Aepli, Mathias Müller, and Phillip Ströbel. 2016. Building a parallel corpus on the world's oldest banking magazine. In *KONVENS*. s.n.

A Languages

Our study covers seven typologically and script-diverse languages. Among Indo-European languages, we include Spanish (ES, Romance), English (EN, Germanic), and Russian (RU, Slavic). Notably, Russian uses the Cyrillic script, in contrast to the Latin script used by the first two ones. We also include Korean (KO, Koreanic), which uses the Hangul script, and Vietnamese (VI, Austroasiatic, Mon-Khmer branch), which is written with a Latin-based script enriched with diacritics. Chinese is represented in two script variants: Simplified (ZH(S)) and Traditional (ZH(T)), both logographic and part of the Sino-Tibetan language family. Finally, Haitian Creole (HT), a French-based creole with significant lexical influences from West African languages, is written in the Latin script.

It is worth noting that Haitian Creole is the lowest-resource language/creole in our study (see Table 2). Although the extracted parallel data for the EN–HT language pair contains less than 100 samples, we decided to include HT in the study because the total number of covered terms (266 for IRS and 211 for CFPB) and unique terms (83 for IRS and 72 for CFPB) compose a robust sample for assessing translation accuracy at the term-level.

B Parallel Data Source

The following list contains the extracted English articles from the IRS repository:

- https://www.irs.gov/publications/p17
- https://www.irs.gov/publications/p334
- https://www.irs.gov/publications/p519
- https://www.irs.gov/publications/p547
- https://www.irs.gov/publications/p596
- https://www.irs.gov/publications/p850
- https://www.irs.gov/publications/p947

To obtain the translated articles, append the corresponding language code at the end of the URL: Spanish→sp, Russian→ru, Korean→ko, Chinese simplified→zhs, Chinese traditional→zht, Vietnamese→vi, Haitian Creole→ht.

C Domain-Specific Annotation Process and Guideline

We categorized all terms in three classes: Domain-Specific (DS), Domain-Contextual (DC) and General (G) as defined in Section 3. First, we obtained automatic labels using three commercial LLMS: OpenAI's GPT-4o (https://openai.com/index/hello-gpt-4o/), Anthropic's Claude Sonnet 3.7 (https://www.anthropic.com/news/claude-3-7-sonnet) and Google's Gemini 2.5 Pro (https://deepmind.google/technologies/gemini/pro/). The prompt used to obtain the annotations in the tax domain is shown in Appx. F. The prompt for the financial education case is similar, but we updated the examples per category. Second, we provided the three model outputs plus the prompt details to a human expert with a background in both linguistics and finance to review disagreements. During the review process, the human annotator was provided the three labels from the LLMs but could select any label from the three classes. All judgments were based on the term alone and it is important to note that the glossary terms provided for annotation did not include sentence context or definitions. This could lead to some ambiguity, especially between General and Contextual terms, where the particular word sense in scope would have been made clearer by the context.

D Parallel Data Extraction

First, with the CRAWL4AI tool (UncleCode, 2024), we transformed all articles from HTML to Markdown. For each set of translated articles, we extracted the indexes (headers provided in the same HTML), and aligned the hierarchy nodes between the English index and each other language. If we found a disagreement between the number of nodes at any hierarchy level, we dropped that branch for the language-pair. After pruning, we searched, in a sequential way, for the text spans between one header and the next one in the main body of the article. This step guaranteed that we extracted the same paragraphs for the article in the language-pair. Besides, we validated that each text span contains the same number of line breaks, and we dropped it otherwise. The total number of segments/paragraphs, plus the number of term pairs and unique term pairs per subset, are included in Table 2.

Afterwards, to validate the overall process, a Spanish and a Russian native speaker checked that

Lang	Total #segments	IRS subset #segments	Term Pairs	Unique P.	CFPB subset #segments	Term Pairs	Unique P.
es	1,188	1,066	4,092	437	995	3,037	278
ht	95	82	266	83	79	211	72
ko	1,522	1,198	3,431	268	779	1,265	90
ru	1,539	668	1,140	156	854	1,533	200
vi	1,539	1,334	4,764	361	1,139	2,838	218
zh(s)	1,577	1,076	2,488	250	-	-	-
zh(t)	1,534	1,067	2,462	242	806	1,465	118
Total	8,994	6,491	18,643	1,797	4,652	10,349	976

Table 2: Statistics of parallel data (each language paired with English) and terminology matches. CFPB terminology does not include Chinese Simplified.

the alignments in the English-Spanish and English-Russian language-pairs were correct in two out of the seven articles, and reviewed a sample of entries from the other five articles.

D.1 Alignment Automatic Evaluation

In addition to the manual validation, we conducted an automatic evaluation of the alignment quality using multilingual embeddings: PARAPHRASE-MULTILINGUAL-MINILM-L12-V2 (Reimers and Gurevych, 2019). For each language pair, we computed cosine similarity between:

- Correctly aligned pairs ($\texttt{en}_\texttt{i} \leftrightarrow \texttt{target}_\texttt{i}$)
- Consecutively misaligned pairs ($\texttt{en}_\texttt{i} \leftrightarrow \texttt{target}_{\texttt{i}\pm\texttt{1},\texttt{i}\pm\texttt{2}}$)

and calculate the Mean Similarity Difference (MSD) as the difference between the two scores. The quality assessment for MSD is as follows: GOOD > 0.1, MODERATE = 0.05-0.1, POOR ≤ 0.05. We also report Cohen's d effect size for statistical significance (Cohen, 2013):

- Cohen's d >0.8 = Large effect
- Cohen's d in 0.5-0.8 = Medium effect
- Cohen's d <0.5 = Small effect

Dataset-Language	MSD	Cohen's d
IRS-Spanish	0.5674	4.2370
CFPB-Spanish	0.5571	4.3680
CFPB-Vietnamese	0.5458	4.2047
CFPB-Russian	0.5445	4.3206
IRS-Vietnamese	0.5393	4.1262
IRS-Simplified Chinese	0.5320	4.1801
IRS-Traditional Chinese	0.5196	3.9632
IRS-Korean	0.5075	3.9422
IRS-Russian	0.5067	4.0115
CFPB-Korean	0.4881	3.9007
CFPB-Traditional Chinese	0.3540	2.3210
IRS-Haitian Creole	0.1152	0.9027
CFPB-Haitian Creole	0.1093	0.8627

Table 3: Alignment quality results. Sorted (desc.) by Mean Similarity Difference (MSD).

Table 3 shows that MSD is greater than 0.1 in all cases, indicating good alignment quality. Meanwhile, Cohen's $d > 0.8$ indicates a **large effect size** (statistically significant) for all languages. Regarding Haitian Creole, while showing the lowest mean difference, this is expected as it is a lower-resource language with limited representation in multilingual embedding spaces. Importantly, our alignment procedure was identical across all language pairs, ensuring methodological consistency.

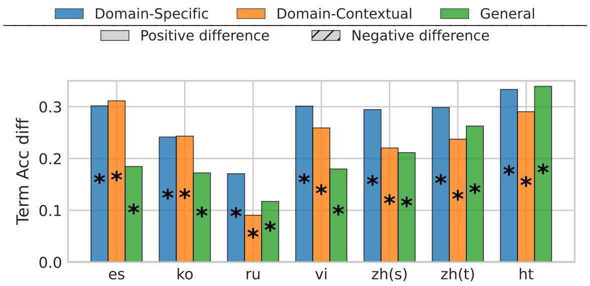
E Terminology-Aided Translation

Token Size per Prompt Type We computed the token size difference between the baseline prompt and the terminology-aided one. Table 4 shows the results.

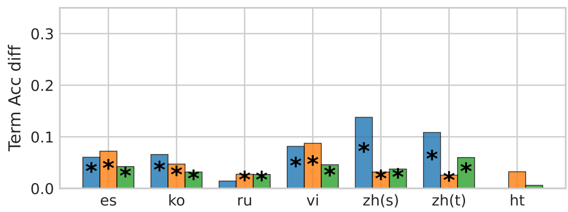
Dataset	Language	MTB	MTT	%Increase
IRS	Spanish	146.9	206.1	40.3%
	Vietnamese	149.8	209.6	39.9%
	Haitian Creole	154.3	208.1	34.8%
	Korean	153.5	202.8	32.1%
	Chinese(T)	160.3	204.8	27.7%
	Chinese(S)	161.3	205.8	27.5%
	Russian	168.5	207.4	23.1%
CFPB	Spanish	151.1	200.7	32.9%
	Haitian Creole	155.3	202.9	30.6%
	Vietnamese	156.3	204.1	30.6%
	Russian	166.2	205.1	23.4%
	Chinese(T)	172.4	211.9	22.9%
	Korean	169.0	206.6	22.3%

Table 4: Token Cost Increase Summary. MTB = Mean tokens for Baseline prompt, MTT = Mean tokens for Terminology-aware prompt.

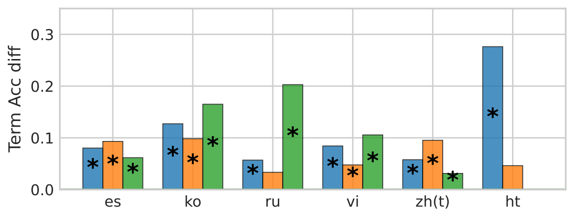
Domain-Specific Analysis We compared the Term Accuracy difference between the terminology-aided translation prompt and the baseline one. Results are shown in Figure 5.



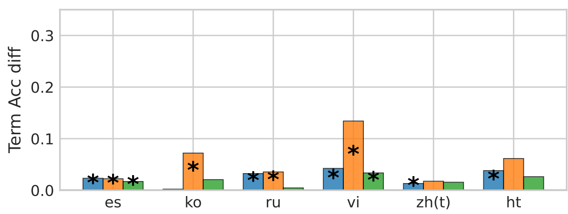
(a) IRS: EN→XX



(b) IRS: XX→EN



(c) CFPB: EN→XX



(d) CFPB: XX→EN

Figure 5: Difference in Term Accuracy scores using GPT-4o with a terminology-aided translation prompt and a baseline one. (*) indicate statistically significant differences. Results are divided by term category.

F Prompts

We provide the prompt templates used in the paper across experiments on different tasks.

Domain-Specificity Annotation (Tax)

```
We need to label and differentiate the domain specificity of the extracted
terminology list. Domain specificity measures how exclusively a term belongs
to the tax domain:
1. Domain-Specific (DS): Terms that have a primary meaning directly related
to taxation, tax procedures, or tax administration (e.g., "nonresident alien",
"withholding allowance", "earned income tax credit")
2. Domain-Contextual (DC): Terms that have specific meanings when used in tax
contexts but also have common meanings in other domains (e.g., "exemption",
"deduction", "filing")
3. General (G): Common words that appear in tax documents but aren't specialized
tax terms (e.g., "housewife", "conversion", "tab", "university").
Annotate the terms provided in the file filename.csv and provide the annotations
as an additional column in out_filename.csv.
```

Translation I: Baseline

```
You are a professional translator in the tax and finance domains.
Translate the following sentence from “{source lang.}” to “{target lang.}”:
“{source lang.}”: “{text}”
Return the translation in a JSON structure: “{target lang.}”: “ ...”
```

Translation II: Terminology-Aided Traslation

```
You are a professional translator in the tax and finance domains.
Translate the following sentence from “{source lang.}” to “{target lang.}”:
“{source lang.}”: “{text}”
Be careful and pay special attention to the following keywords in “{source lang.}”
along with their translations in “{target lang.}”: “{term pairs list}”
Return the translation in a JSON structure: “{target lang.}”: “ ...”
```

Parallel Term Extraction

```
You are a translation and word alignment specialist in financial and tax domains.
Given a sentence in a source language and its translation to a target language,
return a translation of a given term in a target language. Do not add any extra
statements to the output.
Source language sentence: {source lang. sent.}
Target language sentence: {target lang. sent.}
Term in source language: {source lang. term}
Term in target language:
```

G Examples and Error Analysis

We present example inputs and outputs for each of our tasks, we comment on the types of issues observed.

Example 1

Source Text (EN): For more information on excise taxes, see Pub. 510, Excise Taxes.
Reference Translation (RU): Более подробную информацию об акцизах см. в Публикации № 510 «Акцизные налоги».
Term Pair(s): 'excise taxes': 'акцизные налоги'

GPT-4o Translation (RU): Для получения дополнительной информации о акцизных налогах см. Публикацию 510, Акцизные налоги.
GPT-4o Predicted Term Pair(s): 'excise taxes': 'акцизные налоги'
GPT-4o Terminology-aided Translation (RU): Для получения дополнительной информации об акцизных налогах см. Публикацию 510, Акцизные налоги.

Mistral Translation (RU): Для получения дополнительной информации о налогах на спиртные напитки, ознакомьтесь с Pub. 510, Налог на спиртные напитки.
Mistral Predicted Term Pair(s): 'excise taxes': 'акцизные налоги'
Mistral Terminology-aided Translation (RU): Для получения дополнительной информации о акцизных налогах, ознакомьтесь с Pub. 510, Акцизные налоги.

Comment: (1) The term 'excise taxes' is used twice in the source text, and the reference provides two translations as a noun 'акцизы' and as a phrase 'акцизные налоги';
(2) Both models do well at parallel term extraction;
(3) In the translation task, GPT-4o translates the term correctly (but makes a minor grammatical error ('о')), while Mistral translates it as 'tax on alcohol drinks';
(4) Both models benefit from Terminology-aided Translation, making Mistral provide the correct term, and GPT-4o to fix the grammatical error.

Example 2

Source Text (EN): The IRS will accept a copy of a power of attorney that is submitted by facsimile transmission (fax). If you choose to file a power of attorney by fax, be sure the appropriate IRS office is equipped to accept this type of transmission.
Reference Translation (RU): IRS примет копию доверенности, предоставленной по факсу. Если вы решите послать доверенность по факсу, убедитесь, что соответствующий офис IRS оборудован для приема такого рода отправлений.
Term Pair(s): 'power of attorney': 'доверенность'

GPT-4o Translation (RU): Налоговая служба США (IRS) примет копию доверенности, отправленную по факсу. Если вы решите подать доверенность по факсу, убедитесь, что соответствующее отделение IRS оснащено для приема такого типа передачи.
GPT-4o Predicted Term Pair(s): 'power of attorney': 'доверенность'
GPT-4o Terminology-aided Translation (RU): Налоговая служба США (IRS) примет копию доверенности, отправленную по факсу. Если вы решите подать доверенность по факсу, убедитесь, что соответствующее отделение IRS оборудовано для приема такого типа передачи.

Mistral Translation (RU): IRS принимает копию полномочий, подготовленную передачей по факсу (факс). Если вы решите подать полномочия по факсу, убедитесь, что соответствующий офис IRS оборудован для приема такого типа передачи.
Mistral Predicted Term Pair(s): 'power of attorney': 'доверенность'
Mistral Terminology-aided Translation (RU): Бюро внутренних доходов (IRS) примет копию доверенности, которую вы отправите по передаче факса (факсу). Если вы решите отправить доверенность по факсу, убедитесь, что соответствующий офис IRS оборудован для приема этого типа передачи.

Comment: (1) The terms in Russian Language can be conjugated depending on their role in the sentence, and the models, performing parallel term extraction may predict the standard form or extract the conjugated form, which migth affect the evaluation; (2) Mistral translates the term incorrectly initially (as полномочий), which also leads to the incorrect word usage, but recovers given the term translation in the prompt.

Example 3

Source Text (EN): If you operate your own business or have other self-employment income, such as from babysitting or selling crafts, see the following publications for more information. Pub. 334, Tax Guide for Small Business. Pub. 225, Farmer's Tax Guide. Pub. 587, Business Use of Your Home.
Reference Translation (RU): Если вы имеете собственный бизнес или иной доход от самостоятельной занятости, например, от присмотра за детьми или продажи ремесленной продукции, для получения дополнительной информации см. следующие публикации. Публикация № 334 «Руководство по налогам для малого бизнеса». Публикация № 225 «Налоговый справочник фермера». Публикация № 587 «Использование дома для предпринимательской деятельности».
Term Pair(s): 'income': 'доход', 'tax': 'налоговый'

GPT-4o Translation (RU): Если вы ведете собственный бизнес или имеете другой доход от самозанятости, например, от няни или продажи изделий ручной работы, ознакомьтесь с следующими публикациями для получения дополнительной информации. Публ. 334, Налоговое руководство для малого бизнеса. Публ. 225, Налоговое руководство для фермеров. Публ. 587, Использование вашего дома в бизнесе.
GPT-4o Predicted Term Pair(s): 'income': 'доход', 'tax': 'налог'
GPT-4o Terminology-aided Translation (RU): Если вы ведете собственный бизнес или имеете другой доход от самозанятости, например, от присмотра за детьми или продажи изделий ручной работы, ознакомьтесь с следующими публикациями для получения дополнительной информации. Публ. 334, Налоговый справочник для малого предприятия. Публ. 225, Налоговый справочник для фермеров. Публ. 587, Использование вашего дома в бизнесе.

Mistral Translation (RU): Если вы ведете собственный бизнес или имеете доход от самозанятости, например, от попечительства или продажи ремесел, ознакомьтесь с следующими публикациями для получения дополнительной информации. Pub. 334, Налоговая инструкция для малого бизнеса. Pub. 225, Налоговая инструкция для фермеров. Pub. 587, Бизнес-использование вашей квартиры.
Mistral Predicted Term Pair(s): 'income': 'доход', 'tax': 'налог'
Mistral Terminology-aided Translation (RU): Если вы ведете собственный бизнес или имеете другие доходы от самозанятости, например, от попечения о ребенке или продажи ремесел, ознакомьтесь с следующими публикациями для получения дополнительной информации. Pub. 334, Налоговая инструкция для малого бизнеса. Pub. 225, Налоговая инструкция для фермеров. Pub. 587, Бизнес-использование вашей квартиры.

Comment: Whilst both models get all the terminology translation correctly, they both make mistakes around translating more general phrases, such as 'babysitting' and 'selling crafts'. Interestingly, the terminology-aided translation prompt helps recover some of those errors. Mistral consistently doesn't translate 'Pub.' ans leaves it as is.

Table 5: Examples of EN to RU inputs and predictions by GPT-4o and Mistral for Translation, Parallel Term Extraction and Terminology-aided Translation tasks. Blue highlights the terms, red - issues with translation, and green - fixed issue.

Lang.	MADLAD	NLLB	AYA23	LLAMA3.1	MISTRAL	GPT4O
ES	51.68	34.56	59.24	63.71	60.92	68.57
HT	40.23	28.45	7.29	34.00	25.40	67.93
KO	31.99	17.42	32.16	22.32	28.93	48.17
RU	48.14	20.57	48.85	50.97	49.66	62.02
VI	55.87	33.39	53.95	55.25	44.70	69.90
ZH(S)	23.93	13.36	23.66	15.88	25.91	35.18
ZH(T)	15.86	7.90	24.25	18.66	22.94	36.19
Avg.	38.24	22.24	35.63	37.25	36.92	55.42

Table 6: chrF++ Scores for IRS dataset (EN→XX)

Lang.	MADLAD	NLLB	AYA23	LLAMA3.1	MISTRAL	GPT4O
ES	0.59*	0.40	0.53	0.57	0.53	0.69
HT	0.35	0.31	0.01	0.13	0.05	0.65
KO	0.50	0.29	0.54*	0.32	0.30	0.74
RU	0.66*	0.38	0.61	0.56	0.56	0.77
VI	0.54	0.42	0.53	0.54	0.37	0.72
ZH(S)	0.55	0.40	0.51	0.59*	0.50	0.72
ZH(T)	0.37	0.21	0.50*	0.37	0.39	0.71
Avg.	0.51*	0.34	0.46	0.44	0.38	0.72

Table 7: Term Accuracy Scores for IRS dataset (EN→XX). * indicates significant improvement over other comparable models (MT systems and open-weights LLMs)

Lang.	MADLAD	NLLB	AYA23	LLAMA3.1	MISTRAL	GPT4O
ES	72.50	62.36	67.25	75.48	69.10	79.99
HT	52.78	57.75	26.52	60.40	52.56	79.84
KO	44.09	49.12	45.16	62.41	55.52	68.88
RU	52.53	52.36	56.92	61.15	62.55	71.75
VI	64.94	56.02	64.08	68.30	61.71	76.25
ZH(S)	59.43	48.58	59.31	32.26	60.62	72.25
ZH(T)	61.71	47.16	62.28	41.12	60.64	72.14
Avg.	58.28	53.33	54.50	57.30	60.38	74.44

Table 8: chrF++ Scores for IRS dataset (XX→EN)

Lang.	MADLAD	NLLB	AYA23	LLAMA3.1	MISTRAL	GPT4O
ES	0.85	0.76	0.74	0.85	0.75	0.93
HT	0.65	0.69	0.35	0.63	0.47	0.98
KO	0.64	0.69	0.63	0.83*	0.74	0.93
RU	0.81	0.80	0.76	0.82	0.89*	0.97
VI	0.72	0.68	0.74	0.80*	0.70	0.90
ZH(S)	0.73	0.66	0.77	0.83*	0.78	0.90
ZH(T)	0.78	0.63	0.82	0.84	0.80	0.91
Avg.	0.74	0.70	0.69	0.80*	0.73	0.93

Table 9: Term Accuracy Scores for IRS dataset (XX→EN). * indicates significant improvement over other comparable models (MT systems and open-weights LLMs)

Lang.	MADLAD	NLLB	AYA23	LLAMA3.1	MISTRAL	GPT4O
ES	51.65	32.43	59.46	63.83	61.14	68.65
HT	40.57	27.60	7.32	33.39	24.82	67.79
KO	30.94	15.65	32.05	26.10	28.45	48.46
RU	49.38	21.52	49.24	51.76	50.67	62.85
VI	55.91	32.13	54.60	55.30	44.55	70.11
ZH(T)	15.52	7.18	24.34	18.32	22.72	36.16
Avg.	40.66	22.75	37.84	41.45	38.73	59.01

Table 10: chrF++ Scores for CFPB dataset (EN→XX)

Lang.	MADLAD	NLLB	AYA23	LLAMA3.1	MISTRAL	GPT4O
ES	0.66	0.50	0.76	0.81*	0.75	0.88
HT	0.47*	0.35	0.03	0.28	0.14	0.83
KO	0.53	0.30	0.62*	0.41	0.33	0.83
RU	0.70*	0.39	0.65	0.64	0.65	0.85
VI	0.76*	0.52	0.70	0.71	0.44	0.91
ZH(T)	0.61	0.27	0.68	0.80*	0.66	0.93
Avg.	0.62*	0.39	0.57	0.61	0.49	0.87

Table 11: Term Accuracy Scores for CFPB dataset (EN→XX). * indicates significant improvement over other comparable models (MT systems and open-weights LLMs)

Lang.	MADLAD	NLLB	AYA23	LLAMA3.1	MISTRAL	GPT4O
ES	71.85	62.90	66.54	74.99	68.74	79.54
HT	51.75	56.46	25.28	60.79	52.25	79.54
KO	42.42	48.01	43.80	62.76	55.98	69.66
RU	53.62	52.96	57.28	33.75	62.73	71.56
VI	64.94	55.50	64.50	67.58	61.82	76.30
ZH(T)	61.50	45.90	61.70	42.13	60.87	71.84
Avg.	57.68	53.62	53.18	57.00	60.40	74.74

Table 12: chrF++ Scores for CFPB dataset (XX→EN)

Lang.	MADLAD	NLLB	AYA23	LLAMA3.1	MISTRAL	GPT4O
ES	0.91	0.82	0.80	0.92	0.90	0.97
HT	0.64	0.69	0.32	0.82*	0.65	0.95
KO	0.63	0.73	0.60	0.86*	0.81	0.95
RU	0.80	0.80	0.76	0.50	0.89*	0.95
VI	0.82	0.77	0.79	0.83	0.74	0.91
ZH(T)	0.90	0.72	0.88	0.91	0.92	0.98
Avg.	0.78	0.75	0.69	0.80	0.82	0.95

Table 13: Term Accuracy Scores for CFPB dataset (XX→EN). * indicates significant improvement over other comparable models (MT systems and open-weights LLMs)