Culture-aware machine translation: the case study of low-resource language pair Catalan-Chinese

Xixian Liao, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari
Javier García Gilabert, Miguel Claramunt Argote, Ella Bohman, Maite Melero
Barcelona Super Computing Center (BSC)
{xixian.liao,carlos.escolano,audrey.mash,francesca.delucafornaciari,
javier.garcia1,miguel.claramunt,ella.bohman,maite.melero}@bsc.es

Abstract

High-quality machine translation requires datasets that not only ensure linguistic accuracy but also capture regional and cultural nuances. While many existing benchmarks, such as FLORES-200, rely on English as a pivot language, this approach can overlook the specificity of direct language pairs, particularly for underrepresented combinations like Catalan-Chinese. In this study, we demonstrate that even with a relatively small dataset of approximately 1,000 sentences, we can significantly improve MT localization. To this end, we introduce a dataset specifically designed to enhance Catalan-to-Chinese translation by prioritizing regionally and culturally specific topics. Unlike pivot-based datasets, our data source ensures a more faithful representation of Catalan linguistic and cultural elements, leading to more accurate translations of local terms and expressions. Using this dataset, we demonstrate better performance over the English-pivot FLORES-200 *dev* set and achieve competitive results on the FLORES-200 *devtest* set when evaluated with neural-based metrics. We release this dataset as both a human-preference resource and a benchmark for Catalan-Chinese translation. Additionally, we include Spanish translations for each sentence, facilitating extensions to Spanish-Chinese translation tasks.

1 Introduction

In recent years, the field of neural machine translation (NMT) has seen substantial progress in the development of multilingual models, which can translate across multiple languages as a single unified model (e.g., Zhang et al., 2020; Siddhant et al., 2020; Fan et al., 2021; Costa-jussà et al., 2022; Kudugunta et al., 2024), as well as the creation of human-translated multilingual benchmark datasets (e.g., Costa-jussà et al., 2022; Federmann et al., 2022). These advancements have pushed the boundaries of many-to-many translation capabilities. However, practical applications often require systems to be tailored to specific cultural and regional contexts (e.g., Naveen and Trojovskỳ, 2024). One particularly challenging area is the translation of texts that contain entity names, as cultural-related references can vary significantly across languages (Conia et al., 2024). Translating names between languages with different scripts, such as Latin and logographic (e.g., Chinese), also involves transliteration to maintain ease of pronunciation and closeness to the original sound. Sometimes, the same name can even yield different transliterations based on the source language's pronunciation. For example, the name *José* is transliterated as 若泽(ruò zé) from Portuguese to Chinese, but from Spanish, it becomes 何塞(hé sài). Therefore, we need to adapt the many-to-many system to be more language- and culture-specific.

This study focuses on the Catalan-to-Chinese (CA→ZH) translation, a relatively underexplored area despite its growing relevance given the deepening economic and cultural connections between Catalonia and China. Chinese speakers form one of the five largest immigrant communities in Catalonia, where Catalan is an official language.[1] Besides, China is also Catalonia's third-largest non-European investor and the top source of non-European, non-English-speaking tourists.[2] These growing interactions underline the urgent need for effective translation tools to facilitate communication and foster collaboration between Catalan and Chinese speakers. Despite its significance, developing robust CA-ZH MT systems remains challeng-

[1] https://www.idescat.cat/novetats/?id=4815&lang=en. Accessed January 3, 2025.
[2] https://catalonia.com/w/catalan-government-launches-china-desk-to-promote-chinese-investment-and-strengthen-economic-ties#. Accessed January 3, 2025.

Proceedings of Machine Translation Summit XX Volume 1, pages 150–161, June 23-27, 2025

ing due to the limited availability of high-quality parallel datasets.

In this study, we address the problem of adapting multilingual NMT models to CA→ZH for more region-specific translation. More specifically, the contributions of our work are as follows:

- Human-crafting a Catalan-Chinese parallel dataset containing 1,022 sentences sourced from Catalan/Spanish Wikimedia, translated directly to Mandarin Chinese. This dataset captures cultural and linguistic nuances more specific to Catalonia and Spain than existing benchmark datasets, which are more English-centric.[3]

- Demonstrating the benefits of using preference data with more region-specific content in Contrastive Preference Optimization (CPO) to align the model with human preferences, especially for cultural-specific terms. This approach better enhances the model's ability to handle both region-specific content and English-centric data. Notably, with only 1,022 sentences, we achieve good improvements in MT localization.

2 Related work

2.1 Research on Catalan-Chinese machine translation

Research on CA-ZH MT remains limited, with most previous efforts focusing on creating and mining parallel corpora for this low-resource language pair. Early work by Costa-Jussa et al. (2019) first addressed the lack of resources by creating a pseudo-parallel corpus via pivot translation, with Spanish as the intermediary language. Later, Zhou (2022) created human-selected CA-ZH parallel corpora by mining and validating bitexts from Wikipedia. Their efforts resulted in two datasets: CA-ZH 1.05 (110k sentence pairs) and CA-ZH 1.10 (65k higher-quality pairs). Using these datasets, Liu (2022) fine-tuned the M2M-100-418M multilingual model (Fan et al., 2021). Their full fine-tuning improved translation performance for both CA→ZH and ZH→CA directions, achieving BLEU score gains of +0.3–0.5 with the larger CA-ZH 1.05 corpus and +0.1–0.2 with the smaller, higher-quality CA-ZH 1.10 corpus. More recently, Chen et al. (2024) combined pivot translation (using Spanish) with multilingual training to leverage both synthetic and authentic data. Using the FLORES-200 benchmark (Costa-jussà et al., 2022), their findings showed that fine-tuning M2M-100-418M on the authentic CA-ZH dataset from Zhou (2022) only marginally improved the spBLEU score from 22.0 to 22.4. However, combining pseudo-parallel CA-ZH and Spanish-Chinese (ES-ZH) data alongside authentic CA-ZH and ES-ZH data yielded a significant improvement, increasing the spBLEU score to 26.7.

In this study, we take a different approach by creating a much smaller dataset of authentic CA-ZH data, consisting of 1,022 sentence pairs. Despite the dataset's small size, we demonstrate meaningful improvements in translation performance.

2.2 Contrastive Preference Optimization

Reinforcement Learning from Human Feedback (RLHF) has proven effective in aligning large language models (LLMs) with human preferences (Christiano et al., 2017; Ouyang et al., 2022). However, RLHF relies on a complex training pipeline, requiring first the training of a reward model based on human preference data. To simplify the training, recent work has proposed contrastive preference learning methods, such as Direct Preference Optimization (DPO) (Rafailov et al., 2024), which tune models directly on human preference data without explicitly training a reward model. The primary objective of these methods is to increase the likelihood gap between preferred and dispreferred responses.

Building on DPO, Contrastive Preference Optimization (CPO) was originally developed for machine translation tasks. CPO trains models to consistently favor preferred translations and avoid generating adequate but not perfect outputs. It has demonstrated significant improvements in translation quality. For example, in Spanish-to-Aranese translation tasks using only the FLORES-200 *dev* split, CPO outperformed both supervised fine-tuning and 5-shot fine-tuning, achieving a 1.9 BLEU score improvement with a Qwen2-0.5B-based (Yang et al., 2024) distillation model evaluated on the FLORES-200 *devtest* split (Hu et al., 2024).

In this study, we apply CPO to CA→ZH translation using a preference dataset that captures cultural and linguistic nuances more specific to Catalonia and Spain, in contrast to the more common English-pivot approach (using FLORES-200). We then compare the results to assess the impact of this lo-

[3]The dataset is available upon request from the authors.

calization. In the following section, we describe the construction of the preference datasets.

3 Dataset Construction

CPO requires a preference dataset, consisting of a "prompt", a "chosen" completion, and a "rejected" completion. The objective is to train the model to prefer the "chosen" response over the "rejected" response.

This section describes the construction of our preference datasets. Specifically, we create two datasets to assess the effects of using different types of data in CPO:

- `CPO FLORES DEV`: Based on the *dev* split of the FLORES-200 dataset, which is an English-pivot multilingual dataset including Catalan and Chinese.
- `CPO CA-ZH`: Built by sourcing sentences from Catalan and Spanish Wikimedia resources, and subsequently directly translated to Chinese.

Below, we describe the data sourcing process for `CPO CA-ZH`, the translation methodology, and a more detailed composition of the two preference datasets.

3.1 Sourcing sentences

Original Source. Following the methodology of FLORES-200 , all source sentences were extracted from Wikimedia resources, which are publicly available under permissive licensing. To ensure that the selected data did not overlap with parallel datasets already included in the models, we verified that none of the chosen Wikimedia pages had corresponding versions in Chinese.

The dataset was divided into three (roughly) equal parts to ensure diversity and coverage across different domains. Approximately one-third of the sentences were collected from Catalan *Wikinews*[4], a collection of news articles, with content selected from ten distinct topics. These topics, chosen to maintain balance and variety, include science and technology, culture and leisure, law, economy, sports, environment, obituaries, politics, health, and incidents. The second portion of the dataset was drawn from Catalan *Wikipedia*, a general-purpose encyclopedia containing a wide range of topics. The final third was sourced from *Wikivoyage*, a travel guide featuring articles on travel tips, cuisine, and destinations worldwide. Since Catalan *Wikivoyage* is still under development and, as of January 3, 2025, contains only 31 articles, this portion was instead sourced from Spanish *Wikivoyage*, which is significantly more developed and includes 3,347 articles.

Sentence Selection. Sentences were selected using a systematic approach to ensure diversity. Articles were selected from each source domain by randomly generating URLs using the *requests* library.[5] Following the methodology of FLORES-200 , between 3 and 5 contiguous sentences were extracted from each selected article, avoiding very short or malformed sentences. For Catalan *Wikinews* and Catalan *Wikipedia*, sentences were chosen equally from the beginning, middle, and end of each article to capture varied contexts. For Spanish *Wikivoyage*, selected sentences represented different topics, such as "drinking and nightlife", "climate", "shopping", and "flora and fauna" (see Appendix B for detailed dataset statistics).

Each selected sentence was annotated with metadata, including the article ID, sentence ID, URL and topic. On average, 3.5 contiguous sentences were extracted per article, with URLs included to allow access to the full document, which can be useful for document-level translation.

3.2 Translation

We used GPT-4 (OpenAI et al., 2024), which has demonstrated performance comparable to junior translators (Yan et al., 2024), to translate Catalan sentences into Spanish and Chinese. For sentences sourced from Spanish *Wikivoyage*, GPT-4 was used to translate them into Catalan and Chinese (see Appendix A for the specific GPT-4 prompt). Given the linguistic similarities between Spanish and Catalan, high-quality translations are assumed for this pair. For the Chinese translations, a native Chinese-speaking translator conducted post-editing and revisions of the machine-translated sentences to ensure naturalness and accuracy.

3.3 Two preference datasets

The `CPO CA-ZH` dataset consists of 1,022 triplets. Each Catalan sentence sourced from Wikimedia

[4] `https://ca.wikinews.org/wiki/Portada`.

[5] For example, a GET request to `https://es.wikivoyage.org/wiki/Especial:Aleatoria` redirects to a random article on Spanish *Wikivoyage*.

served as the *prompt*. Machine-translated Chinese sentences from GPT-4 were used as the *rejected* translations, while human-revised translations were labeled as the *chosen* sentences. Although GPT-4 translations are of relatively high quality, the goal of CPO is to train the model to recognize and prefer human-revised translations, thereby aligning more closely with human preferences.

In `CPO FLORES DEV`, there are in total 997 triplets. Catalan sentences from the FLORES-200 *dev* split served as the *prompt*. GPT-4 was used to translate the original English sentences into Chinese, producing the *rejected* translations. The original Chinese translations from the FLORES-200 *dev* split, which were also translated from English sentences, served as the *chosen* sentences.

In summary, `CPO CA-ZH` features direct Catalan-to-Chinese translations, while `CPO FLORES DEV` relies on an English pivot for generating Chinese translations. Both datasets use GPT-4 generated translations as *rejected* outputs and human-revised or human-produced translations as *chosen* outputs.

4 Entities in the CPO CA-ZH dataset

To analyze the key entities discussed in our `CPO CA-ZH` dataset, we used *spaCy* (version 3.8.3) to extract proper noun phrases and their corresponding frequencies from the Catalan sentences. These entities were then compared with those in the FLORES-200 and NTREX (Federmann et al., 2022) to assess how the topics in our dataset differ from those in existing datasets.

Overall, the `CPO CA-ZH` dataset is more focused on geographically specific topics, with frequent references to entities such as *Barcelona*, *Espanya* (Spain), and *Catalunya* (Catalonia). These entities are either absent or significantly less prominent in the other datasets. In contrast, the *dev* and *devtest* splits of the FLORES-200 prominently feature *Estats Units* (United States) as the most frequent entity, and the NTREX also tends to focus more often on entities like *Trump* and *USA*. For a complete comparison, see Table 1 and the frequency of each phrase in Appendix C.

This analysis suggests that our `CPO CA-ZH` dataset is more localized and culturally specific, emphasizing topics relevant to the region, whereas the FLORES-200 and NTREX are more focused on the United States and globally oriented topics.

5 Experiments

We applied CPO to the M2M-100-1.2B model using each of the two preference datasets introduced in Section 3.3. To assess the models after CPO training, we evaluated their translation performance on the FLORES-200 *devtest* split, which primarily focuses on topics relevant to the United States and global contexts. In addition, we conducted A/B testing on translations of 100 sentences containing localized terms specific to Catalonia. This allowed us to evaluate and compare the models' capabilities in handling more region-specific translations.

5.1 Training setup

We used the `facebook/m2m100_1.2B` (Fan et al., 2021)[6], a seq-to-seq model trained for multilingual translation, as the base model. It covers 100 languages, including Catalan and Mandarin Chinese.

Fine-tuning was performed using the Hugging Face's CPOTrainer class[7] which is compatible with the M2M-100 encoder-decoder architecture. We adhere to the default β value of 0.1 as suggested by Rafailov et al. (2024). The fine-tuning process involved a total batch size of 5, training for 6 epochs. The learning rate started at 8e-6 and linearly decayed throughout training. Checkpoints were saved every 50 steps and evaluated on the FLORES-200 *devtest* set. Training was conducted on a single NVIDIA H100 GPU with 64GB of RAM and completed in approximately 10 minutes.

5.2 Inference

Inference for all models was conducted using beam search with a beam size of 5, limiting the translation length to 200 tokens.

6 Results

6.1 Evaluation on FLORES devtest

This section reports the evaluation results of the models on the FLORES-200 *devtest* split for the Catalan→Chinese translation direction. The evaluation was conducted using `MT Lens` (Gilabert et al., 2024).[8] To provide a comprehensive assessment, we report a variety

[6]The smaller M2M-100-418M model often generates unknown tokens when translating from Catalan to Chinese (e.g., unknown tokens appear in 15% of translations on the FLORES-200 *devtest* split). To better support our evaluation of the translation of localized terms, we chose the larger 1.2B model, which provides greater vocabulary coverage for our experiments.

[7]https://huggingface.co/docs/trl/en/cpo_trainer

[8]

Our Dataset	FLORES-200 Dev	FLORES-200 Devtest	NTREX
Barcelona	Estats Units (United States)	Estats Units (United States)	Trump
Estats Units (United States)	Terra (Earth)	Terra (Earth)	EUA (USA)
Europa (Europe)	Xina (China)	Austràlia (Australia)	Regne Unit (United Kingdom)
Universitat (University)	EUA (USA)	Alemanya (Germany)	Xina (China)
Espanya (Spain)	Europa (Europe)	França (France)	Kavanaugh
Catalunya (Catalonia)	Àfrica (Africa)	Japó (Japan)	Corea (Korea)
Xina (China)	Sol (Sun)	Europa (Europe)	Palu
França (France)	Itàlia (Italy)	Hong Kong	Nord (North)
Madrid	Alemanya (Germany)	Taiwan	UE (EU)
Alemanya (Germany)	Turquia (Turkey)	Suècia (Sweden)	Washington

Table 1: Top 10 most frequent proper noun phrases across datasets

of metrics: BLEU (version 2.3.1), BLEURT (`lucadiliello/BLEURT-20-D12`), COMET (`Unbabel/wmt22-comet-da`), COMET-Kiwi (`Unbabel/wmt22-cometkiwi-da`), MetricX (`google/metricx-23-xl-v2p0`), MetricX-QE (`google/metricx-23-qe-xl-v2p0`) and statistical significance testing using paired bootstrap resampling (Koehn, 2004).

As shown in Table 2, BLEU scores (Papineni et al., 2002) indicate that `+CPO FLORES DEV` achieved a significant improvement in n-gram overlap between model translations and the reference, while the improvement with `+CPO CA-ZH` was not statistically significant. This result was expected, given the similarities between FLORES-200 *dev* (used for training) and FLORES-200 *devtest*.

In contrast, neural-based metrics such as BLEURT (Sellam et al., 2020), COMET (Rei et al., 2020), and MetricX (Juraska et al., 2023), as well as neural-based reference-free metrics like COMET-Kiwi (Rei et al., 2022) and MetricX-QE, suggest that `+CPO CA-ZH` led to greater improvements in translation quality. This indicates that `+CPO CA-ZH` improves aspects of translation quality such as semantic accuracy and fluency, without necessarily relying on the same n-gram phrases as the reference translation.

6.2 Evaluation on more culture-specific entities and data

In addition to the FLORES-200 *devtest* set, we assessed the models on sentences that contain Catalan- and Spanish-specific topics and culturally significant entities. We randomly selected 100 sentences from the Catalan Entity Identification and Linking dataset (Gonzalez-Agirre et al., 2024)[9] and ensured that most selected sentences contained regionally or culturally specific entities.

We used the two fine-tuned models to generate Chinese translations of these sentences. The translations were then assessed through A/B testing by two annotators: a linguist (the author) fluent in both Catalan and Chinese, and a professional Catalan-Chinese translator with eight years of experience. The annotators evaluated which translation more accurately conveyed the original meaning and sounded more natural. To measure the consistency between the annotators' preferences, we calculated the inter-annotator agreement using Cohen's kappa statistic with the *sklearn* library (version 1.5.2). The kappa score was 0.68, indicating substantial agreement according to the guidelines by Landis and Koch (1977).

Translations produced by `+CPO CA-ZH` were preferred more often (Annotator 1: 59% of the time; Annotator 2: 68%) compared to `+CPO FLORES DEV`. Among the 85 items where both annotators agreed, 56 (66%) favored `+CPO CA-ZH`. These results indicate a general preference for the translations from `+CPO CA-ZH`.

Furthermore, through manual examination, `+CPO CA-ZH` produced more accurate translations for region-specific terms and exhibited better transliteration capabilities from Catalan to Chinese. Examples of these translations are shown in Table 3, with the complete translated sentences available in the Appendix D. Even though these terms have never appeared in our preference dataset, aligning the model with localized data improved its ability to accurately translate and transliterate region-specific terminology. This highlights the effectiveness of incorporating culturally and regionally relevant data into the training process for practical use.

[9]This datset comprises sentences from tweets, news articles, reports, forums, encyclopedias, parliamentary proceedings, and reviews, and was originally designed for Named Entity Recognition.

Models	BLEU ↑	BLEURT ↑	COMET ↑	COMET-Kiwi ↑	MetricX ↓	MetricX-QE ↓
M2M100 1.2B	28.23	0.65	0.82	0.77	3.12	2.71
+ `CPO CA-ZH`	29.15	0.68	0.84 *	0.79 * †	2.51 * †	1.91 * †
+ `CPO FLORES DEV`	29.58 * †	0.67	0.84 *	0.77	2.60 *	2.15 *

* Significant improvement over the baseline M2M100 1.2B ($p < 0.05$).
† Significant difference between the two CPO-tuned models ($p < 0.05$).
Note: Significance testing was not performed for BLEURT as it is currently unsupported by `MT Lens`.

Table 2: The results in CA→ZH for FLORES-200 devtest set.

Catalan phrase in sentences	Explanation	+ `CPO FLORES DEV`	+ `CPO CA-ZH`
Bàsquet Girona	professional basketball club based in Girona	吉罗纳篮球队(Girona Basketball Team)	吉罗纳篮球俱乐部(Girona Basketball Club)
autònoms	self-employed workers or freelancers	自治人(Autonomous People)	自主经营者(Self-Employed)
Sant Feliu de Llobregat	municipality in the province of Barcelona	罗布拉格(Robrag)	圣费利乌·德·卢布雷加特(Sant Feliu de Llobregat)
Blanes	municipality in Catalonia	布莱斯(bù lái sī)	布拉内斯(bù lā nèi sī)
Corredor Mediterrani	Mediterranean Corridor, a major rail transport network in Europe	地中海跑道(Mediterranean Track)	地中海走廊(Mediterranean Corridor)
merder dels okupes	the mess caused by squatters; colloquial	混乱(Chaos)	占领活动的混乱(Chaos of Squatter Activities)

Table 3: Examples of Chinese translation of Catalan and Spanish region-specific terms, with English translations or *pinyin* provided in parentheses.

7 Conclusion

Many existing machine translation benchmarks, such as FLORES-200, rely on English as a pivot language for non-English language pairs. This approach can overlook the linguistic and cultural specificity of direct translations, particularly for language pairs like Catalan-Chinese (CA-ZH), where structural differences, idiomatic expressions, and cultural references may not have direct equivalents in English. To address this gap, we present a CA-ZH parallel dataset containing 1,022 sentences sourced from Catalan and Spanish Wikimedia and directly translated into Mandarin Chinese. Unlike most existing benchmarks, our dataset prioritizes linguistic and cultural authenticity by capturing regional nuances specific to Catalonia and Spain. This localization ensures that translations reflect real-world usage rather than being filtered through a more globalized or English-centric lens. By comparing our dataset to the FLORES-200 *dev* set, we demonstrate the benefits of aligning machine translation (MT) systems with culturally and regionally grounded data. This direct translation approach outperforms English-pivoted methods, which often introduce biases from the English-speaking world. Additionally, our dataset enables more accurate pronunciation mapping and transliteration between Catalan and Chinese, further improving transliteration quality for practical applications. Our work highlights the importance of developing non-English-centric datasets to better serve low-resource language pairs. We hope that the release of this dataset will encourage further research into localized, culturally rich resources and improve MT systems for real-world use.

Limitations

One limitation of our dataset is its relatively small size. While we aimed to create a high-quality dataset, the process of finding linguists and professional translators who are fluent in both Chinese and Catalan, as well as knowledgeable about Catalan culture, is costly. However, this constraint also ensures that the dataset (1,022 sentences) allows for meaningful comparisons with the FLORES dev set (997 sentences), maintaining fairness in evaluation.

That said, the limited number of sentences, combined with the fact that we did not explicitly ensure that every randomly selected document discusses Catalan or Spanish culture during the sentence sourcing process, means that the dataset could have been richer in regionally and culturally specific topics. Future expansions could address this by incorporating more diverse sources that better reflect the cultural and linguistic nuances of Catalan-speaking communities.

Ethical statements

The annotators were fairly compensated at a rate of approximately 20 euros per hour, ensuring ethical payment for their work. Sentences in our dataset were sourced from Wikimedia under a public license, adhering to open data principles and respecting intellectual property rights.

Carbon impact Statement

This work considers the environmental impact of computational resources used in model training. Each CPO training runs in 10min on 1 NVIDIA H100 GPU, and draws 201.47 Wh. Based in Spain, this has a carbon footprint of 34.46 g CO_2e, which is equivalent to 3.76e-02 tree-months, (calculated using green-algorithms.org v2.2 (Lannelongue et al., 2021)). Compared to large-scale deep learning methods, which can emit several metric tons of CO_2e, our approach remains computationally efficient and environmentally sustainable. In fact, the emissions per run are comparable to just a few Google searches, highlighting the low-carbon footprint of this training process while maintaining high model performance.

Acknowledgements

This work has been promoted and financed by the Generalitat de Catalunya through the Aina Project.

This work has been supported by the Spanish project PID2021-123988OB-C33 funded by MCIN/AEI/10.13039/501100011033/FEDER, UE. This work is partially supported by DeepR3 (TED2021-130295B-C32) funded by MCIN/AEI/10.13039/501100011033 and European Union NextGeneration EU/PRTR.

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA.

References

Yongjian Chen, Antonio Toral, Zhijian Li, and Mireia Farrús. 2024. Improving NMT from a low-resource source language: A use case from Catalan to Chinese via Spanish. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 229–245, Sheffield, UK. European Association for Machine Translation (EAMT).

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, Miami, Florida, USA. Association for Computational Linguistics.

Marta R Costa-Jussa, Noé Casas, Carlos Escolano, and José AR Fonollosa. 2019. Chinese-Catalan: A neural machine translation approach based on pivoting and attention mechanisms. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(4):1–8.

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*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Javier García Gilabert, Carlos Escolano, Audrey Mash, Xixian Liao, and Maite Melero. 2024. Mt-lens: An all-in-one toolkit for better machine translation evaluation. *arXiv preprint arXiv:2412.11615*.

Aitor Gonzalez-Agirre, Montserrat Marimon, Carlos Rodriguez-Penagos, Javier Aula-Blasco, Irene Baucells, Carme Armentano-Oller, Jorge Palomar-Giner, Baybars Kulebi, and Marta Villegas. 2024. Building a data infrastructure for a mid-resource language: The case of Catalan. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2556–2566, Torino, Italia. ELRA and ICCL.

Tianxiang Hu, Haoxiang Sun, Ruize Gao, Jialong Tang, Pei Zhang, Baosong Yang, and Rui Wang. 2024. Sjtu system description for the wmt24 low-resource languages of spain task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 943–948.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. 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. *Advances in Neural Information Processing Systems*, 36.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms: quantifying the carbon footprint of computation. *Advanced science*, 8(12):2100707.

Zixuan Liu. 2022. Improving Chinese-Catalan machine translation with Wikipedia parallel corpus. Master's thesis, Universitat Pompeu Fabra, Barcelona.

Palanichamy Naveen and Pavel Trojovskỳ. 2024. Overview and challenges of machine translation for contextually appropriate translations. *Iscience*, 27(10).

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

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.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2827–2835, Online. Association for Computational Linguistics.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *Preprint*, arXiv:2407.03658.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Chenyue Zhou. 2022. Building a catalan-chinese parallel corpus from wikipedia for use in machine translation. Master's thesis, Universitat Pompeu Fabra, Barcelona.

A Prompt for translations

Adhering to the prompt format for translation as utilized by Xu et al. (2024) for GPT models, we use the same prompt for GPT-4 in our study, as shown in Figure 1.

GPT-4 Prompt

System:
You are a helpful translator and only output the result.

User:
Translate this from `<source language>` to `<target language>`, `<source language>`:
`<source sentence>`
`<target language>`:

Figure 1: The prompt employed for GPT-4 to perform translations.

B Statistics of the CPO CA-ZH dataset

The `CPO CA-ZH` dataset includes sentences collected from three primary sources: Catalan Wikinews, Catalan Wikipedia, and Spanish Wikivoyage. Approximately one-third of the sentences come from each source:

Source	Wikinews	Wikipedia	Wikivoyage
n. sent	328	341	353

Table 4: Number of sentences collected from different sources.

For Catalan Wikinews and Catalan Wikipedia, sentences were chosen roughly equally from the beginning, middle, and end of each article to capture varied contexts:

Sentence Position	**Count**
Middle	231
End	222
Beginning	216

Table 5: Distribution of sentence positions for Catalan Wikinews and Catalan Wikipedia.

The statistics for the Wikinews portion of the dataset are shown in Table 6. The topics, along with their English translations, are as follows: Ciència i Tecnologia (Science and technology), Cultura i esplai (Culture and leisure), Dret (Law), Economia (Economy), Esports (Sports), Medi ambient (Environment), Necrologia (Obituaries), Política (Politics), Salut (Health), Successos (Incidents).

C Top 10 most frequent proper noun phrases across datasets

Table 7 shows the top 10 most frequent proper noun phrases and their frequency in our `CPO CA-ZH` dataset, FLORES-200 *dev* split, FLORES-200 *devtest* split, and the NTREX dataset.

D Examples of translation of region-specific terms

In Section 6.2, we have only shown translation of the phrases. Table 8 below shows the translation of the full sentences where these phrases come from.

Topic	**# Articles**	**# Sentences**
Ciència i tecnologia	9	29
Cultura i esplai	9	28
Dret	10	36
Economia	10	35
Esports	10	35
Medi ambient	10	35
Necrologia	9	28
Política	10	37
Salut	10	34
Successos	10	31
Total	97	328

Table 6: Statistics of the Wikinews portion of the dataset.

Our Dataset	**Freq.**	**FLORES-200 Dev**	**Freq.**	**FLORES-200 Devtest**	**Freq.**	**NTREX**	**Freq.**
Barcelona	18	Estats Units	14	Estats Units	17	Trump	64
Estats Units	13	Terra	10	Terra	11	EUA	61
Europa	12	Xina	9	Austràlia	9	Regne Unit	42
Universitat	9	EUA	8	Alemanya	9	Xina	40
Espanya	9	Europa	7	França	8	Kavanaugh	36
Catalunya	8	Àfrica	7	Japó	7	Corea	35
Xina	7	Sol	7	Europa	7	Palu	29
França	7	Itàlia	6	Hong Kong	6	Nord	27
Madrid	6	Alemanya	6	Taiwan	6	UE	26
Alemanya	6	Turquia	6	Suècia	6	Washington	22

Table 7: Top 10 most frequent proper noun phrases and their frequency across datasets

Catalan sentence (keyword)	+ `CPO FLORES DEV`	+ `CPO CA-ZH`
Són també representatius el **Bàsquet Girona** (fundat el 2014 per Marc Gasol), actualment a l'ACB, La Salle Girona, el Vedruna Girona, el Sant Narcís, el Club Bàsquet Onyar o el CESET	吉罗纳篮球队(由马克·加索尔(Marc Gasol) 于2014 年创立)也代表了这支球队,目前属于ACB 的La Salle Girona、维德鲁纳·吉罗纳(Vedruna Girona)、圣纳基斯(San Narcís)、奥尼亚尔篮球俱乐部(Club Bàsquet Onyar)和CESET。	吉罗纳篮球俱乐部(由 马克·加索尔于2014年创办)也代表着它,目前它属于ACB、拉萨尔·吉罗纳俱乐部、维德鲁纳·吉罗纳俱乐部、圣纳西斯俱乐部、奥尼亚尔篮球俱乐部以及塞塞斯特俱乐部。
Més de 6.500 **autònoms** sol·liciten l'ajut del Govern i Treball estudia ampliar-lo i fer-lo complementari amb l'estatal.	超过6500 名自治人向政府申请了援助,劳工正在考虑扩展援助,并将该援助与州政府互补。	超 过6500名自主经营者申请了政府的援助,劳工部门正在考虑扩大援助计划并将其与州政府的援助互补。
Les obres per soterrar les vies a **Sant Feliu de Llobregat** començaran aquest 2019.	罗布拉格 (Sant Feliu de Llobregat) 的道路埋葬工作将于2019 年开始。	圣费利乌·德·卢布雷加特(Sant Feliu de Llobregat)的道路埋葬工程将于2019年开始。
Per al passeig Marítim de la **ciutat de Blanes** se li va encarregar el monument homenatge a l'escriptor Joaquim Ruyra (1958) amb ocasió del centenari del seu naixement, ho va representar com el frare caputxí «sant Sadurní de Croïlles» protagonista de la seva obra Les coses benignes, la intenció de l'escultor va ser expressar la gran unió que existeix entre la literatura de Ruyra i l'esperit franciscà.	对 于布莱斯(Blanes)市的海上漫游活动,他在1958 年为作者约阿基姆·鲁伊拉(Joaquim Ruyra) 诞生100 周年纪念日提供纪念碑,他将其描绘成"克鲁伊莱斯圣萨德尼"(Saint Sadurní de Croïlles),其著作《善良的事物》(The Good Things)的主角。雕塑家意图是表达鲁伊拉(Ruyra)文学与法兰西斯(Franciscan)精神之间存在的巨大联系。	对于布拉内斯市的海上散步,在1958年出生百周年之际,他委托了作家约阿基姆·鲁伊拉(Joaquim Ruyra)的纪念碑。他将其描绘成"圣萨杜尼·德·克鲁伊莱斯"(Saint Sadurní de Croïlles)作为他的作品主角。雕塑家的目的是表达鲁伊拉文学与弗朗西斯克精神之间深远的联系。
De la Serna assenyala que el ritme inversor al **Corredor Mediterrani** no depèn d'un problema econòmic sinó administratiu.	德尔·塞尔纳(De la Serna)指出,地中海跑道的变速不是因为经济问题,而是因为行政问题。	德 拉·塞 尔 纳 指出,地中海走廊的投资步伐并非因为经济问题,而是因为行政问题。
Aquest **merder dels okupes** a Barcelona i en extensió a tota Catalunya va ser propiciat per la Colau, que oblidem molt ràpid les coses.	巴塞罗那的这一混乱,以及整个加泰罗尼亚的混乱,是由"科洛" (La Colau) 造成的,我们很快就忘了这些事情。	这场在巴塞罗那以及整个加泰罗尼亚的占领活动的混乱,是 由劳拉·科劳(La Colau)推动的。我们很快就忘记了这些事情。

Table 8: Examples of translation of Catalan and Spanish region-specific terms in sentences.