SiniticMTError: A Machine Translation Dataset with Error Annotations for Sinitic Languages

Hannah Liu[1], Junghyun Min[2], En-Shiun Annie Lee[1,3], Ethan Yue Heng Cheung[1], Shou-Yi Hung[1], Elsie Chan[1], Shiyao Qian[1], Runtong Liang[1], Kimlan Huynh[1], Wing Yu Yip[1], York Hay Ng[1], TSZ Fung Yau[1], Ka Ieng Charlotte Lo[1], You-Wei Wu[4], Richard Tzong-Han Tsai[4]

[1]University of Toronto, [2]Georgetown University
[3]Ontario Tech University, [4]National Central University, Taiwan
hannahhere.liu@mail.utoronto.ca

Abstract

Despite major advances in machine translation (MT) in recent years, progress remains limited for many low-resource languages that lack large-scale training data and linguistic resources. In this paper, we introduce SiniticMTError, a novel fine-grained dataset that builds on existing parallel corpora to provide error span, error type, and error severity annotations in machine-translated examples from English to Mandarin, Cantonese, and Wu Chinese, along with a Mandarin-Hokkien component derived from a non-parallel source. Our dataset serves as a resource for the MT community to fine-tune models with error detection capabilities, supporting research on translation quality estimation, error-aware generation, and low-resource language evaluation. We also establish baseline results using language models to benchmark translation error detection performance. Specifically, we evaluate multiple open source and closed source LLMs using span-level and correlation-based MQM metrics, revealing their limited precision, underscoring the need for our dataset. Finally, we report our rigorous annotation process by native speakers, with analyses on pilot studies, iterative feedback, insights, and patterns in error type and severity.

Keywords: machine translation, error annotation, Sinitic language

1. Introduction

Machine translation (MT) systems have made significant advancements in recent years both through supervised systems (Luong et al., 2015; Lakew et al., 2018; Liu et al., 2020; Wang et al., 2022; Liu, 2022; Park et al., 2023) and through large language models (LLMs) (Zhu et al., 2024; Freitag et al., 2024). However, such systems often focus on higher-resource languages, and progress remains limited with low-resource languages (Ranathunga et al., 2023), where fine-tuned models suffer from poor performance (Lee et al., 2022; Shliazhko et al., 2024) and LLMs output noise (Iyer et al., 2024; Levine et al., 2025).

In addition to Mandarin, we focus on three major yet low-resource Sinitic variants, Cantonese[1], Wu Chinese[2], and Hokkien[3]. They remain underserved despite having more than 80, 83 and 47 million speakers respectively, across southern and eastern China and various diasporas across the world (Chappell, 2015; Eberhard et al., 2023). The limited progress is attributable to the dominance of Mandarin as *lingua franca* in these regions (Norman, 1988; Li, 2006), scarcity in publicly available parallel corpora (Xiang et al., 2024), the lack of standardization in writing systems (Pan et al., 1991; Tang et al., 2002; Kwan-hin and Bauer, 2002; Snow, 2008), and their status as primarily

Source: The weather is beautiful today.

MT: 今天 (*Today*) 我觉得 (*I think*) 天气很 (*the weather is*) 漂亮。 (*beautiful*)

Reference: 今天 (*Today*) 天气很 (*the weather is*) 好。 (*beautiful*)

Annotation Spans:

- **Text:** "我觉得"
  Error type: Addition
  Severity: Major
  Start: 2 **End:** 5
- **Text:** "漂亮"
  Error type: Mistranslation
  Severity: Minor
  Start: 8 **End:** 10

Figure 1: Sample Mandarin entry. `mt` looks fluent, but contains subtle semantic errors: an unwarranted subjective phrase (Addition) and a lexical mistranslation (Mistranslation). While 漂亮 *piao4liang* directly translates to *beautiful*, it usually describes people or objects and 好 *good* is more natural when used to describe the weather.

[1]Also known as Yue (Eberhard et al., 2023)
[2]Whose most well-known dialect is Shanghainese (Eberhard et al., 2023)
[3]Also known as Min Nan (Southern Min; Eberhard et al., 2023)

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8662–8684
11–16 May 2026.

vernacular (i.e. spoken rather than written) languages (Pan et al., 1991; Snow, 2004; Li, 2006).

In this paper, we add to the Sinitic MT literature by presenting SiniticMTError, a novel fine-grained suite of datasets that mainly build on FLORES+ (Goyal et al., 2022; NLLB Team et al., 2024; Yu et al., 2024) to provide erroneous machine translation examples and detailed span-level error annotations including error type and severity for Mandarin, Cantonese, Wu Chinese, and Hokkien (Figure 1).

We report dataset statistics across 2,009 Mandarin and 1,402 Cantonese sentences, 68 Wu Chinese and 154 Hokkien pilot annotations. Our results suggest distinctive error distributions across languages. The currently available annotated portions of the dataset can be found in our GitHub Repository.

2. Background

We discuss relevant annotation frameworks, resources for multilingual and Sinitic machine translation, and clarify the language terminology we use in this paper.

2.1. Literature Review

Annotation frameworks. Annotation frameworks like Multidimensional Quality Metrics (MQM; Burchardt, 2013) and Error Span Annotations (ESA; Chen et al., 2020; Kocmi et al., 2024, *inter alia*) represent important steps toward more equitable multilingual NLP by introducing standardized methods for evaluating translation quality. MQM and ESA facilitate identifying and localizing errors to improve translation pipelines (Chen et al., 2020; Zhang et al., 2025a), and explicit fine-tuning or prompting to improve the quality of generative model output (Kocmi and Federmann, 2023). SiniticMTError annotations and design choices, including annotation instructions, error type and data structure are based on MQM (Burchardt, 2013) and AfriCOMET, an adaptation of the MQM framework for African languages (Wang et al., 2024).

Multilingual resources. While MQM and ESA offer a promising foundation, existing datasets built on them have largely focused on high-resource languages such as English, French, and Mandarin Chinese (Freitag et al., 2021; Sellam et al., 2021). Efforts to adapt these tools to lower-resource languages include Singh et al. (2024); Wang et al. (2024); Li et al. (2025), which focus on low-resource Indic and African languages.

FLORES-101, FLORES-200, and FLORES+ (Goyal et al., 2022; NLLB Team et al., 2024) are initiatives that have introduced multilingual benchmarks covering over 100 languages, many of which are low-resource, to support equitable evaluation of MT systems. Building upon the developmental and test splits of FLORES-200, AfriCOMET (Wang et al., 2024) is an annotation effort aiming to bridge the gap between low-resource languages and MT systems by providing annotated datasets for underrepresented African languages. Similar work has emerged for Bambara (Dou and Neubig, 2022), Amharic and Tigrinya (Shapiro et al., 2023), and Spanish languages (Perez-Ortiz et al., 2024), offering parallel corpora and baseline models.

Sinitic resources. Despite related efforts, publicly available resources in Cantonese, Wu Chinese, and Hokkien remain scarce. Proprietary LLMs offer commercial service in Cantonese (OpenAI et al., 2024; Team et al., 2024) and several Cantonese–Mandarin or Cantonese–English parallel corpora exist in the form of subtitles (Wong and Zhang, 2017), dictionaries (Mair and DeFrancis, 2003), or government transcripts (Lee, 2011). However, resources are limited in scale or accessibility as surveyed by Xiang et al. (2024). As a result, previous work in Cantonese MT has relied on synthetic data augmentation (Liu, 2022; Hong et al., 2024). To the best of our knowledge, the recent addition to FLORES+ (Yu et al., 2024) represents the sole publicly available MT resource in Wu Chinese. Hokkien resources comprise of a code-mixing corpus (Lu et al., 2022), a speech-to-speech translation dataset and model (Chen et al., 2023), and orthography standardization for machine translation (Lu et al., 2024a). Beyond Cantonese, Wu, and Hokkien, resources in Hakka (Hung and Huang, 2022; Lai et al., 2024) have been compiled.

2.2. Language Terminology

We note that we use the terms Cantonese, Wu (Chinese), and Hokkien to describe the Sinitic languages our dataset covers. Some may describe the dataset to cover Yue, Shanghainese, or Southern Min; we acknowledge that names and boundaries between languages in China are often fuzzy (Chappell, 2015), with varying conventions across fields. We clarify that this work discusses annotations in the language rather than dialects specific to the city or a region, although the annotations may reflect the prestige dialect that is spoken in Shanghai, Pearl Delta Region, and Taiwan respectively (Chappell, 2015; Eberhard et al., 2023).

Yue or Cantonese? The distinction between Yue and Cantonese can be unclear. While Eberhard et al. (2023) describes Cantonese as an al-

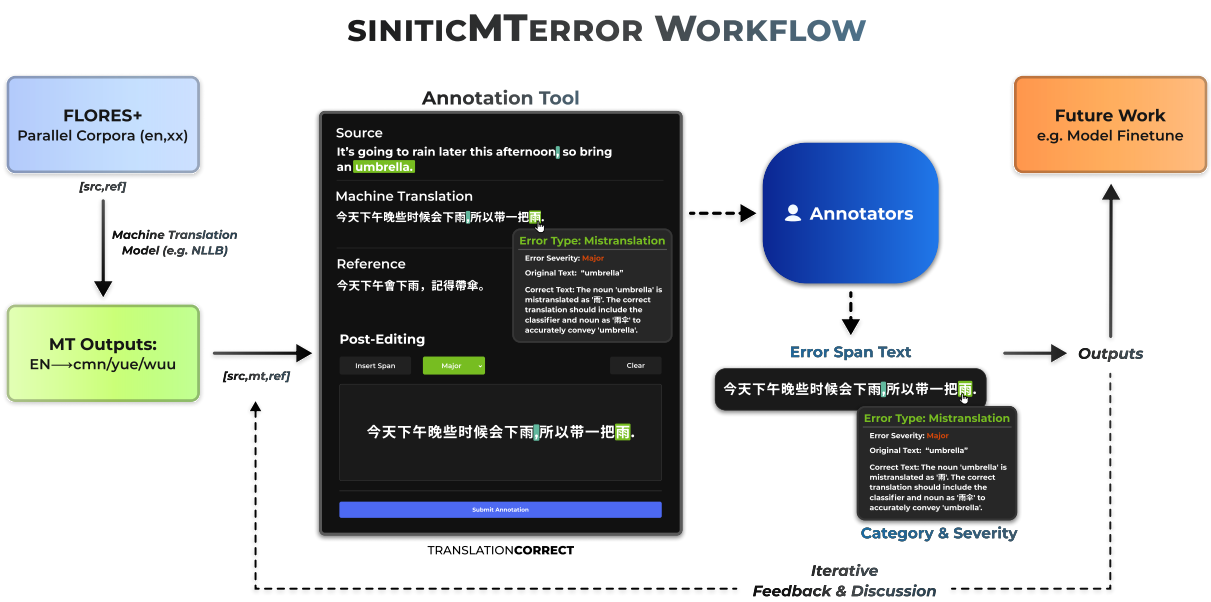

Figure 2: Overview of our annotation pipeline. We input English sentences from FLORES+ to generate mt outputs (e.g., from NLLB) into Sinitic languages (Mandarin, Cantonese, Wu).

ternate name for Yue, some use it to describe the Guangzhou variant of Yue (e.g. Matthews and Yip, 2011). In this work, we follow Ethnologue (Eberhard et al., 2023) and previous work in Cantonese MT (Liu, 2022; Hong et al., 2024) to refer to the entire Yue language as Cantonese. We note that NLLB Team et al. (2024) use "Yue Chinese" to describe what we describe as Cantonese in this paper.

Wu or Shanghainese? The distinction between Wu and Shanghainese is much clearer–Shanghainese is a dialect of Wu (Eberhard et al., 2023). In MT, the sole work in Wu Chinese (Yu et al., 2024) does not use the term Shanghainese. We follow such prior work to discuss our annotations in Wu Chinese, rather than Shanghainese.

Hokkien or Min Nan? Hokkien is a variant of Min Nan or Southern Min, spoken primarily in Fujian and Taiwan, as well as in Chinese diasporas in southeast Asia (Eberhard et al., 2023). Previous NLP and language resource work are described as work on Taiwanese Hokkien (Lu et al., 2022; Chen et al., 2023; Lu et al., 2024a; Chu et al., 2025). We note that one exception (Zhang et al., 2025b) uses the term Hokkien. Following the annotators' preference, we use the term Hokkien.

3. Dataset annotation

We follow Han orthographic standardization of Yu et al. (2024) for Wu Chinese, where phonetic transliterations follow the pronunciation of the Chongming dialect. We also follow NLLB Team et al. (2024)'s choice of traditional Han orthography for Cantonese.

Each example in the dataset consists of a triplet of sentences: a source sentence (src), a machine-translated sentence (mt), and a reference translation (ref). For en-zh, en-yue, and en-wu, the src and ref pairs are drawn from FLORES+, which is an extension of the FLORES-200 dataset (NLLB Team et al., 2024). For Hokkien, the pairs are drawn from The Dictionary from the Ministry of Education[4]. The mt sentences are generated using the 600M NLLB-200 model (Team et al., 2022) for Mandarin and Cantonese, Qwen2.5 Max (Team, 2024) for Wu Chinese, and Taigi-LLaMA-2-Chat-7B (Lu et al., 2024b) for Hokkien.

The constructed dataset is then provided to annotators for error span annotation. For each FLORES+ language pair (English-Mandarin, English-Cantonese, English-Wu Chinese), we recruit at least three bilingual annotators. The Hokkien component of SINITICMTERROR represents a preliminary extension distinct from the other language pairs, and translates from Mandarin to Hokkien.

The annotators are native speakers of their respective target languages and highly proficient in English. They are familiar with both the linguistic conventions of their target language and the goals of the annotation task.

We use the TRANSLATIONCORRECT tool (Wasti et al., 2025) as our annotation interface.

[4] https://sutian.moe.edu.tw/zh-hant/

3.1. `mt` Generation

In selecting the model to generate `mt` sentences, we consider several factors. The ideal `mt` model would be reproducible, accessible, error-prone, but also reliably able to generate plausible sentences in the target language. We select the 600M NLLB model (Team et al., 2022) for Cantonese and Mandarin, Qwen 2.5 Max (Team, 2024) for Wu, and Taigi-LLaMA-2-Chat-7B (Lu et al., 2024a) for Hokkien. We describe our `mt` model selection and generation process below.

3.1.1. Cantonese

For accessibility and reproducibility, we consider open-source models with less than 7B parameters that also have Cantonese proficiency. To verify their reliability in Cantonese generation and error-prone generation, we manually review Cantonese translations of the first 10 English sentences from `FLORES+` (Yu et al., 2024) by each model as a preliminary sanity check.

We determine whether the output was in Cantonese and free of language confusion by checking for traditional Han orthography and Cantonese-specific characters. Then, we evaluate `mt` quality by comparing them to their respective `ref` sentences, using two metrics: SacreBLEU (Post, 2018) and ChrF++ (Popović, 2017).

Out of 600M NLLB-200 (Team et al., 2022), 1.5B Qwen 2.5 Instruct (Team, 2024), 1B Llama 3.2 Instruct (Grattafiori et al., 2024), 1.1B Bloomz (Muennighoff et al., 2023), 1.2B mT0 Large (Muennighoff et al., 2023), 7B Qwen Chat (Bai et al., 2023), 8B Llama 3.1 (Grattafiori et al., 2024), and 8B Aya Expanse (Dang et al., 2024), only NLLB-200, Aya Expanse, and Llama-8B were able to reliably output Cantonese. NLLB-200 and Aya Expanse are described as having been trained on Cantonese data; we were unable to determine whether Llama and Qwen's training data included Cantonese. Other models do not explicitly report Cantonese data in their training corpora.

The SacreBLEU (Post, 2018) and ChrF++ (Popović, 2017) scores of the three models evaluated were as shown in Table 1. We select NLLB-200 as our model to generate Cantonese `mt` sentence due to its lowest average SacreBLEU and ChrF++ scores, small size, and easy accessibility.

Model # Params	**NLLB-200** 600M	**Aya Expanse** 8B	**Llama** 8B
SacreBLEU	74.5	79.9	82.1
ChrF++	75.0	72.7	82.0

Table 1: Comparison of Translation Models Using SacreBLEU and ChrF++.

3.1.2. Mandarin

For consistency across languages, we used the same model for Mandarin as we did for Cantonese. As discussed in 3.1.1, 600M NLLB (Team et al., 2022) was selected based on a thorough analysis of model quality and parameter size. Using a single model allows us to maintain consistency in prompts and output formatting across languages. Moreover, NLLB shows decent performances in Mandarin, thus making it an ideal choice for producing the `mt` outputs.

3.1.3. Wu Chinese

For Wu Chinese, we chose Qwen 2.5 Max (Team, 2024) for producing `mt` sentences as it remains the only publicly accessible LLM with stable Wu Chinese ability. While DeepSeek is also able to reliably generate Wu output, we are unable to use DeepSeek due to institutional restrictions. We found that other models including Llama and NLLB-200, which were unable to produce usable Wu Chinese output even after prompt engineering.

3.1.4. Hokkien

To our knowledge, Hokkien is not reliably supported by any of the general multilingual generative models. Machine translation outputs were produced using Taigi-Llama-2-Chat-7B (Lu et al., 2024b), a 7-billion-parameter derivative of LLaMA-2 chat model additionally trained on Taiwanese Hokkien data. The model is not a general multilingual model; it was trained primarily on Hokkien, allowing it to reliably generate Hokkien `mt` sentences.

3.2. Annotation Guidelines

Annotators are instructed to examine the `mt` sentence with reference to both `src` and `ref`, and identify any translation errors by highlighting spans directly on the `mt` sentences. For each error span, annotators categorize the **severity** and **error type**, while recording the erroneous **span indices** in the `mt` sentence. The label spaces for **severity** and **type** are adopted from MQM guidelines (Burchardt, 2013) and the AfriCOMET framework (Wang et al., 2024), shown in the tables in Section 3.4.

Error types and severity. We adapted our error severity and category definitions based on MQM guidelines (Burchardt, 2013) and the AfriCOMET framework (Wang et al., 2024), with modifications informed by language-specific characteristics of Sinitic Languages. After multiple rounds of pilot annotation and qualitative analysis as described in

Error Category	Definition
Addition	The highlighted span in the translation corresponds to information that does not exist in the source text.
Omission	The highlighted span corresponds to content manually inserted by the annotator into the translation, representing information present in the source text but missing from the original MT output.
Mistranslation	The highlighted span in the translation does not have the exact same meaning as the corresponding span in the source segment.
Untranslated	The highlighted span in the translation is a copy of the corresponding span in the source segment, but should have been translated into the target language.
Grammar	The highlighted span corresponds to issues related to grammar or syntax in the translated text, excluding spelling and orthography.
Spelling	The highlighted span corresponds to spelling issues. Mistranslations of names (e.g., locations, people) are also categorized as spelling errors.
Typography	The highlighted span corresponds to issues related to punctuation or diacritics, except omission of punctuations.
Unintelligible	The exact nature of the error cannot be determined, indicating a major breakdown in fluency.
Register	Characteristic of text that uses a level of formality higher or lower than required by the specifications or general language conventions.

Table 2: Definitions of 9 error categories used for error annotations.

Severity Level	Definition
Major	The error introduced causes a significant change in the meaning of the translated sentence.
Minor	The error does not change the core meaning of the translated sentence, but introduces a slight issue affecting fluency or readability.

Table 3: Definitions of severity levels used for error annotations.

Section 3, we refined the labels to better capture common translation issues observed in our data. Our refinement results in 9 error type labels as outlined in Table 2, and 2 severity labels as outlined in Table 3. More detailed information on the guidelines and modifications can be found in Appendix A.

Granularity. When identifying errors, the annotators are asked to be as fine-grained as possible. For example, if a sentence contains two words that are each mistranslated, two separate mistranslation error spans should be recorded. If a single text segment contains multiple errors, the leftmost span with highest severity is to be recorded.

Additional information. One error type is omission, where content in the source sentence is missing from the translation. In such case, annotators are instructed to insert the missing information the post-editing box and highlight the inserted span with the appropriate type (omission), span, and severity labels.

3.3. Annotation Workflow

First, annotators were trained on using the annotation tool, then provided with detailed guidelines, including definitions of error categories and severity levels, along with examples. The training was then followed by a multi-stage pilot setup.

Mandarin pilot. Mandarin annotators had two rounds of pilot studies. In the first round, each annotator completed 50 examples, which were then reviewed by a language lead. The lead provided feedback and held group discussions to ensure consistency, before annotators completed a second round of 50 examples. Qualitative analysis showed clear improvement in annotation consistency.

Cantonese pilot. Cantonese annotators followed a more granular setup consisting of 4-rounds. The initial round included 50 examples and the others 10 examples each, completing 80 examples in total. Additional rounds of 10 examples were added to ensure sufficient agreement before proceeding.

Main annotation phase. During the main annotation phase, annotators worked in batches of 50 to 200 sentences. With each batch, we employed a similar iterative process where annotations were reviewed, recurring issues discussed, and relevant guidelines fine-tuned. During such iterative process, guidelines on annotation granularity and categorizing missing punctuation as omission were

established.

Quality assurance. After completing the first round of annotations (main annotator round), annotators proceed to the Quality Assurance (QA) stage. In this phase, each annotator will be assigned sentences previously annotated by another team member. Each participant reviews these annotations and discusses any discrepancies with the original annotator, should there be any, to improve consistency across annotators. We plan to perform inter-annotator agreement analysis in multiple stages.

Wu pilot. We select a total of 68 sentences from the Wu subset of FLORES+ (Yu et al., 2024). For Wu annotations, we follow a similar setup to that of Cantonese, with a granular setup consisting of 4 rounds. One cycle of annotations was conducted to finish the pilot annotations.

Hokkien pilot. A total of 154 sentences were collected for this pilot. We use a custom annotation interface[5]. We recruit two annotators, both certified Taiwanese Hokkien teachers and native speakers. The annotation uses a similar MQM-derived schema to Mandarin, Cantonese, and Wu annotation schemes, but with a slightly different set of categories. Nonetheless, our analysis of Hokkien pilot annotations use a mapping between Hokkien categories and those used for Mandarin, Cantonese, and Wu for comparability and consistency across language. We discuss original Hokkien annotation categories in greater detail in Appendix A.

This pilot differs substantially from the Mandarin, Cantonese, and Wu Chinese components in both methodology and scope. Nonetheless, the adoption of SiniticMTError framework in Hokkien annotations confirms that our framework and data schema can generalize to additional low-resource Sinitic varieties.

3.4. Dataset Statistics

We report the distribution of error types in each language on 2009 Mandarin sentences and 1402 Cantonese sentences. We also report error type distribution on our pilots–Wu with 68 sentences and Hokkien with 154 sentences. Tables 4, 5, 6, 7, 8 present the distribution of error categories and severity levels in our current annotation data. On average, each Mandarin sentence contains 1.93 annotation spans, while each Cantonese sentence contains 6.38 spans, showing a much higher span density than Mandarin.

[5]We did not use TranslationCorrect tool for Hokkien

Type	Count	Proportion	Freq / 1k
Mistranslation	2,306	61.2%	1,148
Omission	534	14.2%	266
Grammar	267	7.1%	133
Unintelligible	198	5.3%	99
Typography	133	3.5%	66
Untranslated	128	3.4%	64
Spelling	114	3.0%	57
Addition	86	2.3%	43
Total	**3,766**	**100%**	**1,875**

Table 4: Mandarin (2009 sentences)

Type	Count	Proportion	Freq / 1k
Mistranslation	2,992	33.8%	2,134
Typography	2,329	26.3%	1,661
Omission	1,914	21.6%	1,365
Addition	600	6.8%	428
Grammar	450	5.1%	321
Spelling	326	3.7%	233
Untranslated	242	2.7%	173
Unintelligible	2	0.0%	1
Total	**8,855**	**100%**	**6,316**

Table 5: Cantonese (1402 sentences)

In Mandarin annotations, mistranslation, omission, and grammar errors are the most frequent error types. For Cantonese, the most common error types are mistranslation, typography, and omission. The high frequency of omission and mistranslation error types in both Mandarin and Cantonese, compared to surface level errors like spelling, untranslated, typography, and grammar, reinforces that although autoregressive language models are capable of generating well-formed and plausible sentences, they often struggle with incomplete representations of semantic and syntactic structure, with high sensitivity to surface form (Berglund et al., 2024; Kitouni et al., 2024) and poor semantic generalization to rare constructions (Scivetti et al., 2025). Cantonese outputs also record a much higher number of error spans per sentence compared to Mandarin, which indicates that the model performs worse in Cantonese. This likely represents the limited availability of high-quality Cantonese resources (Xiang et al., 2024). Finally, an analysis of the number of major versus minor errors shows that many errors are minor semantic or felicity issues rather than complete misinterpretations, attesting to the powerful multilingual adaptability of transformer-based language models (e.g. Liu et al., 2020; Zhu et al., 2024). Overall, these results show both shared and language-specific challenges in Sinitic machine translation.

We emphasize that these observations reflect

Type	Count	Proportion	Freq / 1k
Mistranslation	46	37.4%	676
Register	42	34.1%	618
Omission	25	20.3%	368
Addition	4	3.3%	59
Spelling	3	2.4%	44
Grammar	2	1.6%	29
Typography	1	0.8%	15
Total	**123**	**100%**	**1,809**

Table 6: Wu Chinese (68 sentences)

Type	Count	Proportion	Freq / 1k
Mistranslation	107	30.5%	695
Register	38	10.8%	247
Omission	29	8.3%	188
Addition	26	7.4%	169
Grammar	14	4.0%	91
Untranslated	14	4.0%	91
Other	123	35.0%	799
Total	**351**	**100%**	**2,279**

Table 7: Hokkien (154 sentences)

trends within the specific MT outputs used to construct SiniticMTError. Since each translation direction is represented by a single model, the reported distributions and comparisons should be interpreted as dataset-level analyses rather than general claims about English–Chinese machine translation performance. Different models may exhibit different error profiles, and future work incorporating multiple systems would allow for more system-level generalization.

4. Annotation Insights

Our span-level annotations reveal insights with implications in both linguistics and natural language processing. Due to the small size of Wu and Hokkien pilots, we discuss insights from Mandarin and Cantonese annotations. We discuss challenges in MT systems that were also outlined in concurrent work.

Differences in error type distribution. Beyond lexical differences, Cantonese and Mandarin differ in several ways, e.g. word order and function word inventory (Zhang, 1998). The suite of aspect and feature markers and sentence-final particles differ, with Cantonese boasting a richer inventory that encodes more fine-grained nuance in tense, speaker stance or attitude (Yap and Chor, 2011; Lee, 2019). Cantonese also allows serial verb constructions (e.g. *go buy eat rice* as a sequence) to a greater extent than in Mandarin (Matthews, 2006). Such

Severity	Mandarin	Cantonese
Minor	2,127	7,346
Major	1,639	1,509

Table 8: Error severity counts in Mandarin and Cantonese `mt` outputs. Minor errors are more frequent in all languages, though major errors remain substantial.

differences may be reflected in Table 4 and Table 5, where the grammar error type is much more frequent per sentence in Cantonese than in Mandarin. The difference in functional word inventory is likely a major source of such error; we observe erroneous particle uses in Cantonese `mt` sentences, many of which are valid particles in Mandarin yet ungrammatical in Cantonese. Examples include erroneous use of 才 *cai2* in place of Cantonese particle 先 *sin1 only after*, Mandarin possessive 的 de in place of Cantonese 嘅 *ge3*; and Mandarin 在 *zai4* in place of Cantonese copula 喺 *hai2 be at*.

Translationese. Compared to English, Chinese languages typically have simpler sentence structure and clause segmentation (Morbiato, 2017), relies much more heavily on particles and pro-drop (Li and Thompson, 1979; Paul, 2014), and has different headedness principles (Levy and Manning, 2003). However, many machine translation outputs were constructed in an English clause structure (i.e. translationese; Riley et al., 2020), which results in unnatural or even ungrammatical phrasing. For example, subordinate clauses were often translated as long embedded segments, instead of being split into multiple short sentences, which is more natural in Chinese languages.

Lack of standardization. As discussed in Section 1, machine translation in Cantonese and Wu Chinese face unique difficulties as a primarily a spoken language with only a short history of writing (Snow, 2004, 2008). Written Cantonese most often appears in informal contexts like texting, where conventions are inconsistent. As a result, both MT systems and Cantonese annotators face the challenge of the lack of an accepted written norm. There are several instances in the annotations where multiple written forms correspond to the same spoken word, such as 噉樣 *gam2joeng2*, 咁樣 *gam2joeng2 like this, this way*. Annotators accept both variants: 噉樣 is considered the standard form, while 咁樣 is more widely used in practice.

However, such lack of standardization is not to say that there is no systematicity in Cantonese or Wu orthography and by extension, machine translation error annotation. As noted in previous work (Lu et al., 2024b), orthographies can reliably be

converted from one to another; each error we annotate is similarly annotatable across orthography. In SINITICMTERROR, we annotate using Hanzi and Hanzi-derived systems and their general conventions with which annotators are most familiar. We do not annotate acceptable alternative orthography (*cf.* color and colour in English).

Language confusion. We also observe language confusion; some machine translations in Wu or Cantonese unexpectedly output material from other languages. While it may have been unsurprising to observe language confusion within the Sinitic family (e.g. Cantonese 下 *haa6* for 落 *lok6 down*, following Mandarin 下 *xia4 down*), we observe one instance where *flu* was erroneously translated into Japanese インフルエンザ *infuruenza*, a Katakana adoption of Italian-derived-English word *influenza*. This behavior superficially resembles code mixing observed in bilingual speakers (Lanza, 1997; Muysken, 2000). However, in multilingual NLP systems it is more commonly interpreted as language confusion or interference (Wang et al., 2020; Yu et al., 2024; Lee et al., 2025), which has been attributed to high temperature, language mismatch between representation learning and preference tuning, and undertraining (Marchisio et al., 2024).

These behaviors illustrate the challenge of stable and accurate generation in a low-resource language in multilingual machine translation. More broadly, these findings call for more attention and resources for low-resource Sinitic languages, which lack representation in current multilingual NLP research.

5. Language Model Baselines

In addition to dataset statistics, we report language model baselines on SINITICMTERROR. To benchmark the effectiveness of different LLMs in detecting translation errors, we adopt two evaluation schemes inspired by **MQM-APE** (Lu et al., 2025).

5.1. Error Span Evaluation

Following MQM-APE, we evaluate the span-level agreement between LLM-generated error annotations and human references using **Span Precision (SP)** and **Major Precision (MP)**. More formally, for any single error span e, we define $P(e) = \{i, i+1, ..., j\}$, where i is the start of the marked error span and j is the end of the error span. For a collection of error spans $E = \{e_1, ..., e_n\}$, we also define $P(E) = \bigcup_{j=1}^{n} P(e_j)$. Then, SP and MP are defined as follows:

Model	en-zh		en-yue	
	SP	MP	SP	MP
GPT-4o	40.5	26.8	51.7	14.9
Gemini-2.5-pro	40.9	25.9	**59.7**	**19.8**
Gemma-3-12B-it	28.1	25.7	41.4	18.9
Microsoft/Phi-4	29.3	22.8	48.2	12.2
Qwen3-14B	**44.3**	**29.7**	57.8	14.0

Table 9: Span Precision (SP) and Major Precision (MP) across en-zh and en-yue. **Bold** indicates best performance under a metric.

$$\text{SP} = \frac{P(E) \cap P(\hat{E})}{P(\hat{E})} \quad (1)$$

$$\text{MP} = \frac{P(E_{maj}) \cap P(\hat{E}_{maj})}{P(\hat{E}_{maj})} \quad (2)$$

where E is a collection of gold error spans, $\hat{E}$ is a collection of LLM generated error spans, and E_{maj} denotes the subset of errors that are identified as major errors.

SP measures precision across all error spans predicted by the model, whereas MP focuses exclusively on major error spans, which have the greatest impact on translation quality under the **Multidimensional Quality Metrics (MQM)** (Burchardt, 2013) framework. This metric measures the quality of error span alignment, as it verifies if the indices of the identified error spans align with the gold error spans' indices.

Independently of the span index matching, we also compare LLM and human-derived segment-level MQM scores using correlation metrics.

We convert annotations to numeric MQM via the standard weighting scheme (e.g., Minor=1, Major=5) (Freitag et al., 2021) using the official converter from mt-metrics-eval[6], consistent to previous works (Lu et al., 2024c; Kocmi and Federmann, 2023; Fernandes et al., 2023).

We then compute segment-level alignment between human and LLM-derived MQM scores via Pearson's r, Spearman's ρ, and Kendall's τ, which are standard correlation-based metrics widely used in QE and MT evaluation research (Freitag et al., 2021; Stefanik et al., 2021).

5.2. Index Correctness within LLMs

To ensure that the indices of LLM-predicted error spans are consistent with the character positions in the `mt` string, we introduce a lightweight *tag-to-span* generation protocol (`tag2span`). The model outputs a tagged version of the machine translation together with a structured error list, following

[6] https://github.com/google-research/mt-metrics-eval

Model	Task	r	ρ	τ
GPT-4o	en-zh	0.43	0.43	0.33
	en-yue	0.19	0.21	0.16
Gemini-2.5-pro	en-zh	**0.45**	**0.46**	**0.36**
	en-yue	**0.35**	**0.32**	**0.24**
Gemma-3-12B-it	en-zh	0.34	0.34	0.27
	en-yue	0.07	0.07	0.05
Microsoft/Phi-4	en-zh	0.29	0.27	0.21
	en-yue	0.00	0.02	0.02
Qwen3-14B	en-zh	0.40	0.41	0.32
	en-yue	0.15	0.09	0.01

Table 10: Segment-level MQM score correlation coefficients (Pearson's r, Spearman's ρ, and Kendall's τ) across two language pairs. **Bold** indicates best performance under a metric. Language codes en, zh, and yue correspond to English, Mandarin, and Cantonese, respectively.

a simple JSON schema. Each error region in the translation is enclosed by a unique identifier tag such as `<spanN> ... </spanN>`. We then deterministically parse `<spanN> ... </spanN>` boundaries to recover (*start_index*, *end_index*) pairs in the original `mt`. This enforces an unambiguous alignment between model-declared error regions and character offsets, eliminating off-by-one and tokenization-related drift during post-processing.

We also employ the `zero-denominator policy` to ensure correctness. Since SP and MP are precision measures, denominators can be zero. If $\left|P(\hat{E})\right| = 0$, then SP is *undefined*, similar for MP. In such cases we report the score as **N/A**. We do not impute 0 for undefined precision to avoid conflating "predicts nothing" with "predicts many positions but entirely wrong."

5.3. Interpretation

As shown in Tables 9 and 10, across all evaluated models on the *en-zh* and *en-yue* setting, both span-level and correlation-based metrics remain modest. SP and MP values range mostly between 0.25–0.45, while segment-level correlations (r, ρ, τ) between LLM-derived and human MQM scores stay below 0.5. These consistently limited scores indicate that current open-source LLMs, such as *Gemma-3-12B-it*, *Phi-4*, and *Qwen3-14B*, lack the specialized training required for fine-grained translation error analysis. Furthermore, performance on closed-source LLMs seems limited as well, as both segment-level analysis and MQM score correlation scores are in similar ranges across language pairs.

In addition to this domain-specific limitation, another contributing factor is the low-resource nature of the benchmarked language pairs. Languages such as Cantonese are severely underrepresented in existing MT datasets and LLM pretraining corpora, leading to weaker cross-lingual representations and reduced sensitivity to fine-grained translation schemes. This scarcity of high-quality bilingual data makes it difficult for models to accurately identify context-dependent or culturally specific translation errors.

Together, these findings underscore the necessity of our newly constructed dataset. By providing high-quality human annotations for underrepresented language pairs and explicit error span annotations, our dataset fills a gap in the current multilingual translation and evaluation ecosystem.

It not only enables the systematic study of translation error patterns in low-resource settings but also provides a foundation for fine-tuning and aligning future LLMs toward more accurate, human-consistent translation quality assessment.

6. Conclusion

In this paper, we introduce SiniticMTError, a dataset of machine translation errors, comprising parallel English–Mandarin, English–Cantonese, and English–Wu Chinese splits, as well as a separate Mandarin–Hokkien component. Each entry contains an erroneous machine translation text, erroneous spans, their respective error type and severity, and a gold-label translation. The Mandarin split contains 2,009 annotated sentences; Cantonese 1,402, Wu 68, and Hokkien 154.

SiniticMTError is one of the first human-annotated span-level MT error resources for Wu Chinese, complementing other emerging resources such as the Shanghainese UD dataset (Yang, 2025). It also serves as a significant addition to a small collection of Cantonese and Hokkien human-annotated resources. In addition, our baseline experiments show that even the strongest open and closed-source LLMs achieve modest span-precision and correlation scores, particularly. These results demonstrate both the difficulty of automatic error detection in low-resource Sinitic MT and the potential of SiniticMTError to serve as a training and evaluation benchmark for future models. Beyond the dataset's immediate role in error analysis and MT, we anticipate that the parallel dataset can support a wide range of downstream applications. The dataset may be adapted for several natural language understanding tasks, including language detection and linguistic acceptability judgment (e.g. Min et al., 2025). Moreover, the parallel nature of the dataset makes it a promising resource for transfer learning, which allows models trained on high-resource languages to be adapted more effectively to low-resource Sinitic varieties.

Limitations and Future Work

While we present datasets in four languages, two splits (Wu and Hokkien) are pilot annotations with only a small number of annotated sentences. Future work may benefit from an extension of these datasets.

In addition, the dataset builds on `FLORES+` (Goyal et al., 2022; NLLB Team et al., 2024; Yu et al., 2024), whose source sentences are in English. Although the Cantonese and Wu sentences are parallel, the erroneous machine translations are attempts to translate from English using MT systems and LLM, and may not represent real-world use cases of translating to and from Cantonese and Wu Chinese, whose speakers may often be limited proficiency bilinguals fluent in Mandarin (Li, 2006).

Finally, our work only spans Mandarin, Cantonese, Wu Chinese, and Hokkien among over a dozen of Sinitic languages (Tang and van Heuven, 2007; Chappell, 2015), each with varying numbers of native and bilingual speakers (Norman, 1988; Eberhard et al., 2023). Future work may bootstrap our annotation guidelines and tools to expand to other Sinitic languages as well.

Ethics Statement

Our work introduces a dataset built from the publicly available `FLORES+` (Goyal et al., 2022; NLLB Team et al., 2024; Yu et al., 2024). The annotators are native speakers of the target language (Mandarin, Cantonese or Wu Chinese), knowledgeable about their work and the dataset's downstream use. We internally review the annotations to ensure it does not contain any sensitive or personally identifiable information.

The `mt` outputs on which error spans were annotated were generated by publicly available LLMs and may contain unintended biases or stereotypes. While we ensure that they do not explicitly contain sensitive material, we acknowledge that they may include token distributions that does not represent our values, or opinions. We encourage responsible and context-aware use of our dataset in downstream applications.

Bibliographical References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2024. The reversal curse: LLMs trained on "a is b" fail to learn "b is a". In *The Twelfth International Conference on Learning Representations*.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Hilary Chappell. 2015. *Diversity in Sinitic languages*. Oxford University Press.

Mengyun Chen, Tao Ge, Xingxing Zhang, Furu Wei, and Ming Zhou. 2020. Improving the efficiency of grammatical error correction with erroneous span detection and correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7162–7169, Online. Association for Computational Linguistics.

Peng-Jen Chen, Kevin Tran, Yilin Yang, Jingfei Du, Justine Kao, Yu-An Chung, Paden Tomasello, Paul-Ambroise Duquenne, Holger Schwenk, Hongyu Gong, Hirofumi Inaguma, Sravya Popuri, Changhan Wang, Juan Pino, Wei-Ning Hsu, and Ann Lee. 2023. Speech-to-speech translation for a real-world unwritten language. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4969–4983, Toronto, Canada. Association for Computational Linguistics.

Yun-Hsin Chu, Shuai Zhu, Shou-Yi Hung, Bo-Ting Lin, En-Shiun Annie Lee, and Richard Tzong-Han Tsai. 2025. ATAIGI: An AI-powered multimodal learning app leveraging generative models for low-resource Taiwanese hokkien. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 11–19, Albuquerque, New Mexico. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj

Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier.

Zi-Yi Dou and Graham Neubig. 2022. Unsupervised machine translation of low-resource languages: A case study on bambara. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 335–346. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2023. *Ethnologue: Languages of the World*, 26 edition. SIL International, Dallas.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang,

Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao

Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Kung Hong, Lifeng Han, Riza Batista-Navarro, and Goran Nenadic. 2024. CantonMT: Cantonese to English NMT platform with fine-tuned models using real and synthetic back-translation data. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 590–599, Sheffield, UK. European Association for Machine Translation (EAMT).

Yi-Hsiang Hung and Yi-Chin Huang. 2022. A preliminary study on Mandarin-Hakka neural machine translation using small-sized data. In *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, pages 307–315, Taipei, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Vivek Iyer, Bhavitvya Malik, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. 2024. Quality or quantity? on data scale and diversity in adapting large language models for low-resource translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1393–1409, Miami, Florida, USA. Association for Computational Linguistics.

Ouail Kitouni, Niklas Nolte, Diane Bouchacourt, Adina Williams, Mike Rabbat, and Mark Ibrahim. 2024. The factorization curse: Which tokens you predict underlie the reversal curse and more. In *Advances in Neural Information Processing Systems*, volume 37, pages 112329–112355. Curran Associates, Inc.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Cheung Kwan-hin and Robert S Bauer. 2002. The representation of cantonese with chinese characters. *Journal of Chinese Linguistics Monograph Series*, pages i–489.

Yen-Chun Lai, Yi-Jun Zheng, Wen-Han Hsu, Yan-Ming Lin, Cheng-Hsiu Cho, Chlh-Chung Kuo, Chao-Shih Huang, and Yuan-Fu Liao. 2024. Construction of large language models for taigi and hakka using transfer learning. In *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.

Surafel M. Lakew, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. 2018. Transfer learning in multilingual neural machine translation with dynamic vocabulary. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 54–61, Brussels. International Conference on Spoken Language Translation.

Elizabeth Lanza. 1997. *Language mixing in infant bilingualism: A sociolinguistic perspective*. Oxford University Press.

En-Shiun Annie Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Ifeoluwa Adelani, Ruisi Su, and Arya D. McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Nahyun Lee, Yeongseo Woo, Hyunwoo Ko, and Guijin Son. 2025. Controlling language confusion in multilingual LLMs. In *Proceedings of the*

63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1026–1035, Vienna, Austria. Association for Computational Linguistics.

Peppina Po-lun Lee. 2019. *Focus Manifestation in Mandarin Chinese and Cantonese: A Comparative Perspective*. Routledge Studies in Chinese Linguistics. Taylor & Francis.

Thomas H. C. Lee. 2011. A bilingual corpus of legislative texts in hong kong: The hong kong hansard corpus. *Language Resources and Evaluation*, 45(2):123–139.

Lauren Levine, Junghyun Min, and Amir Zeldes. 2025. Building UD cairo for Old English in the classroom. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 97–104, Ljubljana, Slovenia. Association for Computational Linguistics.

Roger Levy and Christopher D. Manning. 2003. Is it harder to parse Chinese, or the Chinese treebank? In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 439–446, Sapporo, Japan. Association for Computational Linguistics.

Charles N Li and Sandra A Thompson. 1979. Third-person pronouns and zero-anaphora in chinese discourse. *Syntax and semantics*, 12(01).

David C. S. Li. 2006. Chinese as a lingua franca in greater china. *Annual Review of Applied Linguistics*, 26:149–176.

Senyu Li, Jiayi Wang, Felermino D. M. A. Ali, Colin Cherry, Daniel Deutsch, Eleftheria Briakou, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. Ssa-comet: Do llms outperform learned metrics in evaluating mt for under-resourced african languages?

Evelyn Kai-Yan Liu. 2022. Low-resource neural machine translation: A case study of Cantonese. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 28–40, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Bo-Han Lu, Yi-Hsuan Lin, Annie Lee, and Richard Tzong-Han Tsai. 2024a. Enhancing Taiwanese hokkien dual translation by exploring and standardizing of four writing systems. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6077–6090, Torino, Italia. ELRA and ICCL.

Bo-Han Lu, Yi-Hsuan Lin, Annie Lee, and Richard Tzong-Han Tsai. 2024b. Enhancing Taiwanese hokkien dual translation by exploring and standardizing of four writing systems. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6077–6090, Torino, Italia. ELRA and ICCL.

Qingyu Lu, Liang Ding, Kanjian Zhang, Jinxia Zhang, and Dacheng Tao. 2025. MQM-APE: Toward high-quality error annotation predictors with automatic post-editing in LLM translation evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5570–5587, Abu Dhabi, UAE. Association for Computational Linguistics.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024c. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816, Bangkok, Thailand. Association for Computational Linguistics.

Sin-En Lu, Bo-Han Lu, Chao-Yi Lu, and Richard Tzong-Han Tsai. 2022. Exploring methods for building dialects-Mandarin code-mixing corpora: A case study in Taiwanese hokkien. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6287–6305, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Victor H. Mair and John DeFrancis. 2003. Abc cantonese-english dictionary corpus. Data derived from the ABC Cantonese-English dictionary.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677, Miami, Florida, USA. Association for Computational Linguistics.

Stephen Matthews. 2006. On serial verb constructions in cantonese. *Serial verb constructions: A cross-linguistic typology*, 2.

Stephen Matthews and Virginia Yip. 2011. *Cantonese: A Comprehensive Grammar*, 2nd edition. Routledge, London. EBook published 23 May 2013.

Junghyun Min, York Hay Ng, Sophia Chan, Helena Shunhua Zhao, and En-Shiun Annie Lee. 2025. Cantonlu: A benchmark for cantonese natural language understanding. *arXiv preprint arXiv:2510.20670*.

Anna Morbiato. 2017. *Word order and sentence structure in Mandarin Chinese: new perspectives*. Ph.D. thesis. Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Linguistics, Faculty of Arts and Social Sciences, The University of Sydney and Doctor of Philosophy in Asian and African Studies Ca' Foscari University of Venice.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Pieter Muysken. 2000. *Bilingual speech: A typology of code-mixing*. Cambridge university press.

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. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Jerry Norman. 1988. *Chinese*. Cambridge University Press.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Hui-

wen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card.

Wuyun Pan, S.F. Zhengzhang, R.J. You, and Lien Chinfa. 1991. An introduction to the wu dialects. *Journal of Chinese Linguistics Monograph Series*, (3):235–291.

Geon Woo Park, Junghwa Lee, Meiying Ren, Allison Shindell, and Yeonsoo Lee. 2023. VARCO-MT: NCSOFT's WMT'23 terminology shared task submission. In *Proceedings of the Eighth Conference on Machine Translation*, pages 919–925, Singapore. Association for Computational Linguistics.

Waltraud Paul. 2014. Why particles are not particular: Sentence-final particles in chinese as heads of a split cp. *Studia linguistica*, 68(1):77–115.

Juan Antonio Perez-Ortiz, Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Aaron Galiano Jimenez, Antoni Oliver, Claudi Aventín-Boya, Alejandro Pardos, Cristina Valdés, Jusèp Loís Sans Socasau, and Juan Pablo Martínez. 2024. Expanding the FLORES+ multilingual benchmark with translations for Aragonese, aranese, Asturian, and Valencian. In *Proceedings of the Ninth Conference on Machine Translation*, pages 547–555, Miami, Florida, USA. Association for Computational Linguistics.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11).

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

Wesley Scivetti, Tatsuya Aoyama, Ethan Wilcox, and Nathan Schneider. 2025. Unpacking let alone: Human-scale models generalize to a rare construction in form but not meaning. *arXiv preprint arXiv:2506.04408*.

Thibault Sellam, Surafel M. Lakew, Marcos Zampieri, Barry Haddow, and Alexandra Birch. 2021. The multilingual evaluation landscape: A survey of datasets for multilingual machine translation. In *Proceedings of the 2021 Conference on Machine Translation (WMT)*, pages 958–973. Association for Computational Linguistics.

Sam Shapiro, Fesseha Ghidey, Shammur Chowdhury, et al. 2023. Lesan: A multilingual and multimodal platform for low-resource languages in ethiopia. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mgpt: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Anushka Singh, Ananya Sai, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh Khapra. 2024. How good is zero-shot MT evaluation for low resource Indian languages? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 640–649, Bangkok, Thailand. Association for Computational Linguistics.

Don Snow. 2004. *Cantonese as written language: The growth of a written Chinese vernacular*, volume 1. Hong Kong University Press.

Don Snow. 2008. Cantonese as written standard? *Journal of Asian Pacific Communication*, 18(2):190–208.

Michal Stefanik, Vít Novotný, and Petr Sojka. 2021. Regressive ensemble for machine translation quality evaluation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1041–1048, Online. Association for Computational Linguistics.

Chaoju Tang and Vincent J van Heuven. 2007. Mutual intelligibility and similarity of chinese dialects: Predicting judgments from objective measures. *Linguistics in the Netherlands*, 24(1):223–234.

Sze-Wing Tang, Fan Kwok, Thomas Hun-Tak Lee, Caesar Lun, Kang Kwong Luke, Peter Tung, and Kwan Hin Cheung. 2002. Guide to lshk cantonese romanization of chinese characters. *Hong Kong: Linguistic Society of Hong Kong*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe,

Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh,

George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Rus-

Ian Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei,

Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

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.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2022. Progress in machine translation. *Engineering*, 18:143–153.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Hassan Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Abdi Mohamed, Hassan Ayinde, Oluwabusayo Olufunke Awoyomi, Lama Alkhaled, Sana Al-azzawi, Naome A. Etori, Millicent Ochieng, Clemencia Siro, Njoroge Kiragu, Eric Muchiri, Wangari Kimotho, Lyse Naomi Wamba Momo, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Nasir Iro, Saheed S. Abdullahi, Stephen E. Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Raphael Ogbu, Sam Ochieng', Verrah Akinyi Otiende, Chinedu Emmanuel Mbonu, Sakayo Toadoum Sari, Yao Lu, and Pontus Stenetorp. 2024. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Mexico City, Mexico. Association for Computational Linguistics.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Syed Mekael Wasti, Shou-Yi Hung, Christopher Collins, and En-Shiun Annie Lee. 2025. Translationcorrect: A unified framework for machine translation post-editing with predictive error assistance.

Kam-Fai Wong and Xiaodong Zhang. 2017. Building a parallel corpus for english-cantonese machine translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT)*, pages 85–90. Association for Computational Linguistics.

Rong Xiang, Ming Liao, and Jing Li. 2024. Cantonese natural language processing in the transformers era. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 69–79, Bangkok, Thailand. Association for Computational Linguistics.

Qizhen Yang. 2025. ShUD: the first shanghainese Universal Dependency treebank. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 186–193, Ljubljana, Slovenia. Association for Computational Linguistics.

Foong Ha Yap and Winnie Chor. 2011. Asymmetry in grammaticalization –the case of directional particles in cantonese. 2011 The 5th Conference on Language, Discourse and Cognition (CLDC-5) ; Conference date: 29-04-2011 Through 01-05-2011.

Hongjian Yu, Yiming Shi, Zherui Zhou, and Christopher Haberland. 2024. Machine translation evaluation benchmark for Wu Chinese: Workflow and analysis. In *Proceedings of the Ninth Conference on Machine Translation*, pages 600–605, Miami, Florida, USA. Association for Computational Linguistics.

Lily H Zhang, Hamid Dadkhahi, Mara Finkelstein, Firas Trabelsi, Jiaming Luo, and Markus Freitag. 2025a. Learning from others' mistakes: Finetuning machine translation models with span-level error annotations. In *Forty-second International Conference on Machine Learning*.

Tai Zhang, Lucie Yang, Erin Chen, Karen Riani, Jessica Zipf, Mariana Shimabukuro, and En-Shiun Annie Lee. 2025b. Learning low-resource languages through NLP-driven Flashcards: A case study of hokkien in language learning applications. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 303–312, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiaoheng Zhang. 1998. Dialect MT: A case study between Cantonese and Mandarin. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*.

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

A. Additional Details on Annotation Guidelines

Hokkien Category	Aligned Category
Acc./Mistranslation	Mistranslation
Acc./Omission	Omission
Acc./Addition	Addition
Fluency/Grammar	Grammar
Fluency/Spelling	Spelling
Fluency/Punctuation	Typography
Fluency/Register	Grammar
Fluency/Inconsistency	Mistranslation
Terminology/Inappropriate	Mistranslation
Terminology/Inconsistent	Mistranslation
Style/Awkward	Grammar
Locale/*	Mistranslation
Non-translation	Unintelligible
Purity	*No mapping*

Table 11: Proposed alignment between 7 Hokkien error categories 17 subcategories and standard SINITICMTERRORerror categories.

While we base our annotation guidelines on prior work in MQM (Burchardt, 2013) and AfriCOMET (Wang et al., 2024), we make several adjustments during the pilot and main annotation stages described in Section 3.3, resulting in error categories described in Table 2 and severity levels described in Table 3. We describe such adjustments in detail below.

Inappropriate Proper Nouns. For proper nouns such as specific names of people and places, if their translations are not the same as in the reference sentences, then annotators should classify them as Spelling errors, instead of Mistranslation. This guideline is based on the assumption that the reference translations from FLORES+ (Goyal et al., 2022; NLLB Team et al., 2024; Yu et al., 2024) are of high quality, as they were created by professional human translators. The purpose is to distinguish between general semantic mistranslation and inappropriate translations of the proper nouns, which may vary historically or regionally. Labeling such differences as Spelling errors helps the future work, as models could better learn about the boundaries between semantic-level errors and surface-level variations.

Full-width form punctuations. In Sinitic writing systems, full-width punctuation marks are commonly used. However, the 600M NLLB-200 model (Team et al., 2022) consistently outputs sentences with punctuations in half-width forms. The annotators were instructed to skip such annotation cases, as the goal is to distinguish between misuse of punctuations and incorrect forms of them. In the future QA stage, we plan to consult linguistic experts and may introduce a new error category, if this issue will significantly affect translation quality according to the experts.

Omission of quality score evaluations. In our dataset, we focus exclusively on the error span positions, error type, and severity, rather than assigning overall quality scores to the machine translations. Our design reflects our goal of helping the models better identifying and classifying errors, instead of performing quality assessments. Omitting quality scores also improves the translation efficiency and helps avoid subjectivity and disagreement over score interpretation, thus improving consistency and efficiency in the annotation process.

Hokkien annotations. In Section 3.3, we describe Hokkien annotations as substantially different from other components in SINITICMTERRORin both methodology and scope. One notable difference is in the error category schema. Hokkien error categories are hierarchical, with 7 categories,

and 17 subcategories in total; they are outlined in Table 11. Hokkien category 'Locale' contains 5 sub-categories: 'Currency', 'Time', 'Name', 'Date', and 'Address'. Purity is a separate category that annotates language confusion—Mandarin lexical items in an otherwise Hokkien sentence. It is an artifact due to structural and lexical differences specific to Mandarin-Hokkien transfer, which are not directly comparable to English-Mandarin annotation schemes.