Benchmarking Multilingual LLM Translation Accuracy for Fuzhounese

Sue Zheng[♦], **Jelke Bloem**[♣♠]

♦ Kunstmatige Intelligentie, University of Amsterdam
♣ Institute for Logic, Language and Computation, University of Amsterdam
♠ Data Science Centre, University of Amsterdam
s.y.zheng@students.uu.nl, j.bloem@uva.nl

Abstract

Multilingual large language models are known to perform very well on high-resource languages, while their ability to process severely under-resourced languages remains underexplored. We investigate multilingual LLM translation performance on Fuzhounese, an under-resourced Sinitic language without a standardized orthography and almost no digital presence. Having adopted some methodological insights from the HKCanto-Eval benchmark, this paper presents a bidirectional translation framework based on a dataset of 305 sentences (300 constructed English sentences and 5 additional reference translations), that assesses the comprehension and generation of Fuzhounese, evaluated using automatic metrics and human Likert-scale judgments. The results reveal poor performance on Fuzhounese in both translation directions: BERTScore and chrF++ values consistently stay low when models are faced with comprehension tasks, while for generation tasks, scores are generally more than twofold lower than those for Mandarin or Cantonese. These findings highlight structural biases in multilingual LLMs toward high-resource languages and stress the need for resource-aware modeling and evaluation approaches in multilingual NLP systems.

Keywords: multilingual LLMs, under-resourced languages, Fuzhounese, machine translation, evaluation

1. Introduction

Multilingual large language models (LLMs) have been developing rapidly, raising questions about how effectively these systems handle linguistic diversity. Recent studies show a substantial difference in performance when comparing high-resource languages to low-resource languages (Joshi et al., 2020), which challenges the inclusiveness and generalizability of current NLP technologies. LLMs are primarily trained on high-resource languages, performing far more reliably on well-represented languages than under-resourced ones (Qin et al., 2025), while nearly 90% of the world's languages do not have sufficient data for robust NLP modeling (Joshi et al., 2020). This underscores a structural bias inherent in the design and training processes of modern multilingual LLMs.

Fuzhounese represents a particularly neglected case in multilingual NLP research. Despite its large speaker population of 10 million people (Norman, 1988; Eberhard et al., 2023), it has no standardized orthography and minimal publicly available corpora (Maclay and Baldwin, 1929), making it a perfect example of a severely under-resourced language. A recent survey by Liu and Best (2025) reveals that the Fuzhounese language or its variants has never been mentioned in the 73,285 papers in the ACL Anthology up until 2022 (ACL OCL corpus, 1952-2022, Rohatgi et al., 2023) and subsequently, that survey was the only mention.

Given this lack of attention in the community, and the lack of any systematic evaluation that we are aware of, it is unclear how well current multilingual LLMs process this language. Within the Sinitic landscape, Cantonese research provides the closest available comparison. The HKCanto-Eval benchmark (Cheng et al., 2025) assesses the comprehension, translation, and idiomatic fluency of multilingual LLMs when faced with Cantonese, a comparatively well-documented Chinese variety. It is consistently found that models often default to Mandarin-centric grammar and vocabulary, leading to the assumption that if robust models already struggle with Cantonese, the performance drop on Fuzhounese is likely to be even steeper.

We address this gap by evaluating Fuzhounese machine translation when paired up with Cantonese, Mandarin or English. Two different translation directions are assessed: Fuzhounese-to-English/Cantonese/Mandarin and English-to-Fuzhounese/Cantonese/Mandarin. Mandarin and Cantonese serve as reference languages and represent different levels of linguistic standardization and resource availability within the Sinitic language family. In doing so, we also contribute the first parallel dataset for Fuzhounese machine translation benchmarking. The gold translations were constructed by a native Fuzhounese speaker and subsequently reviewed by additional native and heritage speakers to ensure linguistic validity.[1]

[1] All data, prompts, and evaluation scripts are available at https://github.com/fzhkch/fuzhounese-translation-benchmark.

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 198–209
11–12 May 2026.

We hypothesize that multilingual LLMs will exhibit clear performance disparities between the languages. Based on prior research into cross-lingual transfer and data imbalance (Conneau et al., 2020; Wu and Dredze, 2020), we expect that, when Fuzhounese serves as the input language, Mandarin, considering its high-resource status and Sinitic affiliation, will be the most accurate, followed by English and, lastly, Cantonese. When English serves as the input language for translation, Mandarin translations are, likewise, expected to be the most accurate, followed by Cantonese, with Fuzhounese showing the lowest performance.

2. Background

2.1. Multilingual Large Language Models

Multilingual large language models (LLMs) are trained on data from multiple languages using shared tokenization schemes and shared embedding spaces (Xu et al., 2025). As a result, languages are not learned in isolation but through joint representations that enable the application of knowledge from one language to another, called cross-lingual transfer.

High-resource languages tend to develop stronger and more stable representations than under-resourced languages, because training data is highly imbalanced across languages. Subsequently, as a result of cross-lingual transfer, under-resourced languages are more likely to be processed through proxy mappings from dominant languages, which introduces systematic errors and bias (Joshi et al., 2020; Conneau et al., 2020).

2.2. Under-resourced Language Processing

The lack of standardized writing systems, robust digital corpora and consistent annotation are all factors that play a big role in under-resourced NLP (Pakray et al., 2025), complicating both model behavior and its evaluation: models may fail to recognize non-standard writing, and the lack of corpora makes multilingual LLMs rely heavily on proxy transfer from dominant languages, resulting in incorrect or distorted outputs. Automatic evaluation metrics might underestimate quality when outputs differ from references (Doddapaneni et al., 2025), as their embedding spaces are often not trained to capture linguistic variation in low-resource languages.

2.3. Cross-Lingual Transfer in Sinitic Languages

Sinitic languages, despite displaying substantial variation in lexicon, syntax and phonology, are often clustered together by multilingual LLMs. Lower-resourced varieties are put in the same group as Mandarin, whose dominant resource status far outweighs that of the other languages. This results in the incorrect mapping of non-Mandarin varieties onto Mandarin-like structures (Cheng et al., 2025), which may lead to significant errors for Sinitic varieties that are structurally divergent from Mandarin, such as Fuzhounese.

2.4. Evaluating LLMs through Translation

Translation tasks require models to demonstrate several kinds of linguistic capabilities, such as lexical access, grammatical mapping, and semantic preservation (Papineni et al., 2002). For under-resourced languages, these tasks are especially revealing as they show whether a multilingual LLM relies on genuine linguistic representations or defaults to structures from better-represented languages (Shani and Basirat, 2025). Moreover, translation tasks highlight asymmetries between comprehension and generation, providing insight into how well a model can process a language that is effectively absent from its training data (Cheng et al., 2025). For these reasons, translation tasks are widely used as a primary diagnostic tool for assessing multilingual LLM performance on severely under-resourced languages, such as Fuzhounese.

2.5. Linguistic and Sociotechnical Background of Fuzhounese

Fuzhounese, as a member of the Eastern Min (Min Dong) branch, split from other branches of the Sinitic language family prior to the major phonological and lexical developments that shaped most modern Chinese languages (Norman, 1991). The language is primarily spoken in and around the city of Fuzhou, in a region characterized by mountains and rivers. This limited mobility has fostered a strong local identity (Ramsey, 1989) which, in combination with the early divergence from other Sinitic branches, allowed Fuzhounese to evolve relatively independently from neighbouring language variants. This resulted in substantial phonological and lexical differences from most Sinitic languages (Branner, 2000). Presently, the language is also spoken by a diaspora community from this region.

Phonologically, Fuzhounese has a complex tonal system with extensive tone sandhi chains, in which tones change across entire phrases rather than only in isolated pairs (Chen, 2000). Another property that sets Fuzhounese apart from other Sinitic languages is its lexical distinctiveness: many common Fuzhounese words have no direct equivalents in other Sinitic varieties (Norman, 1991).

Fuzhounese does not have a standardized orthography. There is no official system that pre-

scribes the correct way to write Fuzhounese, resulting in ambiguous and inconsistent written forms (Branner, 2000). This study adopts Bàng-uâ-cê (Foochow Romanized), which is a long-established romanization system designed to reflect the phonological structure of the language (Baldwin, 1871), as the primary representation of Fuzhounese. Although not officially standardized, Bàng-uâ-cê provides the most transparent and linguistically faithful representation currently available. It directly encodes tonal and segmental distinctions that may get obscured or lost in character-based approximations, making it well suited for translation-based evaluation.

In relation to other Sinitic varieties, Fuzhounese shares more typological properties with other Southern Chinese languages, in particularly Cantonese. While Fuzhounese and Cantonese don't belong to the same branch of the Sinitic Family, both preserve rich tonal inventories and complex tonal realizations: this is a feature that has been shown to contribute to perceived similarity across Southern Sinitic languages, even when lexical overlap is limited (Tang and van Heuven, 2007). Thus, tonal complexity appears to provide perceptual cues that facilitate partial comprehension. This relative tonal closeness stands in contrast to Mandarin's simplified tonal system, which differs substantially in both tonal inventory and sandhi behavior (Duanmu, 2007), and motivates the inclusion of both Mandarin and Cantonese as complementary comparison languages in our evaluation.

Excluding Hong Kong, where Cantonese is widely supported and used in public and institutional domains, Mandarin is the only standardized language of formal instruction and governance in mainland China (Norman, 1988). This policy limits support for regional varieties, making Fuzhounese one of many Chinese languages that are largely confided to informal and spoken domains. As Fuzhounese does not hold an official language status in China, the availability of linguistic resources becomes limited: the production of textual materials is constrained due to not having a standardized orthography (Branner, 2000), and the language's primary use in informal spoken contexts further limits its presence in formal written communication (Norman, 1991). Overall, Fuzhounese is widely considered a low-resource language in both linguistic and computational contexts. Existing written resources consist mostly of fragmented and sparse data, which are often outdated or lacking consistency in spelling and representation (Norman, 1991).

Therefore, Fuzhounese may be a particularly challenging case for multilingual large language models. The absence of standardized orthography and limited textual data affects model exposure, increasing the reliance on cross-lingual transfer from dominant Sinitic languages (Jiang et al., 2025). Such characteristics provide a clear motivation for evaluating LLM performance on Fuzhounese through translation-based tasks. To date, there are no known large language models that have been pre-trained or fine-tuned on Fuzhounese corpora (Liu and Best, 2025).

2.6. Related Work

NLP researchers have previously examined how multilingual language models perform across languages that vary in levels of linguistic resources, revealing that inequalities in the distribution of training data directly influences model performance. Qin et al. (2025) provide empirical evidence showing that multilingual models consistently favor languages with bigger corpora over low-resource languages, as they produce more accurate and stable outputs.

The HKCanto-Eval benchmark (Cheng et al., 2025) reveals that even state-of-the-art multilingual LLMs often mistranslate idiomatic expressions or default to Mandarin vocabulary and syntax in translating Cantonese. Medium-resource languages studies, such as Armengol-Estapé et al.'s (2021) work on Catalan, reveal the same issues. Studies on low-resource machine translation show that multilingual LLMs struggle with under-resourced languages lacking parallel data (Kumar et al., 2025; Mamasaidov et al., 2025). Furthermore, it is argued that dialects, which mostly do not have standardized orthographies, are significantly underevaluated (Joshi et al., 2020).

3. Methodology

To examine how LLMs handle Fuzhounese, our methodology draws inspiration from HKCanto-Eval (Cheng et al., 2025) by adopting a selection of their components. Although HKCanto-Eval was originally developed for Cantonese, the benchmark offers a useful structural reference for assessing the performances of other Sinitic languages in LLMs. As Fuzhounese is represented using Bàng-uâ-cê, we adapt the evaluation into a translation-focused benchmark suited to the linguistic and resource constraints of Fuzhounese.

3.1. Dataset Construction

The dataset used in this study consists mainly of 300 newly prepared English source sentences, manually constructed to cover a range of linguistic phenomena, including basic syntax, grammatical variation, idiomatic expressions, and narrative structures. Additionally, an extra five sentences

with previously documented Fuzhounese translations are added to complement the dataset, providing a reference point for the qualitative assessment of translation accuracy. While the dataset size is modest, it reflects the practical constraints of constructing expert-validated Fuzhounese data under severe resource scarcity.

Category	Short	Medium	Long	Total
Everyday	60	48	12	120
Grammar	40	32	13	85
Idioms	30	24	6	60
Narratives	10	10	10	30
Imperatives	10	0	0	10
Total	150	114	41	305

Table 1: Sentence distribution by category and length – short (2-5 words), medium (6-11 words), long (≥12 words).

Table 1 shows the sentence distribution across five semantic categories and three different sentence lengths. The categorization follows the same scheme as HKCanto-Eval (Cheng et al., 2025), ensuring coverage of multiple language functions, including everyday usage, grammatical constructions, idiomatic expressions, narrative descriptions, and imperative forms. Sentences are further grouped into short (2-5 words), medium (6-11 words), or long (≥12 words), which allows analysis of how sentence complexity influences translation performance.

To form a parallel multilingual dataset, we manually translated each English sentence into gold-standard Cantonese and Mandarin translations. As previously mentioned, these languages serve as comparative baselines, possessing different levels of larger digital corpora as well as greater representations in the training data of multilingual LLMs. This makes them suitable comparison candidates for evaluating the relative difficulty that multilingual LLMs encounter when processing Fuzhounese.

Furthermore, as the absence of a standardized writing system and the extremely limited digital footprint make it difficult to source any suitable written corpora, we create a full set of manual Fuzhounese gold-translations in Bàng-uâ-cê for the 300 English-source sentences in the dataset. These translations were created in a two-step procedure.

In the first step, initial translations were formulated by a native Fuzhounese speaker, in consultation with historical materials to confirm romanization accuracy and lexical validity. Specifically, The Manual of the Foochow Dialect (Baldwin, 1871) and the Dictionary of the Foochow Dialect (Maclay and Baldwin, 1929) were consulted, which are some of the few detailed descriptions of the dialect's phonology and vocabulary and provide good reference information for tone categories and grammatical markers. At this stage, the translations were also cross-checked with a publicly available transliteration tool for Eastern Min, which outputs rough Bàng-uâ-cê romanizations based on Mandarin tokens[2], to have a more modern reference.

In the second step, five Fuzhounese speakers from mainland China and overseas communities review the sentences to help uncover and correct potential issues, including tone sandhi errors, unnatural lexical choices, and grammatical inconsistencies. This ensures that the translations reflect natural and correct Fuzhounese.

The resulting dataset is a four-way parallel corpus of gold translations. This structure enables detailed comparative analyses across languages and categories, helping the study evaluate translation performance in both directions and allowing the outputs of multilingual LLMs to be compared against the reference translations.

3.2. Adaptations from HKCantoEval

Rather than reproducing the full HKCantoEval framework (Cheng et al., 2025), this study adapts 3 main elements that are transferable to a low-resource, romanized linguistic setting:

3.2.1. Bidirectional Translation

The main component of this evaluation is a bidirectional translation task that examines both Fuzhounese to English and English to Fuzhounese comprehension, as multilingual LLMs often show asymmetric capabilities. Including both translation directions allows the study to capture these differences and evaluate whether the model's linguistic representations are robust or merely pattern-based approximations.

3.2.2. Zero-Shot Evaluation Setting

Multilingual LLM performance is evaluated in a zero-shot setting, meaning that no Fuzhounese examples or task demonstrations are provided prior to generation. Zero-shot testing is essential given the absence of training corpora for Fuzhounese, making fine-tuning or supervised alignment infeasible. Therefore, this setup reflects real-world usage conditions and allows model performance to be interpreted as a direct indicator of existing multilingual generalization rather than the result of task-specific guidance.

We use the same prompt template as HKCantoEval (Cheng et al., 2025), modified slightly to

[2] https://transliterationtools.blogspot.com/2023/11/eastern-min-conversion-tool.html

accommodate romanized Fuzhounese input and output. A single standardized translation prompt is used across all models and languages to ensure comparability. Using an identical instruction template prevents prompt-engineering bias, which allows performance differences to be attributed to model competence rather than instruction quality. The prompt templates used in this study are provided in Appendix A.

3.3. Model Set-Up

The evaluation is conducted on a set of five different language models, developed across two different research environments. GPT-4o, GPT-4o mini (OpenAI), and Gemini-2.5 Flash (Google DeepMind) originate from Western research environments, whereas DeepSeek (DeepSeek AI) and Qwen (Alibaba Cloud) come from Chinese research and industry contexts. While information on the representation of particular languages in the training or instruction tuning data of these models is not available, and all of these models are designed as general-purpose multilingual systems, their stated language support and development in different research environments provides a rationale for examining whether performance differs between Chinese and non-Chinese language pairs.

All models are accessed through their respective APIs using a consistent prompting format, with each sentence being evaluated using one single API call per model. All generation parameters are kept at their API default values: the temperature parameter is set to 0.0 to minimize randomness and maximize determinism in model output. Top-p is set to 1.0, and there is no frequency penalty or presence penalty. No post-editing is applied to the model outputs.

All prompts, API versions, dataset details, and evaluation scripts are publicly available in the aforementioned online supplementary materials on GitHub.

3.4. Evaluation Framework and Metrics

The evaluation framework consists of two translation directions, which are assessed using a combination of automatic scoring with established metrics and a manual evaluation tailored to the linguistic properties of Fuzhounese.

3.4.1. Automatic Evaluation Metrics

Following recommendations from multilingual machine-translation research and the HKCanto-Eval Framework (Cheng et al., 2025), we apply two automatic evaluation metrics: BERTScore (Zhang et al., 2020) and chrF++ (Popović, 2015).

BERTScore evaluates translation quality based on semantic similarity, using contextualized embeddings. Scores range from 0 to 1, with prior work showing that high-quality translations typically achieve scores around 0.80, while values below 0.55 indicate poor semantic alignment (Zhang et al., 2020). A clear limitation of this metric in our setting is that BERT models are not trained to represent Bàng-uâ-cê or other orthographic conventions of Fuzhounese, which may lead to unreliable or uninterpretable similarity scores for the language. Therefore, we use BERTScore only to evaluate translations in English, Mandarin, and Cantonese, where pretrained BERT models do provide meaningful semantic representations (Joshi et al., 2020).

chrF++ compares character-level n-gram sequences rather than word tokens, making the metric more robust for low-resource and romanized languages like Fuzhounese. chrF++ has been shown to perform reliably on Chinese-family languages due to shared morphophonological patterns and overlapping character sequences (Ma et al., 2019). Scores range from 0 to 100, with values around 45 being common in high-quality translations, and values below 20 reflecting minimal overlap with the reference translation (Popović, 2015).

3.4.2. Likert Scale for Translation Comparison

Due to the limitations of BERTScore for evaluating Bàng-uâ-cê, automatic evaluation alone cannot provide a complete comparison of model performance with respect to translation accuracy across Fuzhounese, Mandarin, and Cantonese generation. Thus, we additionally use a human Likert-scale evaluation to evaluate English-source translations, an approach adopted in other work on under-resourced languages (Graham et al., 2013; Kumar et al., 2025). We use a five-point Likert scale to evaluate translation quality across multiple dimensions: Adequacy, Fluency, and Word Validity.

Every produced translation is rated independently by three reviewers that are fluent in all three languages, including native and heritage speakers of Fuzhounese. They are instructed to prioritize semantic equivalence over literal word matching and to disregard spelling or character variation where meaning is clear. This allows the direct comparison of Fuzhounese, Mandarin, and Cantonese translations on a shared qualitative scale. When a model's output is clearly invalid or unintelligible, the translation will be considered a failure. Such outputs are automatically assigned the minimum score of 1 without further subjective judgment, ensuring consistency in the handling of faulty translations. Ratings are averaged across annotators and across the three dimensions to obtain final scores. To assess the reliability of the ratings and contextual-

ize the result, we present separate inter-annotator agreement values per dimension and per language.

4. Results

4.1. Fuzhounese-Source Translation Tasks

In Table 2 we observe that in most cases, English generated translations obtain the highest BERTScore values, with Cantonese predictions ranking second and Mandarin predictions consistently performing worst. This pattern is observed across almost all evaluated models. The BERTScores range from 0.151 to 0.508, meaning that even the highest-scored model translations indicate very limited comprehension of Fuzhounese.

The chrF++ scores exhibit a similar pattern: English scores notably higher across all models, while absolute values are much lower for Mandarin and Cantonese. Like BERTScore, the highest chrF++ score is also suboptimal, being a mere 30.59, once again indicating very limited overlap between the model outputs and gold translations.

As for models, we see that Gemini-2.5 Flash achieves the highest scores across all categories. Among these models, there doesn't appear to be a clear advantage for models developed in a Chinese context. Despite Qwen-2.5-72b coming from a Chinese context, it achieves higher scores when translating Fuzhounese into English than into Cantonese or Mandarin. The relative ranking of models remains consistent across languages.

4.2. English-Source Translation Tasks

Clear performance differences across the target languages can be observed in Table 3. Unsurprisingly, English sentences that are translated into Mandarin consistently achieve the highest scores across both human and automatic metrics, obtaining near-ceiling Likert scores and strong BERTScores (≈ 0.81–0.83) which indicate fluent and semantically correct output. Cantonese exhibits a similar but slightly weaker pattern, with Likert scores above 4.1 and chrF++ values in the 27–45 range, suggesting that multilingual LLMs are largely capable of producing acceptable Cantonese translations, albeit with reduced lexical and syntactic similarity compared to Mandarin. Fuzhounese generated translations score substantially lower, with Likert ratings being consistently poor and several models even receiving scores close to the minimum.

Between models, the largest performance disparities emerge in Fuzhounese translations, which is the most challenging task. All models except Gemini-2.5 Flash fail to produce somewhat intelligible translations, with Likert scores clustered near the minimum (1.0–2.1). Its performance, however, still lags far behind its own results for Cantonese and Mandarin. In both translation directions, translation quality does appear to be slightly influenced by model capacity, but a persistent gap between Fuzhounese on the one hand and Cantonese and Mandarin on the other hand remains.

4.3. Stratified evaluation

We subdivide our results by two factors that may affect translation difficulty. For these analyses, we average scores across all models and only distinguish translation directions. Detailed evaluation results per model, including breakdowns by sentence length and category, are reported in Appendix B.

Table 4 shows that the BERT and Likert scores decrease with sentence length, indicating worse semantic overlap, but the surface-level chrF++ metric shows the opposite trend. This indicates that surface-level character overlap does not always reliably correspond to perceived translation quality. Table 5 shows that imperative sentences achieve the highest scores across sentence categories, reflecting their structural simplicity. Idioms receive the lowest scores, indicating that models struggle with non-literal and culturally specific expressions in Fuzhounese. Patterns for sentence categories are largely the same across both translation directions. Overall, even under the best conditions, average scores remain far below those observed for high-resource languages.

4.4. Inter-annotator agreement

To assess the reliability of our Likert-scale ratings and gain insight into the difficulty of the rating task, we computed inter-annotator agreement for the ratings of translations generated for the three languages, split by the three different dimensions they were rated on. As our agreement metric, we use Krippendorff's α (ordinal, Krippendorff, 2004). We choose this metric as there are more than two raters (three raters) which excludes Cohen's kappa, and the data is ordinal, while Fleiss' kappa would treat it as nominal.

Table 6 shows that agreement is very high for Fuzhounese and high for Mandarin. For Cantonese, agreement is high for Adequacy but lower for Fluency and Validity ratings. This difference for Cantonese may be explained by the fact that raters are from an area where Cantonese isn't a standard language, so it is more difficult for raters to assess Validity in particular (e.g. was the correct register used). The exceptionally high agreement for Fuzhounese is likely due to the fact that it's easy to agree on low ratings for this task – a bad translation is clearly bad to any native speaker – and the lowest rating of 1 was very common for Fuzhounese.

Model	Cantonese		English		Mandarin	
	BERTScore	chrF++	BERTScore	chrF++	BERTScore	chrF++
Gpt-4o	0.297	6.66	0.305	14.45	0.251	5.55
Gpt-4o mini	0.225	4.10	0.302	12.49	0.197	4.02
Gemini-2.5 Flash	**0.508**	**17.08**	**0.507**	**30.59**	**0.496**	**20.23**
DeepSeek-v3.2	0.395	10.61	0.388	22.36	0.382	13.73
Qwen-2.5-72b	0.199	3.19	0.249	11.61	0.151	2.91

Table 2: Automatic evaluation results per model and language with Fuzhounese as the source language.

Model	Fuzhounese			Cantonese			Mandarin		
	Likert	chrF++	BERT	Likert	chrF++	BERT	Likert	chrF++	BERT
GPT-4o	1.18	3.96	–	**4.73**	**44.81**	**0.792**	**4.80**	**51.88**	**0.829**
GPT-4o mini	1.10	7.46	–	4.33	27.05	0.681	**4.80**	50.49	0.823
Gemini-2.5 Flash	**2.76**	**20.86**	–	4.68	41.68	0.778	4.76	48.08	0.810
DeepSeek-V3.2	2.07	12.66	–	4.70	40.97	0.764	4.78	48.26	0.808
Qwen-2.5-72B	1.09	8.06	–	4.17	29.86	0.712	4.72	47.83	0.818

Table 3: Translation performance into Fuzhounese, Cantonese and Mandarin from English sentences.

Length	Fzh Source		Fzh Target	
	BERT	chrF	Likert	chrF
Short (2-5)	**0.36**	12.54	**1.71**	9.66
Med (6-11)	0.29	10.60	1.60	11.32
Long (≥12)	0.28	**13.69**	1.50	**12.03**

Table 4: Evaluation results by sentence length for Fuzhounese as source and target language.

Category	Fzh Source		Fzh Target	
	BERT	chrF	Likert	chrF
Everyday	0.33	11.97	1.60	10.51
Grammar	0.35	13.16	1.76	10.98
Idioms	0.28	9.62	1.55	9.65
Imperatives	**0.38**	**16.14**	**1.88**	**12.75**
Narrative	0.30	11.91	1.59	11.06

Table 5: Evaluation results by sentence category for Fuzhounese as source and target language.

Overall, these results indicate that our Fuzhounese and Mandarin Likert ratings are reliable, and for Cantonese, at least translation adequacy is agreed upon.

4.5. Error Analysis

When Fuzhounese serves as the source language, failures most commonly take the form of explicit translation refusals. Models respond with statements indicating that they do not understand the sentence or are unable to translate it, declining to produce a translation. When Fuzhounese is the target language, the dominant failure mode is script mismatch. Instead of producing output in Bàng-uâ-cê romanization, models will generate translations

Language	Adequacy	Fluency	Validity
Fuzhounese	0.932	0.926	0.920
Cantonese	0.792	0.392	0.199
Mandarin	0.840	0.727	0.600

Table 6: Inter-rater agreement for generated translations across three dimensions and between three annotators. Values are Krippendorff's α (ordinal).

in Chinese characters. These outputs do not constitute written Fuzhounese in characters, but instead reflect a fallback to Mandarin translation. Although these outputs may be semantically related to the English source sentence, they are invalid under the task definition and are therefore annotated as translation failures.

Table 7 shows clearly distinct failure patterns across models. Qwen-2.5-72B and GPT-4o mini, which exhibit the highest number of explicit translation refusals, also consistently achieve the lowest BERT- and chrF++ scores for Fuzhounese-input translations 2. In contrast, GPT-4o, despite exhibiting 181 script mismatches for Fuzhounese-output translations, still attains higher Likert scores than several models with substantially fewer mismatches, indicating that its translations are generally of higher quality when the output is produced in the correct script (Table 3).

We observe further error types through manual examination of the data. To illustrate this, Table 8 presents a representative example of semantic distortion.

When Fuzhounese serves as the source language, model outputs are generally fluently written, but misrepresent the meaning of the input. Sim-

Model	Refusal	Script	N
DeepSeek-v3.2	4	56	60
Gemini-2.5 Flash	3	5	8
GPT-4o	9	**181**	**190**
GPT-4o mini	24	1	25
Qwen-2.5-72b	**48**	52	100

Table 7: Frequency of translation failures by model – translation refusals, script mismatches, totals.

Input	Gold	Model Output
I don't feel well	Nguài mâ su hŭk.	Nguāi sĭng-tāi mâ hō .

Table 8: Example of semantic distortion in English-to-Fuzhounese translation.

ilarly, when Fuzhounese is the target language, outputs may appear structurally plausible, yet fail to convey the intended meaning. While the latter does not happen as frequently, surface-level fluency may mask semantic errors in both cases.

Moreover, repetition and degeneration are also frequently observed across models. They are most prominent, however, when GPT-4o mini is tasked with English-Fuzhounese translation. In such cases, the model outputs repeated syllables or short strings without semantic content, often by generating the same token multiple times.

Finally, nonsensical outputs are also a common failure pattern. Translations with nonsensical output contain strings that resemble Bàng-uâ-cê romanization, but do not correspond to real Fuzhounese words. This renders the output unintelligible to native speakers.

5. Discussion

Looking at the results, it is clear that a systematic degradation in translation quality takes place as language resources decrease. This holds across models and evaluation methods, suggesting that current multilingual LLMs remain highly sensitive to data availability and orthographic standardization.

The consistently low performance scores for Fuzhounese-source tasks across all languages suggest that models do not necessarily benefit from the larger amount of English-based representations nor from cross-Sinitic mappings when processing Fuzhounese. Remarkably, translations into Cantonese did achieve higher scores than those into Mandarin, suggesting that the closer genealogical proximity between Fuzhounese and Cantonese has a more positive influence on translation quality than Mandarin's higher resource status in this setting. The overall low scores, however, indicate that similarity alone is insufficient without sufficient exposure in training data.

For English-source tasks, the contrast is sharper. Models generate high-quality Mandarin and Cantonese translations, but more often than not fail to produce valid Fuzhounese in Bàng-uâ-cê. Human ratings showed that many Fuzhounese outputs were judged as largely unintelligible, reflecting limited lexical knowledge and instability in romanized generation. While Gemini-2.5 Flash achieved higher Fuzhounese scores than others, it is still far below its Cantonese and Mandarin performance, indicating that even larger multilingual models struggle with severely under-resourced language varieties.

In our error analysis, the explicit refusals and script mismatches can be interpreted as responses to low model certainty under severe data scarcity: models either decline the task when comprehension is weak, or fall back to more familiar orthographic systems when uncertain over Bàng-uâ-cê generation. This interpretation is further supported by the observation that translation failures are mostly systematic responses to extreme data scarcity and unfamiliar orthographic constraints.

Lastly, the stratified evaluation shows that sentence length and type do have a modest effect on translation quality. These findings align with prior studies on machine translation and the results from HKCanto-Eval, where short directive sentences were also the easiest category and idiomatic expressions the hardest (Cheng et al., 2025; Isabelle et al., 2017).

5.1. Limitations

Despite these findings, our study has several methodological limitations. The dataset primarily consists of English source sentences. As they are manually translated into gold-standard Fuzhounese translations, the amount of Fuzhounese-specific grammatical constructions is likely minimal. The usage of natively Fuzhounese source sentences could have resulted in a broader range of Fuzhounese-specific structures, but this was not feasible due to the severe scarcity of written Fuzhounese resources.

Furthermore, for the same reason, the gold-standard Fuzhounese sentences are manually translated with the assistance of native and heritage speakers. This is a limitation as well, as it reduces reproducibility and verifiability, making it less ideal than gold standards derived from verified parallel written corpora.

Lastly, another limitation is that the human evaluation is conducted by only a small group of annotators. While all annotators are fluent in the evaluated languages and follow a standardized rubric, subjective judgment, especially when rating severely degraded or partially intelligible outputs, cannot be entirely eliminated. Two of the rated dimensions

had lower inter-annotator agreement for Cantonese, making our Cantonese ratings potentially less reliable.

Despite all of the previous points, this study provides a systematic and comparative analysis of multilingual LLM performance on a severely under-resourced Sinitic variety, offering insights into current model capabilities and failure modes.

6. Conclusion

We investigated how effectively multilingual large language models translate the severely under-resourced Fuzhounese language, using Cantonese and Mandarin as higher-resource baselines. To address this research question, we constructed a novel bidirectional translation benchmark and used it to evaluate five multilingual LLMs using automatic metrics and human judgments. As far as we are aware, this is the first study addressing Fuzhounese capabilities of language models.

Our results show that multilingual LLMs demonstrate inconsistent translation performance across Fuzhounese, Cantonese, and Mandarin when dealing with English-source translations. While models perform well on Mandarin and reasonably on Cantonese, translation quality deteriorates sharply for Fuzhounese across all evaluated LLMs. Gemini-2.5 Flash showed the best results out of the tested models, though its performance is still rather limited.

Furthermore, we observe that rather than affecting model performance, translation direction leads to different types of model-failure. Models frequently fail to generate valid Bàng-uâ-cê when translating into Fuzhounese, and translations are often refused or semantically distorted when translating from Fuzhounese. This indicates that neither direction of Fuzhounese translation is reliably supported in current multilingual LLMs, with both directions involving Fuzhounese exhibiting poor performance.

Ultimately, the results have shown that a broader resource bias is present in multilingual LLMs, with multilingual LLMs performing well on higher-resource languages but crumbling under the low-resource conditions of Fuzhounese. These findings show that current multilingual LLMs remain strongly constrained by data availability and do not generalize reliably to Fuzhounese.

6.1. Future Directions

Given the lack of attention for Fuzhounese in the NLP literature so far, future work can extend these findings in many directions. For example, future benchmarking efforts could include more Fuzhounese-native source sentences to better capture language-internal structures and discourse patterns that may not surface through English-based sentence design. Additionally, the benchmark could be expanded beyond translation to additional tasks, such as summarization, paraphrasing, question answering, or dialogue in Bàng-uâ-cê, which would simultaneously evaluate whether model failures generalize across tasks.

To address the performance gap, future studies could experiment with lightweight adaptation methods such as parameter-efficient fine-tuning (e.g., LoRA), using small amounts of curated Fuzhounese data to explore whether targeted adaptation could yield meaningful performance gains under extreme data scarcity.

7. Bibliographical References

Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. 2021. Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946.

Caleb C Baldwin. 1871. *A manual of the Foochow dialect*. Methodist Episcopal Mission Press.

David Prager Branner. 2000. *Problems in comparative Chinese dialectology: The classification of Miin and Hakka*, volume 123. Walter de Gruyter.

Matthew Y Chen. 2000. *Tone sandhi: Patterns across Chinese dialects*, volume 92. Cambridge University Press.

Tsz Chung Cheng, Chung Shing Cheng, Chaakming Lau, Eugene Lam, Wong Chun Yat, Hoi On Yu, and Cheuk Hei Chong. 2025. HKCanto-Eval: A benchmark for evaluating Cantonese language understanding and cultural comprehension in LLMs. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 1–11.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 8440–8451.

Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Dilip Venkatesh, Raj Dabre, Anoop

Kunchukuttan, and Mitesh M Khapra. 2025. Cross-lingual auto evaluation for assessing multilingual LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29297–29329.

San Duanmu. 2007. *The phonology of standard Chinese*. Oxford University Press.

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

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 33–41.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496.

Jiyue Jiang, Pengan Chen, Liheng Chen, Sheng Wang, Qinghang Bao, Lingpeng Kong, Yu Li, and Chuan Wu. 2025. How well do LLMs handle Cantonese? Benchmarking Cantonese capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4464–4505.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 6282–6293.

Klaus Krippendorff. 2004. *Content analysis: An introduction to its methodology*. Sage publications.

Deepak Kumar, Kshetrimayum Boynao Singh, and Asif Ekbal. 2025. Tackling low-resource NMT with instruction-tuned LLaMA: A study on Kokborok and Bodo. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1215–1221.

Shuheng Liu and Michael Best. 2025. A survey of NLP progress in Sino-Tibetan low-resource languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7804–7825.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90.

R. S. Maclay and C. C. Baldwin. 1929. *Dictionary of the Foochow Dialect*. Methodist Episcopal Mission Press. Accessed via Wikisource transcription.

Mukhammadsaid Mamasaidov, Azizullah Aral, Abror Shopulatov, and Mironshoh Inomjonov. 2025. Filling the gap for Uzbek: Creating translation resources for Southern Uzbek. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1081–1087.

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

Jerry Norman. 1991. The Mǐn dialects in historical perspective. *Journal of Chinese Linguistics Monograph Series*, (3):323–358.

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

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

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1).

S Robert Ramsey. 1989. *The languages of China*. Princeton University Press.

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The ACL OCL corpus: Advancing open science in computational linguistics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10348–10361.

Nadav Shani and Ali Basirat. 2025. Language dominance in multilingual large language models. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 137–148.

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.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A. Prompt Design

This appendix provides the full set of prompt templates used in the evaluation. The prompts specify the source and target languages, as well as output constraints, and are kept consistent across all models and translation directions to ensure comparability and minimize prompt-engineering effects.

HKCanto-Eval (Cheng et al., 2025) makes use of structured translation prompts that explicitly specify the source language, target language, and output constraints. Consequently, each prompt in this study likewise specifies these elements. This design ensures comparability across models, languages, and translation directions while minimizing prompt-engineering effects.

A.1. Fuzhounese-to-Target Language Prompts

For translations from Fuzhounese (in Bàng-uâ-cê) into English, Standard Mandarin, and Cantonese, the following prompt templates are used. <Text> represents the source sentence inserted into the prompt during evaluation:

Fuzhounese to English

> Fuzhounese (Bàng-uâ-cê) Text to Translate:
> <Text>
>
> Ensure the translation is accurate and natural,
> preserving the original meaning and using concise and fluent English expression.
>
> ONLY RETURN THE TRANSLATION. DO NOT EXPLAIN.

Fuzhounese to Standard Mandarin

> Fuzhounese (Bàng-uâ-cê) Text to Translate:
> <Text>
>
> Ensure the translation is accurate and natural,
> preserving the original meaning and using concise and fluent Standard Mandarin expression.
>
> ONLY RETURN THE TRANSLATION. DO NOT EXPLAIN.

Fuzhounese to Written Cantonese

> Fuzhounese (Bàng-uâ-cê) Text to Translate:
> <Text>
>
> Ensure the translation is accurate and natural,
> preserving the original meaning and using
> concise and fluent written Cantonese expression.
>
> ONLY RETURN THE TRANSLATION. DO NOT EXPLAIN.

A.2. English-to-Target Language Prompts

For translations from English into Fuzhounese (in Bàng-uâ-cê), Standard Mandarin, and Cantonese, the following prompt templates are used. <Text> represents the source sentence inserted into the prompt during evaluation.

English to Fuzhounese

> English Text to Translate:
> <Text>
>
> Ensure the translation is accurate and natural,
> preserving the original meaning and using concise and fluent Fuzhounese romanization
> (Bàng-uâ-cê) expression.
>
> ONLY RETURN THE TRANSLATION. DO NOT EXPLAIN.

English to Standard Mandarin

> English Text to Translate:
> <Text>
>
> Ensure the translation is accurate and natural,
> preserving the original meaning and using concise and fluent Standard Mandarin expression.
>
> ONLY RETURN THE TRANSLATION.
> DO NOT EXPLAIN.

English to Written Cantonese

> English Text to Translate:
> <Text>
>
> Ensure the translation is accurate and natural,
> preserving the original meaning and using concise and fluent written Cantonese expression.
>
> ONLY RETURN THE TRANSLATION.
> DO NOT EXPLAIN.

This study maintains consistent prompt structures across all translation directions and target languages, ensuring that observed performance differences can be attributed to model capabilities rather than prompt variation.

B. Extended Evaluation Results

This appendix presents additional detailed evaluation results that complement the main analysis in the paper. Specifically, it includes model performance broken down by sentence length and sentence category, providing a more fine-grained view of how different linguistic factors affect translation quality.

B.1. Performance per Sentence Length

Table 9 reports the translation performance per model across different sentence lengths (short, medium, and long). For each model, scores are shown separately for Fuzhounese input translations (BERTScore and chrF++) and Fuzhounese output translations (Likert ratings and chrF++). This table illustrates how sentence length affects performance per model.

Model	Length	Fuzhounese Input		Fuzhounese Output	
		BERTScore	chrF++	Likert	chrF++
GPT-4o	Short	0.345	10.02	1.31	4.32
	Medium	0.241	7.60	1.07	3.21
	Long	0.185	8.31	1.04	4.71
GPT-4o mini	Short	0.293	7.43	1.19	6.96
	Medium	0.204	6.03	1.03	7.80
	Long	0.157	7.16	1.01	8.35
Gemini-2.5 Flash	Short	0.506	23.23	2.70	18.21
	Medium	0.481	20.09	2.86	23.27
	Long	0.556	27.52	2.69	23.87
DeepSeek-V3.2	Short	0.414	16.18	2.20	11.05
	Medium	0.359	13.78	2.02	14.30
	Long	0.380	18.54	1.76	13.96
Qwen-2.5-72B	Short	0.246	5.79	1.16	7.75
	Medium	0.165	5.60	1.03	8.02
	Long	0.124	6.91	1.01	9.27

Table 9: Effect of sentence length on translation quality per model.

B.2. Performance per Sentence Category

Table 10 presents the translation performance per model across different sentence categories. The table contrasts Fuzhounese input and output performance using human and automatic evaluation metrics, highlighting the influence of different linguistic constructions on multilingual LLM behavior.

Model	Category	Fuzhounese Input		Fuzhounese Output	
		BERTScore	chrF++	Likert	chrF++
GPT-4o	Everyday	0.293	8.95	1.10	4.08
	Grammar	0.305	10.21	1.28	4.59
	Idioms	0.245	7.19	1.11	2.26
	Imperatives	0.362	10.21	1.67	7.79
	Narrative	0.248	7.82	1.24	3.81
GPT-4o-mini	Everyday	0.246	6.95	1.12	7.40
	Grammar	0.269	7.73	1.14	7.23
	Idioms	0.223	5.99	1.04	7.23
	Imperatives	0.241	6.68	1.12	8.27
	Narrative	0.180	5.93	1.05	8.52
Gemini-2.5 Flash	Everyday	0.515	22.75	2.67	20.07
	Grammar	0.527	23.71	2.96	22.11
	Idioms	0.422	17.71	2.59	19.40
	Imperatives	0.549	32.02	3.19	25.15
	Narrative	0.542	25.81	2.73	21.99
DeepSeek-V3.2	Everyday	0.386	15.30	1.99	12.62
	Grammar	0.445	18.03	2.36	13.84
	Idioms	0.312	11.76	1.96	11.15
	Imperatives	0.524	26.15	1.83	14.37
	Narrative	0.349	14.06	1.91	11.89
Qwen-2.5-72B	Everyday	0.198	5.86	1.09	8.37
	Grammar	0.222	6.26	1.07	7.15
	Idioms	0.186	5.36	1.05	8.20
	Imperatives	0.240	5.62	1.60	8.15
	Narrative	0.156	5.90	1.02	9.08

Table 10: Effect of sentence category on translation quality per model.