

# Enhancing Taiwanese Hokkien Dual Translation by Exploring and Standardizing of Four Writing Systems

Bo-Han Lu<sup>1</sup>, Yi-Hsuan Lin<sup>1</sup>, En-Shiun Annie Lee<sup>2,3</sup>, Richard Tzong-Han Tsai<sup>1,4</sup>

<sup>1</sup>National Central University, Taiwan <sup>2</sup>University of Toronto <sup>3</sup>Ontario Tech University <sup>4</sup>Academia Sinica, Taiwan



國立中央大學  
National Central University



UNIVERSITY OF  
TORONTO

 OntarioTech  
UNIVERSITY



中央研究院  
ACADEMIA SINICA

# Outline

- Motivation
- Background
- Datasets
- Experiment Setup
- Experiment Results
- Conclusion

# Outline

- **Motivation**
- Background
- Datasets
- Experiment Setup
- Experiment Results
- Conclusion

# Taiwanese Hokkien

- A unique variant of the Southern Min dialects, influenced by:
  - Indigenous languages
  - Dutch and Japanese colonial legacies
- Despite being spoken by over 30% of Taiwan's population, Taiwanese Hokkien faces significant challenges:
  - Decreasing proficiency among younger generations.
  - Predominantly used in spoken form, with limited written resources.

# LLM Training Data Highly Concentrated on English

- Most Large Language Models (LLMs) are trained predominantly on English data, leading to subpar performance in low-resource languages.

Language	Percent	Language	Percent
en	89.70%	uk	0.07%
unknown	8.38%	ko	0.06%
de	0.17%	ca	0.04%
fr	0.16%	sr	0.04%
sv	0.15%	id	0.03%
zh	0.13%	cs	0.03%
es	0.13%	fi	0.03%
ru	0.13%	hu	0.03%
nl	0.12%	no	0.03%
it	0.11%	ro	0.03%
ja	0.10%	bg	0.02%
pl	0.09%	da	0.02%
pt	0.09%	sl	0.01%
vi	0.08%	hr	0.01%

Source: Llama 2: Open Foundation and Fine-Tuned Chat Models

# Our Approach

- We developed a machine translation model for bidirectional translation between Taiwanese Hokkien and both English and Mandarin Chinese using large language models.
- Our research investigates strategies to enhance machine translation performance by utilizing on limited language resources within the LLM training paradigm.

# Outline

- Motivation
- **Background**
- Datasets
- Experiment Setup
- Experiment Results
- Conclusion

# Taiwanese Hokkien Writing Systems

- There are four writing systems in Taiwanese Hokkien

Language	Writing	Abbreviation	Script	Example
Hokkien	Tâi-lô	TL	Latin	Tsit kuí kang lóng sī pháinn-thinn.
Hokkien	Pèh-ōe-jī	POJ	Latin	Chit kúi kang lóng sī pháin <sup>n</sup> -thin <sup>n</sup> .
Hokkien	Hàn-lô	HL	Hybrid	這幾工 lóng 是歹天。
Hokkien	Han	HAN	Chinese Character	這幾工攏是歹天。
Mandarin	Chinese	ZH	Chinese Character	這幾天都是壞天氣。
English		EN	Latin	It's been bad weather these days.

# Taiwanese Hokkien Writing Systems

- Previous research most focused on a single script system without exploring the potential benefits of integrating all available corpora.
- Given the shared vocabulary between Chinese and Hokkien Han characters, we base our model on Traditional Chinese LLM for continued pre-training in Taiwanese Hokkien.

# Outline

- Motivation
- Background
- **Datasets**
- Experiment Setup
- Experiment Results
- Conclusion

# Datasets – Training Data

- Monolingual Corpora
  - Recitation contests, song lyrics, religious texts, Wikipedia entries, textbooks, TV program subtitle and news articles.
  - Totaling 78MB.
- Parallel Dataset (Augmented the Hokkien-English dataset by translating Chinese to English using GPT-3.5-turbo)
  - 15,195 example sentences from the Hokkien dictionary, including ZH, EN, POJ and HAN.
  - 2,677 Technical terms, including ZH, EN, POJ and HAN.
  - 17,872 Religious texts (only HL and POJ)

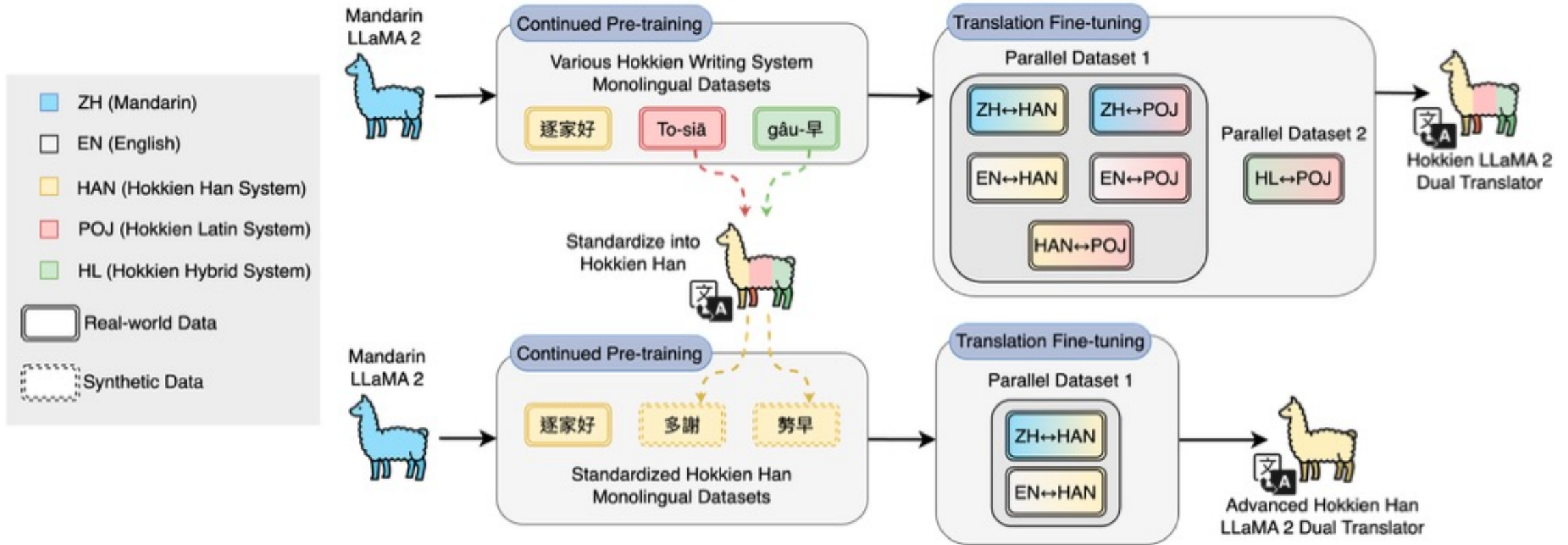
# Datasets – Testing Data

- iCorpus-100
  - iCorpus is a parallel news dataset in Hokkien (HAN, POJ) and ZH.
  - We augmented it by translating ZH to EN.
  - To address issues with incorrect HAN characters:
    1. Instances ordered by term frequency in HAN for difficulty
    2. Top 100 most frequent instances sampled
    3. Manual lexicon correction based on official orthography
- TAT (Taiwanese Across Taiwan) speech recognition dataset.
  - Using 2,661 parallel sentences
  - Focused on ZH-HAN and EN-HAN language pairs

# Outline

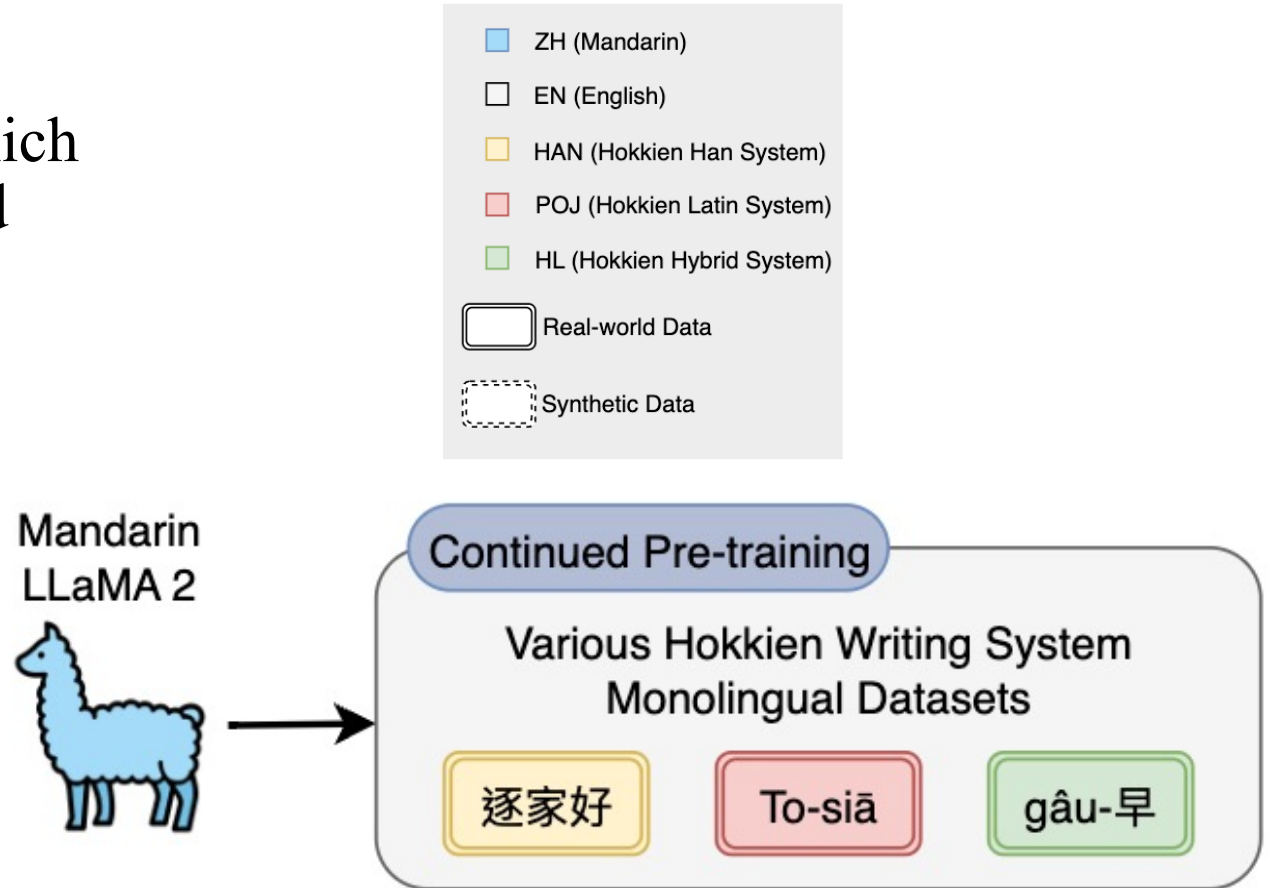
- Motivation
- Background
- Datasets
- **Experiment Setup**
- Experiment Results
- Conclusion

# Training Process - Overview



# Training Process – Continued Pre-training

- Base Model
  - Leveraged LLaMA 2 model, which undergone continued pre-trained on Mandarin Chinese corpus.<sup>1</sup>
- Extend Model Vocabulary
  - 130 Han characters
  - 1876 Latin scripts

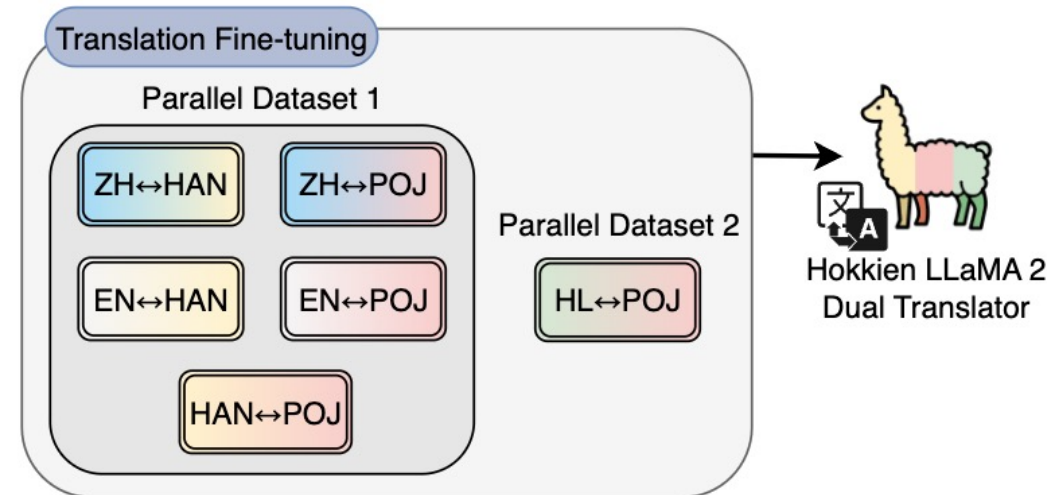
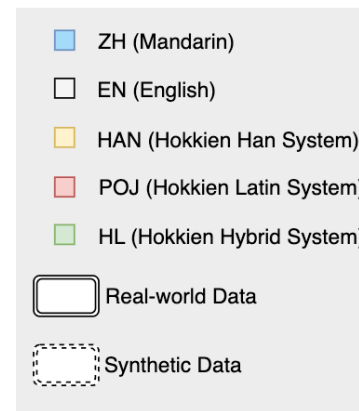


<sup>1</sup>TAIDE: <https://taide.tw/index>

# Training Process – Translation Fine-tuning

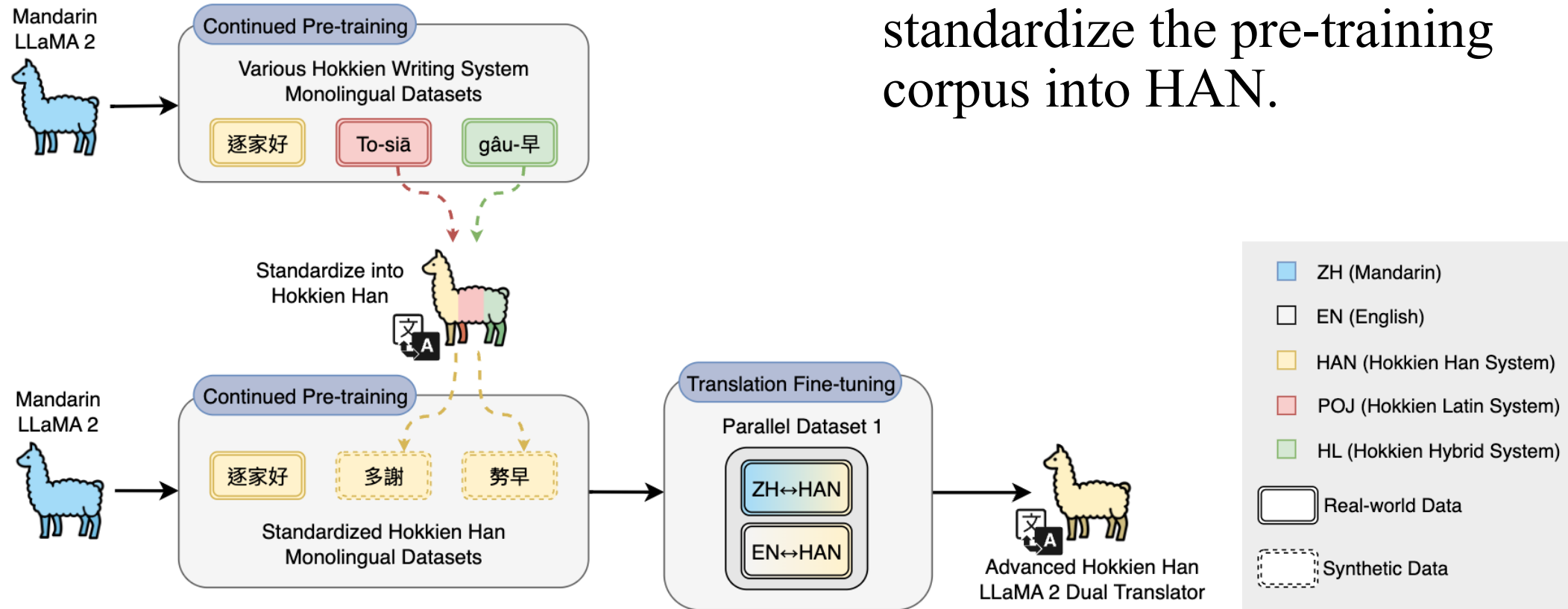
- Fine-tuning prompt template:

```
[TRANS]
{source_sentence}
[/TRANS]
[/{target_language}]
{target_sentence}
[//{target_language}]
```



# Training Process – Pre-training Corpus Script-Standardization

- Using the translator to further standardize the pre-training corpus into HAN.



# Evaluation Methods

- Metrics
  - BLEU, chrF++: Traditional n-gram based translation metrics.
  - GPT4-based evaluation: Score model output with reference by GPT-4, score range if from 0~100.

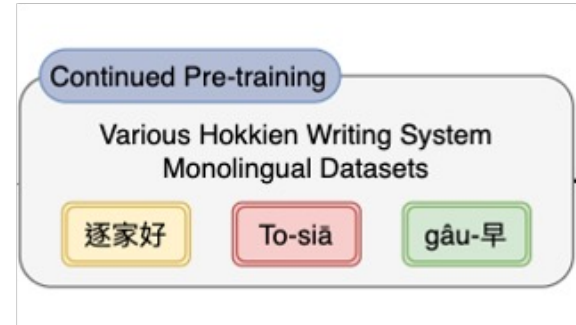
		Target Language	
		ZH, EN (Familiar to GPT-4)	Hokkien (Not familiar to GPT-4)
Source Language	ZH, EN (Familiar to GPT-4)	No this case.	Employing back-translation into the source language for comparison with the original sentence.
	Hokkien (Not familiar to GPT-4)	Direct comparison with the reference sentence.	Our metrics are not applicable, chrF++ is sufficiently effective in this case.

# Outline

- Motivation
- Background
- Datasets
- Experiment Setup
- **Experiment Results**
- Conclusion

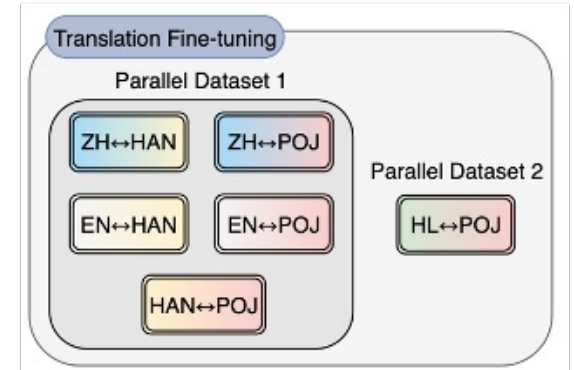
# Continued Pre-training Stage

Continued Pre-train Data and Vocabulary Extension	BLEU	chrF++	GPT-4 Score*	GPT-4 Accuracy	BLEU	chrF++	GPT-4 Score*	GPT-4 Accuracy	BLEU	chrF++	GPT-4 Score*	GPT-4 Accuracy
	ZH↔HAN				ZH↔POJ				HAN↔POJ			
LLaMA-2 (EN)	23.84	26.21	41.45	8.5	6.41	22.60	14.10	0.5	19.20	39.35	-	-
LLaMA-2 (ZH)	30.12	32.67	65.68	42.5	8.68	23.06	22.20	2.0	24.93	41.71	-	-
+ HAN	30.22	32.33	70.68	52.0	9.94	23.38	21.95	3.0	24.31	41.18	-	-
+ ALL	<u>31.99</u>	<u>33.90</u>	<u>73.85</u>	<u>59.0</u>	<u>15.36</u>	<u>29.00</u>	<u>39.15</u>	<u>10.5</u>	<u>28.97</u>	<u>44.84</u>	-	-
+ ALL ( <i>EXT_VOCAB</i> )	31.96	33.69	72.60	53.5	14.38	<u>29.34</u>	37.03	8.0	26.53	44.17	-	-
	EN↔HAN				EN↔POJ							
LLaMA-2 (EN)	9.52	26.52	53.48	20.5	1.05	18.82	5.58	0.0				
LLaMA-2 (ZH)	12.87	29.76	59.45	23.0	1.77	20.72	11.18	0.0				
+ HAN	14.49	30.80	64.53	36.5	1.57	20.94	8.85	0.0				
+ ALL	<u>15.64</u>	<u>31.77</u>	<u>66.28</u>	<u>39.0</u>	<u>4.01</u>	<u>25.21</u>	<u>22.13</u>	<u>2.0</u>				
+ ALL ( <i>EXT_VOCAB</i> )	14.74	30.98	62.95	33.5	3.65	24.89	<u>24.30</u>	<u>2.0</u>				

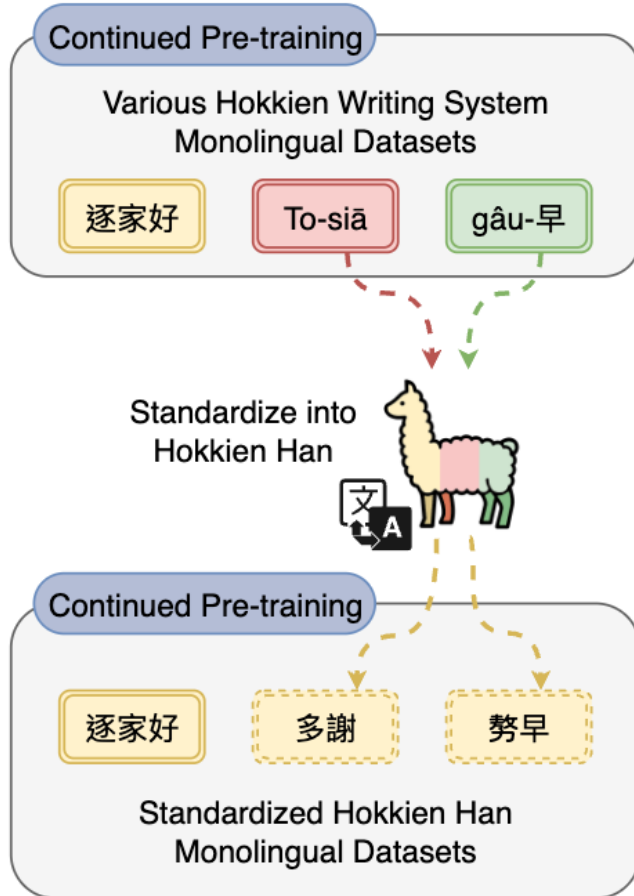


# Translation Fine-tuning Stage

Fine-tuning Data	BLEU	chrF++	GPT-4 Score*	GPT-4 Accuracy	BLEU	chrF++	GPT-4 Score*	GPT-4 Accuracy	BLEU	chrF++	GPT-4 Score*	GPT-4 Accuracy
	<b>ZH↔HAN</b>				<b>ZH↔POJ</b>				<b>HAN↔POJ</b>			
BASELINE (ICL)	16.79 <sup>†</sup>	20.57 <sup>†</sup>	64.78 <sup>†</sup>	48.0 <sup>†</sup>	0.71 <sup>†</sup>	3.75 <sup>†</sup>	10.55 <sup>†</sup>	3.5 <sup>†</sup>	0.53 <sup>†</sup>	4.25 <sup>†</sup>	-	-
ZH, HAN	29.02	31.24	76.70	63.5	1.36 <sup>†</sup>	4.03 <sup>†</sup>	7.70 <sup>†</sup>	0.0 <sup>†</sup>	1.48 <sup>†</sup>	4.20 <sup>†</sup>	-	-
ZH, HAN, EN	29.39	31.67	<u>76.88</u>	<u>66.5</u>	2.41 <sup>†</sup>	4.64 <sup>†</sup>	7.95 <sup>†</sup>	1.0 <sup>†</sup>	2.95 <sup>†</sup>	5.20 <sup>†</sup>	-	-
ZH, HAN, EN, POJ	28.47	30.75	75.63	57.5	11.61	24.25	33.03	8.5	18.25	33.23	-	-
ZH, HAN, EN, POJ, HL	<u>31.99</u>	<u>33.90</u>	73.85	59.0	<u>15.36</u>	<u>29.00</u>	<u>39.15</u>	<u>10.5</u>	<u>28.97</u>	<u>44.84</u>	-	-
	<b>EN↔HAN</b>				<b>EN↔POJ</b>							
BASELINE (ICL)	6.06 <sup>†</sup>	18.78 <sup>†</sup>	42.70 <sup>†</sup>	14.5 <sup>†</sup>	0.11 <sup>†</sup>	5.78 <sup>†</sup>	1.95 <sup>†</sup>	0.0 <sup>†</sup>				
ZH, HAN	6.12 <sup>†</sup>	9.45 <sup>†</sup>	29.00 <sup>†</sup>	21.0 <sup>†</sup>	0.00 <sup>†</sup>	0.37 <sup>†</sup>	6.90 <sup>†</sup>	0.0				
ZH, HAN, EN	17.04	<u>32.97</u>	<u>71.20</u>	<u>47.0</u>	0.79 <sup>†</sup>	9.73 <sup>†</sup>	5.20 <sup>†</sup>	0.0				
ZH, HAN, EN, POJ	16.24	31.99	68.33	41.5	3.58	23.49	21.43	1.0				
ZH, HAN, EN, POJ, HL	15.64	31.77	66.28	39.0	<u>4.01</u>	<u>25.21</u>	<u>22.13</u>	<u>2.0</u>				



# Pre-training Corpus Script-Standardization



Continued Pre-train Data	BLEU	chrF++	GPT-4	
			Score*	Accuracy
ZH↔HAN				
CP_HAN (w/o Standardized Data)	32.92	34.86	74.27	61.65
CP_ALL	33.67	35.97	<u>76.04</u>	<u>65.26</u>
CP_HAN (w/ Standardized Data)	<u>34.67</u>	<u>35.98</u>	76.03	64.66
EN↔HAN				
CP_HAN (w/o Standardized Data)	22.27	35.00	65.69	46.03
CP_ALL	22.24	34.99	66.33	46.41
CP_HAN (w/ Standardized Data)	<u>22.68</u>	<u>35.61</u>	<u>68.22</u>	<u>49.68</u>

# Outline

- Motivation
- Background
- Datasets
- Experiment Setup
- Experiment Results
- **Conclusion**

# Conclusion

## 1. Continued Pre-training

- Substantial improvements in Taiwanese Hokkien training using LLM pre-trained on Chinese.
- Best translation results from pre-training across all writing systems.

## 2. Vocabulary Extension

- Limited impact on translation quality due to corpus scarcity.

## 3. Translation Supervised Fine-tuning

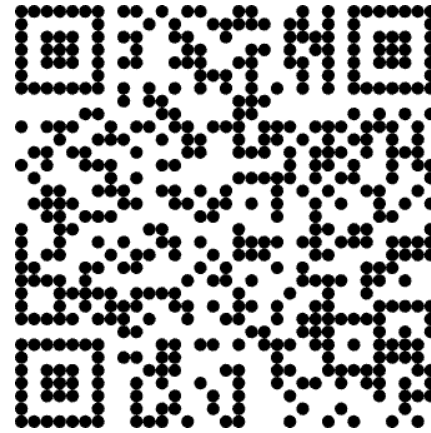
- Best results for HAN translation when training with parallel data in ZH-HAN and EN-HAN.
- Incorporating Latin script parallel data proves ineffective.

## 4. Pre-training Corpus Script-Standardization

- Standardizing to HAN can yield benefits on HAN translation.

# Thank you!

Paper Link



Model Link



**Bo-Han Lu**

qwqsqx30@gmail.com



**Yi-Hsuan Lin**

109502543@cc.ncu.edu.tw



**En-Shiun Annie Lee**

annie.lee@cs.toronto.edu



**Richard Tzong-Han Tsai**

thtsai@g.ncu.edu.tw

Professor Annie Lee is recruiting graduate students, please check out more at her website <https://www.cs.toronto.edu/~ealee/>