White Paper

Evaluation of Translation Ability of Language Models for Education-related Communication

Dipankar Srirag[1] Aditya Joshi[1] Caroline Thompsett[2] Dan Hart[2]
[1] University of New South Wales, Sydney, Australia
[2] NSW Department of Education, Sydney, Australia
aditya.joshi, dipankar.srirag@unsw.edu.au, caroline.thompsett, dan.hart2@det.nsw.edu.au

Contents

Executive Summary

The goal of this project is to assist the NSWDepartment of Education in achieving robust multilingual communication within educational settings through the deployment of Large Language Models (LLMs) (a.k.a., language models). For various communications of interest (emails, conversations, etc.), we were provided a dataset of LLM-generated translations from their new tool named NSWEduChat, alongside manually post-edited translations across multiple domains. Utilising this data provided by NSW Department of Education, we evaluated the performance of language models tasked with translating education-related communications across the 22 languages. We assessed the provided translations against both sentence-level and paragraph-level post-edited translations using established metrics including chrF, BLEU, and ROUGE. Due to restricted availability of post-edited translations, we also performed a reference-less evaluation via multilingual sentence embeddings. We investigated recurrent translation issues for Hindi and Khmer. Finally, we conducted a cost–accuracy analysis, grouping translations into three categories based on a custom metric to determine the viability of automated translation for each language and use case. Our findings shed light on the expected performance of current LLMs for the specific use case of the Department, and provide a methodology for similar evaluations in the future. The code is available at: https://github.com/unswnlp/det-eval-mt/

Acknowledgments

The authors of this report would like to acknowledge the Traditional custodians of the lands on which this report has been written, reviewed and produced. We pay our respects to their Elders past, present and future.

This research was funded by the NSW Department of Education and conducted in the School of Computer Science and Engineering at UNSW Sydney. We thank Multicultural NSW for their assistance in creating in post-edited translations.

Contact

For further information, please contact Dr. Aditya Joshi or Dan Hart.

1 Introduction

The NSW Department of Education (DET) oversees public schools across New South Wales. It is interested in AI-based products to provide administrative assistance and maintaining data security through department-owned, secure GenAI tools. NSWEduChat is a secure, department-owned generative AI tool designed specifically for the NSW education environment, ensuring privacy and equity for all users. It exists to support staff and students by providing curriculum-aligned content, easing workload pressures, and helping build AI skills in a safe, school-appropriate setting. Running in a Sydney-based cloud environment, NSWEduChat combines several algorithms and large language models (LLMs) enhancing them using the NSW curriculum and policies to generate tailored educational material while keeping data secure.

DET identified a use-case for NSWEduChat that can allow teachers to translate communications to students. Reliable translations using NSWEduChat can help educators communicate effectively with multicultural communities. In this project, we evaluate multilingual communication within the New South Wales Department of Education's topics of interest. We received datasets comprising English source texts, LLM-generated translations (as provided by EduChat, the tool used to create the translations), and human post-edited references. We examined the translation performance across 22 languages by utilising both sentence-level and paragraph-level post-edited translations as reference points. For cases lacking dedicated reference texts, we computed sentence similarity using NSWEduChat translations. We employed established evaluation metrics and performed robust error analysis, further informing our cost–accuracy analysis to guide the deployment of translation systems.

This report describes our findings from the project. Section 2 presents an exploratory data analysis of the dataset provided to us. Section 3 discusses the evaluation methodologies used for reference-based, reference-less and back-translation-based evaluation. Section 4 outlines the qualitative results, while Section 5 discusses a two-tier error analysis. Acknowledging the costs associated with proprietary LLMs, we present a cost-accuracy analysis in Section 6. Section 6 concludes the report and provides pointers for future work.

2 Dataset

The dataset consists of 60 LLM-generated translations and 6 post-edited references spanning five distinct domains: Emails, Circulars, Conversations, Lessons, and Miscellaneous texts. The majority of translations pertain to email communications. Our preliminary analysis revealed issues including missing or incomplete translations (notably for Swahili) and a scarcity of post-edited references for languages such as Korean and Tamil. Table 2.1 summarises the dataset statistics, while Figure 2.1 displays the average word count per language as computed by a basic whitespace tokeniser.

Note

The dataset is small relative to typical translation evaluation datasets. This can limit the reliability of statistical evaluations.

Domain	NSWEduChat Translations	Post-edit Pairs
Circulars	10	1
Conversations	10	1
Emails	20	2
Lessons	10	1
Miscellaneous	10	1
Total	60	6

Table 2.1: Dataset statistics grouped by domain.

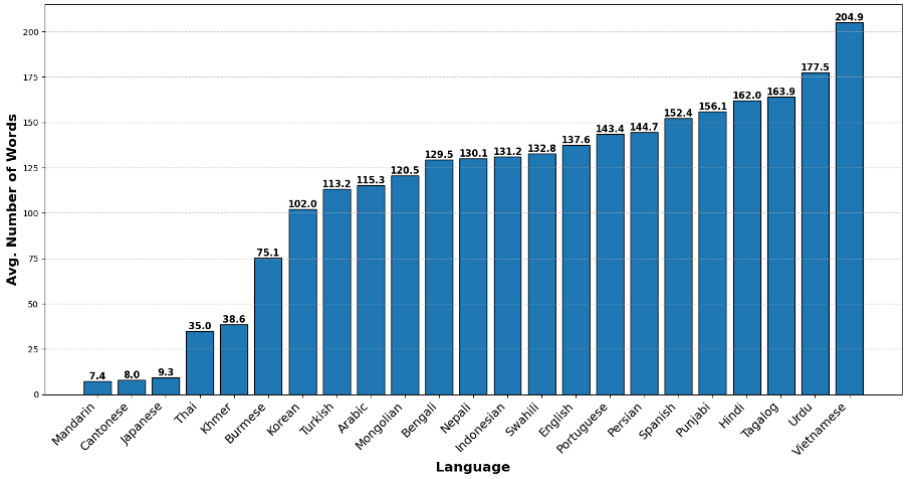

Figure 2.1: Average word count per language.

3 Evaluation Methodology

Reference-based evaluation is when a reference translation and a generated translation are available. We utilised established machine translation metrics to compare candidate translations against post-edited references. The post-edited references are assumed to be accurate.

chrF We computed the character-level F-score (chrF; Popović, 2015) as

$$\text{chrF} = \frac{2 \cdot \text{chrP} \cdot \text{chrR}}{\text{chrP} + \text{chrR}};$$

where *chrP* is the percentage of hypothesis character n-grams present in the reference and *chrR* is the percentage of reference character n-grams present in the hypothesis.

BLEU We measured Bilingual Evaluation Understudy (BLEU; Papineni et al., 2002) using

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right);$$

where p_n denotes the n-gram precisions, w_n are weights summing to 1, and BP is the brevity penalty defined by

$$\text{BP} = \begin{cases} 1, & \text{if } c > r, \\ \exp\left(1 - \frac{r}{c}\right), & \text{if } c \leq r; \end{cases}$$

with c and r representing the lengths of the candidate and reference texts respectively.

ROUGE We computed Recall-Oriented Understudy for Gisting Evaluation (ROUGE; Lin, 2004) as

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{reference}\}} \sum_{\text{gram}_n \in S} \text{count\_match}(\text{gram}_n)}{\sum_{S \in \{\text{reference}\}} \sum_{\text{gram}_n \in S} \text{count}(\text{gram}_n)};$$

where gram_n refers to the n-grams and $\text{count\_match}(\text{gram}_n)$ denotes the maximum number of overlapping n-grams.

For cases without dedicated reference texts, we utilised a reference-less evaluation based on sentence similarity. We computed the cosine similarity between sentence

embeddings of the translated and original English texts using the LaBSE Feng et al. (2022) model, which supports 109 languages within a shared vector space.

Note

Evaluation metrics provide statistical techniques to compare machine translations against human-edited text. They serve as indicators but do not capture all nuances, such as style and cultural context.

To provide further insight into translations with the poorest performance, we used GPT-4 (OpenAI Team, 2024) to back-translate them into English. We then compared the back-translated content with the source to identify significant meaning shifts, omissions, or additions. While back-translation cannot always capture subtle grammatical or stylistic deficits, it helps highlight major semantic divergences and clarifies which segments are most error-prone.

4 Results and Discussion

4.1 Reference-based Evaluation

Paragraph-level Performance: We report aggregate paragraph-level evaluation scores in Table 4.1. We observe that Indonesian translations achieve the highest overall performance ($\mu = 0.89$), while Khmer translations require substantial post-editing ($\mu = 0.35$).

Sentence-level Performance: Table 4.2 presents the evaluation for sentence-level translations. We report that Swahili and Indonesian translations perform strongly (with μ of 0.91 and 0.90, respectively), whereas Khmer translations again exhibit low performance.

Table 4.3 reports the Pearson correlation coefficients between paragraph-level and sentence-level scores across the evaluated metrics. The strong correlations (ranging from 0.82 to 0.90, all significant at $p < 0.05$) confirm the consistency of our evaluation measures.

4.2 Reference-less Evaluation

We computed sentence similarity scores between the translations and the original English texts using the cosine similarity of LaBSE embeddings. Table 4.4 presents these scores grouped by domain. We report high similarity scores (typically exceeding 0.90 for Arabic, Cantonese, Mandarin, and Spanish), which underscore the model's effectiveness in preserving semantic content. Nonetheless, we note that these scores do not fully capture syntactic variations.

Language	chrF	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	μ
Indonesian	0.93	0.85	0.92	0.85	0.91	0.89
Turkish	0.90	0.79	0.92	0.85	0.91	0.87
Tagalog	0.86	0.75	0.86	0.74	0.84	0.81
Portuguese	0.85	0.73	0.85	0.74	0.84	0.80
Spanish	0.82	0.64	0.84	0.73	0.84	0.77
Vietnamese	0.76	0.66	0.84	0.75	0.81	0.76
Swahili	0.71	0.58	0.81	0.78	0.81	0.74
Cantonese	0.78	0.60	0.66	0.67	0.67	0.68
Hindi	0.83	0.72	0.60	0.48	0.55	0.64
Mandarin	0.79	0.35	0.66	0.67	0.67	0.63
Thai	0.78	0.36	0.64	0.63	0.64	0.61
Japanese	0.70	0.00	0.83	0.65	0.82	0.60
Urdu	0.76	0.60	0.55	0.44	0.50	0.57
Mongolian	0.70	0.49	0.62	0.50	0.55	0.57
Arabic	0.78	0.60	0.47	0.45	0.48	0.56
Punjabi	0.70	0.53	0.56	0.48	0.53	0.56
Bengali	0.88	0.76	0.16	0.16	0.16	0.42
Persian	0.85	0.74	0.16	0.16	0.16	0.41
Nepali	0.81	0.63	0.14	0.11	0.14	0.37
Burmese	0.58	0.17	0.38	0.27	0.37	0.35
Khmer	0.55	0.16	0.36	0.31	0.37	0.35

Table 4.1: Paragraph-level evaluation metrics comparing LLM-generated translations with human post-edited references. The overall mean, μ, is computed across all metrics.

Language	chrF	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	μ
Swahili	0.93	0.88	0.95	0.83	0.95	0.91
Indonesian	0.92	0.83	0.94	0.86	0.94	0.90
Turkish	0.89	0.81	0.92	0.83	0.92	0.87
Vietnamese	0.77	0.73	0.87	0.82	0.87	0.81
Tagalog	0.85	0.74	0.83	0.70	0.82	0.79
Portuguese	0.84	0.72	0.83	0.71	0.83	0.79
Spanish	0.82	0.63	0.84	0.66	0.84	0.76
Cantonese	0.77	0.78	0.40	0.31	0.40	0.53
Hindi	0.82	0.69	0.40	0.17	0.39	0.49
Mandarin	0.70	0.71	0.35	0.16	0.35	0.45
Persian	0.89	0.79	0.20	0.07	0.21	0.43
Thai	0.77	0.47	0.38	0.16	0.39	0.43
Urdu	0.74	0.62	0.36	0.08	0.35	0.43
Bengali	0.88	0.76	0.18	0.08	0.18	0.42
Arabic	0.77	0.60	0.32	0.08	0.33	0.42
Punjabi	0.70	0.57	0.34	0.15	0.33	0.42
Mongolian	0.67	0.47	0.39	0.12	0.38	0.41
Nepali	0.80	0.63	0.12	0.04	0.12	0.34
Japanese	0.66	0.00	0.40	0.23	0.39	0.34
Burmese	0.55	0.20	0.21	0.11	0.21	0.26
Khmer	0.53	0.23	0.21	0.09	0.21	0.25

Table 4.2: Sentence-level evaluation metrics for LLM-generated translations compared with human post-edited references. The overall mean, μ, is computed across all metrics.

Metric	r
chrF	0.86
BLEU	0.90
ROUGE-1	0.87
ROUGE-2	0.82
ROUGE-L	0.87

Table 4.3: Pearson correlation coefficients between paragraph- and sentence-level metrics.

Language	Circulars	Conversations	Emails	Lessons	Miscellaneous	μ
Spanish	0.97	0.94	0.95	0.94	0.95	0.95
Arabic	0.96	0.92	0.95	0.95	0.95	0.95
Mongolian	0.94	0.95	0.93	0.95	0.94	0.95
Persian	0.96	0.93	0.95	0.94	0.94	0.95
Korean	0.96	0.95	0.92	0.96	0.95	0.95
Urdu	0.96	0.96	0.93	0.95	0.91	0.95
Punjabi	0.94	0.94	0.93	0.94	0.93	0.94
Cantonese	0.90	0.93	0.92	0.93	0.91	0.92
Khmer	0.91	0.92	0.91	0.94	0.92	0.92
Japanese	0.90	0.93	0.90	0.92	0.94	0.92
Portuguese	0.96	0.93	0.91	0.94	0.85	0.92
Nepali	0.91	0.93	0.88	0.93	0.93	0.92
Swahili	0.92	0.92	0.90	0.91	0.91	0.91
Bengali	0.89	0.93	0.86	0.91	0.94	0.91
Tagalog	0.88	0.93	0.88	0.89	0.93	0.90
Mandarin	0.89	0.91	0.89	0.89	0.90	0.90
Indonesian	0.84	0.94	0.85	0.90	0.92	0.90
Vietnamese	0.86	0.94	0.86	0.91	0.93	0.90
Burmese	0.89	0.91	0.88	0.92	0.83	0.89
Turkish	0.89	0.90	0.90	0.90	0.80	0.88
Thai	0.83	0.92	0.81	0.89	0.92	0.88
Hindi	0.82	0.90	0.84	0.87	0.80	0.85

Table 4.4: Sentence similarity scores (cosine similarity of LaBSE embeddings) between the translations and the original texts, grouped by domain. The overall mean, μ, is computed across all domains.

5 Error Analysis

We undertook a two-tier error analysis, first comparing machine translations directly with **post-edited** texts, and then comparing machine translations with **back-translated** NSWEduChat translations. This dual approach sheds light on different classes of errors.

5.1 Error Analysis vs. Post-Edited Translations

When post-edited references exist, they represent the ground truth for lexical choice, syntax, and style. Our manual inspection identified three primary error categories:

1. **Terminology Mismatch:** Words related to school regulations, attire (e.g., uniform policies), or health guidelines are incorrectly translated. For example, the generated translation 'ក្បាលកន្ទុយ (កណ្តុរ)' describes an infestation but the post-edited translation instead uses 'ប្រភេទសត្វល្អិត (ពងចៃ)' to accurately indicate the presence of lice.
2. **Omissions or Additions:** Certain phrases referencing key instructions (like "please park only in legal spots") are missing, or spurious text is introduced. The generated translation mentions a group under a name – 'អកកេះ', that might be misinterpreted, but the post-edit corrects it to 'អ្នកទះដៃសប្បាយរីករាយ' (Happy Clapper) and provides clearer details about their performance.
3. **Language Style, Tone, and Clarity:** Adopting a more formal and engaging tone suitable for parent and school communications and clarifying ambiguous phrases. The generated translation uses repeated simple greetings like 'សួស្តី, សួស្តី, សួស្តី', while the post-edited translations expand the greetings to include variations such as 'ហេ, អរុណសួស្តី, សាយណ្ណសួស្តី' to provide an engaging tone.

We perform a similar manual analysis on the Hindi translation pairs and present examples for the types of correction in Table 5.1.

5.2 Error Analysis vs. Back-Translations of NSWEduChat Translations

For certain segments lacking post-edited references, we performed **back-translation** checks. Specifically, we used GPT-4 to translate the target-language text back into English, then compared that output with the original English. This method highlights large

Type	Generated Translation	Transliteration	Post-edited Translation	Transliteration	Notes
Terminology Inaccuracies	नट्स (जूं के अंडे)	Nits (ju ke ande)	नट्स (जुओं के अंडे)	Nits (juon ke ande)	Minor change; generated translation is acceptable.
Grammar Corrections	जो 20 स्टेज 2 और 3 के छात्रों से बना है	Jo 20 stage 2 aur 3 ke chhatron se bana hai	जिसमें स्टेज 2 और 3 के 20 छात्र शामिल हैं	Jisme stage 2 aur 3 ke 20 chaatr shaamil hain.	Minor variation; grammar is incorrect and meaning is affected.
Language Style, Tone, and Clarity	यदि आपके पास कोई प्रश्न या चिंता हो, तो कृपया मुझसे या स्कूल नर्स से संपर्क करने में संकोच न करें।	Yaadi aapke paas koi prashn ya chintha ho, toh kripiya mujhse ya school nurse se sampark karne mein sankoch na karein.	यदि आपके कोई प्रश्न या चिंताएँ हैं, तो कृपया मुझसे या स्कूल नर्स से संपर्क करें।	Yadi aapko koi prashn ya chinthayein hai, toh kripiya mujhse ya school nurse se sampaark karein.	Minor variation; meaning is effectively conveyed.

Table 5.1: Example Hindi translation pairs illustrating various error correction types.

semantic deviations as seen in a few cases where the back-translation reverts to discussing a completely different subject (e.g., a note about dogs on campus becomes a note about littering). This signals a severe error where the translator replaced entire clauses.

Example:

- **Original English**:

  "Dear Parents and Carers, As we strive to ensure the safety and well-being of all students, please park legally when dropping off and picking up your child. Stopping in no-parking zones poses a safety risk."

- **Target Language**:

  "អត្ថបទ ១ ប្រធានបទ: ការប្រកាសសំខាន់: មិនអនុញ្ញាតឱ្យមានក្រុមឆ្កែនៅក្នុងទីសាលារៀន។ ឪពុកម្តាយ និងអ្នកថែទាំជាទីគោរព, សង្ឃឹមថាអ៊ីមែលនេះនឹងស្លាប់អ្នកនៅក្នុងសុខភាពល្អ។ ..."

- **Back-translation (via GPT-4)**:

  "Topic: Important Notice: No Dogs Allowed on School Grounds. Respected Parents, We wish to remind you that bringing dogs onto the campus is prohibited. ..."

- **Analysis:** The entire core message about "parking legally" is replaced by a statement about banning dogs on campus. This is a major semantic error.

Note

Back-translation is not a perfect proxy for post-editing but is valuable in cases where references are missing. It exposes large content shifts, omissions, or additions by letting us see how the target text re-appears in English. In the example above, the original mention of "parking legally" is entirely replaced by a new topic. This approach can highlight major semantic problems and aid in diagnosing repeated mistranslations of specific school-related terminology.

6 Cost–Accuracy Analysis

We performed a cost–accuracy analysis to evaluate the expense of using the GPT-4 model for translating English correspondence into various target languages. Since the overall cost is directly related to the number of prompt and generated tokens, Figure 6.2 illustrates the token counts as determined by the GPT-4 tokenizer. To quantify the trade-off between translation accuracy and cost, we compute the Economical Prompting Index (EPI; McDonald et al., 2025) as follows:

$$\text{EPI}(A, C, T) = A \cdot \exp\left(-C \cdot T\right),$$

where:

- A is the accuracy (measured via sentence similarity),
- C is the cost factor, and
- T is the total token count (including prompt tokens, English input tokens, and output tokens in the target language).

For these translations, we assume the following prompt:

Given an English text, translate it to the target language: {language}.

After calculating the EPI for each translation across all languages and use cases, we derive a cost–accuracy index (CI) by applying a multiplier based on the quartile distribution of EPI values. Specifically:

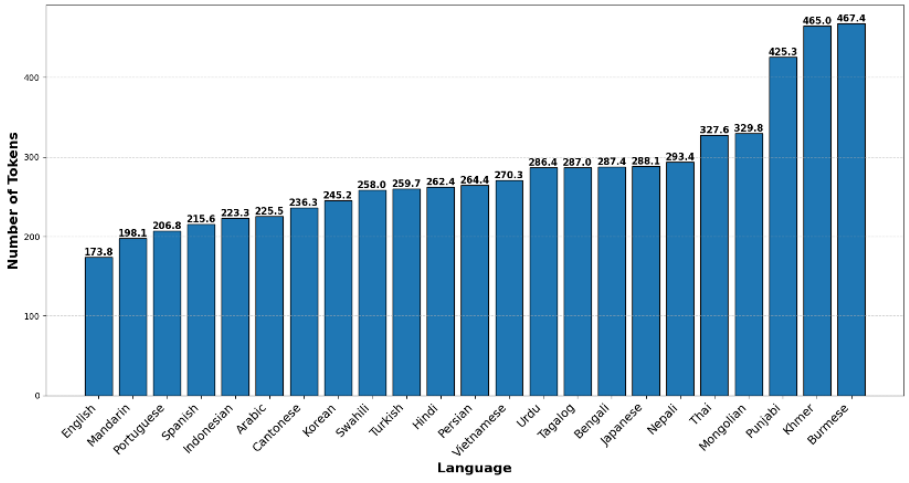

Figure 6.1: Average token count per language.

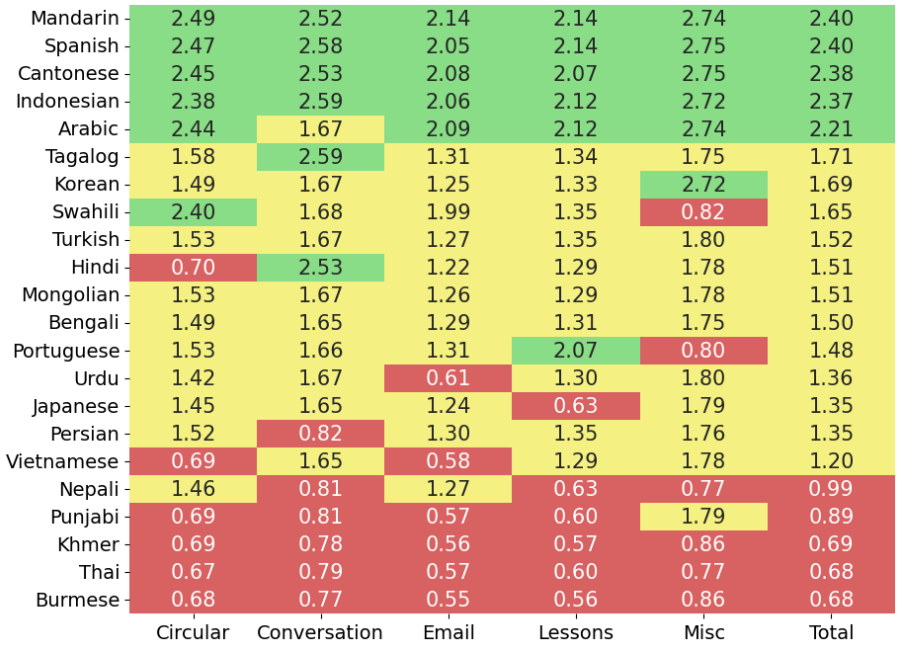

Figure 6.2: Cost–accuracy trade-off. Green indicates acceptable translations for a particular use case, yellow denoting reasonable translations, and red signifying unacceptable results.

- **Acceptable:** For translations with EPI values above the 75th percentile, CI is computed as $3 \times$ EPI.
- **Reasonable:** For translations with EPI values between the 25th and 75th percentiles, CI is computed as $2 \times$ EPI.
- **Unacceptable:** For translations with EPI values below the 25th percentile, CI is computed as $1 \times$ EPI.

Figure 6.2 illustrates this categorisation, with green indicating acceptable translations, yellow denoting reasonable translations, and red signifying unacceptable results.

Conclusion & Future Directions

We evaluated the dataset for 22 languages using reference-based and reference-less metrics. We followed it up with error analysis and cost-benefit analysis.

Based on the findings, we make the following recommendations:

1. Figure 3 provides a detailed map for acceptable, reasonable and unacceptable outputs from LLMs. These may be displayed to the users of NSWEduChat as labels to warn them about potential mis-translations.
2. In the current form, we recommend that domain-specific terminology may not be translated using LLMs but be retained as it is, even in English. This may be done by maintaining a dictionary of domain-specific terms and appending the terms to the LLM-generated output.
3. For future versions of NSWEduChat and similar systems, our code will be useful to replicate the evaluation undertaken as a part of the project.

As extensions of the project, we see the following opportunities for future work:

1. Development of LLMs for the education context: Continual pre-training Ke et al. (2023) and task-based fine-tuning of LLMs can be expected to improve the performance for the translation task.
2. Ensembling LLMs may be useful for improved output. Similarly, we point to advanced techniques in prompt engineering (specifically, System 2 attention Weston and Sukhbaatar (2023)) for improved outputs.
3. While the project deals with translation as a use-case, the project team that worked on the report has the capability of expanding to other use-cases and look beyond evaluation of outputs from LLMs.

Bibliography

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022.
Language-agnostic BERT sentence embedding.
In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023.
Continual pre-training of language models.
In *Proceedings of The Eleventh International Conference on Learning Representations (ICLR-2023)*.

Chin-Yew Lin. 2004.
ROUGE: A package for automatic evaluation of summaries.
In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tyler McDonald, Anthony Colosimo, Yifeng Li, and Ali Emami. 2025.
Can we afford the perfect prompt? balancing cost and accuracy with the economical prompting index.
In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7075–7086, Abu Dhabi, UAE. Association for Computational Linguistics.

OpenAI Team. 2024.
Gpt-4 technical report.
Preprint, arXiv:2303.08774.

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

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, Lisbon, Portugal. Association for Computational Linguistics.

Jason Weston and Sainbayar Sukhbaatar. 2023.
System 2 attention (is something you might need too).
arXiv preprint arXiv:2311.11829.

FOR MORE INFORMATION

aditya.joshi@unsw.edu.au
unswnlp.github.io

UNSW SYDNEY

NSW GOVERNMENT | Education