Proverbs Run in Pairs: Evaluating Proverb Translation Capability of Large Language Model

Minghan Wang, Viet-Thanh Pham, Farhad Moghimifar, Thuy-Trang Vu
Department of Data Science & AI, Monash University
{minghan.wang,thanh.pham1,trang.vu1}@monash.edu

Abstract

Despite achieving remarkable performance, machine translation (MT) research remains underexplored in terms of translating cultural elements in languages, such as idioms, proverbs, and colloquial expressions. This paper investigates the capability of state-of-the-art neural machine translation (NMT) and large language models (LLMs) in translating proverbs, which are deeply rooted in cultural contexts. We construct a translation dataset of standalone proverbs and proverbs in conversation for four language pairs. Our experiments show that the studied models can achieve good translation between languages with similar cultural backgrounds, and LLMs generally outperform NMT models in proverb translation. Furthermore, we find that current automatic evaluation metrics such as BLEU, CHRF++ and COMET are inadequate for reliably assessing the quality of proverb translation, highlighting the need for more culturally aware evaluation metrics.[1]

1 Introduction

Translating multi-word figurative expressions, particularly idioms and proverbs, have long been a challenge in MT due to their meanings diverging from the literal interpretation of individual words (Constant et al., 2017; Zaninello and Birch, 2020). Previous research in neural machine translation (NMT) has primarily focused on translating idiomatic expressions to capture their figurative meanings in the source language and accurately convey them in the target language (Isabelle et al., 2017; Fadaee et al., 2018; Avramidis et al., 2019). The development of large language models (LLMs) for MT has demonstrated a reduction in overly literal translations, improving the quality of idiomatic translations (Raunak et al., 2023). However, the translation of proverbs, which are short popular sayings conveying cultural beliefs, has received comparatively less attention.

Translating proverbs presents additional challenges beyond simply preserving figurative meaning. Proverbs are often deeply rooted in cultural contexts, and their translation requires careful cultural adaptation. This involves translating culture-specific terms, paraphrasing, and considering widely accepted versions of the proverb that resonate in the target language (Shehab and Daragmeh, 2014; Newmark, 2003). Recent research by Liu et al. (2024) has shown that while LLMs possess some degree of knowledge about proverbs, their capability of reasoning with proverbs, especially in dealing with figurative proverbs, remains limited. In light of these challenges and recent developments, using proverbs as a proxy for cultural common grounds, our study examines the ability of current NMT and LLM-based MT models in proverb translation to better understand how existing MT models handle cross-cultural elements. In particular, we aim to answer the following research questions: (i) *Can current MT methods, particularly LLM-based MT systems, handle proverb translation?*; (ii) *What are the roles of conversation contexts and prompts in proverb translation ability of LLM-based MT?*; and (iii) *Can current automatic evaluation metrics measure the accuracy of translating cultural nuances?*

To address these research questions, we first expand the existing multicultural proverbs and sayings dataset (Liu et al., 2024) into an English-centric proverb translation dataset. As proverbs can also appear in conversations, we further mine the Proverb in Conversation (PiC) dataset by extracting proverb usage from movie subtitles. These datasets include five languages representing diverse geographical areas: English, German, Bengali, Indonesian, and Mandarin Chinese. We then conduct extensive evaluations of state-of-the-art NMT systems and multiple LLM families on these datasets

[1]Dataset and code are available at https://github.com/yuriak/LLMProverbMT.

Findings of the Association for Computational Linguistics: ACL 2025, pages 1646–1662
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

to assess their proverb translation capabilities.

Our experiments reveal that current MT models demonstrate a certain level of proficiency in handling proverb translations, with notably better performance observed when translating between languages from similar cultural areas. Furthermore, we observe that the reliability of existing automatic evaluation metrics is insufficient for accurately assessing the cultural nuance in proverb translation. Specifically, we make the following contributions

- We construct a proverb translation and proverb in conversation translation dataset in four translation pairs to facilitate the future research in figurative language translation.
- Our experiments provide insights into the roles of context and prompting in the performance of proverb translation tasks. Larger models already learn the meaning of proverbs, hence, adding the proverb interpretation or example does not help. On the other hand, conversation context in dialogue format plays a more important role.
- We also find that the current automatic evaluation metrics such as BLEU and COMET as well as the LLM-as-a-judge are unreliable and very sensitive to small lexical changes when evaluating figurative translation quality.

2 Proverb Translation Dataset

2.1 Standalone Proverb Translation Dataset

Proverbs are fixed expressions that convey traditional wisdom and deeply rooted in lived experiences and socio-cultural contexts. This makes proverbs an excellent lens through which to evaluate how well the current MT model captures and translates cultural information. To facilitate such analysis, we extend the MAPS dataset (Liu et al., 2024) which is a collection of proverbs from multiple languages, including English (EN), German (DE), Bengali (BN), Mandarin Chinese (ZH) and Indonesian (ID).[2] These languages come from diverse geographical regions and exhibit varying linguistic structures and resource availability according to Joshi et al. (2020). Each proverb in MAPS dataset is companied with its explanation, the machine translation into English, and a label indication whether the proverb is figurative or literal. The figurative proverbs have meanings that differ from their literal expressions, while the literal ones convey meaning directly. This dataset allows us to evaluate not only the translation quality but also how well models can handle figurative language. Table 1 provides detailed statistics.

Language	#Prov.	#Fig.	Region
English (EN)	424	232	Western Europe
Bengali (BN)	340	272	South Asia
German (DE)	334	183	Western Europe
Indonesian (ID)	341	267	Southeast Asia
Chinese (ZH)	334	143	East Asia

Table 1: MAPS dataset statistics including number of proverbs (#Prov.), number of figurative proverbs (#Fig.) and geography region.

To ensure the quality of the dataset, we recruit annotators to verify and post-edit the machine-translated proverbs. These annotators were fluent in both English and native speakers of Chinese, Indonesian, or Bengali, with experience in either professional or volunteer translation work. For German-to-English translations, although the annotators were not native German speakers, they were proficient in both German and English. The annotation process was structured to ensure the cultural and linguistic accuracy of the translations.

Annotators were presented with proverbs in their native language alongside the machine-generated English translations. Their primary task was to evaluate the correctness of these translations and post edit where necessary. Additionally, they are also tasked with identifying context-dependent proverbs whose meanings can shift based on the situation in which they are used. For English proverbs, the annotators are also asked to provide equivalent proverbs in their native languages that convey similar meanings, when possible. The native proverb provided will serve as a translation reference during the evaluation of from English translation. The details of annotation protocols and addition analysis can be found in Appendix A.

2.2 Proverb in Conversation Mining

OpenSubtitles (Lison and Tiedemann, 2016) is a multilingual parallel corpora of movie and TV subtitles. It contains culturally rich conversation and spans multiple languages, making it a good source to mine the parallel corpus of proverb usage. However, as OpenSubtitles is often included in LLM pretraining corpus, we perform data contamination

[2] We omit Russian (RU) in our study due to lack annotators.

Lang	From-En P1	From-En P2	From-En P3	To-En P1	To-En P2	To-En P3
BN	89	69	42	511	76	8
DE	7028	2000	1540	1903	1755	1129
ID	2459	1969	1214	51	44	13
ZH	3456	2000	272	1498	1488	827

Table 2: The statistics of the mined subtitles in each phrase: **P1: Initial Mining §2.2.1**; **P2: Fine-grained Filtering §2.2.2**; **P3: Human Evaluation §2.2.3**.

analysis in §5.3 and find that contamination is not significant enough to bias the evaluation.

2.2.1 Initial Mining with Source-side Proverb

We first collect potential translation pairs which containing proverbs from OpenSubtitles dataset. In this step, we preprocess both proverbs and translation pairs with lemmatization[3]. We then use an edit-distance-based string matching library[4] to search for translation pairs where proverbs are contained in the source sentence. When a source sentence contains an exact match of the given proverb, its matching score will be 1.0, any replacement of characters will reduce the score. We set a threshold as 0.8 during the search.

2.2.2 Fine-grained Filtering

Although lemmatization can solve some of the matching errors caused by morphological inflection, the search results will still contain a large number of errors. Therefore we propose two fine-grained methods for further filtering through the semantics of the mined sample.

LLM-based Proverb Usage Filtering We use an LLM[5] to filter those translation pairs that contain the proverb by asking the model: "Whether the proverb is contained in the sentence". This ensures that the filtering is performed through the semantic meaning of the given text and thus is more accurate. To make sure that the LLM's output is reliable, we set the temperature to 0 and constrain the model's output as "Yes" and "No". For samples labeled as "NO", we remove them from our candidate set.

Filtering with Quality Estimation Another filtering process aims to ensure the translation of the collected pairs is good enough, as there are cases where source and target texts are mismatched in the subtitles due to reordering of utterances which have to be excluded. In this step, we use both LLM (LLM-QE) and a dedicated Quality Estimation with Direct Assessment (DA-QE) model to score the collected translation.

- For LLM-QE, we let the model to score the translation from 1-5 and replace the order of source and target, then, score it again to reduce the influence of the order; an average of two values is computed as the label of the candidate pair.
- For DA-QE, we use "Unbabel/wmt23-cometkiwi-da-xxl" as the dedicated DA scorer to score the translation pair in a range between 0 to 1. Then, we compute the overall score for each pair as $\text{score} = \text{score}_{\text{LLM-QE}} + \text{score}_{\text{DA-QE}} \times 5$.

Filtering is conducted separately across language pairs as the quality of the mined corpus in each direction differs largely. Specifically, we set the maximum required sample size for each direction as **2000**, and use it to compute the minimum quantile as ($q_{\min} = max(0, 1 - \frac{2000}{|\mathcal{D}_{s\to t}|})$, where $|\mathcal{D}_{s\to t}|$ stands for the number of samples for a specific direction e.g. En→De) and the corresponding overall score ($max(\text{score}_{q_{\min}}, 4)$, where $\text{score}_{q_{\min}}$ is the corresponding score of the quantile $q_{\min}$, 4 is the minimum score threshold we assigned for all language pairs) as the threshold. Finally, we used this threshold to filter qualified samples. Detailed statistics in each step are presented in Table 2.

Conversation Context Retrieval While proverbs and sayings are self-contained, they are typically used in conversation. As we aim to study whether the provided context could influence the translation of a sentence containing a proverb, we need to retrieve the prior and proceeding sentences for each filtered translation pair. In this step, we retrieve a maximum of 5 sentences for each direction (prior and proceeding).

2.2.3 Human Evaluation

To validate the collected translation data, the annotators are presented with a proverb in source language, and ask to evaluate the translation in target language given the preceding and following context. Additionally, they are asked whether the proverb is correctly used in the given sentences. The details of annotation protocols and addition analysis can be found in Appendix A. Finally, we collect the parallel pairs which contains correct

[3] https://spacy.io/

[4] https://docs.python.org/3/library/difflib.html

[5] meta-llama/Meta-Llama-3.1-70B-Instruct

	Proverbs		PiC	
	#lit.	#fig.	#lit.	#fig.
en -> bn	130	163	29	13
bn -> en	68	272	2	6
en -> de	180	214	1191	349
de -> en	151	183	614	515
en -> id	162	183	886	328
id -> en	71	262	3	10
en -> zh	134	161	227	45
zh -> en	191	143	386	441

Table 3: Data statistics of the standalone proverb translation (Proverbs) and proverb in conversation translation (PiC). The number of samples contains literal and figurative proverbs are denoted by **#lit** and **#fig** respectively.

proverb usage and correct translation. The data statistics are shown in Table 3. Since the sample size of EN->BN, BN->EN, and ID->EN in our final PiC dataset is relatively small, we omit these three translation directions in our experiments.

3 Experimental Setup

In this paper, we investigate the ability of the current MT system, with a particular focus on LLM-based MT models, in translating proverbs. More specifically, we aim to answer the following research questions (**RQs**):

- **RQ1**: To what extent can existing MT methods accurately translate proverbs?
- **RQ2**: What are the roles of conversation contexts and prompts in the proverb translation ability of LLM-based MT?
- **RQ3**: Are current automatic evaluation metrics reliable and effective in measuring the accuracy of proverb translation?

3.1 Models

We study the proverb translation ability of the current MT methods, including

- **State-of-the-art multilingual NMT** We experiment with NLLB which are trained on translation data of 200 languages and available with different sets of parameters: 600M, 1.3B and 3.3B (Costa-jussà et al., 2022).[6]
- **Instruction-following LLMs** We evaluate instruction-following LLMs with different model parameter size from multiple model families: MISTRAL (7B) (Jiang et al., 2023), QWEN2 (7B) (Yang et al., 2024), LLAMA-3.1 (8B, 70B) (Dubey et al., 2024), GEMMA2 9B (Team et al., 2024) and GPT-4O MINI.[7]
- **LLM-based MT** In addition to NMT and off-the-shelf LLMs models, we also evaluate fine-tuned LLM model for MT tasks. Particularly, we consider ALMA-R 13B which is based on `Llama2 13B` further fine-tuning with contrastive preference optimization on high quality translation data (Xu et al., 2024b).[8]

3.2 Prompts

We design 5 types of prompt templates to evaluate the performance of LLM under different conditions (see Figure 6 in the Appendix). All 5 prompt templates are used in the evaluation on the PiC test set, but only the zero- and one-shot prompts are used in proverb standalone translation.

- **Zero-Shot** We only provide a simple system message to set a role (a professional translator) for the LLM and instruct it to translate the given source sentence without returning any irrelevant content.
- **One-Shot** As smaller LLMs may have limited instruction-following capability, we add an example of translating the sentence “Good morning" in the first round of dialogue. This allows LLM to have access to the dialogue history. [9]
- **Proverb Explanation** In the system message, we first signal the LLM that the proverb may contained in the given source text, and the explanation of the proverb, followed by the source sentence.
- **Contextualization through Dialogue** To study the role of the conversation context in translation performance, we consider previous subtitle sentences as contexts and place

[6]Model signatures: `facebook/nllb-200-distilled-600M`, `facebook/nllb-200-1.3B`, and `facebook/nllb-200-3.3B`

[7]Signatures: `mistralai/Mistral-7B-Instruct-v0.3`, `Qwen/Qwen2-7B-Instruct`, `meta-llama/Meta-Llama-3.1-8B-Instruct`, `meta-llama/Meta-Llama-3.1-70B-Instruct`, `gpt-4o-mini-2024-07-18`

[8]Model signature: `haoranxu/ALMA-13B-R`.

[9]We use the simple sentence for the one-shot case to prevent introducing any bias to the actual translation.

	BLEU				CHRF++				COMET			
	lit.		fig.		lit.		fig.		lit.		fig.	
	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot
From English translation												
NLLB-600M	9.42		8.23		23.14		21.49		67.36		60.30	
NLLB-1.3B	10.66		8.97		24.12		21.88		67.94		60.37	
NLLB-3.3B	11.23		9.17		24.73		22.43		67.99		60.60	
Alma-R 13B	6.55		5.68		17.97		16.93		62.55		55.56	
Qwen2	9.57	10.37	8.17	7.94	22.42	22.93	20.71	20.70	67.13	67.86	60.55	60.86
Mistral	6.44	6.92	5.24	5.06	18.91	18.74	16.79	16.57	63.79	64.46	56.07	56.86
Gemma2 9B	13.27	14.22	10.23	10.40	24.89	25.54	22.52	22.69	71.13	70.88	63.32	63.38
Llama-3.1 8B	10.51	11.08	8.22	8.87	23.40	23.68	20.67	21.33	69.44	70.17	62.28	62.76
Llama-3.1 70B	**15.16**	**16.83**	**13.53**	**13.97**	**27.59**	**28.68**	**25.59**	**26.12**	**72.49**	**73.09**	**65.24**	**65.55**
GPT-4o mini	13.61	12.82	11.20	10.64	26.97	26.62	24.61	24.04	71.62	71.65	63.45	63.37
To English translation												
NLLB-600M	11.42		11.42		28.63		28.35		60.79		57.91	
NLLB-1.3B	14.30		14.12		30.83		30.70		62.37		59.38	
NLLB-3.3B	15.84		14.41		31.41		31.35		62.71		59.69	
Alma-R 13B	13.98		17.42		29.45		32.48		61.33		61.19	
Qwen2	15.66	15.07	16.47	16.66	32.25	32.26	33.79	33.50	66.46	66.24	63.45	63.23
Mistral	13.03	13.58	14.83	14.46	27.98	28.91	30.47	30.34	61.75	62.75	60.39	60.89
Gemma2 9B	17.78	17.43	18.93	18.41	34.97	34.30	35.98	35.33	67.78	67.50	64.35	64.21
Llama-3.1 8B	15.66	16.72	16.20	16.77	30.84	31.67	32.03	32.69	64.89	65.64	62.37	63.24
Llama-3.1 70B	**20.85**	**20.10**	**22.35**	**21.55**	36.73	**36.57**	**38.22**	**37.26**	68.77	**68.74**	**65.29**	**65.00**
GPT-4o mini	18.84	17.14	18.47	18.22	**37.15**	36.83	38.05	38.10	**68.88**	69.02	65.20	65.42

Table 4: Results on standalone proverb translation. **Bold** highlights the best score in each column. The better score among one-shot and zero-shot are underlined.

them in the dialogue history, up to 5 rounds with source and target sentences acting as user input and model responses (essentially becoming a maximum of 5-shot form).

- **Contextualization through Concatenation** Although using a dialogue format to provide previous sentences as contexts to the model is an intuitive approach, it may introduce noise when source and target sentence pairs in the context containing reordering. To this end, we design another approach for contextualization by concatenating all source and target sentence pairs in the context into one user input and model response, making it into a one-shot form. Through this, reordered contexts can be placed in the same round and naturally recovered to the correct alignment.

3.3 Evaluation

Evaluation Metrics We evaluate the translation quality with lexical-overlap metrics including BLEU (Papineni et al., 2002) and CHRF++ (Popović, 2017) using SacreBLEU (Post, 2018),and neural evaluation metric such as COMET (Rei et al., 2020).[10]

Inference We use sampling with the default decoding parameters for GPT-4o mini, and beam search with a beam size of 5 for NLLB and open-source LLMs.

4 Main Results

4.1 Standalone Proverb Translation

NMT vs LLMs Table 4 presents the performance of various models on both literal and figurative proverb translation with zero-shot and one-shot prompting. Notably, despite being a LLM specialized for MT, Alma-R 13B consistently underperforms compared to other LLMs across the board and even falls behind the smallest NLLB model in from-English translation direction. We speculate that it can be attributed to the absence of Bengali and Indonesian languages in its fine-tuning data. Similarly, Mistral also struggles on this task due to its limited language support and relatively smaller size. In contrast, other LLMs outperform NLLB models, with Llama-3.1 70B emerging as the strongest model.

[10]COMET signature: `Unbabel/wmt22-comet-da`.

Zero-shot vs One-shot Prompting One-shot prompting generally outperforms zero-shot prompting, especially in from-English translation direction. Interestingly, our strongest model LLAMA-3.1 70B achieves slightly higher score in zero-shot prompting in to-English translation tasks.

Literal vs Figurative Proverbs Overall, the performance on figurative proverbs is consistently lower than on literal proverbs across all metrics in from-English translation directions. It is expected as the figurative proverbs have an underlying meaning different to their literal wording which makes them more challenging to translate. However, in the to-English translation direction, we notice an unexpected trend across different metrics. While NLLB models generally score higher on literal proverbs than figurative one for all metrics, the opposite trend is observed with LLMs. Specifically, LLMs achieve higher BLEU and CHRF++ scores when translating figurative proverbs, but they tend to score better COMET scores on literal proverbs.

Figure 1 breaks down the performance of 4 models on literal and figurative proverb translation in each translation direction. All models perform reasonably well in DE-EN pairs because both languages are high-resource and belong to similar cultural regions. On the other hand, BN-EN are the most challenging tasks. Overall, all models perform better on literal translation in all translation directions, except the ID->EN direction. This leads to the average performance on to-English figurative proverb translation is higher than literal ones.

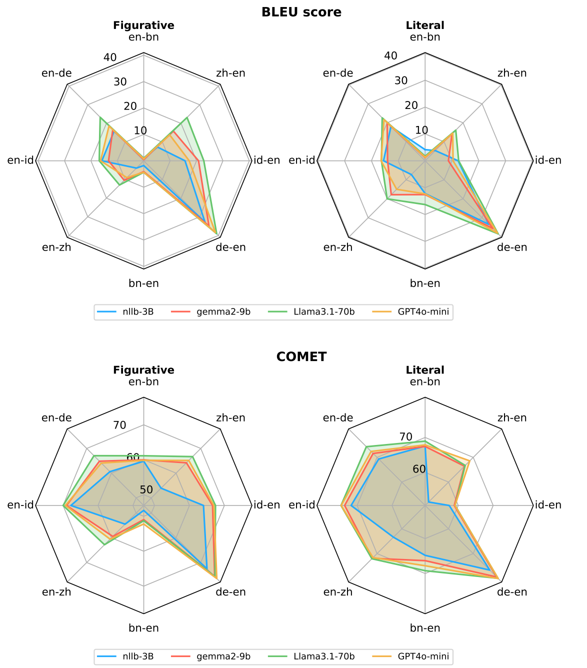

Figure 1: Result on proverb translation of each translation direction.

4.2 Proverb in Conversation Translation

NMT vs LLMs Table 5 reports COMET scores of various models on the translation of proverbs in conversational contexts. Detailed results in BLEU and CHRF++ can be found in the Appendix. Consistent with earlier findings in proverb translation, MISTRAL 7B lags behind other models, while GPT-4O MINI stands out as the strongest LLM model, following by LLAMA-3.1 70B. In this particular task, interestingly, NLLB models show highly competitive results to the LLMs, especially in the literal proverb subset. However, NLLB models fall short in translating figurative proverbs.

Roles of Context and Prompting Among the different prompting strategies, one-shot prompting consistently improves upon the performance of zero-shot prompting. However, we do not observe any notable improvement when providing explanations of proverbs, particularly in the case of literal proverbs. This may be because LLMs have already exposed to the meanings of proverbs during their pre-training, allowing them to capture these meanings without additional explanation. On the other hand, incorporating conversational context significantly enhances translation performance. This is likely due to the extra context clues in the conversations, which help the models generate more accurate translations. Furthermore, framing the context as a dialogue proves to be more effective than simple concatenation, as it aligns more naturally with the conversational nature of LLMs.

5 Analysis

5.1 Limitations of Evaluation Metrics on Proverb Translation

As proverbs are often highly culture-specific, standard evaluation metrics like BLEU, CHRF, or COMET may fall short in certain cases. In this section, we illustrate the weaknesses of these metrics regarding proverb translation. With all of the hypotheses from different models and prompt templates, we compute the cosine similarity between the sentence embedding[11] of the hypothesis and

[11]We use `sentence-transformers/all-mpnet-base-v2` model to obtain the embeddings.

	Literal					Figurative				
	0-shot	1-shot	EXPL.	DIALOG	CONCAT	0-shot	1-shot	EXPL.	DIALOG	CONCAT
From English translation, incl. EN-DE, EN-ID, EN-ZH										
NLLB-600M	84.93					78.87				
NLLB-1.3B	86.04					80.18				
NLLB-3.3B	85.34					80.32				
ALMA-R 13B	84.94					80.41				
QWEN2	84.50	84.95	84.31	<u>85.74</u>	<u>85.74</u>	80.63	81.40	81.24	<u>83.08</u>	82.71
MISTRAL	81.63	81.54	81.48	<u>83.38</u>	82.88	76.37	76.10	76.66	<u>77.94</u>	77.34
GEMMA2 9B	85.51	85.31	85.12	<u>86.19</u>	86.05	82.05	82.20	82.54	<u>82.82</u>	82.62
LLAMA-3.1 8B	83.72	84.69	83.30	<u>85.75</u>	85.59	79.16	78.82	78.37	<u>80.96</u>	80.57
LLAMA-3.1 70B	86.29	86.42	86.37	<u>87.30</u>	86.71	84.26	83.92	**84.81**	<u>85.14</u>	84.44
GPT-4O MINI	**87.22**	**87.36**	**86.91**	**<u>88.06</u>**	**87.95**	**84.60**	**84.70**	84.75	**<u>85.59</u>**	**84.61**
To English translation, incl. DE-EN, ZH-EN										
NLLB-600M	60.07					61.61				
NLLB-1.3B	63.16					63.17				
NLLB-3.3B	62.99					63.64				
ALMA-R 13B	64.83					65.18				
QWEN2	64.32	64.04	64.78	<u>65.86</u>	65.45	66.41	64.98	66.26	<u>68.64</u>	67.74
MISTRAL	60.54	62.86	61.33	<u>64.83</u>	64.15	63.14	62.46	62.88	<u>65.90</u>	65.10
GEMMA2 9B	64.87	65.93	66.05	<u>67.83</u>	67.96	66.22	66.98	66.81	<u>68.98</u>	68.37
LLAMA-3.1 8B	62.83	65.17	62.67	<u>67.22</u>	66.53	63.98	65.55	65.15	<u>66.44</u>	66.21
LLAMA-3.1 70B	65.76	66.28	**66.93**	**<u>68.83</u>**	68.01	65.59	67.86	**67.25**	<u>68.48</u>	68.44
GPT-4O MINI	**66.30**	**66.32**	66.30	<u>68.30</u>	**68.17**	**66.65**	**68.03**	66.79	**<u>68.87</u>**	**68.46**

Table 5: Subtitle Translation (COMET score) beam search. **Bold** highlights the best score in each column. The better score among different prompting methods is <u>underlined</u>.

Example	BLEU	CHRF	COMET
Reference: “Distance determines the stamina of a horse.”			
Hypothesis 1: “Distance reveals the strength of a horse”	26.27	46.19	82.85
Hypothesis 2: “A long journey reveals the strength of a horse”	19.06	35.42	71.58
Reference: “The face of a tiger, the heart of a mouse.”			
Hypothesis 1: “The face of a tiger, the heart of a mouse”	89.32	95.05	93.23
Hypothesis 2: “The look of a tiger, the heart of a rat.”	71.03	80.81	81.99
Reference: “Spare the rod and spoil the child.”			
Hypothesis 1: “Spare the rod and spoil the child.”	100.00	100.00	95.22
Hypothesis 2: “Discipline brings forth filial children.”	0.0	13.93	60.10

Table 6: Evaluation of Hypotheses with BLEU, CHRF, and COMET Scores. Red highlights the words that make the metrics score the hypotheses lower.

its reference. We focus on cases where the cosine similarity difference of two hypotheses against the same reference is low (< 0.05), while the difference is above a threshold of 10.0, 5.0 and 10.0 for COMET, BLEU and CHRF++, respectively. From the total of 5M hypothesis pairs, we find that 22,704 pairs satisfy these thresholds. Notably, specialized NMT models (NLLB and ALMA-R 13B) have the fewest appearances in these cases, accounting for 1,962 pairs. This may indicate that LLMs may produce more creative translations that the evaluation metrics may have neglected.

Qualitative Analysis We manually check the detected cases to identify problems of evaluation metrics. Table 6 shows some cases where the metrics are unreliable for evaluating proverb translation. In the first and second examples, the hypothesis 2 scores lower on all metrics due to the usage of different phrases to the reference, even though it conveys the same idiomatic meaning ("a long journey" instead of "distance", and "rat" instead of "mouse"). This indicates that the metrics are overly sensitive to surface-level lexical differences. The third example, "Spare the rod and spoil the child," show-

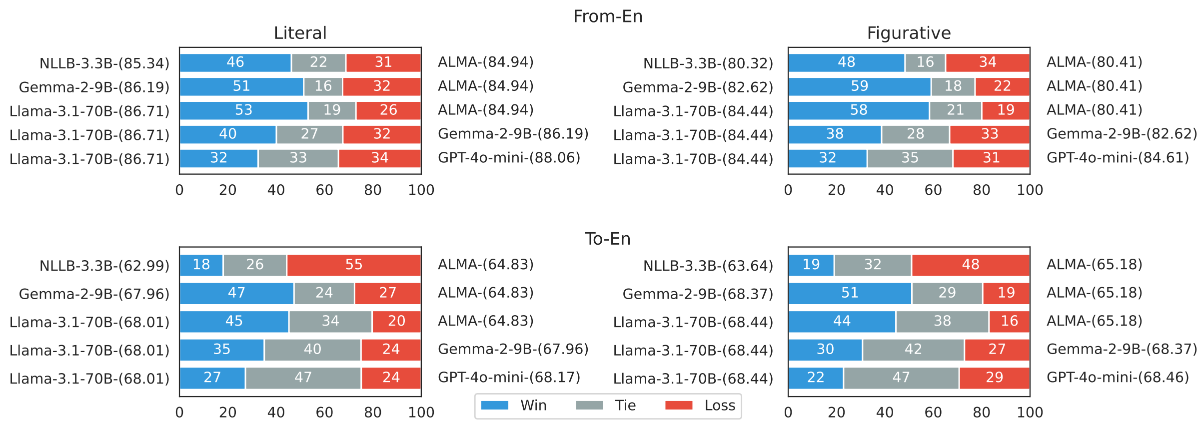

Figure 2: We present the win rate of 5 pairs of models evaluated by the GPT-4O MINI. Results are separated with From/To-En and Figurative/Literal, with corresponding COMET scores indicated near the model name.

cases the metrics' inability to handle significant paraphrasing. Despite conveying the same core meaning, the metrics fail to recognize the equivalence due to their reliance on word overlap rather than deeper semantic understanding. These examples demonstrate that standard metrics can be inadequate when evaluating translations involving idioms and non-literal expressions, as they tend to penalize valid translations that do not strictly adhere to the reference's lexical choices. Even COMET, while more robust in capturing meaning, struggles with cases of metaphorical language and significant rephrasing.

5.2 LLM-as-a-Judge Evaluation

Realizing the ineffectiveness of traditional evaluation metrics, we further use the **LLM-as-a-judge** method for evaluation. Here, we primarily evaluate the PiC translation results and follow the setup of Table 5. We selected five representative model pairs and compared their translation results using GPT-4O MINI, calculating the win rates. For NLLB and ALMA-R 13B, we use zero-shot results, while for other LLMs, we use results generated with dialogue prompts. The prompt used for evaluation is shown in Figure 7 in the Appendix. We prompted the model to compare the translations comprehensively based on three aspects: translation accuracy, fluency, and cultural appropriateness. Additionally, we provided the complete contextual information before and after the sample, along with the corresponding reference to assist the model in its judgment. To avoid the bias caused by the positioning of hypothesis A and B, we randomly assigned the results of the two models as A and B, thus eliminating the influence of position. Finally, GPT-4O MINI will return one of three results: A, B, or tie.

	% samples with $\gamma > 0.9$				
Models	*En*	Bn	De	Id	Zh
LLAMA-3.1 8B	4.5	0.0	3.8	1.1	3.8
GEMMA2 9B	1.3	0.0	1.9	0.4	1.7
QWEN2	4.7	0.0	1.5	0.1	1.2
MISTRAL	1.0	0.0	0.7	0.3	0.8

Table 7: In this table, we present the percentage of samples with $\gamma > 0.9$ as the measurement of contamination.

In Figure 2, we present the win rates along with the COMET scores of the two groups of models. It can be seen that the win rates and COMET scores generally follow a consistent trend (models with higher COMET scores usually have higher win rates). However, the gap in COMET scores does not strictly correlate linearly with the win rate. For instance, a larger difference in COMET scores does not necessarily mean a higher win rate (e.g. LLAMA-3.1 70B vs ALMA-R 13B, LLAMA-3.1 70B vs GEMMA2 9B). This suggest that evaluation with LLM-as-judge cannot fully solve the limitations of traditional evaluation methods.

5.3 Data Contamination Analysis

Following (Liu et al., 2024), we measure the memorization rate of the Proverb-in-Conversation dataset by assessing the model ability to complete an utterance based on partial context. This is quantified using the longest common subsequence (LCS) rate between the model's prediction and the actual utterance, given the preceding context. However, since proverbs are often well-known phrases, models might accurately predict them even without relying on the provided context. To account for this, we introduce the contamination rate γ as the difference between the LCS of the model's prediction and the reference with and without context, normalized by the utterance length. We provide more details of this metric in Appendix B.1.

A higher value of γ indicates a greater likelihood that the model has been affected by data contamination for a given sample. We present the percentage of samples with $\gamma > 0.9$ in Table 7 across each language. A relatively higher correlation between the contamination rate and the resource-level of the language can be found, but the correlation to the translation performance is not significant. This suggests that the contamination issue is not biasing our evaluation.

6 Related works

Figurative Expression in MT Multi-word expressions (MWEs), including idioms, phrasal verbs, and multi-word named entities, present a unique challenge in natural language processing due to their non-compositional nature, where the meaning of the whole expression cannot be easily inferred from the meanings of its individual words (Constant et al., 2017). Previous research has largely focused on understanding and paraphrasing MWEs, particularly in English (Liu and Hwa, 2016; Wada et al., 2023). In NMT literature, much focuses has been on idiomatic and slang translation, primarily targeting European languages (Fadaee et al., 2018; Sun et al., 2022; Baziotis et al., 2023). However, a key obstacle to further progress in this area is the absence of standardized evaluation benchmarks and metrics. Our work addresses this gap by constructing a dataset on the translation of proverbs for four language pairs, each representing distinct geographical and cultural regions.

LLM-based MT Several studies have explored the application of LLMs for translation tasks, highlighting their impressive performance across multiple high-resource language pairs (Xu et al., 2024a,b; Wu et al., 2024). One notable advantage of LLMs over traditional neural machine translation (NMT) systems is their ability to generate more controlled and nuanced translations, particularly when dealing with idiomatic expressions that require less literal interpretation (Manakhimova et al., 2023; Stap et al., 2024). In this work, we focus on evaluating the capabilities of LLMs in proverb translation, a challenging task due to the cultural and figurative nature of proverbs.

7 Conclusion

We curate a proverb and its usage in conversation to investigate the ability of LLMs on proverb translation. Our experiments reveal that LLMs generally outperform NMT model on this task, showcasing the advantage of LLMs in translating figurative expressions, especially between high-resource languages and languages from similar cultural region. Additionally, our analysis also reveals that current automatic evaluation metrics are unreliable in measuring translation with figurative languages.

Limitation

Although we have constructed the Proverb in Conversation dataset and conducted systematic evaluations of different models, we acknowledge the following two limitations of this study:

- The scale of our dataset is currently relatively small. This is partly due to the limited number of proverbs in each language, which constrains the size of the samples that can be collected. Additionally, our strict filtering and human annotation process further excluded low-quality samples, which may have led to excessive discarding and thus limited the dataset size. In future work, we will consider expanding data sources and ensure that the scale of the collected data meets the needs for more comprehensive evaluations.

- Another limitation lies in the relatively limited variety of models and prompts tested. Although most of them are commonly used models, they still cannot fully represent the capabilities of other models, especially those with more than 70B or fewer than 7B parameters. Therefore, in our future work, we will further expand the selection of models and the variety of prompts to enhance the comprehensiveness of the evaluation.

Acknowledgments

The authors are grateful to the anonymous reviewers for their helpful comments. This research/work was supported by Monash eResearch capabilities, including m3.

References

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, and Hans Uszkoreit. 2019. Linguistic evaluation of German-English machine translation using a test suite. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 445–454, Florence, Italy. Association for Computational Linguistics.

Christos Baziotis, Prashant Mathur, and Eva Hasler. 2023. Automatic evaluation and analysis of idioms in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3682–3700, Dubrovnik, Croatia. Association for Computational Linguistics.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword Expression Processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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, Copenhagen, Denmark. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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, Online. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Changsheng Liu and Rebecca Hwa. 2016. Phrasal substitution of idiomatic expressions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 363–373, San Diego, California. Association for Computational Linguistics.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. 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 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can ChatGPT outperform NMT? In *Proceedings of the Eighth Conference on Machine Translation*, pages 224–245, Singapore. Association for Computational Linguistics.

Peter Newmark. 2003. A textbook of translation.

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. 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.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023. Do GPTs produce less literal translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1041–1050, Toronto, Canada. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ekrema Shehab and Abdelkarim Daragmeh. 2014. A context-based approach to proverb translation: The case of arabic into english translation. *Translation Review*, 90(1):51–68.

David Stap, Eva Hasler, Bill Byrne, Christof Monz, and Ke Tran. 2024. The fine-tuning paradox: Boosting translation quality without sacrificing LLM abilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6189–6206, Bangkok, Thailand. Association for Computational Linguistics.

Zhewei Sun, Richard Zemel, and Yang Xu. 2022. Semantically informed slang interpretation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5213–5231, Seattle, United States. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Takashi Wada, Yuji Matsumoto, Timothy Baldwin, and Jey Han Lau. 2023. Unsupervised paraphrasing of multiword expressions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4732–4746, Toronto, Canada. Association for Computational Linguistics.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *arXiv preprint arXiv:2401.06468*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *Forty-first International Conference on Machine Learning*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Andrea Zaninello and Alexandra Birch. 2020. Multiword expression aware neural machine translation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3816–3825, Marseille, France. European Language Resources Association.

A Data Annotation

We use Label Studio[12] as our annotation platform. For Bengali, Indonesian and Mandarin Chinese, we recruit four annotators per language. For German-English, two annotators are recruited.

Figure 3 shows the instruction to annotate proverb translation dataset. Figure 4 shows the Label Studio interface for the task.

B Additional Experiment Details

B.1 Data Contamination Analysis

We measure the contamination rate of the LLM using a method similar to Liu et al. (2024). Specifically, for a given sentence and its preceding context $(s_t, s_{<t})$, we create an input prefix by truncating s_t to a certain proportion of its words (τ), denoting the prefix as $s_t^{<\tau}$ and the remaining part as the suffix $s_t^{\geq\tau}$. Two types of prompts are then created, one with context ($X_c = [s_{<t}; s_t^{<\tau}]$) and one without context ($X_\emptyset = s_t^{<\tau}$), where $[;]$ represents string concatenation.

We use the base version of the LLM[13] to complete the given prefix in both forms, resulting in hypotheses denoted as $\hat{Y}^c$ (with context) and $\hat{Y}^\emptyset$ (without context). During the generation, the greedy search strategy (temperature set as 0) is used to ensure the reproduction of the result.

Finally, we compute the length of the longest common subsequence (LCS) between the model's predictions (with and without context) and the reference suffix, denoted as $|\text{LCS}(\hat{Y}^c, s_t^{\geq\tau})|$ and $|\text{LCS}(\hat{Y}^\emptyset, s_t^{\geq\tau})|$. The contamination ratio γ is then estimated as:

$$\gamma = \max(0, \frac{|\text{LCS}(\hat{Y}^c, s_t^{\geq\tau})| - |\text{LCS}(\hat{Y}^\emptyset, s_t^{\geq\tau})|}{|s_t^{\geq\tau}|}) \tag{1}$$

A higher value of γ indicates a greater likelihood that the model has been affected by data contamination for a given sample. The reasoning behind this is as follows: *(i)* A large γ occurs only when the first term ($|\text{LCS}(\hat{Y}^c, s_t^{\geq\tau})|$) is significantly larger than the second term ($|\text{LCS}(\hat{Y}^\emptyset, s_t^{\geq\tau})|$), indicating that the model can predict the suffix accurately when given context but struggles without it. *(ii)* When both terms are large, it likely means the sample is a proverb or well-known phrase, which the model can predict accurately even without context, resulting in a small γ. *(iii)* When both terms are small, it suggests that the model lacks sufficient knowledge to predict the suffix, with or without context.

B.2 Prompt Template for Translation

B.3 Prompt Template for LLM-based Evaluation

[12] https://labelstud.io/

[13] To reduce costs, we evaluated four LLMs at the 7-9B scale.

Objective
Your task is to analyse a given proverb in a language other than English and its provided English translation. You will evaluate the context dependency of the proverb, assess the accuracy of the translation, and suggest corrections if necessary.

- Please be as clear and concise as possible in your explanations.
- If you are unsure about the context dependency or translation accuracy, provide your best judgement and include a note explaining your reasoning.
- Use the space provided for each question to write your answers.

Steps to Follow

1. Read the Proverb and Translation:
  - Carefully read the provided proverb in the original language.
  - Read the given English translation of the proverb.

2. Context Dependency:
  - Determine whether the meaning of the proverb changes based on different contexts or situations.
  - Question to Answer: Is the meaning of this proverb context-dependent?
    - Options: Yes / No

3. Translation Accuracy:
  - Assess the accuracy of the provided English translation.
  - Question to Answer: Is the provided English translation accurate?
    - Options: Yes / No
    - If you select "No," please provide the correct translation of the proverb in English.

Figure 3: Proverb Translation Annotation Instruction

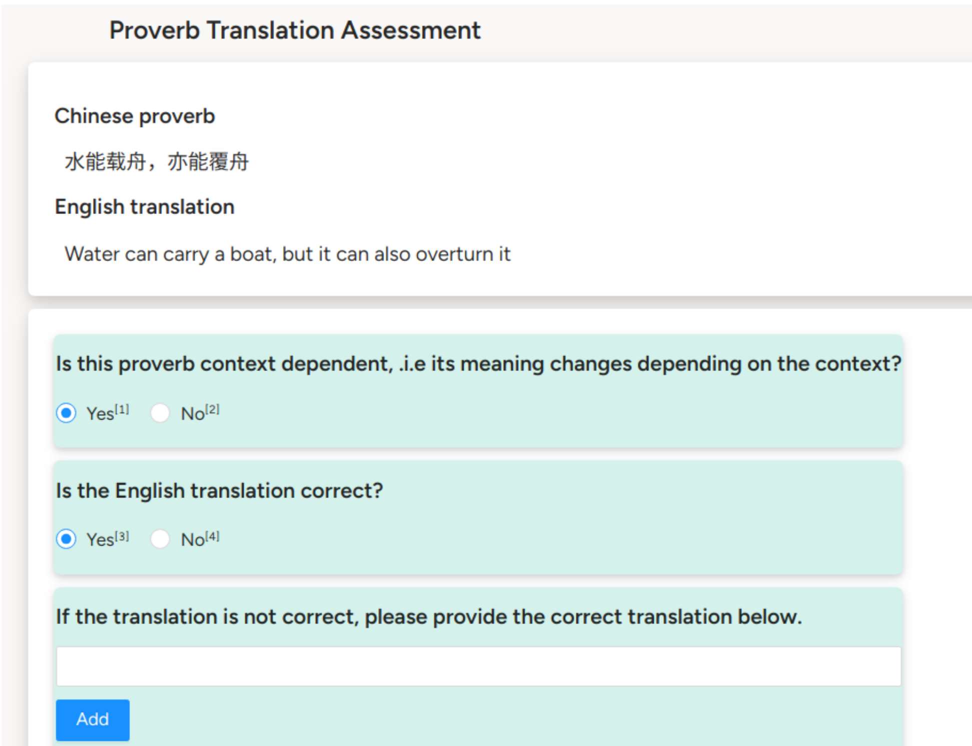

Figure 4: Proverb Translation Annotation Interface

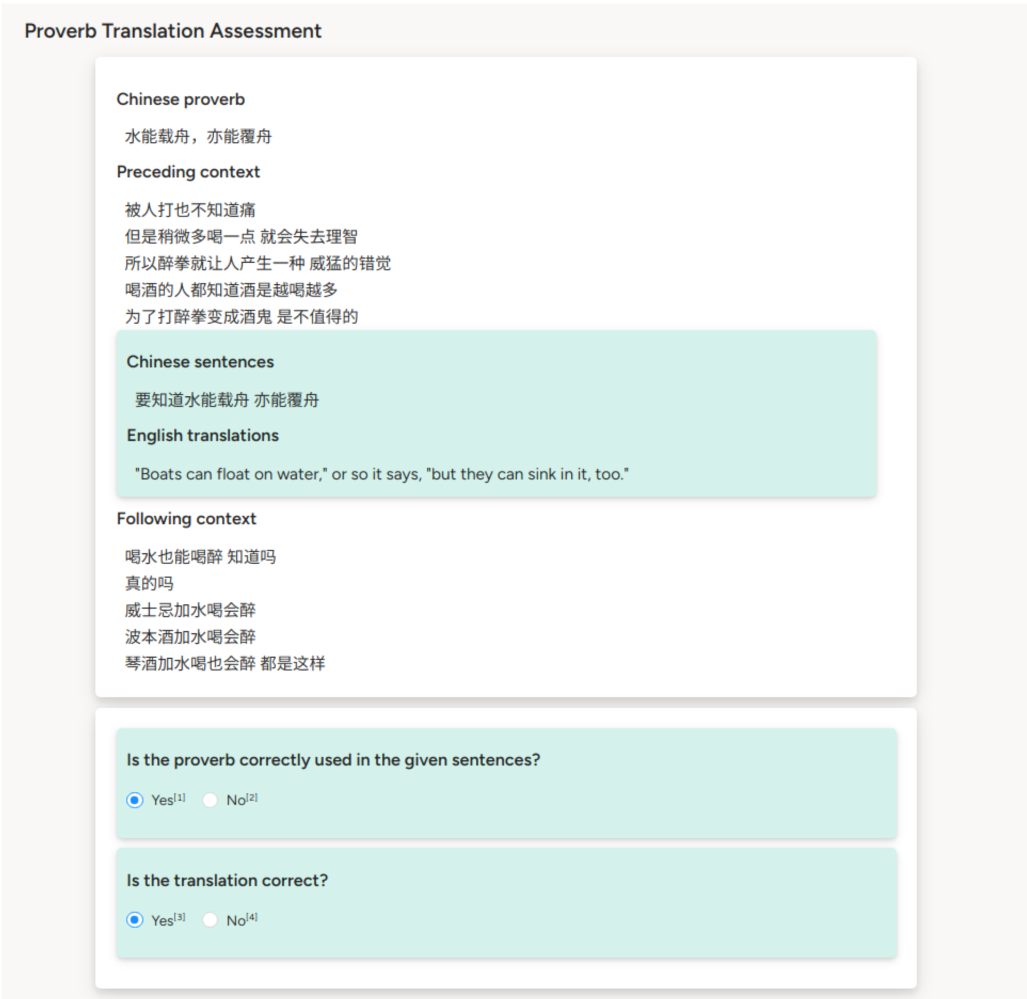

Figure 5: OpenSubtitle Translation Annotation Interface

	Literal					Figurative				
	0-shot	1-shot	EXPL.	DIALOG	CONCAT	0-shot	1-shot	EXPL.	DIALOG	CONCAT
From English translation, incl. EN-DE, EN-ID, EN-ZH										
NLLB-600M	27.33					27.37				
NLLB-1.3B	29.85					29.38				
NLLB-3.3B	**30.70**					31.01				
ALMA-R 13B	20.14					23.05				
QWEN2	23.63	24.87	22.11	27.31	26.89	26.12	25.88	26.00	30.72	29.08
MISTRAL	17.87	18.66	16.12	22.33	22.09	19.80	19.74	18.70	22.23	21.99
GEMMA2 9B	27.93	28.24	26.85	30.94	30.88	32.01	31.08	31.37	33.81	33.17
LLAMA-3.1 8B	23.85	26.62	23.26	28.99	29.13	25.64	26.09	24.98	29.45	29.05
LLAMA-3.1 70B	28.59	29.19	**28.68**	**33.30**	32.40	35.61	**35.72**	**35.07**	**37.97**	**36.54**
GPT-4O MINI	29.92	**30.37**	27.86	32.34	**32.57**	**34.27**	34.51	33.91	36.47	35.03
To English translation, incl. DE-EN, ZH-EN										
NLLB-600M	13.96					18.96				
NLLB-1.3B	18.28					23.05				
NLLB-3.3B	**19.31**					**24.16**				
ALMA-R 13B	16.99					22.42				
QWEN2	14.77	15.56	13.51	20.05	18.06	22.27	22.44	19.85	26.03	24.48
MISTRAL	12.23	13.40	11.55	19.76	17.89	18.56	19.18	17.52	23.45	21.64
GEMMA2 9B	18.17	19.23	17.19	24.51	23.66	22.97	23.34	22.40	27.29	26.66
LLAMA-3.1 8B	16.52	18.42	16.04	23.65	22.48	20.31	20.89	20.91	24.89	24.54
LLAMA-3.1 70B	19.28	**20.25**	**19.56**	**25.96**	**24.07**	23.53	**24.13**	**23.89**	**28.45**	**28.08**
GPT-4O MINI	18.07	17.78	17.25	22.48	21.88	22.72	22.83	21.68	26.08	24.89

Table 8: BLEU scores on Proverb in Conversation Translation. **Bold** highlights the best score in each column. The better score among different prompts are underlined.

	Literal					Figurative				
	0-shot	1-shot	EXPL.	DIALOG	CONCAT	0-shot	1-shot	EXPL.	DIALOG	CONCAT
From English translation, incl. EN-DE, EN-ID, EN-ZH										
NLLB-600M	40.39					41.81				
NLLB-1.3B	42.45					43.31				
NLLB-3.3B	**43.66**					44.17				
ALMA-R 13B	34.29					37.36				
QWEN2	37.51	38.27	36.39	40.93	39.43	40.40	40.18	40.05	46.46	41.46
MISTRAL	31.65	31.71	31.74	36.79	35.52	34.60	34.34	33.94	40.53	40.16
GEMMA2 9B	39.74	39.47	38.78	43.54	41.76	44.59	44.27	42.58	44.97	44.56
LLAMA-3.1 8B	36.24	38.35	35.62	41.26	41.38	38.76	39.91	38.64	42.02	41.35
LLAMA-3.1 70B	41.38	41.41	41.23	45.15	43.74	**46.77**	**46.79**	**46.61**	**47.97**	**47.16**
GPT-4O MINI	42.82	**42.98**	**41.77**	**45.89**	**44.69**	46.15	46.19	46.18	47.93	46.81
To English translation, incl. DE-EN, ZH-EN										
NLLB-600M	29.52					34.35				
NLLB-1.3B	33.11					36.83				
NLLB-3.3B	33.82					37.89				
ALMA-R 13B	33.00					37.72				
QWEN2	31.84	32.07	31.43	35.65	34.61	38.19	38.39	36.72	41.01	39.91
MISTRAL	27.26	29.73	27.55	34.19	33.24	34.08	34.68	33.54	38.39	36.94
GEMMA2 9B	34.03	34.96	33.61	38.98	38.74	38.65	38.86	38.35	42.31	41.75
LLAMA-3.1 8B	30.71	33.47	30.96	37.50	36.86	34.94	35.59	35.66	39.06	39.03
LLAMA-3.1 70B	**34.64**	**35.33**	**35.86**	**40.26**	**38.94**	38.09	38.66	**39.40**	**42.39**	**42.42**
GPT-4O MINI	34.57	34.47	34.15	37.87	37.75	**39.48**	**39.35**	38.41	42.04	41.36

Table 9: CHRF++ scores on Proverb in Conversation Translation. **Bold** highlights the best score in each column. The better score among different prompts are underlined.

System Message

You are a professional translator. Your task is to translate the user input from {src_lang_name} to {tgt_lang_name}. Remember! Don't return any irrelevant content except from the translation!

User Message

{src}

Zero-shot Prompt

System Message

You are a professional translator. Your task is to translate the user input from {src_lang_name} to {tgt_lang_name}. Remember! Don't return any irrelevant content except from the translation!

User Message

{src_sample}

Assistant Message

{tgt_sample}

User Message

{src}

One-shot Prompt

System Message

You are a professional translator. Your task is to translate the user input in the curly brackets from {src_lang_name} to {tgt_lang_name}. The user input may contain the proverb: "{proverb}", where the explanation of that proverb is given as a context for the support of your translation: "{proverb_explanation}". Remember! Don't return any irrelevant content except from the translation!

User Message

{src}

Proverb-Explanation Prompt

System Message

You are a professional translator. Your task is to translate the user input from {src_lang_name} to {tgt_lang_name}. Remember! Don't return any irrelevant content except from the translation!

For i in {0...5}:

User Message

{src_utterance_history[i]}

Assistant Message

{tgt_utterance_history[i]}

End For

User Message

{src}

Contextualization-through-Dialogue Prompt

System Message

You are a professional translator. Your task is to translate the user input from {src_lang_name} to {tgt_lang_name}. Remember! Don't return any irrelevant content except from the translation!

User Message

{' '.join(src_utterance_history[i] for i in range(5))}

Assistant Message

{' '.join(tgt_utterance_history[i] for i in range(5))}

User Message

{src}

Contextualization-through-Concatenation Prompt

Figure 6: The prompt template of subtitle translation.

You are an expert bilingual translation evaluator in {src_lang} and {tgt_lang}, with deep knowledge of both cultures, especially in the use of proverbs and idioms. Your task is to evaluate two translations ({src_lang}->{tgt_lang}) that may contain proverbs or culturally specific content.

Evaluation Criteria:

1. **Accuracy**: Does the translation accurately convey the meaning and intent of the source text, considering the context?
2. **Fluency**: Is the translation grammatically correct and natural-sounding in the target language?
3. **Cultural Appropriateness**: Has the translation appropriately handled cultural references, proverbs, and idiomatic expressions? Specifically, are proverbs translated into equivalent proverbs or expressions in the target language to maintain cultural resonance?

Context:

- **Preceding Context ({src_lang}):**

  {src_pre_context}

- **Preceding Context ({tgt_lang}):**

  {tgt_pre_context}

- **Following Context ({src_lang}):**

  {src_post_context}

- **Following Context ({tgt_lang}):**

  {tgt_post_context}

Source Text ({src_lang}):

{src}

Target Reference ({tgt_lang}):

{ref}

Translation A:

{hyp1}

Translation B:

{hyp2}

Instructions:

- Analyze both translations based on the criteria above, taking into account the context and the reference provided.
- Conclude by stating which translation is better overall according to the evaluation criteria, or if they are equally good or bad.
- Return your answer in a JSON object: {{"result":"A"}} or {{"result":"B"}} or {{result:"tie"}}.

Question:

Which translation is better according to the evaluation criteria above?

Figure 7: The prompt template of LLM-based evaluation.