

Benchmarking Machine Translation on Chinese Social Media Texts

Kaiyan Zhao¹, Zheyong Xie², Zhongtao Miao¹, Xinze Lyu², Shaosheng Cao^{2,3*}

¹The University of Tokyo, ²Xiaohongshu Inc., ³Tsinghua University

caoshaosheng@xiaohongshu.com

Abstract

The prevalence of rapidly evolving slang, neologisms, and highly stylized expressions in informal user-generated text, particularly on Chinese social media, poses significant challenges for Machine Translation (MT) benchmarking. Specifically, we identify two primary obstacles: (1) data scarcity, as high-quality parallel data requires bilingual annotators familiar with platform-specific slang, and stylistic cues; and (2) metric limitations, where traditional evaluators like COMET often fail to capture stylistic fidelity and non-standard expressions. To bridge these gaps, we first introduce CSM-MTBench, an expert-curated benchmark covering five Chinese→foreign language directions and comprising 10,915 total instances, organized into two subsets: **Fun Posts**, featuring context-rich content with frequent slang and neologisms, and **Social Snippets**, focusing on concise, emotion- and style-driven expressions. Next, we propose tailored evaluation approaches for each subset, including slang translation success rate and style-preservation metrics combining embeddings and LLM-as-a-judge evaluation. Experiments on over 20 models reveal substantial variation in how current MT systems handle semantic fidelity, slang translation, and stylistic nuances. We further look into model behaviors through prompt sensitivity analysis and qualitative case studies, highlighting the challenges of translating these informal expressions. CSM-MTBench thus serves as a rigorous testbed for advancing MT systems capable of mastering real-world Chinese social media texts.¹

1 Introduction

Machine Translation (MT) has long been one of the central tasks in Natural Language Processing (NLP), as it enables communication and

knowledge exchange across languages (Bahdanau et al., 2015; Newman-Griffis et al., 2021). Despite recent advances with Large Language Models (LLMs) (Wang et al., 2023; Zhu et al., 2024; Xu et al., 2024a; Wu et al., 2024), translating informal user-generated content on social media remains challenging (Qian et al., 2024; Guo et al., 2025).

Unlike the formal texts (e.g., news or Wikipedia) commonly used in existing MT benchmarks (Guzmán et al., 2019; Goyal et al., 2022), social media texts exhibit distinct linguistic characteristics, including rapidly evolving slang, informal expressions, and highly stylized language (Al Amer et al., 2025). This phenomenon is particularly prominent on Chinese social media platforms, where users frequently create neologisms, abbreviations, and novel expressions through creative recombinations of existing characters, many of which evolve rapidly over time and are rarely captured by static MT benchmarks (Miao et al., 2026; Zhao et al., 2026). Figure 1 illustrates an example where MT models translate two semantically equivalent Chinese sentences of distinct styles, formal versus social media, into Japanese. While the models successfully handle the formal sentence, they fail to preserve the slang and the tone specific to social media, highlighting the challenges of translating these informal and creatively recombined² user-generated content.

To build a benchmark for Chinese social media MT, we identify two main challenges: data scarcity and evaluation limitations. Existing Chinese social media resources such as BOLT (Song et al., 2014) and HADQAET (Qian et al., 2023) largely reflect earlier-stage social-media languages that can be easily handled by recent LLMs. Meanwhile, constructing high-quality annotations for up-to-date social media data requires bilingual annotators who

*Corresponding author.

¹We release the dataset at <https://github.com/KYuto1006/CSM-MTBench>.

²For example, the Chinese phrase “无语到家了” (really speechless) in Figure 1 creatively combines “无语” (speechless) with “到家了” (back to home), forming a new slang.

	 Formal Tone 我对这次活动感到很失望	v.s.	 Social Media Style 这活动就这？我真是无语到家了😞
Qwen3-8B	このイベントにとてものがっかりしました。✔️ <i>I feel disappointed in this event.</i>		このイベントってこれ？俺、本当に言葉も出ないくらい無理やりだ😞。❌ <i>This event, this? I, really, to the point that I can't even speak, is forced 😞.</i>
Google Translate	私はこのイベントに非常に失望しました。✔️ <i>I feel disappointed in this event.</i>		イベントはこれだけです？言葉が出ません😞。❌ <i>Is this the only event? I'm speechless 😞.</i>
Gold label	私は今回のイベントにとても失望しました。 <i>I feel disappointed in this event.</i>		このイベントってこれだけ？本当に呆れたわ😞 <i>This event? That's all? I'm honestly speechless 😞.</i>

Figure 1: MT into Japanese using Qwen3-8B and Google Translation for two Chinese source sentences expressing the same meaning (“I feel disappointed in this event.”) in formal tone versus social media style. English references are shown in *italics*. Social-media-specific tone is highlighted in **red**, and slang is highlighted in **blue**.

are familiar with social media expressions in both languages, a skill set that is not widely available. In addition, traditional automatic metrics including BLEU (Papineni et al., 2002) and chrF (Popović, 2015) and COMET (Rei et al., 2020) are often insufficient for evaluating translations of informal user-generated content where meaning may be conveyed through special expressions.

To address these challenges, we introduce a new benchmark, CSM-MTBench, comprising 10,915 total translation samples for systematically evaluating MT performance on authentic Chinese social media texts. The up-to-date source data are collected from a real Chinese social media platform and manually translated into Spanish (es), French (fr), Japanese (ja), Korean (ko) and Russian (ru) by expert human annotators, ensuring high-quality multilingual references for evaluation. Specifically, CSM-MTBench consists of two complementary datasets: **Fun Posts** and **Social Snippets**. Fun Posts contain longer, content-rich user posts describing events or experiences and often include slang and neologisms. In contrast, Social Snippets capture short, highly emotional comments that exhibit distinctive tones and stylistic features. Together, these datasets reflect the diverse linguistic styles in Chinese social media.

To better evaluate the MT performance, we combine standard evaluation metrics such as XCOMET (Guerreiro et al., 2024) with targeted evaluation methods tailored to the distinct challenges of each dataset. For Fun Posts, we focus on the accurate translation of neologisms and slang, leveraging a carefully curated slang dictionary and fuzzy matching to determine whether candidate translations appear in the model output. For Social Snippets, we evaluate the preservation of social-media-specific tone and style by combining style

embeddings (Patel et al., 2025), emotion embeddings (Poulaei et al., 2025), and sentiment embeddings (tabularisai et al., 2025), complemented by LLM-as-a-judge (Gu et al., 2025) method to assess tone and style consistency.

We evaluate over 20 different models on CSM-MTBench, including both MT-specific models and general-purpose LLMs. Our results reveal several notable trends. While LLMs achieve strong overall performance on Fun Posts according to standard metrics, their translations of neologisms and slang remain inconsistent. On Social Snippets, models often fail to preserve the tone and style of the source, particularly in emotionally charged content. These findings highlight the challenges of translating real-world Chinese social media and the need for benchmarks and evaluation methods that better capture these linguistic phenomena. We further conduct prompt sensitivity experiments and qualitative analyses to better understand model behaviors, and find that even simple sentences become difficult for models to translate when additional emotional features, such as emojis or repeated exclamation marks are present.

2 Related Work

2.1 Machine Translation with LLMs

Initial attempts of applying general-purpose LLMs to MT, including models such as GPT3 (Brown et al., 2020) and BLOOM (BigScience Workshop et al., 2023), demonstrate strong multilingual capabilities and promising zero-shot and few-shot performance of LLMs. Subsequent work shows that prompting plays a crucial role in unlocking LLMs’ translation ability (Vilar et al., 2023; Zhang et al., 2023; Ghazvininejad et al., 2023), while studies such as Chitale et al. (2024) and Lee et al. (2025)

further establish the effectiveness of in-context learning (ICL) for MT. The Aya family (Üstün et al., 2024; Aryabumi et al., 2024; Dang et al., 2024) represents a line of LLMs optimized explicitly for multilingual abilities.

In parallel, recent research has explored reinforcement learning for MT specific models (Xu et al., 2024b; Wu et al., 2024; Guo et al., 2025). The Tower series models (Alves et al., 2024) are proposed to deal with multiple tasks present in translation workflows. GemmaX2 (Cui et al., 2025) incorporates data-mixing strategies during continual pretraining to enhance multilingual performance, while Hunyuan-MT (Zheng et al., 2025) employs carefully engineered pre-training and post-training pipelines tailored specifically for MT.

2.2 MT on Social Media Texts

Prior work has explored MT for informal text in several languages. The NUS SMS Corpus (Chen and Kan, 2012) contains short message data, reflecting earlier chat-style informal texts. Open-Subtitles (Lison and Tiedemann, 2016) provides conversational parallel data derived from movie subtitles, which exhibit spoken language characteristics. Michel and Neubig (2018) introduce MTNT, a parallel corpus of noisy social media text collected from Reddit, designed to evaluate English, French and Japanese MT performance. MMTC (McNamee and Duh, 2022) is constructed from multilingual Twitter posts and focuses on translation of user-generated microblog content.

Closest to our setting, several datasets have been proposed for Chinese social media MT. Ling et al. (2013) rely on automatic mining methods for microblogs data. BOLT (Song et al., 2014) collects informal chat data from online messaging platforms. Qian et al. (2024) construct a Weibo³-based dataset, with evaluation restricted to the zh→en direction. Guo et al. (2025) propose Redtrans-Bench, which contains 2858 zh→en SNS (Social Networking Service) translation pairs. However, the dataset is limited to a single language pair, and evaluation is only performed using XCOMET (Guerreiro et al., 2024), without consideration for SNS-specific features.

CSM-MTBench differs from existing resources in four important aspects: (1) it is constructed from up-to-date Chinese social media data, capturing the rapidly evolving slang and stylistic conventions on modern platforms; (2) it covers multiple target

languages beyond the commonly studied zh→en direction; (3) a more careful curation process with human annotations; and (4) targeted evaluation metrics for social-media texts besides traditional ones.

2.3 Domain-specific MT

For Chinese-specific domains, recent research has explored MT challenges arising from culturally rich or linguistically unique content. Several works focus on Chinese idiom translation (Li et al., 2024; Fu et al., 2025; Yang et al., 2025b), a domain where fixed expressions and metaphorical meanings pose difficulties for both neural MT systems and LLMs. WYWEB (Zhou et al., 2023) introduces a benchmark for classical ancient Chinese, covering tasks such as sentence classification and MT. Similarly, Chen et al. (2025) propose PoetMT for translating classical Chinese poetry into English.

For culturally related domains more broadly, Yao et al. (2024) evaluate MT on culture-specific entities and concepts across languages. NEO-BENCH (Zheng et al., 2024) examines model robustness to neologisms, with MT evaluation on en→zh samples. NewTerm (Lerner and Yvon, 2025) focuses on scientific neologisms which covers en→fr translation.

3 CSM-MTBench

In this section, we first describe the construction of CSM-MTBench, followed by detailed explanations of the two subsets: Fun Posts and Social Snippets, and their corresponding evaluation metrics.

3.1 Dataset Construction

The source data of CSM-MTBench are collected from Xiaohongshu⁴, a popular Chinese social media platform where users publish posts and interact through comments. We first collect a large pool of candidate data from January 2025 to July 2025, including over 100k posts and comments.

Unlike prior work that typically relies on random sampling through platform APIs, we adopt an engagement-aware data construction strategy to better capture representative social-media language. For posts, we select 1,500 posts as candidate source texts by considering exposure level, view counts, and the number of likes, aiming to prioritize widely viewed and interacted posts that reflect commonly encountered language patterns.

³A mainstream Chinese social platform.

⁴Also known as RedNote, see <https://en.wikipedia.org/wiki/Xiaohongshu>

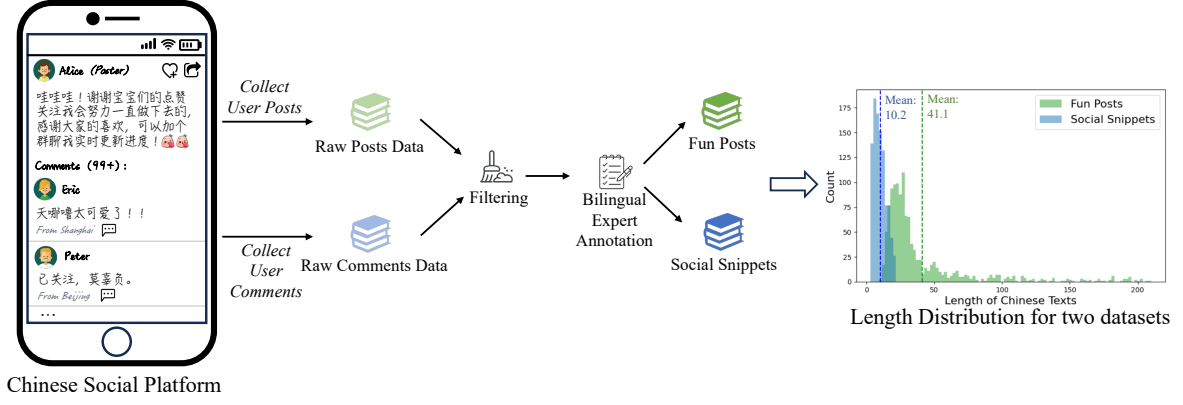


Figure 2: Construction of CSM-MTBench. We collect real-world data from a Chinese social platform and split them into Fun Posts and Social Snippets. All data undergo filtering and bilingual expert translation.

Subset	Pair num.	Total num.	Total Samples
Fun Posts	1,183	5,915	10,915
Social Snippets	1,000	5,000	

Table 1: Statistics for CSM-MTBench. Pair nums stands for the number of samples per source-target pair.

For comments, based on engagement signals (e.g., likes, replies) and expressiveness of the text, we initially selected 5,000 candidate comments.

We then separate user posts and comments into two subsets: Fun Posts and Social Snippets (Figure 2; English translations of examples are in Appendix, Figure 4). The candidate data undergo filtering before being translated into Spanish, French, Japanese, Korean, and Russian by bilingual experts. After filtering, 1,183 samples are retained in Fun Posts and 1,000 in Social Snippets. Dataset statistics are summarized in Table 1, and Figure 3 visualizes the t-SNE figure for the two subsets using Qwen3-Embedding-8B. The two subsets exhibit partial separation with limited overlap, indicating noticeable distributional differences. Quantitatively, the cosine similarity between the subset centroids is approximately 0.44, and the normalized centroid distance is 0.24. These values suggest a meaningful semantic distribution shift rather than merely minor stylistic variation. Additional details on the data processing pipeline are provided in Appendix A.1 and A.2.

3.2 Descriptive Side: Fun Posts

As shown in the right part of Figure 2, Fun Posts consist of longer and more coherent user posts that resemble short narratives or mini-blogs, with an average length of 41.1 Chinese characters. These sentences are generally complete and information-

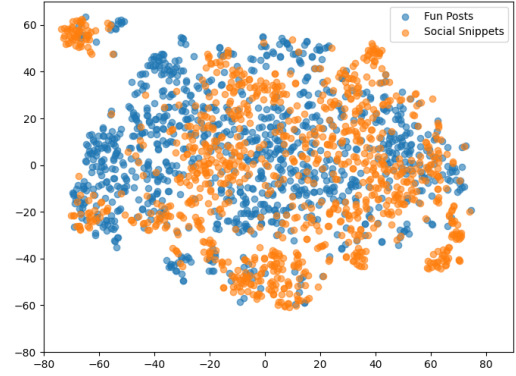


Figure 3: t-SNE illustration for two subsets.

rich, often containing neologisms and slang (see Appendix, Figure 7). Fun Posts thus represent the descriptive side of social media, where successful translation requires preserving semantic nuances.

To quantify the preservation of these expressions, we introduce the Slang Success Rate (SSR), following Semenov et al. (2025). Given the i_{th} annotated source-target pair (S_i, T_i) , we first prompt GPT-5 to identify slang or neologistic expressions s in the Chinese source sentence and locate their corresponding translations t in the gold target sentence. This produces a set of slang-translation pairs:

$$\mathcal{P}_i = [(s_1, t_1), \dots, (s_k, t_k)], \quad (1)$$

where k denotes the number of slang items in the i_{th} source-target pair.

Based on these slang-translation pairs, we further prompt GPT-5 to augment each pair into a candidate dictionary, where multiple plausible target-language expressions are generated under the given

context. For each slang item s_j , GPT-5 generates a list of valid translation candidates, denoted as $\mathcal{C}_j$:

$$\mathcal{C}_j = [\text{candi}_j^1, \dots, \text{candi}_j^m], \quad (2)$$

where m is the number of generated candidate expressions. We then define the full set of acceptable translation candidates for s_j as $\mathcal{C}_j^+ = [t_j, \text{candi}_j^1, \dots, \text{candi}_j^m]$. The prompt employed to identify slang-translation pairs $\mathcal{P}$ and generate these candidates is presented in Figure 8, Appendix. Subsequently, human annotators review the generated candidates to filter out invalid entries. The final data structure for Fun Posts is illustrated in the upper part of Figure 7.

After obtaining the model translation $\hat{T}_i$, we apply fuzzy matching⁵ to determine whether any expression in $\mathcal{C}_j^+$ appears in the output. The score for slang item s_j is defined as:

$$\text{Score}_{ij} = \begin{cases} 1, & \text{if } FM(\hat{T}_i, \mathcal{C}_j^+) \geq \text{threshold} \\ 0, & \text{otherwise} \end{cases}, \quad (3)$$

where $FM(\cdot)$ denotes the fuzzy matching function. Based on these scores, the Slang Success Rate is defined as:

$$SSR = \frac{1}{N} \sum_{i=1}^N \frac{1}{k} \sum_{j=1}^k \text{Score}_{ij}, \quad (4)$$

where N denotes the number of source–target pairs containing slang or neologisms.

In total, 619 of the 1,183 Fun Posts samples contain detected slang or neologisms, and each slang term has 6.6 candidate translations on average.

3.3 Reactive Side: Social Snippets

Social Snippets, by contrast, capture the reactive and highly informal side of social media communication. These entries are typically short, sometimes only a few characters, and often rely on tone markers and stylized expressions that convey meaning beyond the literal words. Much of the content is emotion- or attitude-driven, expressing reactions such as laughter, surprise, or frustration rather than conveying factual information (see Appendix, Figure 7). As a result, Social Snippets provide a challenging testbed for MT systems, where standard metrics such as XCOMET (Guerreiro et al., 2024) may not fully capture the informal and nuance-rich characteristics of the text.

⁵<https://github.com/rapidfuzz/RapidFuzz>

Unlike Fun Posts, the lexical simplicity of Social Snippets makes it difficult to isolate interpretable slang or neologisms for evaluation (see Appendix A.3.2). To address this issue, we adopt an embedding-based evaluation that measures whether translations preserve stylistic and emotional properties of the source. Specifically, we combine three types of embeddings: style embeddings (Patel et al., 2025), emotion embeddings (Poulaei et al., 2025), and sentiment embeddings (tabularisai et al., 2025).

Given a source sentence S and model translation $\hat{T}$, we encode them using the embedding models and compute the cosine similarity between them:

$$e_S, e_{\hat{T}} = \text{EmbeddingModel}(S, \hat{T}), \quad (5)$$

$$\text{sim} = \text{CosineSim}(e_S, e_{\hat{T}}). \quad (6)$$

The final Embedding Similarity (ES) score is obtained by averaging the cosine similarities across the three embedding types, providing a holistic measure of how well the translation preserves the source’s style, emotion, and sentiment.

To further leverage our annotated translations, we also adopt a GEMBA-based evaluation method (Kocmi and Federmann, 2023). Specifically, we use the GEMBA-stars prompting framework (see Appendix Figure 6), where an LLM judges whether the tone and style of the source sentence are preserved in the translation.

4 Experiments

4.1 Experimental Settings

4.1.1 Evaluated Models

We evaluate a broad set of models, covering both traditional MT architectures and modern LLMs. The full list of 22 evaluated models is as follows:

- **Closed-source LLM APIs:** GPT-4o (OpenAI et al., 2024), GPT-5 (OpenAI, 2025a), Claude-Sonnet-4 (Anthropic, 2025);
- **Open-source general-purpose LLMs:** Deepseek-V3 (DeepSeek-AI et al., 2025), GPT-OSS-120B (OpenAI, 2025b), Aya-101 (Üstün et al., 2024), Aya-Expanse-8B (Dang et al., 2024), three Gemma3-IT (Gemma Team et al., 2025) series models (4B, 12B, 27B), and eight Qwen3 (Yang et al., 2025a) series models (1.7B, 4B, 4B-Instruct, 8B, 32B, 30B-A3B, 30B-A3B-Instruct, 235B-A22B);

Models	$zh \rightarrow es$		$zh \rightarrow fr$		$zh \rightarrow ja$		$zh \rightarrow ko$		$zh \rightarrow ru$		avg.	
	XCOMET	SSR	XCOMET	SSR	XCOMET	SSR	XCOMET	SSR	XCOMET	SSR	XCOMET	SSR
<i>Closed-source LLM APIs</i>												
GPT-4o	86.38	69.14	83.06	68.01	88.21	63.17	85.36	60.58	84.49	63.65	85.50	64.91
GPT-5	86.34	79.32	82.67	81.09	87.64	75.12	<u>85.46</u>	76.58	<u>84.17</u>	76.41	<u>85.26</u>	77.70
Claude-Sonnet-4	86.05	<u>74.15</u>	82.49	<u>73.34</u>	<u>87.81</u>	<u>66.72</u>	85.54	<u>72.21</u>	84.07	<u>70.11</u>	85.19	<u>71.31</u>
<i>Translation-specialized Models</i>												
NLLB-3.3B	70.07	27.95	64.50	27.14	61.52	18.26	64.28	21.16	67.09	23.10	65.49	23.52
google-translate	83.63	58.97	79.87	55.57	82.90	49.43	81.55	48.95	<u>81.72</u>	54.93	81.93	53.57
GemmaX2-9B	<u>84.06</u>	<u>57.03</u>	80.18	<u>55.41</u>	<u>84.94</u>	49.76	<u>82.95</u>	<u>49.60</u>	81.45	<u>51.59</u>	<u>82.72</u>	<u>52.68</u>
Hunyuan-MT-7B	84.35	54.60	<u>80.17</u>	52.18	86.31	<u>49.60</u>	85.05	51.70	82.66	49.92	83.71	51.60
<i>Open-source General-purpose LLMs</i>												
DeepSeek-V3	<u>84.33</u>	70.11	<u>81.05</u>	70.92	87.26	67.04	<u>84.34</u>	64.94	83.50	69.31	84.10	68.46
GPT-OSS-120B	84.25	65.11	<u>81.05</u>	62.04	85.47	58.64	83.80	58.16	81.62	56.22	<u>83.24</u>	60.03
Aya-101	74.26	29.89	69.45	28.59	77.36	30.86	72.89	26.82	72.32	29.08	73.26	29.05
Aya-Expanse-8B	82.31	56.54	78.25	54.28	83.42	51.86	81.64	48.47	78.70	52.10	80.86	52.65
Gemma3-4B	79.72	53.15	74.98	48.79	80.61	42.49	77.67	44.43	76.83	46.37	77.96	47.05
Gemma3-12B	83.21	56.38	78.84	57.35	83.88	51.05	81.25	54.60	81.13	57.18	81.66	55.31
Gemma3-27B	84.49	64.29	80.73	60.90	85.65	56.22	83.13	55.09	<u>82.29</u>	<u>58.16</u>	83.26	58.93
Qwen3-1.7B	68.43	39.26	68.56	37.48	75.58	35.86	65.51	30.86	67.12	31.18	69.04	34.93
Qwen3-4B	79.99	48.30	75.47	46.37	80.24	41.84	74.04	40.39	74.55	40.06	76.86	43.39
Qwen3-4B-Ins	81.96	54.12	77.49	52.02	82.64	46.37	78.03	44.91	77.28	49.11	79.48	49.31
Qwen3-8B	81.35	57.99	78.43	52.83	83.97	49.60	79.32	48.47	78.08	50.88	80.23	51.95
Qwen3-32B	83.90	60.09	79.86	58.64	85.15	55.57	82.02	53.96	80.34	51.86	82.25	56.02
Qwen3-30B-A3B	83.41	58.48	79.27	58.16	83.75	54.93	81.25	53.80	79.54	51.70	81.44	55.41
Qwen3-30B-A3B-Ins	84.25	60.90	80.49	58.18	85.76	54.28	82.16	58.00	80.50	55.90	82.63	57.45
Qwen3-235B-A22B	85.33	<u>65.43</u>	82.39	<u>64.30</u>	<u>86.57</u>	<u>60.26</u>	84.63	<u>60.26</u>	81.59	54.93	84.10	<u>61.04</u>

Table 2: Results for Fun Posts. SSR denotes Slang Success Rate. Within each model category, the best and second-best scores are highlighted in **bold** and underlined font, respectively.

- **Translation-specialized Models:** NLLB-3.3B (NLLB Team et al., 2022), Hunyuan-MT-7B (Zheng et al., 2025), GemmaX2-9B (Cui et al., 2025) and Google Translate⁶.

This selection allows us to examine model performance across a range of architectures and model scales, providing a comprehensive view of MT capabilities on Chinese social media text.

4.1.2 Implementation Details

We use XCOMET-XL⁷ for overall translation quality evaluation. For Fun Posts, the threshold for fuzzy matching is set to 0.8 to tolerate surface variations that should not be penalized (tense, pluralization, abbreviation, details in Appendix A.7.1). For Social Snippets, style embeddings are obtained using mStyleDistance (Qiu et al., 2025)⁸, emotion embeddings are produced using the model from Bianchi et al. (2022)⁹, and sentiment embeddings are produced using tabularisai et al. (2025)¹⁰.

⁶<https://docs.cloud.google.com/translate/docs/reference/rpc/google.cloud.translate.v2>

⁷<https://huggingface.co/Unbabel/XCOMET-XL>

⁸<https://huggingface.co/StyleDistance/mstyledistance>

⁹<https://huggingface.co/MilaNLPProc/xlm-emo-t>

¹⁰<https://huggingface.co/tabularisai/multilingual-sentiment-analysis>

Other implementation details can be found in Appendix A.4.

4.2 Main Experimental Results

4.2.1 Fun Posts

We summarize the results for Fun Posts in Table 2. For clarity, we divide the table into: (1) closed-source LLM APIs, (2) translation-specialized models, and (3) open-source general-purpose LLMs.

Closed-source LLM APIs exhibit strong translation quality across all five directions. On XCOMET, GPT-4o and GPT-5 perform comparably (the difference is not statistically significant between the two models, Table 6), while GPT-5 leads by a large and consistent margin on SSR, suggesting superior handling of slang and neologisms in informal social media text. Claude-Sonnet-4 performs slightly below the OpenAI models on XCOMET but surpasses GPT-4o on SSR, highlighting its strength in capturing colloquial and stylistic aspects of online discourse. This trend aligns with the observations from Kocmi et al. (2025). Across languages, models show complementary strengths. For example, Claude-Sonnet-4 performs particularly well on $zh \rightarrow ko$, while all APIs demonstrate strong results on $zh \rightarrow ja$, reflecting the relatively close language similarity between Chinese

Models	$zh \rightarrow es$		$zh \rightarrow fr$		$zh \rightarrow ja$		$zh \rightarrow ko$		$zh \rightarrow ru$		avg.		
	XCOMET	ES	XCOMET	ES	XCOMET	ES	XCOMET	ES	XCOMET	ES	XCOMET	ES	GEMBA
<i>Closed-source LLM APIs</i>													
GPT-4o	76.67	67.84	71.60	67.82	77.83	67.27	81.77	67.25	74.61	70.83	76.50	68.20	3.65
GPT-5	77.18	67.54	71.91	<u>67.80</u>	<u>77.21</u>	68.16	<u>81.81</u>	<u>67.24</u>	74.63	<u>70.63</u>	76.55	68.27	3.69
Claude-Sonnet-4	75.04	68.83	68.37	67.47	76.81	<u>67.74</u>	81.98	66.35	73.48	70.12	75.14	68.10	3.58
<i>Translation-specialized Models</i>													
NLLB-3.3B	64.00	63.77	54.49	61.55	53.94	59.55	59.01	59.80	60.39	64.03	58.37	61.74	2.13
google-translate	75.68	66.51	71.54	66.04	71.67	64.62	77.64	64.50	<u>72.51</u>	<u>68.39</u>	73.81	66.01	3.17
GemmaX2-9B	<u>76.64</u>	<u>67.13</u>	70.26	<u>66.69</u>	<u>75.98</u>	<u>65.34</u>	<u>79.84</u>	<u>65.98</u>	72.00	68.07	<u>74.94</u>	<u>66.64</u>	<u>3.19</u>
Hunyuan-MT-7B	77.10	67.21	<u>70.59</u>	66.84	76.62	65.35	81.49	66.04	73.86	68.79	75.93	66.85	3.28
<i>Open-source General-purpose LLMs</i>													
DeepSeek-V3	77.32	68.31	72.92	67.51	78.30	67.45	<u>81.66</u>	<u>66.97</u>	74.70	70.17	76.98	68.08	3.64
GPT-OSS-120B	<u>77.03</u>	66.89	<u>72.00</u>	66.97	<u>76.93</u>	65.57	81.69	66.59	<u>73.68</u>	69.21	<u>76.27</u>	67.05	<u>3.57</u>
Aya-101	69.38	61.20	62.73	63.17	69.39	60.51	72.54	63.05	67.52	64.79	68.31	62.54	2.28
Aya-Expand-8B	72.33	66.23	66.42	65.01	71.19	64.99	76.99	65.03	68.65	68.05	71.12	65.86	3.14
Gemma3-4B	71.35	65.19	64.88	64.39	68.85	64.86	70.98	64.33	66.79	67.21	68.57	65.20	2.85
Gemma3-12B	74.77	65.90	68.57	65.58	73.31	65.30	77.99	65.21	70.47	67.37	73.02	65.87	3.20
Gemma3-27B	74.86	66.33	69.22	65.93	74.18	65.75	79.08	65.58	71.90	67.69	73.85	66.26	3.37
Qwen3-1.7B	68.24	65.65	61.38	65.23	67.97	65.09	64.73	64.13	63.65	67.28	65.19	65.48	2.50
Qwen3-4B	72.41	66.01	66.68	66.34	68.36	66.08	71.66	65.21	67.47	69.22	69.32	66.57	2.79
Qwen3-4B-Ins	73.38	67.15	68.16	67.07	72.20	66.83	75.32	65.80	68.83	69.50	71.58	67.27	3.03
Qwen3-8B	74.06	66.53	67.84	66.96	72.75	67.10	76.09	66.20	69.93	69.66	72.13	67.29	3.06
Qwen3-32B	74.35	66.98	69.29	66.93	73.90	67.55	77.67	66.69	70.91	<u>70.23</u>	73.22	67.68	3.24
Qwen3-30B-A3B	74.14	66.33	68.69	66.82	73.10	66.81	76.87	66.78	70.68	69.07	72.70	67.16	3.17
Qwen3-30B-A3B-Ins	75.34	67.27	69.56	67.28	74.75	66.64	78.55	66.88	71.25	69.55	73.89	67.52	3.32
Qwen3-235B-A22B	76.92	<u>67.38</u>	70.94	<u>67.44</u>	74.82	<u>67.51</u>	80.80	67.23	72.23	70.53	75.14	<u>68.02</u>	3.41

Table 3: Results for Social Snippets. ES denotes Embedding Similarity. Within each category, the best and second-best scores are highlighted in **bold** and underlined font, respectively. Full GEMBA results are in Table 5.

and Japanese.

Translation-specialized models reveal a consistent gap compared with closed-source LLM APIs. Traditional encoder-decoder models such as NLLB perform noticeably worse across most metrics. Among other models, Hunyuan-MT-7B stands out with robust XCOMET performance across all language pairs, achieving quality comparable to larger models. GemmaX2-9B also performs well in both semantic correctness and slang preservation. Google Translate achieves the highest SSR within this group, indicating relatively strong preservation of informal social-media expressions. Nevertheless, translation-specialized systems still lag behind closed-source LLMs in SSR.

Open-source general-purpose LLMs present a wide performance spectrum, with larger models generally outperforming smaller ones. DeepSeek-V3 leads this group, achieving the highest XCOMET and SSR among open-source models and surpassing GPT-4o in SSR. Qwen3-235B-A22B also ranks among the strongest systems overall. Within the Gemma3 and Qwen3 families, performance improves consistently with model scale, and recent instruction-tuned variants outperform earlier versions. This trend suggests that both scaling and advanced post-training significantly improve translation robustness (Yang et al., 2025a).

Statistical significance tests on our results are provided in Appendix A.6.

4.2.2 Social Snippets

We summarize the results for Social Snippets, including XCOMET, the combined embedding similarity (ES) and Gemba-stars (GEMBA) in Table 3. Overall, XCOMET scores are noticeably lower than those on Fun Posts, highlighting the difficulty of translating short, informal, and emotionally expressive social media texts.

Closed-source LLM APIs continue to achieve strong performance. Unlike in Fun Posts, GPT-5 surpasses GPT-4o across both XCOMET and ES, suggesting stronger capability in capturing emotional nuance and informal tone. GEMBA evaluations closely follow the ES trend, further confirming GPT-5’s advantage in preserving stylistic aspects of short social-media texts (OpenAI, 2025a).

Translation-specialized Models show similar patterns to those observed on Fun Posts. Hunyuan-MT-7B achieves the strongest performance among this group across all metrics, with GemmaX2-9B ranking slightly behind. Google Translate preserves the meaning of slang reasonably well but struggles with informal tone and stylistic nuances. Traditional encoder-decoder models such as NLLB-3.3B perform substantially worse on this dataset.

Open-source General-purpose LLMs again exhibit clear scaling trends, with larger models consistently outperforming smaller ones. DeepSeek-V3 leads this group across all metrics and achieves performance comparable to closed-source APIs. Other trends observed on Fun Posts, such as the benefits of instruction tuning, remain consistent.

Overall, the combined embeddings provide a reasonable assessment of translations’ style and emotional fidelity, largely consistent with GEMBA-stars evaluations. Considering the computational resources required for using LLMs as evaluators, ES offers an efficient yet effective alternative. More discussion on ES is provided in Appendix A.5, and we provide human judgments on these automatic metrics in Appendix A.7.

4.3 Analysis

4.3.1 Prompt Sensitivity Analysis

Prior work has shown that prompting plays a critical role in enhancing translation performance, for example by explicitly providing definitions of culture-specific words or phrases (Ghazvininejad et al., 2023; Yao et al., 2024). However, for Chinese social-media-specific slang and neologisms, whose meanings are highly implicit and evolving, manually supplying accurate definitions is both difficult and labor-intensive. Instead, we investigate prompt sensitivity on CSM-MTBench through five prompting strategies: (1) Direct Instruction (en), (2) Reminder Prompt, (3) Direct Instruction (zh), (4) Two-stage Reasoning and (5) Few-shot. Details for these prompts are provided in Appendix A.4. In addition, for Social Snippets, we select 100 samples whose original posts can be retrieved and incorporate the corresponding posts as additional context for translating comments (denoted as w/o or w/ context). We specifically choose Qwen3-4B-Ins, Qwen3-8B and GPT-4o for these experiments and the results are reported in Table 4.

We observe different prompt sensitivities for Fun Posts and Social Snippets. For Fun Posts, explicitly alerting LLMs to the existence of slang or neologisms (+Reminder) consistently improves both XCOMET and SSR scores, whereas this strategy provides limited benefit for Social Snippets. In contrast, the two-stage reasoning prompts and the few-shot prompts have little or even negative impact on Fun Posts but improve performance on Social Snippets. Finally, switching the prompt language from English to Chinese does not significantly affect the

Models	Fun Posts		Social Snippets		
	XCOMET	SSR	XCOMET	ES	GEMBA
Qwen3-4B-ins					
+ Direct (en)	79.48	49.31	71.58	67.27	3.03
+ Reminder	79.65	50.01	71.35	67.27	3.04
+ Direct (zh)	79.29	49.47	71.50	67.24	3.03
+ Two-stage	77.54	48.07	71.97	67.31	3.08
+ Few-shot	79.38	49.23	72.17	67.30	3.10
w/o context	-	-	70.69	67.12	3.00
w/ context	-	-	72.34	67.35	3.11
Qwen3-8B					
+ Direct (en)	80.23	51.95	72.13	67.29	3.06
+ Reminder	80.93	53.08	72.13	67.31	3.06
+ Direct (zh)	80.29	51.55	72.07	67.29	3.05
+ Two-stage	78.07	51.05	72.74	67.35	3.13
+ Few-shot	80.15	52.17	73.27	67.40	3.19
w/o context	-	-	72.51	67.20	3.02
w/ context	-	-	73.97	67.37	3.19
GPT-4o					
+ Direct (en)	85.50	64.91	76.50	68.20	3.65
+ Reminder	85.94	67.08	76.65	68.23	3.72
+ Direct (zh)	85.44	65.01	76.57	68.19	3.65
+ Two-stage	85.01	64.33	77.12	68.33	3.77
+ Few-shot	85.97	65.37	77.24	68.35	3.77
w/o context	-	-	76.40	68.21	3.66
w/ context	-	-	77.48	68.40	3.74

Table 4: Average evaluation results under different prompt strategies. Grey indicates unchanged scores; light green and light red denote small changes (<0.5), while green and red denote larger changes (≥ 0.5).

overall performance on either subset. For Social Snippets, incorporating the original post as additional context information can consistently improve the translation quality.

This analysis characterizes how prompting strategies interact with the two subsets rather than establishing prompt-robustness. Paired bootstrap tests against the Direct (en) baseline confirm that the largest effects in each direction are significant ($p < 0.001$): +Reminder on Fun Posts (Qwen3-8B, GPT-4o) and +Few-shot on Social Snippets (Qwen3-8B, GPT-4o). Differences from +Direct (zh) are negligible in magnitude on both subsets. Prompt design thus does not transfer across subsets in a uniform way.

4.3.2 Qualitative Analysis

To better understand the challenges of translating social media texts and how models behave under different prompting strategies, we present qualitative case studies in this section. We focus on the Qwen3-4B-Ins model for the zh→ja direction. Examples for Fun Posts are shown in Figure 9 and Figure 10, Appendix, while examples for Social Snippets are provided in Figure 11, Appendix.

The case studies reflect the performance trends

observed in Table 4. As shown in Figure 9 and Figure 10, the Reminder prompt helps the model better handle polite or stylistic expressions that commonly appear in formal texts. In contrast, the Two-stage prompting strategy occasionally causes the model to generate non-existent Japanese words or add extra content beyond the source sentence, possibly due to over-reasoning. The Few-shot prompt helps the model handle overused polite expressions in Japanese, but occasionally leads to unnatural word choices. The Social Snippets example further shows that when the source sentence contains emojis or repeated exclamation marks, even simple expressions can become difficult for models to translate (e.g., the “田！老！师！我！想！你！了！（Mr! Tian! I! miss! you!）” in Figure 11 can be easily translated if we exclude redundant exclamation marks, but the model fails when given direct instructions). However, this issue can be mitigated through Two-stage reasoning, Few-shot prompting, or by incorporating additional contextual information.

4.3.3 Annotation Analysis

The annotation of social-media-specific texts is inherently challenging. During the peer-review process, a sample is only finalized once agreement is reached among reviewers. If a reviewer disagrees with the previous translation, the sample is passed to another reviewer for further revision. We measure annotation difficulty using the average number of review rounds required before final agreement. For Fun Posts, the average number of rounds is 1.2, indicating that most translations are accepted after a single peer review. In contrast, Social Snippets require an average of 2.7 rounds before agreement is reached. This substantial difference suggests that Social Snippets contain greater ambiguity and are inherently more challenging for both human annotators and models. In addition, agreement on slang or neologism usage in Fun Posts requires an average of 1.4 rounds. Additional discussions on annotations are provided in the Appendix A.2.

Other analyses, including a discussion of the distinction between the two subsets and the agreement between automatic metrics and human judgments, can be found in Appendix A.3 and Appendix A.7.

5 Conclusion

In this work, we introduce CSM-MTBench, a benchmark for Chinese social-media machine translation, curated with bilingual expert translations

and covering complementary scenarios. Fun Posts focus on context-rich, slang- and neologism-heavy content, while Social Snippets emphasize concise, emotion- and style-driven expressions. Our evaluations reveal differences in how models handle semantic fidelity, stylistic nuances, and social-media-specific language. Our targeted metrics such as Slang Success Rate, embedding-based similarity effectively assess these challenges, highlighting areas where current models excel or fall short. CSM-MTBench thus provides a practical tool for diagnosing and guiding the development of MT systems tailored to real-world Chinese social media.

Limitations

While our work provides a comprehensive benchmark for evaluating machine translation on Chinese social media texts, this work has several limitations. First, we only focus on prompting strategies to improve translation quality, such as reminding models of the presence of slang, neologisms, or stylistic requirements. More sophisticated approaches, e.g., slang-aware pretraining, targeted fine-tuning can be investigated in future works. Second, although CSM-MTBench covers five Chinese-foreign language directions, the inclusion of additional directions is constrained by annotation availability and cost. Finally, CSM-MTBench is designed to reveal systematic patterns in how LLMs handle Chinese social-media text, not to produce a definitive system ranking. Several top-tier models score almost identically on some metrics, and we leave the definitive ranking among such closely matched models to future work.

Ethics Statement

CSM-MTBench is constructed with careful filtering to mitigate potential ethical concerns, including the removal of sensitive or inappropriate content. All models used in our experiments are evaluated in accordance with their respective licenses. We use ChatGPT only for grammar refinement and language polishing. Our work does not introduce ethical biases but aims to make new, positive contributions to MT community.

Acknowledgment

Kaiyan Zhao was supported by JST SPRING, Grant Number JPMJSP2108.

References

- Shareefa Ahmed Al Amer, Mark G. Lee, and Phillip Smith. 2025. [Comparative evaluation of machine translation models using human-translated social media posts as references: Human-translated datasets](#). In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 1–9, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.
- Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. [Tower: An open multilingual large language model for translation-related tasks](#). *Preprint*, arXiv:2402.17733.
- Anthropic. 2025. [System card: Claude opus 4 & claude sonnet 4](#).
- Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, and 2 others. 2024. [Aya 23: Open weight releases to further multilingual progress](#). *Preprint*, arXiv:2405.15032.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.
- Federico Bianchi, Debora Nozza, and Dirk Hovy. 2022. XLM-EMO: Multilingual Emotion Prediction in Social Media Text. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics.
- BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Lucioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, and 374 others. 2023. [Bloom: A 176b-parameter open-access multilingual language model](#). *Preprint*, arXiv:2211.05100.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA. Curran Associates Inc.
- Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2025. [Benchmarking LLMs for translating classical Chinese poetry: Evaluating adequacy, fluency, and elegance](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33007–33024, Suzhou, China. Association for Computational Linguistics.
- Tao Chen and Min-Yen Kan. 2012. [Creating a live, public short message service corpus: the nus sms corpus](#). *Language Resources and Evaluation*.
- Pranjal Chitale, Jay Gala, and Raj Dabre. 2024. [An empirical study of in-context learning in LLMs for machine translation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7384–7406, Bangkok, Thailand. Association for Computational Linguistics.
- Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. [Multilingual machine translation with open large language models at practical scale: An empirical study](#). *Preprint*, arXiv:2502.02481.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. [Aya expand: Combining research breakthroughs for a new multilingual frontier](#). *Preprint*, arXiv:2412.04261.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, and Bochao Wu et al. 2025. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. [Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.
- Yicheng Fu, Zhemin Huang, Liuxin Yang, Yumeng Lu, and Zhongdongming Dai. 2025. [CHENGYU-BENCH: Benchmarking large language models for Chinese idiom understanding and use](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2355–2366, Suzhou, China. Association for Computational Linguistics.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard

- Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. [Dictionary-based phrase-level prompting of large language models for machine translation](#). *Preprint*, arXiv:2302.07856.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. [The Flores-101 evaluation benchmark for low-resource and multilingual machine translation](#). *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. [A survey on llm-as-a-judge](#). *Preprint*, arXiv:2411.15594.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. [xCOMET: Transparent machine translation evaluation through fine-grained error detection](#). *Transactions of the Association for Computational Linguistics*, 12:979–995.
- Hongcheng Guo, Fei Zhao, Shaosheng Cao, Xinze Lyu, Ziyang Liu, Yue Wang, Boyang Wang, Zhoujun Li, Chonggang Lu, Zhe Xu, and Yao Hu. 2025. [Redefining machine translation on social network services with large language models](#). *Preprint*, arXiv:2504.07901.
- Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc’Aurelio Ranzato. 2019. [The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.
- Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakouga, Jessica Lundin, Kenton Murray, Masaaki Nagata, and 9 others. 2025. [Preliminary ranking of wmt25 general machine translation systems](#). *Preprint*, arXiv:2508.14909.
- Tom Kocmi and Christian Federmann. 2023. [Large language models are state-of-the-art evaluators of translation quality](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.
- Dohyun Lee, Seungil Chad Lee, Chanwoo Yang, Yujin Baek, and Jaegul Choo. 2025. [Exploring in-context example generation for machine translation](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26554–26568, Vienna, Austria. Association for Computational Linguistics.
- Paul Lerner and François Yvon. 2025. [Towards the machine translation of scientific neologisms](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 947–963, Abu Dhabi, UAE. Association for Computational Linguistics.
- Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. [Translate meanings, not just words: Idiomkb’s role in optimizing idiomatic translation with language models](#). In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’24/IAAI’24/EAAI’24*. AAAI Press.
- Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. 2013. [Microblogs as parallel corpora](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 176–186, Sofia, Bulgaria. 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).
- Paul McNamee and Kevin Duh. 2022. [The multilingual microblog translation corpus: Improving and evaluating translation of user-generated text](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 910–918, Marseille, France. European Language Resources Association.
- Zhongtao Miao, Kaiyan Zhao, Masaaki Nagata, and Yoshimasa Tsuruoka. 2026. [Neoamt: Neologism-aware agentic machine translation with reinforcement learning](#). *Preprint*, arXiv:2601.03790.
- Paul Michel and Graham Neubig. 2018. [MTNT: A testbed for machine translation of noisy text](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.
- Denis Newman-Griffis, Jill Fain Lehman, Carolyn Rosé, and Harry Hochheiser. 2021. [Translational NLP: A new paradigm and general principles for natural language processing research](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*:

- Human Language Technologies*, pages 4125–4138, Online. Association for Computational Linguistics.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, and Aditya Ramesh et al. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- OpenAI. 2025a. [Gpt-5 system card](#). Technical report, OpenAI.
- OpenAI. 2025b. [gpt-oss-120b & gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- 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.
- Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2025. [StyleDistance: Stronger content-independent style embeddings with synthetic parallel examples](#). 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 8662–8685, Albuquerque, New Mexico. 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.
- Mohammad Sadegh Poulaei, Mohammad Erfan Zare, Mohammad Reza Mohammadi, and Sauleh Eetemadi. 2025. [YNWA_PZ at SemEval-2025 task 11: Multilingual multi-label emotion classification](#). In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 508–521, Vienna, Austria. Association for Computational Linguistics.
- Shenbin Qian, Constantin Orasan, Félix Do Carmo, and Diptesh Kanojia. 2024. [Evaluating machine translation for emotion-loaded user generated content \(TransEval4Emo-UGC\)](#). In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 43–44, Sheffield, UK. European Association for Machine Translation (EAMT).
- Shenbin Qian, Constantin Orasan, Felix do Carmo, Qiliang Li, and Diptesh Kanojia. 2023. [Evaluation of chinese-english machine translation of emotion-loaded microblog texts: A human annotated dataset for the quality assessment of emotion translation](#). *Preprint*, arXiv:2306.11900.
- Justin Qiu, Jiacheng Zhu, Ajay Patel, Marianna Apidianaki, and Chris Callison-Burch. 2025. [mStyleDistance: Multilingual style embeddings and their evaluation](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16917–16931, Vienna, Austria. 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.
- Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025. [Findings of the WMT25 terminology translation task: Terminology is useful especially for good MTs](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 554–576, Suzhou, China. Association for Computational Linguistics.
- Zhiyi Song, Stephanie Strassel, Haejoong Lee, Kevin Walker, Jonathan Wright, Jennifer Garland, Dana Fore, Brian Gainor, Preston Cabe, Thomas Thomas, Brendan Callahan, and Ann Sawyer. 2014. [Collecting natural SMS and chat conversations in multiple languages: The BOLT phase 2 corpus](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 1699–1704, Reykjavik, Iceland. European Language Resources Association (ELRA).
- tabularisai, Samuel Gyamfi, Vadim Borisov, and Richard H. Schreiber. 2025. [multilingual-sentiment-analysis \(revision 69afb83\)](#).
- David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. [Prompting PaLM for translation: Assessing strategies and performance](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.
- Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. [Document-level machine translation with large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.
- Qiyu Wu, Masaaki Nagata, Zhongtao Miao, and Yoshimasa Tsuruoka. 2024. [Word alignment as preference for machine translation](#). In *Proceedings of the 2024*

Conference on Empirical Methods in Natural Language Processing, pages 3223–3239, Miami, Florida, USA. Association for Computational Linguistics.

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 *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, and Binyuan Hui et al. 2025a. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

Cai Yang, Yao Dou, David Heineman, Xiaofeng Wu, and Wei Xu. 2025b. [Evaluating LLMs on chinese idiom translation](#). In *Second Conference on Language Modeling*.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. [Benchmarking machine translation with cultural awareness](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. [Prompting large language model for machine translation: A case study](#). *Preprint*, arXiv:2301.07069.

Kaiyan Zhao, Zhongtao Miao, Zheyong Xie, Shaosheng Cao, and Yoshimasa Tsuruoka. 2026. [Cneo-bench: Diagnosing large language models on chinese neologisms](#). *Preprint*, arXiv:2608.28053.

Jonathan Zheng, Alan Ritter, and Wei Xu. 2024. [NEO-BENCH: Evaluating robustness of large language models with neologisms](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13885–13906, Bangkok, Thailand. Association for Computational Linguistics.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. [Hunyuan-mt technical report](#). *Preprint*, arXiv:2509.05209.

Bo Zhou, Qianglong Chen, Tianyu Wang, Xiaomi Zhong, and Yin Zhang. 2023. [WYWEB: A NLP evaluation benchmark for classical Chinese](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3294–3319, Toronto, Canada. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. [Multilingual machine translation with](#)

[large language models: Empirical results and analysis](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. [Aya model: An instruction finetuned open-access multilingual language model](#). *Preprint*, arXiv:2402.07827.

A Appendix

A.1 Dataset Construction Details

Our candidate data undergo strict filtering to remove: (1) personal information, including names, phone numbers, ID numbers, bank card numbers, email addresses, and physical addresses; (2) violent or graphic content; (3) politically sensitive content; (4) hate speech targeting individuals or groups; and (5) pornographic, inappropriate; or (6) any other potentially sensitive content. The filtering prompt we used is shown in Figure 5. After the LLM-based filtering, all data additionally undergo security screening via internal automated safety APIs to remove any potentially sensitive or risky content.

For Fun Posts, all the 1183 source sentences are shared across all five target languages. For Social Snippets, due to filtering, since short sentences are more difficult to reliably pass quality control, a portion of the source sentences is duplicated across language pairs. We ensure that all data collection, processing, and usage comply with relevant legal and ethical standards, and no private or personally identifiable information is included in the released dataset. Therefore, the dataset can be safely and legally used for research purposes.

Following this filtering process, bilingual expert annotators translate the remaining Chinese texts into Spanish, French, Japanese, Korean, and Russian.

A.2 Annotation

Annotators (employees) are assigned with the translation tasks as part of their regular work responsibilities, and no additional payment is provided beyond their standard compensation. They are proficient in both source and target languages and familiar with Chinese social media expressions, including slang and culture-specific references.

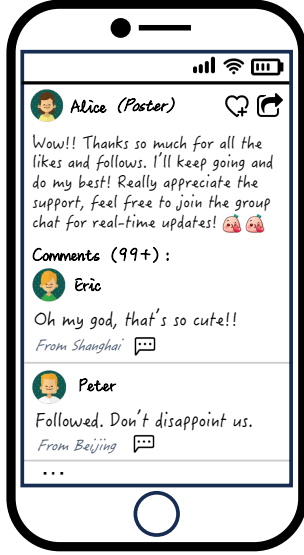


Figure 4: English translation for examples in Figure 2.

All annotators undergo training before the formal annotation process. We first sample 100 instances for pilot translation and quality checking, during which we refine our translation instructions. This calibration phase also help align annotators’ understanding of how to handle slang, tone preservation, and stylistic differences across subsets.

Translation guidelines instruct them to preserve the tone, style, and intent of the original text. Ambiguous expressions or neologisms are carefully interpreted in context to ensure faithful and natural translations. All translations undergo quality control, including peer review and adjudication, to form the final two datasets.

Guidelines for human translators: “Translate the Chinese sentence into XXX. Give appropriate translations for potential Chinese slang or neologisms, and preserve the original tone and style as much as possible.”

We intentionally adopt a unified set of translation guidelines for both subsets. The guidelines explicitly emphasize preserving slang expressions, informal tone, and pragmatic nuances whenever possible. The training phase ensures that annotators develop a clear understanding of these requirements.

Guidelines for peer reviewers: “Check whether the translation is accurate and complete with respect to the Chinese source sentence. Revise any incorrect, missing, or inappropriate translations.”

System Prompt

You are a data security reviewer. Your task is to determine whether a given text contains any of the following types of sensitive information, and output **only “yes” or “no.”**
Sensitive categories include:
Personal information (such as name, phone number, ID number, bank card number, email, address, etc.);
Violent or graphic content;
Politically sensitive content;
Hate speech (attacking a group or an individual);
Pornographic or inappropriate content;
Any other potentially sensitive information.
Be strict in your judgment—if there is any suspicion, classify it as sensitive. Only output **yes** or **no**.

User Prompt

{source_Chinese_text}

Figure 5: The prompt we used for strict filtering. All samples labeled as “yes” are excluded from our dataset. Text enclosed in italic braces {...} indicates placeholders for the source Chinese content.

Models	zh→es	zh→fr	zh→ja	zh→ko	zh→ru	avg.
<i>Closed-source LLM APIs</i>						
GPT-4o	3.50	3.54	3.68	3.87	3.65	3.65
GPT-5	3.56	3.64	3.73	3.84	3.70	3.69
Claude-Sonnet-4	3.46	3.38	3.58	3.84	3.66	3.58
<i>Translation-specialized Models</i>						
NLLB-3.3B	2.19	2.14	1.97	2.17	2.18	2.13
google-translate	3.17	3.20	3.00	3.23	3.27	3.17
GemmaX2-9B	3.12	3.11	3.27	3.31	3.15	3.19
Hunyuan-MT-7B	3.20	3.21	3.33	3.45	3.23	3.28
<i>Open-source General-purpose LLMs</i>						
DeepSeek-V3	3.58	3.55	3.65	3.79	3.63	3.64
GPT-OSS-120B	3.46	3.52	3.51	3.74	3.60	3.57
Aya-101	2.20	2.21	2.53	2.29	2.18	2.28
Aya-Expans-8B	3.04	3.12	3.14	3.27	3.13	3.14
Gemma3-4B	2.85	2.82	2.88	2.92	2.78	2.85
Gemma3-12B	3.16	3.11	3.24	3.38	3.13	3.20
Gemma3-27B	3.26	3.31	3.38	3.54	3.37	3.37
Qwen3-1.7B	2.46	2.56	2.71	2.36	2.41	2.50
Qwen3-4B	2.85	2.83	2.76	2.80	2.73	2.79
Qwen3-4B-Ins	3.05	3.02	3.05	3.08	2.93	3.03
Qwen3-8B	3.05	3.01	3.10	3.14	3.00	3.06
Qwen3-32B	3.21	3.23	3.24	3.36	3.14	3.24
Qwen3-30B-A3B	3.13	3.11	3.20	3.28	3.12	3.17
Qwen3-30B-A3B-Ins	3.28	3.28	3.37	3.44	3.24	3.32
Qwen3-235B-A22B	3.32	3.34	3.44	3.59	3.35	3.41

Table 5: Full Gemba-stars results for Social Snippets.

A.3 Distinction of Fun Posts and Social Snippets

Beyond text length, different belongings (posts and comments), and the distribution difference introduced in Section 3.1, we separate CSM-MTBench into two subsets for the following reasons.

A.3.1 Different Roles of Slang

From a linguistic perspective, slang expressions play different roles in the two subsets. In Fun Posts, slang and neologisms typically appear as identifiable lexical items embedded in longer narrative descriptions. This makes explicit slang identification a meaningful evaluation task.

In contrast, Social Snippets are short reaction-style comments where meaning is often conveyed through overall tone, emotion, and pragmatic style rather than isolated lexical slang units. In such contexts, the stylistic and emotional alignment between source and translation becomes more important than detecting individual slang tokens. Therefore, we evaluate Social Snippets using the ES score, which measures style and emotional consistency between source and target texts.

This distinction is further reflected in the Section [A.3.2](#).

A.3.2 Different Difficulty in Slang and Identification

We also observe different levels of difficulty in identifying slang and neologisms across the two subsets. We try to identify the slang and neologisms for social snippets, but GPT-5 frequently label the entire sentences as slang or neologisms (in over 50% of sampled cases), suggesting higher ambiguity in short comment-style texts.

As explained in Section [4.3.3](#), we perform human agreement for annotation. During pilot slang identification annotation, we find that the agreement on slang or neologism in Social Snippets often required multiple rounds of discussion (average 3.1 rounds), whereas Fun Posts required significantly fewer revisions (average 1.4 rounds). There even exist cases where annotators cannot make agreements (<5%). This indicates that short comments introduce greater ambiguity for both humans and LLMs in determining what constitutes slang or neologism usage and the reason why we don't apply slang identification on Social Snippets. Thus, separating the subsets therefore helps capture differences in annotation difficulty.

A.3.3 Unsuitability for ES on Fun Posts

We also examined the behavior of ES score on the two subsets. While ES reliably distinguishes human and machine translations in Social Snippets, the metric becomes unstable for longer Fun Posts.

One explanation is that long posts often contain multiple emotional tones or discourse intentions within a single sentence. For example, the Chinese sentence in Fun Posts: “这么鲜活又真实的小作文属实难得一见，像是把日记分享给我们看希望宝宝越来越好，能不受情绪左右，勇敢做自己！她的每一套穿搭，用的所有小东西都超级无敌种草！” (*It's honestly rare to see such a vivid and genuine little post, it feels like she's sharing*

her diary with us . I really hope she keeps getting better, stays strong, and can bravely be herself without being swayed by emotions. Every outfit she wears and every little thing she uses makes me want to buy it so badly!) contains multiple and complex emotions. In such cases, embedding-based classifiers tend to collapse these nuanced signals into a single category, which introduces noise in ES measurements. Due to this instability, we find that ES is more suitable for short, focused comment-style text, and therefore restrict its use to Social Snippets.

A.3.4 Different Prompt Sensitivity

Finally, the two subsets exhibit different prompt sensitivities as introduced in Section [4.3.1](#). This difference further supports our decision to evaluate the subsets separately. If we merge two subsets together, we may not observe the different sensitivities.

A.4 Implementation Details

We utilize different prompts depending on model family. For NLLB and Google Translate, we explicitly specify the source and target languages for translation. For Aya-101, we apply a simple translation prompt:

“Translate to XXX: {source text}”.

For the remaining LLMs, we use a unified instruction:

“You are a translation expert. Please translate the following sentence into XXX and output only the translated result: {source text}”.

For models equipped with “thinking” modes (e.g., Qwen3), we disable these features to ensure fair comparison. The maximum number of newly generated tokens is set to 512 across all models. We set temperature to 0.0 for API models and use greedy decoding (do_sample=False) for other LLMs to ensure reproducibility. We use GPT-4o for GEMBA-stars evaluation. Filtering is performed with GPT-5. All experiments are conducted on eight H800 GPUs with a single run.

In Section [4.3.1](#), the five prompting strategies we used are as follows:

- Direct Instruction (en): Exactly the same English prompt as we use in the main experiments.
- Reminder Prompt: For Fun Posts, we use “You are a translation expert. Please translate the following sentence into XXX. **Some**

sentences contain Chinese social media neologisms or slang; when translating, use appropriate expressions. Output only the translated result: *{source text}*". For Social Snippets, we use "You are a translation expert. Please translate the following sentence into XXX, **preserving the tone and style as much as possible**. Output only the translated result: *{source text}*".

- Direct Instruction (zh): The Chinese translation for the English direct instruction.
- Two-stage Reasoning: For Fun Posts, we use "You are a translation expert. **First, identify any Chinese internet slang in the following sentence and provide an appropriate XXX expression for it. Then, translate the full sentence into XXX, making sure to preserve suitable expressions for the Chinese internet slang.** Only output the translated result.". For Social Snippets, we use "You are a translation expert. **First, analyze the theme, tone, and style of the following sentence, then translate it into XXX.** Only output the translation result.".
- Few-shot Prompts: For Fun Posts, we use the following 3-shot examples: "1. 姐妹们，这个面霜真的有被惊艳到！质地很轻薄，上脸一点都不黏，第二天起床皮肤软软的，感觉状态都变好了。2. 第一次自己做蛋糕，没想到居然成功了！虽然卖相一般，但味道真的不错，成就感拉满。3. 本来只是随便试试，结果一试就停不下来，这个真的会上瘾。and their corresponding translations in five languages". For Social Snippets, we use "1. 狠狠共鸣了。2. 这也太真实了吧。3. 看完有点被治愈到，谢谢你分享。and their corresponding translations in five languages".

For Qwen3 series models, the suffix "-ins" is used interchangeably with "Instruct-2507", for example, Qwen3-4B-ins equals to Qwen3-4B-Instruct-2507¹¹.

A.5 Discussion on Embedding Similarity

In general, embedding-based automatic metrics aim to offer a convenient and computationally efficient alternative to LLM-as-a-judge methods.

¹¹<https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507>

GEMBA-stars Style Prompt

Score the following translation from Chinese to *{LANG}* based on their styles and tone, with respect to the human reference with one to five stars.

Do not consider translation accuracy, focus on whether the style and tone of the source sentence are preserved in the target translation.

Where one star means 'Nonsense/No style or tone preserved', two stars mean 'Some style or tone preserved, but not understandable', three stars mean 'Some style and tone preserved and understandable', four stars mean 'Most style and tone preserved with possibly few mistakes', and five stars mean 'Perfect style and tone preserved'.

Only output a number.

Chinese source: *{source}*

{LANG} human reference: *{gold_label}*

{LANG} translation: *{target}*

Stars:

Figure 6: Our GEMBA-stars style prompt. **Bold** font highlights the difference from the original GEMBA-stars usage.

The mStyleDistance (Qiu et al., 2025) we adopt serves as a style embedding model that maps texts with similar stylistic properties to nearby representations while pushing stylistically different texts farther apart, largely independent of semantic content. This design aligns well with our setting, where the goal is to evaluate whether the stylistic characteristics of the source sentence are preserved in the translation (Patel et al., 2025).

The emotion embedding model (Bianchi et al., 2022) classifies sentences into four emotion categories: joy, sadness, anger, and fear. The sentiment embedding model (tabularisai et al., 2025) categorizes sentences into five sentiment classes: Neutral, Positive, Negative, Very Negative, and Very Positive. A good translation should preserve the emotion and sentiment of the source sentence.

For these models, we do not directly compare whether the predicted labels of the source and target sentences match, as social media texts often express mixed or nuanced emotions that fall outside the predefined label sets. Instead, we report the cosine similarity between the source and target sentence embeddings to measure the degree of emotional and sentiment alignment.

However, embedding-based automatic metrics still have limitations. As shown in Table 3, the highest ES score is 68.27 for GPT-5, while the lowest score is 61.74 for NLLB-3.3B, indicating a relatively narrow performance gap across models. When computing ES between the source sentences and human-annotated translations, the average ES score reaches 70.32. Although ES exhibits limited absolute discriminability, it remains effective in re-

Fun Posts Examples (zh→ja)	
Source: 哪吒的童子尿包浓郁的呀 <i>Nezha's child urine pouch is super strong</i>	
Slang Candidates: ["絶対に", "間違いない", "確実に", "に決まってる", "ガチで"]	
Target: 哪吒の童尿はきつと濃厚なんだろうね	
Source: 无删减加速版沉浸式炒雪花酥全过程 今天依旧是我们的团宠 奥利奥草莓 换了个大锅 一锅出9斤没问题🥰 <i>Uncut, fast-forward, immersive full process of making Snowflake Crisp. Today's everyone's favorite again: Oreo Strawberry! Switched to a bigger pot — no problem making 9 pounds in one batch 🥰</i>	
Slang Candidates: ["うちの一番人気", "みんなの寵児", "グループのアイドル", "みんなの推しフレーバー"]	
Target: 無編集加速版没入型雪花酥作り全过程 今日我々の人気商品 オレオストロベリー 大きな鍋に変えて1缶で9斤は問題なし🥰	
Source: 春节倒计时🔥浅浅期待一下今年的年夜饭吧🍷 <i>Countdown to Chinese New Year 🔥 A little looking forward to this year's New Year's Eve dinner 🍷</i>	
Slang Candidates: ["少しでも楽しみにしています", "ちょこっと楽しみにしています", "ほんのり楽しみにしています", "軽く楽しみにしています", "ちょっとだけワクワクしています", "ささやかに楽しみにしています"]	
Target: 春節カウントダウン🔥今年の年越し料理をちょこっと楽しみにしています🍷	
Social Snippets Examples (zh→ja)	
Source: 天啊，我都没经历过 <i>OMG, I've never experienced that</i>	
Target: まさか！私、まだ経験したことないよ！	
Source: 我是中国人哈哈哈 <i>I'm Chinese, lol</i>	
Target: 私は中国人ですははは	
Source: 哇塞老师好萌！ <i>WOW, the teacher is so cute!</i>	
Target: わー、先生がすごく可愛い！	

Figure 7: Examples from Fun Posts and Social Snippets (zh→ja). In Fun Posts, we highlight slang and neologisms in **blue**. **Target** denotes the human-provided translation. **Slang Candidates** list alternative slang or neologism choices other than the gold label one in the **Target**. English references are provided in *italics*.

vealing relative performance differences between models, as its ranking trend closely aligns with GEMBA-stars evaluations. Considering the substantially lower computational cost compared to LLM-as-a-judge methods, we believe ES provides a practical and efficient alternative for large-scale evaluation.

A.6 Statistical Significance

We assess the top-two comparison within each model tier using a paired bootstrap test (10,000 resamples at the segment level, two-sided), with Holm correction applied within each metric. Table 6 reports the resulting p -values on Fun Posts. Two patterns emerge. First, significance is *language-dependent*: Hunyuan-MT-7B is ahead of GemmaX2-9B on zh→{ja, ko, ru} ($p < 0.001$) but the two systems are statistically indistinguishable on zh→{es, fr} ($p = 0.41$ and $p = 0.98$), so an aggregate ranking would obscure a systematic split across target languages. Second, and more consequentially, significance is *metric-dependent*: GPT-4o and GPT-5 cannot be separated under XCOMET on any language pair, yet SSR separates them on all five ($p < 0.001$). XCOMET measures general adequacy while SSR targets social-media-specific phenomena, and the two therefore differ in discriminative power precisely on the phenomena

CSM-MTBench is designed to isolate. The reverse case is equally informative: Hunyuan-MT-7B and GemmaX2-9B are cleanly separated by XCOMET on three pairs but by SSR on none, indicating that dedicated MT models differ in general translation quality without differing in their handling of social-media conventions.

A.7 Automatic Metrics v.s. Human Judgments

In this section, we compare automatic metrics, including SSR (fuzzy matching), Embedding Similarity (ES) and LLM-as-a-judge to human judgments.

A.7.1 SSR (fuzzy matching)

The SSR is calculated using fuzzy matching with a threshold of 0.8 (0-1), where scores ≥ 0.80 is treated as a match (score=1) and < 0.80 as non-match (score=0). The threshold is chosen to tolerate minor surface variations such as tense differences, pluralization, or small lexical changes, which are common in translation outputs and should not be penalized as incorrect. For example, the Japanese translation for “Instagram” sometimes becomes “ins” or “insta” in Japanese (both abbreviations are correct), but if the threshold is set to 100 (fuzzy match becomes exact match), the score becomes 0 for this sample.

Comparison	zh→es	zh→fr	zh→ja	zh→ko	zh→ru
<i>Panel A: XCOMET</i>					
GPT-4o vs. GPT-5	0.890	0.234	0.036	0.750	0.264
Hunyuan-MT-7B vs. GemmaX2-9B	0.406	0.984	<0.001	<0.001	<0.001
DeepSeek-V3 vs. GPT-OSS-120B	0.846	0.960	<0.001	0.222	<0.001
<i>Panel B: SSR</i>					
GPT-4o vs. GPT-5	<0.001	<0.001	<0.001	<0.001	<0.001
Hunyuan-MT-7B vs. GemmaX2-9B	0.984	0.352	0.306	0.132	0.934
DeepSeek-V3 vs. GPT-OSS-120B	<0.001	<0.001	0.010	<0.001	<0.001

Table 6: Paired bootstrap p -values (10,000 resamples, two-sided) for the top-two ranked systems in each model tier on the Fun Posts. **Bold** indicates significance at $\alpha = 0.05$ after Holm correction within each panel (15 tests). Under XCOMET, GPT-4o and GPT-5 are statistically indistinguishable on all five language pairs, whereas SSR separates them on all five.

System Prompt
Internet language (also known as online slang) refers to expressions that originate from or are primarily used in online communication. Such expressions often carry special meanings within specific online contexts. You are an expert in internet linguistics. You will be given a pair of sentences in Chinese and another language; they are translations of each other and may contain internet slang. Your tasks are: 1. Identify any internet slang expressions in the Chinese sentence (be strict—some sentences may contain no slang). 2. Locate the corresponding translation of each slang item in the other-language sentence (simply identify it). 3. Provide several alternative reasonable translations for each identified slang term, based on the translation and the context. 4. Output the final result in JSON format, containing the following fields: - slang (the original Chinese slang) - slang_trans (its translation) - slang_candidates (other possible translations of the slang)
User Prompt
Chinese: {source_Chinese_text}, {LANG}: {gold_label}

Figure 8: Slang-translation pair detection and slang candidates generation prompt.

To assess the reliability of this criterion, we manually inspect 1,000 randomly sampled instances (500 matched and 500 unmatched). We observe a small number of false positives. In several cases, models copy the source sentence instead of translating it, artificially increasing lexical overlap (e.g., leaving “TikTok” untranslated). In other cases, slang expressions are left untranslated, particularly in zh→ja, where shared logographic characters between Chinese and Japanese may make untranslated Chinese slang appear superficially plausible. We also find instances where grammatically incorrect or partially translated outputs were still scored as matches by fuzzy matching.

We also observe a small number of false negatives, where the model produce semantically cor-

rect alternatives that fall below the similarity threshold. Some of these cases are due to limitations of the candidate dictionary, which does not cover all valid alternative expressions, while others result from correct translations that differ substantially in surface form from the reference candidates.

Importantly, common fuzzy-matching failure cases such as word-order variation are rarely observed. This is likely because our evaluation focuses on slang expressions and neologisms, which typically function as lexical units rather than compositional phrases, making surface-form matching relatively robust in this setting.

Overall, manual inspection indicates that approximately 4% of cases show discrepancies between fuzzy matching and human judgment. While imperfect, this relatively low error rate suggests that fuzzy matching provides a reasonably reliable proxy for evaluating slang and neologism translation in our task.

A.7.2 Embedding Similarity (ES)

We conducted a human-correlation study to assess the reliability of ES scores. We sample 500 instances in Social Snippets across the five language directions. Five annotators (one per language direction) rated each translation on three dimensions: style, emotion, and sentiment on a 1–5 Likert scale. We report the Spearman correlation to see the rank of ES scores and human judgments, which results in $\rho = 0.50, p < 0.001$. For reference, neural metrics on related sentence-level evaluation tasks typically report correlations in a similar range (Freitag et al., 2023). Given that style/emotion preservation is a more subjective task than translation quality assessment, our correlation suggests that ES provides a reasonable proxy for human judgment in

Source: 无删减加速版沉浸式炒雪花酥全过程 今天依旧是我们的团宠 奥利奥草莓 换了个大锅 一锅出9斤没问题 🍪
Uncut, fast-forward, immersive full process of making Snowflake Crisp. Today's everyone's favorite again: Oreo Strawberry! Switched to a bigger pot — no problem making 9 pounds in one batch 🍪

Direct: 削除なし・スピードアップ版、雪のスイーツを完全体で体験する全过程 今日私たちの愛されキャラ
 であるオリオストロベリーです。大きな鍋に変更し、一鍋で9斤も作れるのは問題ありません 🍪
 愛されキャラ (loved character), inaccurate slang translation.
 です, ありません are polite expression in Japanese, not needed here.

Reminder: 削除なし・スピードアップ版、雪のスイーツをまるで体験するように作成。今日私たちの愛され
 キアラであるオリオストロベリー。大きな鍋に変更し、一鍋で9斤も作れるよ 🍪
 The polite expressions are fixed.

Two-stage reasoning: 削除なし・加速版・没入型スイーツ作り全過程 今日私たちのツイートキャラ「オリオ・
 ストロベリー」です！大きな鍋に変更し、一鍋で9斤も作れるよ 🍪
 スイーツ (desserts), not accurate as *Snowflake Crisp*.
 ツイートキャラ non-existing words in Japanese.

Few-shot: 無編集・高速版 インパルス体験型 スノーブラウンの調理全过程 今日私たちの愛されキャラ、
 オリオストロベリー。大きな鍋に変更し、一鍋で9斤も調理可能 🍪
 スノーブラウン (snow brown), not accurate as *Snowflake Crisp*.
 The polite expressions are fixed.

Figure 9: Case study (Part 1) for Fun Posts. Different colors highlight translation issues in the model outputs. Explanations are provided below the Japanese translation.

our evaluation setting.

A.7.3 LLM-as-a-judge

We further assess the reliability of LLM-based evaluation (GEMBA-stars) through manual verification. Specifically, we randomly sample 500 translation instances across the five language directions (100 per direction) and compare the scores assigned by the LLM judge with human judgments. The evaluation uses a 1–5 rating scale following the GEMBA-stars protocol.

Overall, we observe a high level of consistency between the LLM judge and human evaluation. In most cases, the difference between the two ratings is within one point on the 1–5 scale. The average absolute score difference is 0.48, and only a small portion of samples exhibit larger deviations.

We then manually inspect the cases with noticeable discrepancies and find that they typically arise when translations contain subtle stylistic differences or model meets extremely short sentences that are difficult to evaluate consistently. Despite these occasional disagreements, the overall alignment suggests that the LLM-as-a-judge provides a reasonably reliable approximation of human evaluation for our task.

Source: 重磅新品今年要来啦
A big new launch/product is coming this year!
Direct: 今年、新作が登場します！
新作 (new product), missing big/huge.
します is polite expression in Japanese, not needed here.
Reminder: 今年、大作が登場しますよ！
Accurate translation for a *big new product*, the polite expression is partially addressed by よ, a spoken word aligns with causal usage.
Two-stage reasoning: 今年の新作、ついに登場します！
ついに (finally), grammatically correct, but not expressed in the source sentence.
Few-shot: 今年、本格的な新製品が登場するぞ！
Perfect translation.

Source: 五官平平，都来试试这个妆！！
If your features are pretty plain, you should definitely try this makeup look!!
Direct: 五感が普通な方は、このメイクを試してみてください！！
五感 (five senses), inaccurate translation.
ください is polite expression in Japanese, not needed here.
Reminder: 五感に満足できないなら、このメイクを試してみよう！
The polite expressions are fixed.
Two-stage reasoning: 顔のパーツが普通で、このメイクを試してみてください！
顔のパーツが普通で (facial features are ordinary), missing your, grammatically incorrect.
Few-shot: 顔のパーツが特に目立たない人でも、このメイクを試してみようぜ！！
Perfect translation.

Figure 10: Case study (part 2) for Fun Posts.

Source: 已关注，莫辜负。 <i>Followed, don't disappoint us.</i>	Source: 天哪太可爱了！！ <i>OMG, that's so cute!</i>
Direct: 注目しました。無駄にしないでください。 注目 (attention), not accurate enough. 無駄 (waste), not accurate enough	Direct: 天啊、もんもんと可愛い！！ 天啊, not translated. もんもんと (feel frustrated), wrong translation
Reminder: 注目しました。無駄にしないでください。 No change compared to direct instruction.	Reminder: 天啊、もんもんと可愛い！！ No change compared to direct instruction.
Two-stage reasoning: 関心を寄せました。無駄にしないでください。 関心を寄せ (focus on you), better but not perfect.	Two-stage reasoning: わー、すごくかわいい！！ Perfect translation.
Few-shot: フォローしました。無駄にしないでください。 Former part is perfect.	Few-shot: やばい、かわいい！！ Perfect translation.
Original post as context: フォローしました！期待に裏切らないでね！ Perfect translation.	Original post as context: ああ、かわいい！！ Perfect translation.
Source: 嘻嘻当然••••• <i>Haha, of course •••••</i>	Source: "田！老！师！我！想！你！了" <i>Mr.! Tian! !! Miss! You!</i>
Direct: もちろん、嘻嘻••••• 嘻嘻, not translated. The order of two phrases is reversed.	Direct: だい！ろう！せん！わたし！おもい！いたい！ Broken words in Japanese. おもい！いたい！ (want to think), grammatically incorrect.
Reminder: 嘻嘻、もちろん••••• The order is fixed.	Reminder: だい！ろう！せい！わたし！おもい！ます！ Almost no change compared to direct instruction.
Two-stage reasoning: ははは、もちろん••••• Perfect translation.	Two-stage reasoning: 田！先生！思い出していました！ 思い出して (think of you), close but imperfect
Few-shot: ほほほ、もちろん••••• Perfect translation.	Few-shot: たん！ろうし！おもしろい！ Broken words in Japanese.
Original post as context: ほほほ、もちろん••••• Perfect translation.	Original post as context: 田！先生！あいたい！ Perfect translation.

Figure 11: Case study for Social Snippets. Different colors highlight translation issues in the model outputs. Explanations are provided below the Japanese translation.