Salute the Classic: Revisiting Challenges of Machine Translation in the Age of Large Language Models

Jianhui Pang[1*] **Fanghua Ye**[2*] **Derek Fai Wong**[1†] **Dian Yu**[3] **Shuming Shi**[3]
Zhaopeng Tu[3] **Longyue Wang**[3†]
[1]University of Macau, Macao, China [2]University College London, United Kingdom
[3]Tencent AI Lab, China
nlp2ct.pangjh3@gmail.com, fanghua.ye.19@ucl.ac.uk, derekfw@um.edu.mo
{yudian, shumingshi, zptu, vinnylywang}@tencent.com

Abstract

The evolution of Neural Machine Translation (NMT) has been significantly influenced by six core challenges (Koehn and Knowles, 2017) that have acted as benchmarks for progress in this field. This study revisits these challenges, offering insights into their ongoing relevance in the context of advanced Large Language Models (LLMs): *domain mismatch*, *amount of parallel data*, *rare word prediction*, *translation of long sentences*, *attention model as word alignment*, and *sub-optimal beam search*. Our empirical findings show that LLMs effectively reduce reliance on parallel data for major languages during pretraining and significantly improve translation of long sentences containing approximately 80 words, even translating documents up to 512 words. Despite these improvements, challenges in domain mismatch and rare word prediction persist. While NMT-specific challenges like word alignment and beam search may not apply to LLMs, we identify three new challenges in LLM-based translation: inference efficiency, translation of low-resource languages during pretraining, and human-aligned evaluation.

1 Introduction

In the Natural Language Processing (NLP) community, one of the most critical and longstanding tasks is Machine Translation (MT), which aims to convert human languages from one form to another (Koehn, 2009; Poibeau, 2017). As the demand for effective translation systems continues to grow, researchers have been striving to develop models that can tackle the inherent challenges of this complex task. In this context, the six challenges about MT proposed by Koehn and Knowles (2017) have been widely recognized and studied by numerous studies, with many efforts revolving around them (Chu and Wang, 2018; Neishi and Yoshinaga, 2019; Garg et al., 2019; Pang et al., 2023).

The emerging Large Language Models (LLMs) have been a significant breakthrough in NLP (Touvron et al., 2023a; OpenAI, 2023; Touvron et al., 2023b). LLMs have demonstrated remarkable capabilities, outperforming traditional approaches and setting new benchmark performance for various applications such as machine translation (Lyu et al., 2023; Zhu et al., 2024; Zhang et al., 2023a; Wang et al., 2023b). LLMs exhibit remarkable translation capabilities for major languages, owing to their extensive pretraining on vast amounts of unpaired data. This implies a significant advancement over conventional techniques. Consequently, recent studies have employed LLMs for translation tasks (Jiao et al., 2023; Alves et al., 2023), achieving remarkable performance. However, it is unclear *how LLMs fare against the six classical challenges,* This intriguing question warrants further investigation and discussion.

To gain insights into LLM-based MT research and identify paths for advancement, we train Llama2 models as German-to-English translation systems and evaluate their abilities in addressing the six classic challenges. Note that English is a high-resource language in the Llama2 pretraining data (Touvron et al., 2023b) and German is a relatively high-resource language, ensuring the model's competence in these languages. Furthermore, we identify two LLM-specific challenges: pretraining resource imbalance and human-like

[*]Work was done when Jianhui Pang and Fanghua Ye were interning at Tencent AI Lab.
[†]Corresponding authors.

Transactions of the Association for Computational Linguistics, vol. 13, pp. 73–95, 2025. https://doi.org/10.1162/tacl_a_00730
Action Editor: Taro Watanabe. Submission batch: 1/2024; Revision batch: 6/2024; Published 1/2025.

Challenges	Experiments	Outcomes		
		Enc2Dec	LLMs	Takeaways
Domain Mismatch	Llama2-7B/13B OPUS datasets	✗	✗	LLMs demonstrate overall advancements yet still face significant domain variance issues.
Amount of Parallel Data	Llama2-7B/13B WMT23 datasets	✗	✓	LLMs diminish the dependence on bilingual data for high-resource pretraining languages.
Rare Word Prediction	Llama2-7B/13B WMT23 datasets	✗	✗	LLMs consistently struggle with predicting infrequent words.
Long Sentence Translation	Llama2-7B/13B WMT23 datasets	✗	✓	LLMs address the task of translating long sentences and display remarkable performance at the document level.
Word Alignment	Llama2-7B WMT23 testsuits	✓	✗	The attention weights in LLMs are unsuitable for word alignment extraction, yet they provide valuable insights into model interpretability.
Inference Efficiency	Llama2-7B WMT23 datasets	✓	✗	Inference efficiency poses a substantial challenge in LLMs, with a 100-fold delay compared to Enc2Dec models in our experiments.
Pretraining Resource Imbalance	Llama2/MLLMs-7B WMT23 datasets	–	✗	LLMs exhibit suboptimal performance in low-resource languages, stemming from the imbalance in pretraining resources.
Evaluation Issues	Llama2-7B WMT23 datasets	–	✗	The issue of automatic evaluation has arisen due to the divergence between human and automated assessments of LLM translation outputs.

Table 1: An overview of revisiting MT challenges in the context of LLMs. The first six lines discuss the classic MT challenges, while the last two lines focus on specific challenges that arise in LLM scenarios. The ‘‘✓’’ symbolizes largely addressed issues, while the ‘‘✗’’ represents persisting unresolved challenges.

evaluation issues. German, Chinese, Ukrainian, and Hebrew, are included for English-to-X low-resource translation tasks to assess the effects of resource imbalance. Table 1 summarizes our key findings, revealing that LLMs have successfully tackled data quantity and long sentence translation challenges but still face unresolved issues.

2 Experimental Setup

2.1 Large Language Models

This paper focuses on decoder-only LLMs, a popular architecture in recent years (OpenAI, 2023). Open-source LLMs are typically trained on datasets dominated by a single major language (Touvron et al., 2023b; Yang et al., 2024). Pretrained Llama2 models are English-centric and highly representative in the LLM community, which have only undergone pretraining without extensive fine-tuning (Touvron et al., 2023b). Therefore, we select pretrained Llama2 models as our base models for most experiments on classic machine translation challenges in Section 3. This choice ensures controlled, reliable comparisons and aligns with current research, enhancing the credibility of our findings. We employ the Llama2-7B and Llama2-13B models, which are open-source language models with 7 billion and 13 billion parameters, respectively (Touvron et al., 2023b). For fine-tuning, we train the model using the Alpaca dataset to enhance its instruction-following capabilities. Below are two training settings based on the input format of paired data.

- **LLM-SFT** undergoes supervised fine-tuning (SFT) using bilingual pairs in conjunction with the Alpaca dataset (Taori et al., 2023), where the bilingual pairs adopt the Alpaca format.
- **LLM-CPT-SFT** involves continuous pretraining (CPT) of the Llama2 model on concatenated translation pairs (Zhu et al.,

2024), followed by fine-tuning using the Alpaca dataset.

Unless stated otherwise, LLM-SFT and LLM-CPT-SFT refer to 7-billion-parameter models.

In Section 4.1, we explore the pretraining resource imbalance, a challenge specific to the multilingual translation of LLMs. To address this, we compare two existing multilingual LLMs, ALMA-7B (Xu et al., 2024a) and TowerInstruct-7B (Alves et al., 2024), as baselines. These models are only evaluated in this section since the focus here is multilingual translation, contrasting with the bilingual nature of the majority of our Llama2-based experiments in Section 3.2.

2.2 Small Encoder-to-Decoder Models

Encoder-to-decoder (Enc2Dec) models are widely recognized as the most effective framework for translation tasks (Vaswani et al., 2017). We utilize the Fairseq[1] toolkit to train Enc2Dec models adhering to the model architectures proposed by Vaswani et al. (2017). Specifically, the base architecture is employed for training models on small bilingual datasets with 500k pairs or fewer, while the large architecture is used for datasets comprising 1M pairs or more (Pang et al., 2023). For each language pair, we use the SUBNMT toolkit[2] to learn byte-pair encoding codes and transform words into subwords (Sennrich et al., 2016a,b).

2.3 Data Conditions

We employ the German-to-English (De2En) parallel data (300 million (M)) procured from the WMT23 translation tasks.[3] Our methodology encompasses training a translation model through random sampling on the dataset, extracting up to 20M translation pairs for each task. To safeguard the robustness of our evaluation process, we utilize the most recent publicly accessible generaltest2023 from WMT23, thereby precluding potential data leakage.[4] To assess LLMs in multi-domain tasks, we conduct an experiment with the multi-domain dataset for German-to-English translation obtained from OPUS (Tiedemann, 2012; Aharoni and Goldberg, 2020).

2.4 Training and Evaluation

Training For LLMs, we train each model with a learning rate of $2e$-5, batch size of 48, over 3 epochs on 32 A100 GPUs, saving checkpoints every 500 steps and reporting the best score. For Enc2Dec models, we use early stopping with a patience of 20, halting training when validation performance plateaus or declines (Yao et al., 2007). We use the Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$ to optimize Enc2Dec models.

Evaluation We employ two widely recognized automatic metrics: BLEU (Papineni et al., 2002), using tokenized and case-sensitive SacreBLEU[5] (Post, 2018), and COMET-DA (Rei et al., 2020), based on the Unbabel/wmt22-comet-da model, a reference-based evaluation approach.

The BLEU metric assesses translation quality by measuring surface-level similarity through n-gram matching between the generated output and reference translations. While BLEU may not fully capture the semantic intricacies of the translation, it provides a reliable indication of fluency and lexical accuracy, ensuring that the basic grammatical structure of the output is maintained (Post, 2018).

To complement BLEU's surface-level evaluation, we utilize COMET-DA to assess deeper semantic-level alignment between the translated output and reference text. COMET-DA leverages pretrained language models to provide a more nuanced assessment of sentence-level translation adequacy and meaning preservation (Rei et al., 2020). By combining both metrics, we ensure a more comprehensive evaluation that accounts for both the syntactic and semantic quality.

3 The Six Classical Challenges

3.1 Domain Mismatch

Domain mismatch has long been a formidable challenge in the field of MT (Wang et al., 2020).

[1] https://github.com/facebookresearch/fairseq.

[2] https://github.com/rsennrich/subword-nmt.

[3] https://www2.statmt.org/wmt23/translation-task.html.

[4] https://github.com/wmt-conference/wmt23-news-systems/tree/master/txt.

[5] https://github.com/mjpost/sacrebleu.

Given the extensive pretraining of LLMs on diverse data, an intriguing question arises: *does the vast knowledge encapsulated in LLMs mitigate the domain mismatch issue in translation tasks?* To explore this, we finetune the LLM using domain-specific parallel data and evaluate its performance on both in-domain (ID) and out-of-domain (OOD) translation tasks. Table 2 presents the results.

- **LLMs excel in-domain but face challenges with domain shifts.** The results indicate that LLM-based translation systems perform exceptionally well on in-domain tasks, as reflected by both surface-level and semantic-level metrics. For instance, in the law domain, the LLM-SFT model achieves a notable BLEU score of 62.0 and a COMET-DA score of 88.0, outperforming Enc2Dec models by approximately 3.0 BLEU points and 4.9 COMET-DA points. However, despite improvements in OOD translation over Enc2Dec models, LLMs still demonstrate significant performance degradation when encountering domain shifts. In the Koran-to-Law OOD translation task, this decline can be as severe as 40.0 BLEU points and 14.4 COMET-DA points. A qualitative analysis is conducted in Table 3.
- **Terminology Mismatch**. A common error in translation systems is the inability to produce accurate domain-specific terminology. For instance, in the medical domain, the term ‘‘Folienverpackung’’ has been inadequately translated as ‘‘slide pack’’, ‘‘sterile packaging’’, and ‘‘slide wraps’’ by the models finetuned with IT, Korean, and Subtitles bitexts, respectively.
- **Style Mismatch**. The LLM fails to generate a hypothesis that accurately matches the out-of-domain style. In the second example, the reference in the IT domain is titled ‘‘Method to calculate forecast’’, while the medical translation system produces ‘‘Prediction method’’.
- **Hallucination**. In the third example, the medical translation system erroneously translates the term ‘‘Tatort’’ as ‘‘accident’’ instead of ‘‘crime’’, where ‘‘accident’’ is a prevalent term in the medical domain. This translation error has been identified as an Input-conflicting Hallucination in the study by Zhang et al. (2023b).

Another concern arises regarding the effective management of domain expertise when a single LLM is tasked with handling multiple domains. Experimental results from the translation system trained on all domains show that, while the LLM performs consistently across various tasks, it falls short compared to domain-specific models. For example, in the Koran translation test, the LLM-SFT model trained on all data lags behind the domain-specific Koran LLM-SFT model by 1.7 BLEU points and 0.8 COMET-DA points.

3.1.1 Scaling Up to Llama2-13B

We utilize the Llama2-13B model to evaluate whether the domain mismatch issue diminishes as the model size increases (Table 2). The results reveal several important trends in multi-domain translation tasks.

First, although the domain mismatch persists, reflected in the performance gap between in-domain and out-of-domain tasks, the 13B model trained on domain-specific data outperforms both the 7B model and Enc2Dec models. It achieves superior results on both sentence-level and semantic-level metrics. For instance, in the law domain, the 13B model attains a BLEU score of 63.9 and a COMET-DA score of 88.4, surpassing the 7B model by 1.9 BLEU points and 0.4 COMET-DA points. This suggests that stronger models can reduce, though not entirely eliminate, the domain mismatch problem. These findings raise the question of whether a more powerful LLM could fully resolve this issue.

Second, the 13B model trained on all data consistently outperforms models trained on specific domains. It falls short by only 0.3 BLEU and COMET-DA points compared to the 13B model trained on Koran data but demonstrates superior translation performance across all other domains. This indicates that a larger model can more effectively learn from multi-domain data due to its increased capacity.

In summary, increasing model capacity enhances overall performance across domains, though the domain mismatch issue remains a significant challenge.

System↓	Law			Medical			IT			Koran			Subtitles		
All Data	64.3	61.5	56.4	61.5	59.7	51.4	50.6	47.5	41.7	20.3	18.6	20.1	28.8	27.0	26.8
Law	63.9	62.0	59.0	38.3	36.1	21.7	31.1	29.6	13.1	14.5	12.0	2.7	22.3	20.4	5.4
Medical	30.2	28.5	18.3	61.4	59.3	56.5	29.2	31.2	11.4	13.6	11.8	1.9	22.0	19.9	4.3
IT	32.8	30.3	9.6	38.5	36.9	14.9	51.0	47.4	43.0	14.4	11.7	2.8	24.2	23.1	8.6
Koran	23.1	22.3	0.2	27.5	28.15	0.1	19.2	16.8	0.2	20.6	20.3	15.9	10.8	10.6	0.5
Subtitles	28.9	27.1	5.5	33.4	33.5	7.9	26.7	26.9	8.5	13.2	11.6	6.4	28.9	28.1	27.3

(a) BLEU scores

System↓	Law			Medical			IT			Koran			Subtitles		
All Data	88.2	87.6	86.3	86.1	85.9	85.1	87.8	87.3	85.1	72.2	71.7	70.4	78.4	77.8	76.5
Law	88.4	88.0	85.9	83.0	82.6	66.0	79.4	78.9	59.4	70.4	68.8	40.6	76.3	75.0	46.5
Medical	76.8	76.3	54.7	85.8	85.7	83.5	78.9	78.4	52.8	69.3	67.7	38.1	75.3	74.2	44.1
IT	81.5	80.1	48.0	82.6	82.1	50.7	88.1	87.5	82.5	70.1	68.1	39.1	76.8	76.2	50.6
Koran	74.0	74.0	33.7	77.3	77.9	32.0	71.1	71.2	37.3	72.5	72.5	58.3	67.4	66.7	41.4
Subtitles	79.5	78.9	47.4	81.1	81.3	54.1	77.8	79.1	57.2	69.3	68.2	51.7	78.4	78.6	74.1

(b) COMET-DA scores

Table 2: Translation quality of multi-domain German-to-English translation tasks, where the system is trained on one domain (rows) and tested on another domain (columns). The black, red, and blue bars refer to the LLM-SFT-13B, LLM-SFT, and Enc2Dec models, respectively. LLMs improve the in- and out-of-domain translation qualities but still suffer from the problem of domain mismatch.

	1: Medical	2: IT	3: Subtitles
Src	Die Pipetten müssen in der intakten Folienverpackung aufbewahrt werden.	Methode die Berechnung der Vorhersage	Du kannst ihr nicht helfen, außer dass du jetzt den Tatort untersuchst.
Ref	Stored pipettes must be kept in the intact foil package.	Method to calculate forecast	You can't do anything but help her by working the crime scene.
All	Pipettes must be stored in the intact foil pouch.	Method to calculate the forecast	You can't help her except by canvassing the scene.
Law	The pipettes must be stored in the intact sheet pack.	Method of calculation of the forecast	You can't help her, except by examining the scene of the crime.
Medical	The pipettes must be stored in the intact foil pouch.	Prediction method	You can't help her now, except by examining the scene of the accident.
IT	The pipettes must be stored in the intact slide pack.	Method for calculating the forecast	You can't help her, except by examining the crime scene.
Koran	The pipettes must be kept in sterile packaging.	A method to calculate the prediction.	If you do not, you will not be able to help her.
Subtitles	The pipettes must be stored in intact slide wraps.	Method for calculating the forecast	You can't help her, except to process the crime scene.

Table 3: Test examples of German-to-English from three domains, which are translated by the LLM-SFT-7B trained on domains listed in the first column. The red, blue, and green color indicates the terminology mismatch, style mismatch, and hallucination phenomena, respectively.

3.2 Amount of Parallel Data

Parallel data is crucial for training encoder-to-decoder translation systems. With the emergence of LLMs, even a small corpus of high-quality parallel data can enhance their translation abilities (Jiao et al., 2023). In this study, using the German-to-English translation task, we examine the impact of varying parallel data sizes, from 10k to 20M, on LLMs and assess the effectiveness of two training strategies for adapting LLMs into translation systems. Our findings suggest:

- **A small amount of parallel data enhances LLM translation performance.** The LLM-SFT curve in Figure 1 shows that supervised fine-tuning with 10k parallel data improves the BLEU score by 2 and the COMET-DA score by 1.1 compared to the Alpaca-only trained model. Moreover, the LLM-SFT model trained 100k parallel data achieves the top BLEU score of 41.6 and the top COMET-DA score of 83.9.

- **An increasing amount of parallel data may degrade LLM translation performance.** Contrary to the belief that more parallel data improves translation quality, LLM models exhibit contrary results. For both LLM-SFT and LLM-CPT-SFT, using large amounts of parallel data (e.g., 5M and 10M) negatively affects performance. Prior research suggests that LLMs acquire most knowledge during the pretraining stage (Zhou et al., 2023), where unintentional exposure to bilingual signals occurs (Briakou et al., 2023). Excessive parallel data might disrupt this acquired knowledge.

- **Supervised fine-tuning outperforms continuous pretraining for utilizing additional parallel data.** Performance curves in Figure 1 reveal that LLM-SFT consistently achieves better results, with a significant increase of up to 2.6 BLEU scores and 1.0 COMET-DA scores in the 100k scenario.

In this section, we evaluate LLM translation systems with varying parallel data amounts. Our

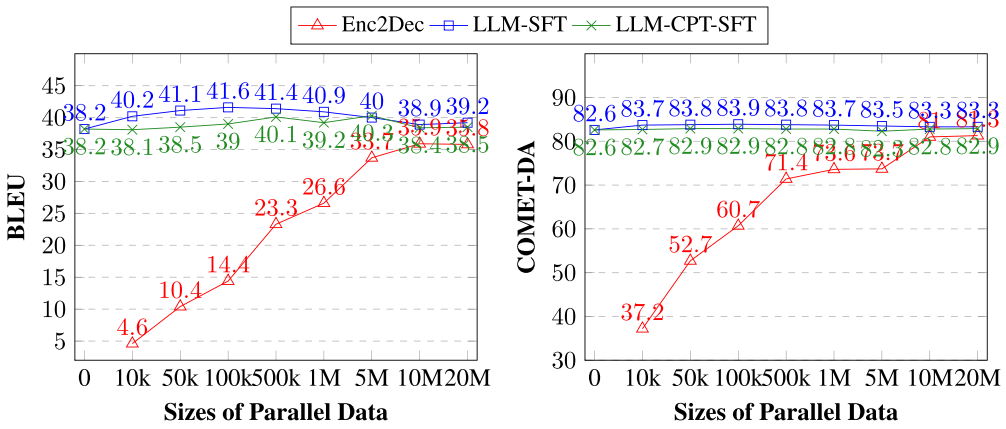

Figure 1: BLEU and COMET-DA scores for German-to-English systems, with ‘‘0’’ on the x-axis indicating models trained exclusively on the Alpaca dataset. LLMs reduce reliance on extensive parallel data.

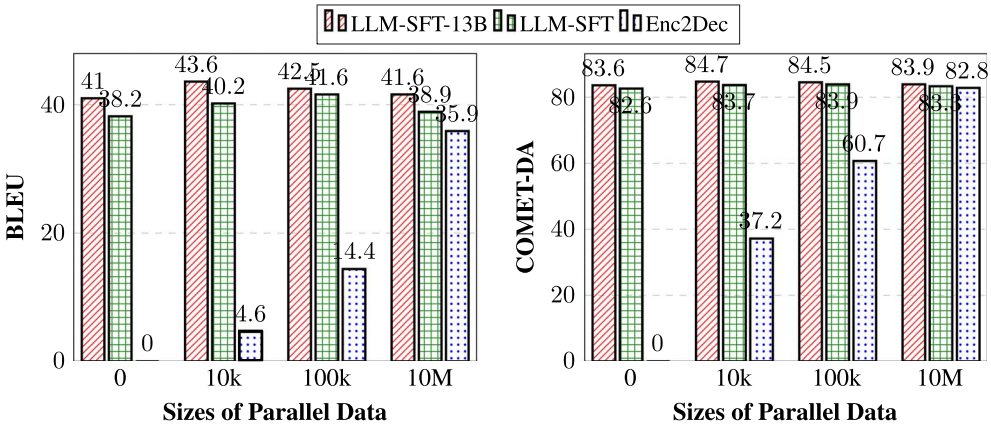

Figure 2: BLEU and COMET-DA scores for German-to-English systems. Increased parallel data adversely affects the performance of the Llama2-13B model.

findings reveal that LLMs do not require extensive translation pair training, unlike conventional Enc2Dec models. This insight encourages researchers to explore efficient ways to utilize parallel data for LLM translation system enhancement, providing a potential direction for future studies to optimize bilingual knowledge and improve machine translation performance using LLMs.

3.2.1 Scaling Up to Llama2-13B

In this study, we utilize a stronger LLM, Llama2-13B, to delve deeper into the influence of parallel data on translation quality. We perform supervised fine-tuning on the parallel data, choosing 10k, 100k, and 10M parallel data as our evaluation points. A noteworthy trend emerges for stronger LLMs, as described below.

For more advanced LLMs, such as Llama2-13B, the degradation point occurs earlier as the volume of parallel data increases. As illustrated in Figure 2, the LLM-SFT-13B model achieves its highest BLEU score of 43.6 and a COMET-DA score of 84.7 when trained on 10k parallel data. However, as the amount of parallel data increases to 100k and 10M, the model’s performance declines, with BLEU scores falling to 42.5 and 40.0, respectively. This finding suggests that LLMs require only a small amount of parallel data to enhance their translation capabilities for major pretrained languages. Conversely, stronger LLMs may be sensitive to performance degradation when exposed to increasing volumes of parallel data.

3.3 Rare Word Prediction

Rare word prediction is crucial in translation, especially for proper nouns, compounds, names, and loanwords referring to specific entities or items (Luong et al., 2015; Sennrich et al., 2016b;

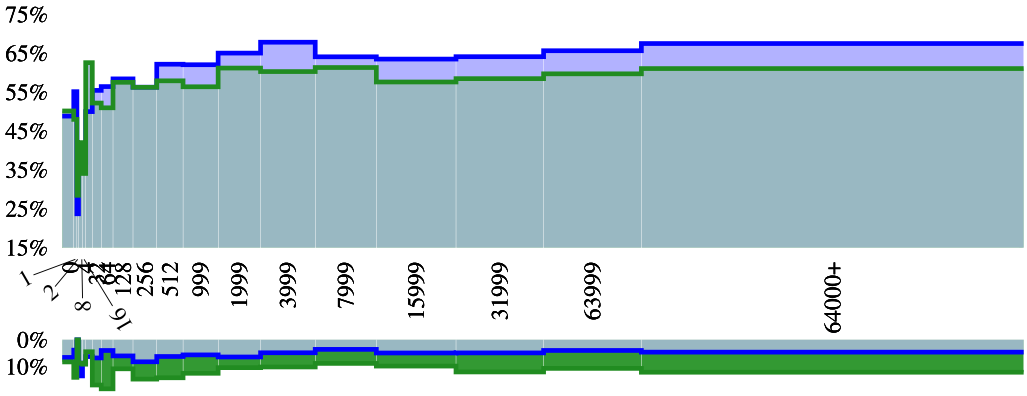

Figure 3: Precision of translation and delete rates by source word frequency. The light blue and dark green indicate the best LLM-SFT and Enc2Dec translation models. The horizontal axis represents the frequency of source word types in the test corpus, where axis labels indicate the upper limit of each frequency range, and the bin width is proportional to the number of word types in that range. The precision and deletion rates are shown on the upper and lower vertical axes respectively. LLMs excel at predicting words that appear more than eight times but perform poorly with rare words.

Wu et al., 2016). Following previous approaches (Koehn and Haddow, 2012; Koehn and Knowles, 2017), we evaluate the precision and deletion rate of rare words in both the LLM and Enc2Dec translation systems. Our analysis includes the best LLM-SFT system trained on 100k language pairs and the top-performing Enc2Dec model trained on 10 million language pairs, as discussed in Section 3.2. Results in Figure 3 reveal that:

- **The LLM translation model demonstrates higher precision and lower deletion rates for frequent words.** Although Enc2Dec models are typically trained on extensive WMT datasets and expected to excel in generating precise in-domain words, the LLM-SFT model outperforms this expectation. The LLM model consistently achieves higher precision for words with a frequency above 16 compared to the Enc2Dec model. Specifically, for words in the frequency bin of (1999,3999], LLM-SFT achieves 67.85% prediction accuracy while that of Enc2Dec is 60.30%. Additionally, the deletion rates are consistently lower than those of the Enc2Dec model.
- **The LLM translation model struggles with infrequent word prediction, leading to translation omissions.** The LLM-SFT system exhibits low precision for words occurring less than 8 times and high deletion rates. For words in the $(2, 4]$ frequency bin, the LLM-SFT reports a word precision of 35.26%, while Enc2Dec achieves 42.03%. Moreover, the LLM deletion rate for the frequency bin $(2, 4]$ is 13.46%, which is 4.76% higher than the Enc2Dec model's 8.70%. Qualitatively, the LLM-SFT fails to translate compound rare words such as ‘‘blätterlos’’ (meaning ‘‘a plant has no leaves’’) and ‘‘lotrechtes’’ (meaning ‘‘perpendicular’’), whereas Enc2Dec successfully generates ‘‘leaflessly’’ as a substitute for ‘‘blätterlos’’. This finding suggests that LLMs struggle with the semantic and morphological complexities of rare or compound words.

Overall, our results show that LLMs achieve reliable word-level translation accuracy, but predicting rare words remains a major challenge that warrants further research.

3.3.1 Scaling Up to Llama2-13B

To further examine the poor performance in predicting unknown and low-frequency words, we conducted an additional evaluation focused on rare word prediction using the Llama2-13B model trained on 10k German-to-English parallel data, as shown in Figure 2. For a clearer comparison, we also present detailed results for the LLM-SFT-7B and Enc2Dec models in Table 4. Based on our findings, we observe that:

A stronger LLM, such as Llama2-13B, more evidently struggles with predicting rare

Freq. Bin	Counts	LLM-SFT-13B		LLM-SFT		Enc2Dec	
		ACC(%)	Delete(%)	ACC(%)	Delete(%)	ACC(%)	Delete(%)
0	80	42.27	12.01	48.84	6.57	**50.17**	8.25
1	22	42.95	15.38	**55.13**	3.85	48.00	14.00
2	11	11.54	7.69	23.72	0.00	**28.47**	0.00
4	28	26.28	9.62	35.26	13.46	**42.03**	8.70
8	22	23.19	8.70	**36.96**	8.70	34.09	9.09
16	46	38.54	14.58	50.00	6.25	**62.59**	4.44
32	62	52.98	6.76	**55.46**	6.76	52.20	16.86
64	82	53.88	8.52	**56.44**	3.98	50.98	18.28
128	139	**59.61**	5.67	58.42	6.00	57.54	10.81
256	165	**56.38**	8.47	56.28	8.20	56.25	14.66
512	183	60.18	9.01	**62.19**	6.22	57.92	14.15
999	244	**62.30**	8.37	62.06	5.62	56.39	12.48
1999	295	63.21	6.46	**65.05**	6.40	61.20	10.38
3999	380	66.33	4.11	**67.85**	4.83	60.30	10.13
7999	425	**66.28**	3.89	64.10	3.56	61.34	8.80
15999	554	**65.31**	4.34	63.51	4.91	57.62	9.80
31999	609	**65.00**	3.77	64.13	4.88	58.48	11.98
63999	680	65.05	3.75	**65.66**	3.97	59.72	10.66
64000	2658	67.29	4.66	**67.52**	4.59	61.07	12.09

Table 4: Precision of translation and delete rates concerning source word types with varying frequencies. ''Freq. Bin'' and ''Counts'' indicate the frequency upper bounds and the count of each source type word. ''ACC'' and ''Delete'' indicate the precision and deletion rates of word prediction.

System	German-to-English
GPT4 (OpenAI, 2023)	39.5
Llama-MT (Du et al., 2023)	39.7
LLM-SFT-0k	22.2
LLM-SFT-100k	36.3

Table 5: d-BLEU scores for TED document-level translation tasks. Llama2-SFT-* are models trained with 0k and 100k German-to-English parallel data in Figure 1, respectively.

words. Despite its high BLEU score, the stronger model consistently underperforms compared to LLM-SFT for rare words occurring less than 16 times. Specifically, for words in the frequency bin (1,2], LLM-SFT-13B achieves a prediction accuracy of 11.54%, while LLM-SFT-7B and Enc2Dec report accuracy scores of 23.72% and 28.47%, respectively. This finding highlights the challenge LLMs face in predicting rare words.

The challenge of accurately generating rare words in LLMs arises from their intrinsic properties and decoding strategies. LLMs, trained using a causal language modeling objective, learn high-frequency words more effectively due to their prevalence in training data (Radford et al., 2018). In contrast, rare words appear less frequently, resulting in a lower level of understanding and mastery by the model. Techniques like greedy decoding or beam search, employed during text generation, favor high-frequency words based on probability distribution (Holtzman et al., 2020). Accordingly, our experiments reveal that more powerful models tend to exacerbate the phenomenon of imprecise generation of rare words.

3.4 Translation of Long Sentences

The length of the source text poses a significant challenge for MT systems due to the need for accurate contextual capture (Wang et al., 2017, 2023b). Given LLMs' extended context window, particularly Llama2's 4,096 maximum input length (Touvron et al., 2023b), we investigate how LLMs handle long sentence translation.

In this section, we evaluate model performance across varying sentence lengths on the general-test2023 test set, segmented following previous

Model ↓	English-to-German		English-to-Chinese		English-to-Ukrainian		English-to-Hebrew	
	BLEU	COMET-DA	BLEU	COMET-DA	BLEU	COMET-DA	BLEU	COMET-DA
LLM-SFT	28.3	84.5	22.5	78.4	**25.2**	**88.8**	**19.8**	**78.6**
ALMA-7B	**30.2**	85.3	29.5	83.0	10.6	88.3	0.7	57.4
TowerInstruct-7B	29.8	**85.7**	**31.4**	**83.5**	12.4	87.0	2.5	44

Table 6: Translation Performance on four English-to-X directions of existing multilingual large language models. LLM-SFT presents the best result in Figure 8. The results suggest that existing multilingual LLMs significantly improve the translation capabilities for pretrained non-English languages. However, they have not yet focused on enhancing the translation performance for low-resource languages.

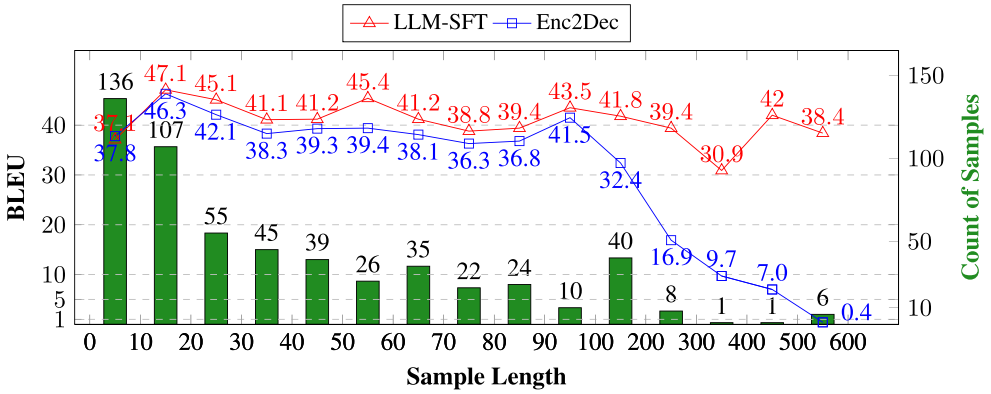

Figure 4: BLEU scores for German-to-English MT systems with varying sample lengths. Sentence-level translation involves lengths below 90 words, while document-level translation concerns longer samples. LLMs improve long-sentence translation and consistently excel in document-level tasks.

settings (Koehn and Knowles, 2017). We categorize the test set into groups based on sentence length, from 1-9 words up to 80-89 words. For document-level translation assessment, we include groups with 100–199 words to 400–499 words, merging with the original documents for a final group of 500-599 words. The test set's longest document contains 582 words. The best models in Section 3.2 are adopted for evaluation, which are the LLM-SFT trained on 100k parallel data and the Enc2Dec model trained on 20M parallel data. Figure 4 illustrates the sentence and document count per group and presents the BLEU scores for each test. Figure 5 presents the COMET-DA score for sentence-level translation cases.

- **The LLM translation system excels at translating long sentences.** Based on the score curves for BLEU and COMET-DA evaluations, the LLM-SFT model outperforms the Enc2Dec model in both surface-level and semantic-level translation tasks for sentences up to 80 words in length. However, for sentences shorter than 10 words, the LLM-SFT model lags behind the Enc2Dec model by 0.7 BLEU points. In contrast, in the semantic-level evaluation, the LLM-SFT model surpasses the Enc2Dec model by 0.5 COMET-DA points. This suggests that the LLM-based model effectively preserves high-quality semantic information in translation tasks.
- **LLMs demonstrate superior document translation capabilities.** As shown in Figure 4, for sentences exceeding 100 words, LLMs maintain consistent high performance, whereas the Enc2Dec model's curve steeply falls to a 0.4 BLEU score. For documents with more than 500 words, LLM-SFT achieves a 38.4 BLEU score, indicating a substantial difference and potential in document-level translation tasks.

The findings presented above empirically demonstrate the proficiency of LLMs in translating sentences of varying lengths and indicate their potential for document-level translation tasks. It

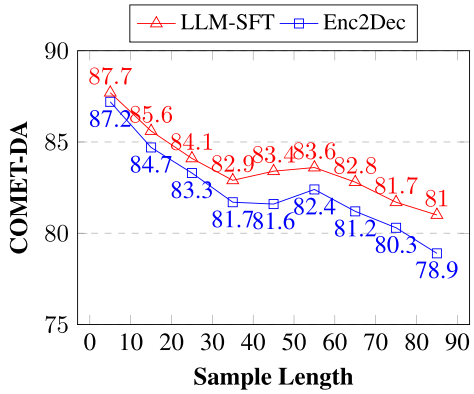

Figure 5: COMET-DA scores for sentence-level translation of German-to-English MT systems with varying sample lengths.

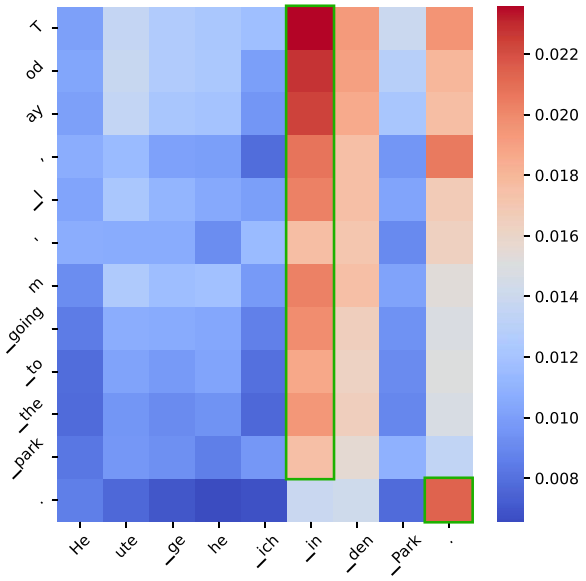

Figure 6: Average attention weight of target-to-source sentences of the LLM-SFT model. The left column is a target English sentence ‘‘Today, I’m going to the park.’’, and the bottom row is the source German sentence ‘‘Heute gehe ich in den Park.’’, tokenized by the Llama2 tokenizer.

is important to note that the LLM-SFT model utilized in Section 3.2 was trained solely on sentence-level parallel data. Despite this limitation, the LLM-SFT model exhibits strong performance for translation tasks involving up to 600 words, highlighting its advantages in handling long context windows and leveraging pretrained knowledge.

3.4.1 Document-Level Translation

We further evaluate our models’ document-level translation competencies by directly testing them on the TED German-to-English document-level translation tasks (Cettolo et al., 2017).[6] For reference, we include the GPT-4 and Llama-MT models (Du et al., 2023) as baselines.

LLMs fine-tuned on sentence-level parallel data demonstrate proficiency in document-level translation tasks. The Llama2-SFT-100k model registers a d-BLEU score of 36.3, significantly surpassing the Llama2-SFT-0k model’s score of 22.2. This model also exhibits competitive performance against the leading GPT-4 and Llama-MT models, with only slightly lower d-BLEU scores. These findings further validate the robustness of LLMs’ pretrained knowledge for document-level translation tasks (Wang et al., 2023b), highlighting their adaptability to more intricate, context-dependent translation scenarios.

3.5 Word Alignment

Previous studies extracted word alignment from the attention matrix within encoder-to-decoder translation models (Garg et al., 2019; Chen et al., 2020; Zenkel et al., 2020) and used it to interpret translation models (Yang et al., 2020, 2021).

In this section, we explore two research questions: 1) *Is it feasible to extract word alignment from LLM attention weights?* and 2) *Can word alignment shed light on the LLM translation process?* To address these questions, we conduct a case study using the LLM-SFT model to process the instruction input and target sentence. We extract attention weights indicating the relationship between source and target words, as shown in Figure 6. Our findings include:

- **Extracting alignments from LLM attention weights is not feasible.** Figure 6 displays the average attention weight across 32 layers in the LLM-SFT translation model. Results reveal that each target sub-token tends to attend the same source token, in this case, ‘‘_in’’, suggesting that attention weights do not explicitly provide word alignment information (Moradi et al., 2021).

- **Aggregated attention weights offer clues for interpreting LLMs.** The observed phenomenon, where target tokens attend the same source tokens, aligns with Wang et al. (2023a)’s findings. They discover that LLMs

[6] https://wit3.fbk.eu/2017-01-d.

tend to aggregate information into one token, and predicted tokens pay the most attention to this token during inference, referred to as the anchor.

To obtain word alignment, methods such as prediction difference (Li et al., 2019) and prompt design offer promising directions for further investigation. However, the most significant challenge lies in interpreting LLMs, for which the insights from this section provide valuable guidance for future research.

3.6 Inference Efficiency

In the realm of inference, two major concerns are inference strategies and efficiency. Beam search and sampling are commonly employed strategies. Beam search predicts the most promising word within a predefined beam size to generate a sentence (Freitag and Al-Onaizan, 2017), while sampling randomly selects the next word based on the word probability distribution. Previous studies have examined the impact of beam size on beam search performance (Koehn and Knowles, 2017), and practitioners in the machine translation field commonly use beam sizes of 4 or 5 (Vaswani et al., 2017). However, due to the extensive size of LLMs, inference efficiency becomes a more challenging issue, with recent works proposing to accelerate the inference process (Li et al., 2023; Alizadeh et al., 2023). In this section, we first analyze the performance difference between these two inference strategies, then discuss inference efficiency. Apart from BLEU, we include the COMET-DA evaluation (Rei et al., 2022) for a semantic-level comparison. We set the beam size to 5 for beam search and vary the temperature with 0.1, 0.5, and 1.0 for sampling.

- **Sampling underperforms beam search in BLEU, but the difference is less pronounced in COMET-DA.** Figure 7 shows that LLM-SFT achieves 41.6 BLEU and 83.9 COMET-DA scores with beam search inference. In contrast, using sampling with a temperature of 0.1, LLM-SFT attains 39.2 BLEU and 83.5 COMET-DA scores. Our results indicate that LLM-SFT, using sampling with a temperature of 0.1, achieves an accuracy of 55.79% in predicting unknown words, but exhibits lower accuracies for frequent words compared to beam search. This suggests that the sampling method's ability to explore more diverse candidate sequences contributes to predicting rare words.

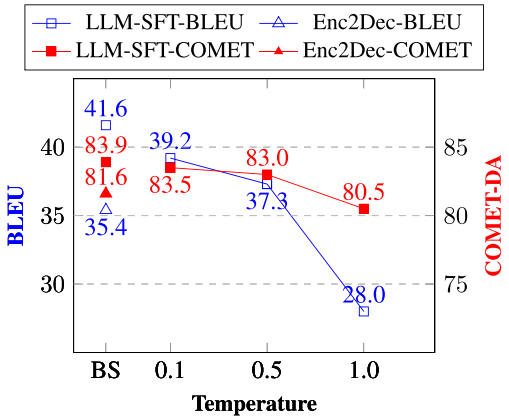

Figure 7: BLEU and COMET-DA scores of the German-to-English systems. ''BS'' indicates the beam search method with a beam size of 5.

- **Compared to Enc2Dec, inference latency poses a significant challenge in utilizing LLMs for MT.** LLMs require an average of 30 seconds for inference, whereas Enc2Dec models require only 0.3 seconds on average, indicating a nearly 100-fold difference. This substantial discrepancy in inference latency between LLMs and Enc2Dec models presents a significant hurdle in the practical application of LLMs for machine translation. The longer inference time of LLMs can be attributed to their large model size and extensive parameters.

Current research is exploring ways to reduce the inference latency of LLMs, such as model compression or hardware acceleration (Dettmers et al., 2022; Agrawal et al., 2023; Frantar and Alistarh, 2023; Dettmers et al., 2023; Lin et al., 2023; Kim et al., 2023). Additionally, further exploration could optimize the performance of sampling methods in handling rare words, potentially enhancing the overall translation quality of LLMs.

4 New Challenges

Within the research field of LLMs, two pressing challenges arise. The first challenge concerns the translation quality for language pairs that are underrepresented during the pretraining

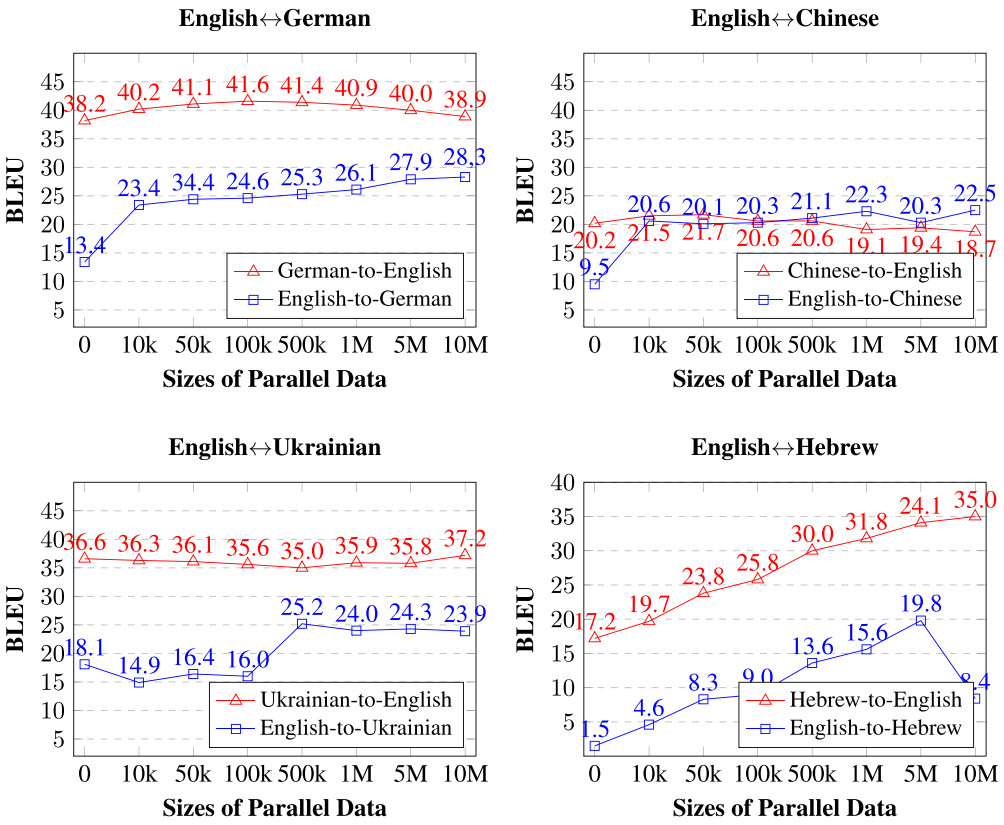

Figure 8: BLEU scores of bi-directional translation results using Llama2-7b based translation models across multiple language pairs, where English, German, Chinese, Ukrainian, and Hebrew are sorted based on the level of available resources, with English having the highest resources and Hebrew the least in the pretraining stage.

stage, possessing significantly less data volume compared to major languages. In our case, approximately 90% of the data in Llama2-7b is English (Touvron et al., 2023b), leading to a highly skewed data distribution. The second challenge involves the evaluation of translation quality. Automatic evaluation metrics, such as BLEU or COMET, may not fully correspond with human evaluation standards (Liu et al., 2023; Mao et al., 2023), complicating the accurate assessment of LLMs.

4.1 Pretraining Resource Imbalance

We conduct an experiment using Llama-7b-based translation models on four translation pairs from the WMT23 tasks: German-to-English, Chinese-to-English, Ukrainian-to-English, and Hebrew-to-English translations. According to the technical report of Llama2 (Touvron et al., 2023b), the distribution of the five languages in the dataset of Llama2 is as follows: 89.70% for English, 0.17% for German, 0.13% for Chinese, 0.07% for Ukrainian, and less than 0.005% for Hebrew. The results are presented in Figure 8, and we observe:

- **Translation performance is significantly impacted by the available resources for each language.** The X-to-English tasks, with the highest resource of English, consistently achieve stable translation performance compared to other directions, as evidenced by a smooth performance curve (except for Hebrew-to-English) with varying amounts of parallel data. Conversely, Hebrew, with the least resource in Llama2 pretraining data, exhibits improved translation performance as parallel data increases. This observation underscores the importance of diverse and balanced datasets for pretraining LLMs to ensure equitable performance across languages.
- **For low-resource pretraining data languages, a substantial change in slope is observed as the amount of parallel data increases.** The English-to-X tasks, except English-to-Hebrew, exhibit fluctuating performance curves as the volume of parallel data increases. Specifically, the curve for

English-to-German experiences a sharp rise from 0k to 10k parallel data, a pattern also observed in English-to-Chinese. Both German and Chinese account for more than 0.1% of the pretraining data distribution. The English-to-Ukrainian curve displays a notable inflection point between 100k and 500k, with Ukrainian constituting 0.07% of the data distribution. In contrast, the English-to-Hebrew curve remains smooth, as Hebrew does not fall within the languages that comprise more than 0.005% of the data distribution. These points of significant increase may serve as indicators of the emergent translation capabilities of LLMs.

The above findings suggest that addressing the challenges of low-resource pretrained languages and identifying inflection points where LLMs' capabilities emerge will be essential to enhance the effectiveness of various translation directions.

4.1.1 Performance of Existing MLLMs

Numerous studies have recently focused on adapting English-centric LLMs into Multilingual LLMs (MLLMs) (Xu et al., 2024b; Qin et al., 2024). We employ two publicly accessible MLLMs, ALMA-7B (Xu et al., 2024a) and TowerInstruct-7B (Alves et al., 2024), and directly evaluate them on the four English-to-X translation routes depicted in Figure 8. Both models consistently utilize monolingual data from various languages for continued pretraining, followed by fine-tuning with high-quality parallel data, based on the Llama2-7B model. Consequently, these two models successfully enhance the translation quality of pretrained languages such as German and Chinese.

MLLMs exhibit a degree of translation generalization. Despite not being pretrained on Ukrainian, ALMA-7B and TowerInstruct-7B achieve high COMET-DA scores of 88.3 and 87.0, respectively, indicating successful English-to-Ukrainian translations with high semantic consistency. We posit that this success is attributable to their training on Russian texts, a linguistically similar language to Ukrainian. Nevertheless, both models demonstrate limited translation capabilities for Hebrew due to insufficient training data in this language. These findings underscore the necessity for further research and development to enhance the LLMs for low-resource languages.

System	MQM	d-BLEU
GPT-4	54.81	43.7
Llama2-MT	28.40	43.1
Google	22.66	47.3

Table 7: The comparison between human (MQM) vs. automatic (d-BLEU) evaluation methods over three representative systems on the Chinese-to-English translation task, with a color scale to denote the ranking. The Pearson correlation coefficient between MQM and d-BLEU is -0.53.

4.2 Evaluation Issues

We conducted an assessment of three representative systems using the WMT2023 Discourse-Level Literary Translation Chinese-English Testset.[7] These include: the *Commercial Translation System*, Google Translate,[8] known for its superior translation performance; the *Commercial LLM Systems*, specifically the GPT-4 (8K) API,[9] recognized for its comprehensive context modeling capabilities (Ouyang et al., 2022; Wang et al., 2023b); and the *Open-sourced LLM Models*, particularly Llama2-7b (4K) (Touvron et al., 2023b), optimized for document-level translation using a 200K general-domain document-level training set. For evaluation, we utilized both automatic and human methods. The former uses document-level sacreBLEU (d-BLEU) (Liu et al., 2020), while the latter adopts multidimensional quality metrics (MQM) (Lommel et al., 2014) to fit the literary translation context. Table 7 presents a comparative performance analysis of these systems.

- **A moderate negative correlation exists between human and automatic evaluations.** A Pearson correlation coefficient of -0.53 indicates a divergence between the two evaluation methods. This discrepancy suggests that human and automatic evaluations can provide

[7] https://www2.statmt.org/wmt23/literary-translation-task.html.
[8] https://translate.google.com.
[9] https://platform.openai.com.

complementary insights into document-level translation quality. The findings underscore the importance of combining both evaluation methods and highlight the potential limitations of current automatic evaluation approaches in assessing machine translation systems.

- **The need for human-aligned evaluation in LLMs.** The application of LLMs for translation tasks emphasizes the need for evaluation methods that accurately reflect human-like or human-preferred translation quality. The observed divergence between human and automatic evaluations in our study suggests that current automatic metrics may not fully capture the nuances appreciated by human evaluators. This calls for further research to develop and refine evaluation methods better aligned with human preferences and expectations in translation, especially as we continue to enhance the complexity and capabilities of language models.

The above findings highlight the importance of advancing evaluation methodologies that align with human preferences for LLM translation.

5 Conclusions

Our research highlights several key findings. On the positive side, LLMs have effectively removed the dependence on bilingual data for translation into major languages. Additionally, despite being fine-tuned solely on sentence-level translation pairs, LLMs demonstrate impressive capabilities in handling long sentences and even document-level translations. However, challenges persist. LLMs face difficulties in adapting to multi-domain tasks and predicting rare words. Our experiments suggest that larger models possess greater potential to acquire multi-domain translation knowledge, offering a promising path to mitigating the domain shift issue. Yet, performance for low-resource languages remains suboptimal, and inference delays pose a significant bottleneck. Addressing these limitations, particularly improving translation for underrepresented languages, will be crucial. Our findings prove that leveraging bilingual corpora more efficiently could offer a viable solution. Furthermore, the need for more robust, human-aligned evaluation metrics remains an urgent area for future research.

Acknowledgments

This work was supported in part by the Science and Technology Development Fund, Macau SAR (Grant No. FDCT/0070/2022/AMJ, the mainland China collaboration project, China Strategic Scientific and Technological Innovation Cooperation Project Grant No. 2022YFE0204900), the Science and Technology Development Fund, Macau SAR (Grant No. FDCT/060/2022/AFJ, the mainland China collaboration project, National Natural Science Foundation of China Grant No. 62261160648), the Multi-year Research Grant from the University of Macau (Grant No. MYRG-GRG2024-00165-FST), and the Tencent AI Lab Rhino-Bird Gift Fund (Grant No. EF2023-00151-FST).

References

Amey Agrawal, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S. Gulavani, and Ramachandran Ramjee. 2023. Sarathi: Efficient llm inference by piggybacking decodes with chunked prefills. *arXiv preprint arXiv:2308.16369*.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.692

Keivan Alizadeh, Iman Mirzadeh, Dmitry Belenko, Karen Khatamifard, Minsik Cho, Carlo C. Del Mundo, Mohammad Rastegari, and Mehrdad Farajtabar. 2023. Llm in a flash: Efficient large language model inference with limited memory. *arXiv preprint arXiv:2312.11514*. https://doi.org/10.18653/v1/2024.acl-long.678

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the*

Association for Computational Linguistics: EMNLP 2023, pages 11127–11148, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.744

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.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.524

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Yun Chen, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. 2020. Accurate word alignment induction from neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 566–576, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.42

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Andrea Cossu, Tinne Tuytelaars, Antonio Carta, Lucia Passaro, Vincenzo Lomonaco, and Davide Bacciu. 2022. Continual pre-training mitigates forgetting in language and vision. *arXiv preprint arXiv:2205.09357*. https://doi.org/10.2139/ssrn.4495233

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332.

Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. 2023. Spqr: A sparse-quantized representation for near-lossless llm weight compression. *arXiv preprint arXiv:2306.03078*.

Zefeng Du, Wenxiang Jiao, Longyue Wang, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek Wong, Shuming Shi, and Zhaopeng Tu. 2023. On extrapolation of long-text translation with large language models.

Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? A comprehensive evaluation. *arXiv preprint arXiv:2304.01746*.

Elias Frantar and Dan Alistarh. 2023. Sparsegpt: Massive language models can be accurately pruned in one-shot. *arXiv preprint arXiv:2301.00774*.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3207

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. 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 4453–4462, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1453

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene

Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pre-training of large language models: How to (re) warm your model? *arXiv preprint arXiv:2308.04014*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. ParroT: Translating during chat using large language models tuned with human translation and feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15009–15020, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.1001

Sehoon Kim, Coleman Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W. Mahoney, and Kurt Keutzer. 2023. Squeezellm: Dense-and-sparse quantization. *arXiv preprint arXiv:2306.07629*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Philipp Koehn. 2009. *Statistical Machine Translation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815829

Philipp Koehn and Barry Haddow. 2012. Interpolated backoff for factored translation models. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers*, San Diego, California, USA. Association for Machine Translation in the Americas.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3204

Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. 2019. On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1293–1303, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1124

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. 2023. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6342–6353, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.391

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. 2023. Awq: Activation-aware weight quantization for llm compression and acceleration. *arXiv preprint arXiv:2306.00978*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.153

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742. https://doi.org/10.1162/tacl_a_00343

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463. https://doi.org/10.5565/rev/tradumatica.77

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing, China. Association for Computational

Linguistics. https://doi.org/10.3115/v1/P15-1002

Chenyang Lyu, Jitao Xu, and Longyue Wang. 2023. New trends in machine translation using large language models: Case examples with chatgpt. *arXiv preprint arXiv:2305.01181*.

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2023. Gpteval: A survey on assessments of chatgpt and gpt-4. *arXiv preprint arXiv:2308.12488*.

Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. 2021. Measuring and improving faithfulness of attention in neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2791–2802, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.243

Masato Neishi and Naoki Yoshinaga. 2019. On the relation between position information and sentence length in neural machine translation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 328–338, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/K19-1031

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Jianhui Pang, Derek Fai Wong, Dayiheng Liu, Jun Xie, Baosong Yang, Yu Wan, and Lidia Sam Chao. 2023. Rethinking the exploitation of monolingual data for low-resource neural machine translation. *Computational Linguistics*, pages 1–22. https://doi.org/10.1162/coli_a_00496

Jianhui Pang, Baosong Yang, Derek F. Wong, Dayiheng Liu, Xiangpeng Wei, Jun Xie, and Lidia S. Chao. 2024a. MoNMT: Modularly leveraging monolingual and bilingual knowledge for neural machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11560–11573, Torino, Italia. ELRA and ICCL.

Jianhui Pang, Fanghua Ye, Derek Wong, Xin He, Wanshun Chen, and Longyue Wang. 2024b. Anchor-based large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4958–4976, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.295

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. https://doi.org/10.3115/1073083.1073135

Thierry Poibeau. 2017. *Machine Translation*. MIT Press. https://doi.org/10.7551/mitpress/11043.001.0001

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. https://doi.org/10.18653/v1/W18-6319

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22:

Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). 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. https://doi.org/10.18653/v1/2020.emnlp-main.213

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2323

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1162

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023a. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9840–9855, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.609

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023b. 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. https://doi.org/10.18653/v1/2023.emnlp-main.1036

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023c. Findings of the WMT 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine Translation*, pages 55–67, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.3

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1301

Yong Wang, Longyue Wang, Shuming Shi, Victor OK Li, and Zhaopeng Tu. 2020. Go from the general to the particular: Multi-domain translation with domain transformation networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9233–9241. https://doi.org/10.1609/aaai.v34i05.6461

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

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*.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024b. A survey on multilingual large language models: Corpora, alignment, and bias. *arXiv preprint arXiv:2404.00929*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Baosong Yang, Longyue Wang, Derek F. Wong, Shuming Shi, and Zhaopeng Tu. 2021. Context-aware self-attention networks for natural language processing. *Neurocomputing*, 458:157–169. https://doi.org/10.1016/j.neucom.2021.06.009

Yilin Yang, Longyue Wang, Shuming Shi, Prasad Tadepalli, Stefan Lee, and Zhaopeng Tu. 2020. On the sub-layer functionalities of transformer decoder. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4799–4811, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.432

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. 2007. On early stopping in gradient descent learning. *Constructive Approximation*, 26:289–315. https://doi.org/10.1007/s00365-006-0663-2

Thomas Zenkel, Joern Wuebker, and John DeNero. 2020. End-to-end neural word alignment outperforms GIZA++. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1605–1617, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.146

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*. https://doi.org/10.2352/EI.2023.35.1.VDA-396

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.

A Detailed Experimental Settings

A.1 Models

In this section, we provide a detailed description of the models and settings used in our experiments.

- For the LLM-SFT models, we employ Llama2-7B and Llama2-13B as the backbone models. During the training process, a learning rate of 3e-4 is used for continued pretraining (CT), while a learning rate of 2e-5 is applied for supervised fine-tuning (SFT) (Touvron et al., 2023b).
- For the Enc2Dec models, we utilize the transformer-base architecture for translation tasks with 500k or fewer parallel texts, and the transformer-big architecture for tasks with more than 500k parallel texts (Pang et al., 2024a). During the training phase, a learning rate of 1e-5 and a batch size of 8,192 maximum tokens are used. We evaluate the model every 1,000 update steps and terminate the training process when the validation performance plateaus or worsens, with a patience of 20 (Pang et al., 2023). The Adam optimizer is employed to optimize model parameters, with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$ (Kingma and Ba, 2014).
- In Section 4.2, Llama2-MT is trained on an open-source, document-level training set (Wang et al., 2023c).[10]
- ALMA-7B and TowerInstruct-7B are obtained from the open-source HuggingFace community.[11] Default settings are used to generate translation hypotheses (Xu et al., 2024a; Alves et al., 2024).

In this section, we have provided a comprehensive overview of the models and settings employed in our study, ensuring that our methodology is clear and reproducible.

A.2 Definitions of LLM Training

In this section, we provide clear definitions of LLM training methods, including instruction tuning, supervised fine-tuning (SFT), and continued pre-training (CPT).

- **Instruction Tuning**: Instruction tuning is a process of fine-tuning large language models to follow specific instructions provided in the input (OpenAI, 2023; Taori et al., 2023). This is done by training the model on a dataset containing examples with both instructions and corresponding correct responses. The goal is to make the model more controllable and useful by enabling it to understand and respond to explicit instructions given by the user.

[10] https://www2.statmt.org/wmt23/literary-translation-task.html.

[11] https://huggingface.co/haoranxu/ALMA-7B, https://huggingface.co/Unbabel/TowerInstruct-7B-v0.2.

- **Supervised Fine-Tuning**: Supervised fine-tuning is the process of adapting a pre-trained language model to a specific task or domain using labeled data. This involves training the model on a dataset containing input-output pairs, where the outputs are the ground truth or correct responses (Jiao et al., 2023). The model learns to generate appropriate responses based on the input by minimizing the difference between its predictions and the ground truth. Supervised fine-tuning helps improve the model's performance on tasks like text classification, sentiment analysis, and question-answering (Fang et al., 2023; Pang et al., 2024b).
- **Continued Pretraining**: Continuous pretraining refers to the ongoing process of training a language model on a large corpus of text, often from diverse sources, to learn general language understanding (Cossu et al., 2022; Gupta et al., 2023). This is done before any fine-tuning or task-specific training. The idea is to keep updating the model's knowledge and understanding of language as new data becomes available, making it more robust and adaptable.

A.3 Instruction Formats

Table 8 showcases two unique data formats that are employed for the purpose of constructing bilingual pairs, which are essential for training LLMs. The first format adheres to the design principles of the Alpaca dataset, as described in the literature (Taori et al., 2023). This format has been widely used and accepted in various research studies and applications. On the other hand, the second format involves concatenating two pairs of text, each representing a different language. To distinguish between the languages in this concatenated format, language tags are incorporated (Zhu et al., 2024).

A.4 Rare Word Precision

First, we calculate the source word type frequency using the entire WMT23 bilingual corpus, which comprises 295,805,439 translation pairs. Subsequently, we concatenate the translation results with the corpus and employ *FastAlign* [12] to determine the alignment information. In this manner, we adhere to the methodologies employed in previous studies (Koehn and Haddow, 2012; Koehn and Knowles, 2017) to compute the word precision and deletion ratio for each source word, as demonstrated in Algorithm 1.

[12] https://github.com/clab/fast_align.

Algorithm 1 Word Precision and Delete Rates

Require: The source texts T_{src}, the target texts T_{tgt}, and the source-to-targ et alignments of hypothesis pairs $\mathcal{A}_{hyp}$, with identical sorted index; The source word frequency $word2freq$;
Ensure: Word precision $\mathcal{P}$ and deletion rates $\mathcal{D}$;

$\mathcal{P}, \mathcal{D} \leftarrow$ Initialized as an zero list;
for each t_{src}, t_{tgt}, a_{hyp} in (T_{src}, T_{tgt}, $\mathcal{A}_{hyp}$) **do**
  for each source word w_s in t_{src} **do**
    if w_s not in a_{hyp} **then**
      $\mathcal{D}_{w_s} += 1$;
      continue;
    end if
    List of target words $W^t_{hyp} = a_{hyp}(w_s)$;
    $c^t_{hyp} \leftarrow$ Length of W^t_{hyp};
    $c^t_{ref} \leftarrow 0$;
    for each rw in W^t_{hyp} **do**
      for each tw in t_{tgt} **do**
        if tw == rw **then**
          $c^t_{ref} += 1$;
        end if
      end for
    end for
    $minc =$ Min(c^t_{ref}, c^t_{hyp});
    $maxc =$ Max(c^t_{ref}, c^t_{hyp});
    $\mathcal{P}_{w_s} +=$ Min($\frac{minc}{maxc}$, 1);
  end for
end for
$\mathcal{P}_w = \mathcal{P}_w / word2freq(w)$;
$\mathcal{D}_w = \mathcal{D}_w / word2freq(w)$;
return $\mathcal{P}, \mathcal{D}$.

A.5 Human Evaluation

This section provide the detail information about the human evaluation in Section 4.2. Following the industry-endorsed criteria of Wang et al. (2023c), the human evaluation was performed by professional translators using an adaptation of the multidimensional quality metrics (MQM) framework (Lommel et al., 2014).

B Further Experimental Resutls

B.1 Alignment

Table 9 showcases the word alignment outcomes derived from each layer of the LLM-SFT translation model, which was trained on 100k

German-to-English datasets. The findings indicate a tendency for target tokens to align with the same source tokens throughout all layers.

B.2 Beam Search

Figure 9 shows the German-to-English translation performance on generaltest2023 of WMT23 with increasing beam size. The results show that increasing the beam size enhances surface-level similarity to the ground truth, but has minimal impact on semantic-level similarity.

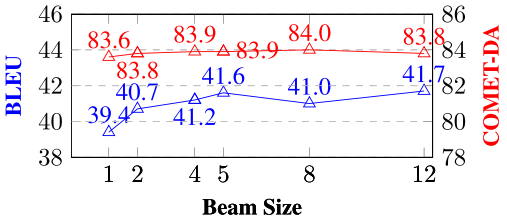

Figure 9: BLEU and COMET-DA scores with beam size for LLM-SFT-100k in Section 3.2.

Type	Data Format
SFT	Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. ### Instruction: Translate the following sentences from German to English. ### Input: Haben Sie einen Blick auf andere Restaurants in der Nähe mit ähnlicher Küche geworfen? ### Response: Have you looked at other nearby restaurants with similar cuisine?
CPT	[German]: Haben Sie einen Blick auf andere Restaurants in der Nähe mit ähnlicher Küche geworfen? [English]: Have you looked at other nearby restaurants with similar cuisine?

Table 8: Data formats utilized for training LLMs, where SFT represents supervised finetuning, and CPT denotes continued pretraining.

Layer	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
T	_in	.	ute	.	.	_in	_den	_ich	_in	_in	.	.	_den	_den	_den	_den
od	he	_Park	ute	_den	ute	_in	_den	_den	_in	_in	.	.	he	_den	_den	_den
ay	he	.	ute	_den	ute	_in	_den	_Park	_in	_in	.	ute	he	_den	_den	_den
,	.	.	ute	_den	ute	_den	_den	_Park	He	_in	.	ute	_in	_den	_den	_den
_I	_in	_den	.	_den	ute	_in	_den	He	_in	_in	.	ute	_den	.	_den	_den
’	_in	He	ute	_den	ute	He	_den	He	He	_in	He	ute	.	.	_den	_den
m	_in	_Park	ute	_den	he	_in	_den	_ge	_in	_in	_in	ute	he	.	_den	_den
_going	_in	.	.	_den	he	_in	_den	_ge	_in	_in	_in	ute	_Park	.	.	_den
_to	_in	_den	ute	_den	he	_in	_den	_ich	_in	_in	.	ute	_den	.	.	_den
_the	_in	_in	ute	_den	.	_den	_den	_den	_in	_in	.	ute	_den	.	.	_den
_park	_in	.	.	_den	.	_in	_den	_Park	_in	_in	_in	ute	_Park	.	.	_den
.	_in	_in	.	He	.	.	_den	_Park	He	.	He	ute	.	.	.	_den

Layer	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
T	_in	_in	_Park	.	_in	_den	_in	.	_in	.	_in	.	_in	_ich	_den	_in
od	_den	_in	_in	.	_in	_den	_in	.	_in	.	_in	.	_in	_ich	.	_in
ay	_den	_in	_in	_den	_in	_den	_in	ute	.	.	_in	.	_in	_ich	.	_in
,	_in	_in	_in	_den	_in	_den	_in	.	.	.	_in	.	_in	.	He	_in
_I	_ich	_in	_in	_den	_in	_den	_in	.	.	.	_in	.	_in	_ich	_in	_in
’	_ich	_in	_ich	.	_in	_den	_in	He	ute	_Park	_in	.	_in	_ich	_den	_in
m	_den	_in	_den	.	_in	_den	_in	.	he	.	_in	.	_in	_ich	_den	_in
_going	_ich	_in	_den	.	_in	_den	_in	.	he	.	_in	.	_in	_ich	_den	_in
_to	_in	_in	_in	.	_in	_den	_in	.	he	.	_in	.	_in	_ich	_den	_in
_the	_in	_in	_in	.	_in	_den	_in	.	_in	.	_in	.	_in	.	He	_in
_park	_ich	_in	_den	_den	_in	_den	_in	.	_Park	.	_in	_in	_in	_ich	He	_in
.	_in	_Park	.	_den	_in	_den	.	.	.	.	_Park	_in	.	_den	He	_in

Table 9: Word alignment induced from target-to-source attention weights for each layer in LLM-SFT German-to-English translation model. The translation model is supervised-finetuned on 100k parallel data. The left row is a target English sentence ‘‘Today, I’m going to the park.’’, and its source German sentence is ‘‘Heute gehe ich in den Park.’’. Both sentences are tokenized by the Llama2 tokenizer. We observe that target tokens tend to attend the same source tokens within each layer.