Source-primed Multi-turn Conversation Helps Large Language Models Translate Documents

Hanxu Hu, Jannis Vamvas, Rico Sennrich
University of Zurich
{hanxu.hu, jannisnikos.vamvas, rico.sennrich}@uzh.ch

Abstract

LLMs have paved the way for truly simple document-level machine translation, but challenges such as omission errors remain. In this paper, we study a simple method for handling document-level machine translation, by leveraging previous contexts in a multi-turn conversational manner. Specifically, by decomposing documents into segments and iteratively translating them while maintaining previous turns, this method ensures coherent translations without additional training, and can fully re-use the KV cache of previous turns thus minimizing computational overhead. We further propose a 'source-primed' method that first provides the whole source document before multi-turn translation. We empirically show this multi-turn method outperforms both translating entire documents in a single turn and translating each segment independently according to multiple automatic metrics in representative LLMs, establishing a strong baseline for document-level translation using LLMs.[1]

1 Introduction

Large language models (LLMs) have demonstrated notable abilities in handling various natural language processing tasks and following diverse instructions effectively. One key application is machine translation, whereas previous approaches relied on specialized encoder-decoder translation models. Recent studies have explored both prompting techniques (Chen et al., 2024; Karpinska and Iyyer, 2023; Lu et al., 2024) and fine-tuning methods (Alves et al., 2024; Wu et al., 2024) for improving LLMs' translation capabilities. While document-level translation remains more challenging than sentence-level tasks (Kocmi et al., 2024), recent works have explored ways to enhance the document translation ability of LLMs through fine-tuning on parallel datasets (Wu et al., 2024) and various prompting techniques (Wang et al., 2023) to improve context awareness. These approaches provide analyses of how current LLMs handle document-level translation.

Multi-turn conversation is a representative way of interacting with LLMs. Previous works have leveraged this ability for following instructions (Zheng et al., 2023) and building agents (Park et al., 2024) for reasoning tasks. More recently, methods have been explored that leverage multiple segments as exemplars in a multi-turn manner to help language models with translation tasks, but these segments tend to be fixed (Kocmi et al., 2024) and do not adapt to subsequent translation data. While retrieval-based approaches have been proposed (Zhang et al., 2023; Zafar et al., 2024), they do not make use of context from the same document. Wang et al. (2023) first explored the setting of inputting a document's content sentence-by-sentence in a multi-turn manner. In this paper, we extend this setting to translate documents and use paragraphs as the translation unit in each turn.

Specifically, we separate the document into multiple segments, and then in each conversational turn, we pass a segment combined with user instructions as input, having the model output the corresponding translated segment. LLMs have access to all previous turns, which helps translate the current segment. In this manner, we merely expand the conversation without modifying the prefix, allowing for caching of previous turns. This strategy has the drawback compared to single-turn strategies in that initial segments are translated with little context. Thus, we propose a new variant that first presents the entire source document before conducting multi-turn translation. This gives the model access to future context that previous multi-turn approaches lack. This approach has the advantage of providing information about the

[1] Code and data are available here: https://github.com/ZurichNLP/multiturn-llm-docmt

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 23702–23712
November 4-9, 2025 ©2025 Association for Computational Linguistics

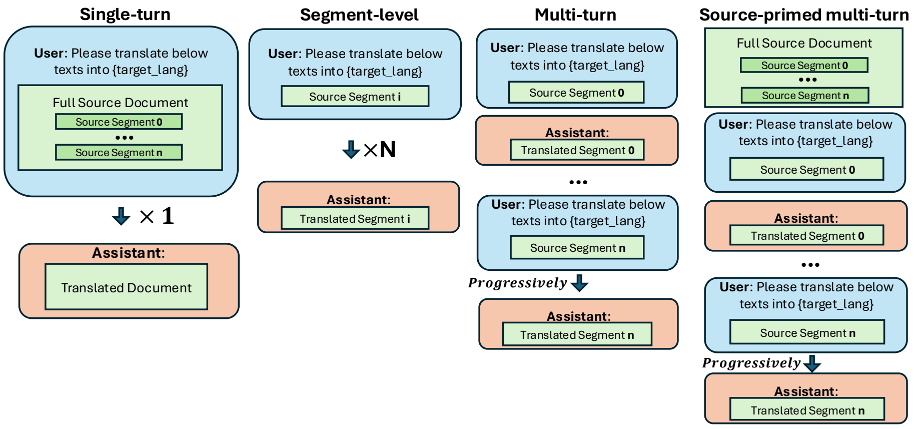

Figure 1: Different settings of document-level translation using LLMs.

document's topic and style, which might help the model generate appropriate tense and formality levels from the start.

This methodology is simple, effective, and requires no additional training, applicable to any chat-capable LLMs. Our experiments on various proprietary and open-weight LLMs demonstrate its effectiveness compared to the original multi-turn setting and two baselines: single-turn document input and context-free segment input. BLEU, COMET and BlonDE (Jiang et al., 2022) scores show that the multi-turn conversational approach outperforms alternatives and can serve as a standard for document translation research. We also analyze how factors like domain type and document length affect this method's performance.

2 Methodology

2.1 Baselines

2.1.1 Single-turn document-level translation

The simplest way of translating a document is to input the whole source document and translate it at once in a single turn. It allows the model to see the context of the entire document, thereby ensuring coherence to a certain extent, but it relies on the model's ability of long-context understanding and generation.

2.1.2 Segment-level translation

Another approach of translating a document is to first split it into segments and then translate each segment independently, as is done in the WMT24 shared task (Kocmi et al., 2024). Segment-level translation has been standard for decades, but it lacks global information of document and might have reduced coherence. There has been a recent trend to use paragraphs instead of sentences as segments (Karpinska and Iyyer, 2023; Kocmi et al., 2024).

2.2 Multi-turn conversational translation

Segment-in-context translation enables LLMs to leverage document-level context while minimizing omission errors. Following Wang et al. (2023), we aim for efficient scalability to long documents, and thus frame document translation as a multi-turn conversation where previous turns are accessible but not overwritten, allowing LLMs to compute and cache hidden states once. This approach utilizes LLMs' existing training on conversations without requiring additional fine-tuning. We also introduce a **'source primed multi-turn'** variant that presents the entire source document before translation, providing access to future context.

3 Experiments

3.1 Models

We used three instruct LLMs to conduct all experiments, due to their general-purpose nature and chat usage. We use the proprietary model GPT-4o-mini (OpenAI, 2024), and the open-weight models Llama-3.1-8B-Instruct (Dubey et al., 2024) and Qwen-2.5-7B-Instruct (Yang et al., 2025). We run the open-weight models using the vLLM framework with greedy search.[2]

3.2 Tasks

We experiment with methods mainly on the WMT 24 General Track for document-level machine translation scenarios (Kocmi et al., 2024),

[2] https://github.com/vllm-project/vllm

Method	**Qwen-2.5-7B-Instruct**		**Llama-3.1-8B-Instruct**		**GPT-4o-mini**	
	dBLEU	COMET_da	dBLEU	COMET_da	dBLEU	COMET_da
Single-turn	20.83	-	20.68	-	31.98	-
Segment-level	18.72	70.36	21.65	76.40	31.82	83.04
Multi-turn	20.37	72.37	22.85	78.40	31.74	84.30
Multi-turn sp (Ours)	**21.46**	**74.23**	**23.22**	**78.95**	**32.51**	**84.38**
Single-turn + ICL	20.66	-	21.79	-	31.95	-
Segment-level + ICL	20.52	71.98	23.07	78.40	32.37	84.08
Multi-turn + ICL	**21.23**	72.96	23.47	79.02	32.50	84.30
Multi-turn sp + ICL (Ours)	21.10	**73.86**	**24.23**	**79.20**	**32.73**	**84.42**

Table 1: Results on WMT-24 (average across all directions), where ICL means adding few-shot in-context learning exemplars as prefix. We do not report COMET in the single-turn setting because we lack segment-level alignments. Multi-turn sp means our source-primed multi-turn variant.

which is segmented and aligned at the paragraph level. It contains 9 En-to-X (English to Chinese, Czech, German, Hindi, Icelandic, Japanese, Russian, Spanish, Ukrainian) directions and 2 X-to-X (Czech to Ukrainian and Japanese to Chinese) directions. It covers various domains (news, literary, speech, social media) and has documents of varying lengths. For further evaluation with the document-level translation metric BlonDE (Jiang et al., 2022), we also use the Chinese-to-English direction from WMT 23 (Kocmi et al., 2023) to evaluate our methods and baselines.

Method / Metrics	**dBLEU**	**BlonDE**
Llama-3.1-8B-Instruct		
Single-turn	23.67	-
Segment-level	24.44	33.67
Multi-turn	25.04	34.35
Multi-turn sp	25.42	37.91
Single-turn + ICL	22.47	-
Segment-level + ICL	24.92	36.61
Multi-turn + ICL	25.49	38.03
Multi-turn sp + ICL	**25.82**	**38.75**

Table 2: Results on WMT-23 Zh-En direction evaluated with dBLEU and BlonDE.

3.3 Evaluation Metrics

We use the COMET-22 default model (Rei et al., 2022) and sacreBLEU (Post, 2018) implementation of BLEU (Papineni et al., 2002) to evaluate translation quality. For COMET, due to its context length limit, we evaluate each segment independently and report average scores; for BLEU, we consider n-gram matches at the document level (Liu et al., 2020). Additionally, we use BlonDE (Jiang et al., 2022) to evaluate document-level translation, which specifically measures the correctness of features that are known to benefit from wider context in Chinese-to-English translation, such as tense correctness, pronouns, transliteration, entities, and connectives.

3.4 Results

We report our main results in Table 1. It shows average scores of document-level BLEU and COMET-22 (Rei et al., 2022) across all directions in the WMT-24 general track. Following the setting of the WMT-24 shared task Kocmi et al. (2024), we also report results with in-context learning (ICL) where we provide the same prompt with 3 exemplars as the WMT-24 shared task did. Detailed results for each language direction are in Appendix Table 7 and Table 8.

The results show that all four multi-turn conversational methods achieve better translation performance compared to both segment-level methods and single-turn methods across both metrics. The setting of multi-turn with first providing the whole source document (Multi-turn sp) achieves the best results in all cases, demonstrating the advantage of our proposed variant.

Additionally, we report WMT-23 Zh-En results using Llama-3.1-8B-Instruct (Table 2). The document-level metric BlonDE shows a clearer performance gap between multi-turn and segment-level translation compared to dBLEU.

We also performed a significance test in Appendix A.2 by comparing Multi-turn sp with other settings. We find that the improvement of the Source-primed Multi-turn approach over the Segment-level baseline in terms of spBLEU is statistically significant at the .05 level for most settings. We observe the same when comparing Source-primed Multi-turn to a Multi-turn approach without source priming.

Domains / Settings	Seg-level ICL	Multi-turn ICL
Literary	21.52	22.22 (+0.70)
News	24.59	25.57 (+0.98)
Social	22.07	20.11 (-1.96)
Speech	24.03	24.25 (+0.22)
Personal	20.37	20.09 (-0.28)
Education	26.11	30.27 (+4.16)
Voice	22.45	22.70 (+0.25)
Official	24.95	25.14 (+0.19)

Table 3: Average dBLEU score across all language directions in different domains on WMT-24.

4 Analysis

4.1 Results for different domains

Table 3 shows Llama-3.1-8B-Instruct results on WMT-24 across domains and language pairs. Multi-turn conversation outperforms single-turn and segment-level translation in Literary, Education, and News domains. The biggest gain (+4.16 dBLEU) occurs in the educational domain (Czech-Ukrainian only), which contains short elementary-school exercises requiring context for interpretation.

4.2 Omission Errors in Long Documents

We visualize different strategies' results on long documents in Figure 2, reporting reference and hypothesis token counts for the top N longest documents. For the longest 10 documents, single-turn translation shows a clear gap between reference and hypothesis lengths, indicating omissions. Both segment-level and multi-turn translation help mitigate this problem.

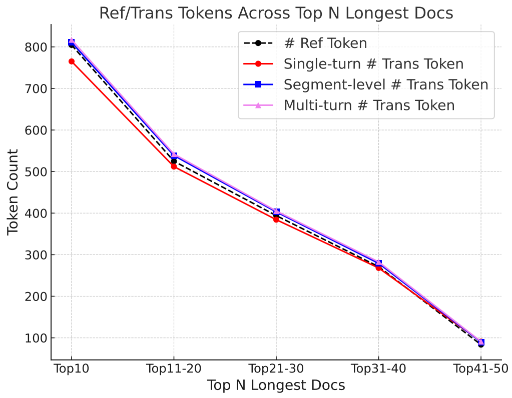

Figure 2: Token number across top N longest docs

4.3 Comparison with Open Submissions of WMT-24 General Task

We evaluate representative open submissions of current state-of-the-art models for WMT-24 general tasks using COMET and dBLEU, and report the results in Appendix Table 4. We used the same prompts as Kocmi et al. (2024). These results primarily demonstrate that we used competitive models, even though we chose smaller models for efficiency reasons.

5 Related Work

5.1 Large Language Models For Translation

LLMs have become widely used in a series of language tasks including machine translation (Zhu et al., 2024). There are various methods to improve the translation performance of LLMs, including prompting, in-context learning techniques (Chen et al., 2024; Lu et al., 2024), and creating instruction tuning data for machine translation to enhance their capabilities (Alves et al., 2024).

5.2 Document Level and Context Aware Machine Translations with LLMs

There are works focusing on using LLMs for document-level machine translation (Wang et al., 2023; Cui et al., 2024; Mohammed and Niculae, 2024; Wu et al., 2024). Jin et al. (2024) uses LLMs to handle chapter-level translation and construct related datasets. Kudo et al. (2024) decodes multiple sentences at each step and selects the most probable one sequentially, which also leverages context information. Luo et al. (2024) uses context information to help document translation, but differs from our multi-turn setting in that their prefix exemplars are not fixed, preventing full KV-cache reuse. Wang et al. (2023) first explores a multi-turn setting similar to ours but focuses on comparing different numbers of sentences in each turn. Karpinska and Iyyer (2023) compares paragraph-level translation and sentence-level translation for literary translation and finds paragraph-level translation to be superior.

5.3 Chat Translation

In the WMT24 Chat shared task (Mohammed et al., 2024), multiple submissions explore segment-in-context prompts (Pombal et al., 2024), with a mix of full document context and sliding window approaches (Yang et al., 2024; Sung et al., 2024). While the latter only requires O(n) tokens to be processed, we aim for full document-level access with our multi-turn framework.

6 Conclusion

Source-primed multi-turn translation: 1) gives structure to the translation task, reducing omission errors; 2) provides access to the full source document context at the beginning compared to original multi-turn translation, enabling LLMs to improve coherence and other document-level aspects of translation; and 3) allows for caching attention keys and values from previous turns, eliminating the efficiency bottleneck of previous segment-in-context prompting strategies.

7 Limitations

In this paper, we only use automatic metrics for evaluating document level machine translation due to lack of resources. Our evaluation includes a document level metric (BlonDE), but human evaluation remains important future work.

Our efficiency argument is a theoretical one based on the ability to re-use prefixes in our multi-turn strategy, compared to other segment-in-context prompts which require re-processing the full prompt. While we did not implement KV caching ourselves, we note that current GPT APIs implement prompt caching, with discounts of 90% compared to uncached inputs for GPT-5.[3]

Acknowledgments

This work was funded by the Swiss National Science Foundation (project InvestigaDiff; no. 10000503).

References

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 181–190, Sheffield, UK. European Association for Machine Translation (EAMT).

Menglong Cui, Jiangcun Du, Shaolin Zhu, and Deyi Xiong. 2024. Efficiently exploring large language models for document-level machine translation with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10885–10897, Bangkok, Thailand. Association for Computational Linguistics.

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

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.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Linghao Jin, Li An, and Xuezhe Ma. 2024. Towards chapter-to-chapter context-aware literary translation via large language models. *Preprint*, arXiv:2407.08978.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz,

[3] https://platform.openai.com/docs/guides/prompt-caching, accessed on 16.09.2025.

Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Keito Kudo, Hiroyuki Deguchi, Makoto Morishita, Ryo Fujii, Takumi Ito, Shintaro Ozaki, Koki Natsumi, Kai Sato, Kazuki Yano, Ryosuke Takahashi, Subaru Kimura, Tomomasa Hara, Yusuke Sakai, and Jun Suzuki. 2024. Document-level translation with LLM reranking: Team-J at WMT 2024 general translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 210–226, Miami, Florida, USA. Association for Computational Linguistics.

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.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024. Chain-of-dictionary prompting elicits translation in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, Florida, USA. Association for Computational Linguistics.

Yuanchang Luo, Jiaxin Guo, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhiqiang Rao, Shaojun Li, Jinlong Yang, and Hao Yang. 2024. Context-aware and style-related incremental decoding framework for discourse-level literary translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 973–979, Miami, Florida, USA. Association for Computational Linguistics.

Wafaa Mohammed, Sweta Agrawal, Amin Farajian, Vera Cabarrão, Bryan Eikema, Ana C Farinha, and José G. C. De Souza. 2024. Findings of the WMT 2024 shared task on chat translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 701–714, Miami, Florida, USA. Association for Computational Linguistics.

Wafaa Mohammed and Vlad Niculae. 2024. Analyzing context utilization of llms in document-level translation. *Preprint*, arXiv:2410.14391.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

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.

Jeongeun Park, Sungjoon Choi, and Sangdoo Yun. 2024. Versatile motion language models for multi-turn interactive agents. *Preprint*, arXiv:2410.05628.

Jose Pombal, Sweta Agrawal, and André Martins. 2024. Improving context usage for translating bilingual customer support chat with large language models. In *Proceedings of the Ninth Conference on Machine Translation*, pages 993–1003, Miami, Florida, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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.

Mingi Sung, Seungmin Lee, Jiwon Kim, and Sejoon Kim. 2024. Context-aware LLM translation system using conversation summarization and dialogue history. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1011–1015, Miami, Florida, USA. 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.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang

Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Xinye Yang, Yida Mu, Kalina Bontcheva, and Xingyi Song. 2024. Optimising LLM-driven machine translation with context-aware sliding windows. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1004–1010, Miami, Florida, USA. Association for Computational Linguistics.

Maria Zafar, Antonio Castaldo, Prashanth Nayak, Rejwanul Haque, and Andy Way. 2024. The SETU-ADAPT submissions to WMT 2024 chat translation tasks. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1023–1030, 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. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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 Appendix

A.1 Comparison With Submissions of WMT-24 shared task

Below are the results of submissions from representative systems on WMT-24 general tasks and the results of multi-turn and segment-level settings in our experiments, evaluated using dBLEU (document-level) and COMET (paragraph-level).

Systems / Metrics	dBLEU	COMET
Claude-3.5	**33.34**	85.01
GPT-4o-mini (sp MTurn ICL)	32.73	84.42
GPT-4o-mini (Seg ICL)	32.37	84.08
Gemini-1.5-Pro	30.76	83.67
GPT-4	30.43	84.05
CommandR-plus	29.51	82.50
Unbabel-Tower70B	28.81	**86.46**
Aya23	28.07	79.85
Llama3-70B	27.30	81.32
Mistral-Large	26.78	80.65

Table 4: Results of representative systems submissions of WMT24.

A.2 Significance Test

We performed a significance test based on spBLEU using flores101's (Goyal et al., 2022) tokenizer, and used the bootstrap resampling method (Koehn, 2004) with 1000 resamples. For simplicity, the significance test is performed on a concatenation of the test sets across all language pairs. We use sacrebleu (Post, 2018) to implement this experiment. The results are shown in Table 6, and most of the comparisons are significant at the .05 level.

A.3 Discussion of Non-General LLMs

We also conducted experiments using the machine translation enhanced LLM Tower-Instruct-7B. We report the detailed results of this model in Table 7 and Table 8. The results indicate that using the multi-turn method with Tower-Instruct-7B is not effective compared to using segment-level translation, which might be due to TowerInstruct having been mostly optimized on single-turn machine translation data (75% zero-shot data).

A.4 Qualitative Results

In the two examples in Table 5 generated by Llama-3.1-8B-Instruct, the strength of the multi-turn ICL setting lies in its accurate use of terminology, natural tone, and coherent sentence structure. In Example 1, it preserves the nuance of "splurging" omitted by single-turn ICL setting, and correctly uses the singular classifier “这副” instead of the awkward plural “这些” used by segment-level ICL. In Example 2, it avoids the referential ambiguity of segment-level ICL, and its omission of ‘‘眼镜” by explicitly repeating ‘‘这副眼镜,” maintaining clarity and alignment with the source. Compared to single-turn ICL and segment-level ICL, which suffer from omission, ambiguity, and literal translation artifacts, multi-turn ICL produces more faithful and fluent outputs.

A.5 Dataset License

In this paper, we use the WMT-24 and WMT-23 test sets, which comply with their respective license rules and can be freely used for research purposes.

A.6 AI Assistance

This work is done with AI assistance, we use Cursor[4] and Claude[5] for writing and coding assistance.

[4] https://www.cursor.com/

[5] https://claude.ai/

	Example 1	Example 2
Source (EN)	I’m splurging on a new set of frames, these red ones I *reeeeally* like.	...I made a decision to get another pair, and that pair are these.
Single_icl (ZH)	我在这里买了一个新的框架，这些红框我特别喜欢。	我决定再买一副眼镜，这副眼镜就是这些。
Seg_icl (ZH)	我在大手大脚地买了一套新框架，这些红色的我真的很喜欢。	我决定再买一双，而这双就是这些。
Mturn_icl (ZH)	我在大手大脚地买了一副新框架，这副红色的我特别喜欢。	我决定再买一副眼镜，这副就是这副。
Analysis	**Single_icl** has omission error where lose the information about "splurging". **Seg_icl** literally carries over “这些”（“these”), which sounds plural and unnatural for eyeglasses in Chinese. **Mturn_icl** infers that “these red ones” refers to a single pair of glasses and therefore uses the classifier “这副,” resulting in a more natural and context-aware translation.	**Seg_icl** again mixes “这双/这些,” introducing plural ambiguity and omitting the noun “眼镜.” **Mturn_icl** explicitly adds “眼镜” and repeats “这副” to mirror the demonstrative emphasis of the English（“that pair are these”), preserving the performative tone and resolving the reference clearly.

Table 5: Qualitative comparison of Single-turn ICL (**Single_icl**), segment-level ICL (**Seg_icl**) and multi-turn ICL (**Mturn_icl**) on Llama-3.1-8B-Instruct.

Model / Settings	**P-Value based on spBLEU**		
	Mturn icl sp vs. Mturn icl	**Mturn icl sp vs. Seg icl**	**Mturn sp vs. Mturn**
GPT-4o-mini	**0.010**	**0.001**	**0.256**
Qwen2.5-7B-Instruct	0.001	**0.001**	**0.008**
Llama-3.1-8B-Instruct	**0.004**	**0.050**	**0.109**

Table 6: Significant test on the spBLEU using FLORES101’s tokenizer. Bold values indicate cases where the left setting performs significantly better than the right setting; non-bold values indicate the opposite.

Model	Setup	ja-zh	cs-uk	en-de	en-zh	en-es	en-hi	en-ru	en-is	en-ja	en-cs	en-uk
TowerInstruct-7B-v0.2	Single-turn	12.52	4.15	27.90	28.39	34.59	1.54	19.45	1.37	5.56	8.89	8.47
	Segment-level	21.87	7.28	30.85	37.01	39.94	4.87	22.46	4.53	10.22	14.30	9.26
	Multi-turn	22.03	8.13	29.96	36.34	40.04	3.15	21.17	2.69	7.20	13.06	7.47
	Single-turn + ICL	12.98	10.90	24.27	25.45	30.11	1.71	17.57	0.89	5.35	7.96	10.93
	Segment-level + ICL	21.79	11.40	30.37	37.08	40.71	5.54	21.62	4.50	14.80	13.23	13.03
	Multi-turn + ICL	22.36	9.69	27.67	35.90	37.18	4.03	21.04	3.04	13.31	12.88	12.37
Llama-3.1-8B-Instruct	Single-turn	11.94	23.23	27.79	31.06	38.90	18.37	17.41	6.09	15.89	19.99	16.84
	Segment-level	11.47	21.30	27.24	33.26	39.12	19.59	19.37	6.80	21.53	20.64	17.89
	Multi-turn	17.32	23.51	27.94	34.87	39.69	20.50	19.70	7.85	21.90	19.83	18.22
	Multi-turn sp	22.78	24.06	26.72	32.21	39.75	20.63	20.21	8.72	21.17	20.02	19.18
	Single-turn + ICL	20.06	23.33	28.14	32.15	39.12	17.63	20.20	6.38	14.58	19.89	18.16
	Segment-level + ICL	23.02	23.02	27.79	36.80	39.97	20.03	19.65	8.54	15.76	20.80	18.40
	Multi-turn + ICL	23.34	23.26	28.08	37.18	40.39	20.48	19.95	6.68	20.04	20.02	18.70
	Multi-turn sp + ICL	26.49	24.10	26.94	37.41	40.13	20.96	19.74	8.93	22.31	20.80	18.67
Qwen-2.5-7B-Instruct	Single-turn	30.10	14.90	25.80	41.02	37.98	7.71	16.69	4.53	21.46	15.57	13.33
	Segment-level	27.85	9.58	21.98	36.65	33.43	10.34	15.39	6.43	16.80	14.49	13.01
	Multi-turn	28.37	13.83	24.46	37.92	36.75	10.81	16.25	7.14	18.55	16.07	13.91
	Multi-turn sp	30.31	16.81	25.91	41.66	37.84	10.83	14.91	4.25	22.61	17.28	13.62
	Single-turn + ICL	29.24	16.46	25.56	41.22	36.96	8.12	16.24	4.11	20.50	15.37	13.51
	Segment-level + ICL	30.06	15.07	23.54	38.10	35.45	10.42	16.21	7.01	20.91	15.10	13.82
	Multi-turn + ICL	30.50	16.30	24.43	39.57	37.16	10.69	16.69	7.08	20.47	16.16	14.48
	Multi-turn sp + ICL	30.51	16.15	24.12	41.43	37.65	10.95	14.53	4.64	22.21	17.27	12.68
GPT-4o-mini	Single-turn	33.99	32.48	34.10	45.11	45.75	26.79	23.85	20.91	29.55	29.94	29.28
	Segment-level	33.52	33.17	34.31	44.35	45.65	26.64	23.59	21.12	29.17	29.05	29.48
	Multi-turn	34.51	33.31	34.60	44.59	46.39	19.68	24.25	22.04	29.98	29.98	29.82
	Multi-turn sp	34.60	33.41	34.71	45.06	45.84	27.19	24.65	21.94	29.93	30.13	30.20
	Single-turn + ICL	33.65	32.19	35.09	45.14	45.80	26.93	22.78	21.28	29.04	30.10	29.44
	Segment-level + ICL	34.04	32.89	34.87	44.51	46.29	27.42	24.16	22.04	29.91	29.85	30.11
	Multi-turn + ICL	34.45	33.23	34.82	44.86	46.28	27.30	24.16	22.18	30.07	29.96	30.09
	Multi-turn sp + ICL	34.49	33.37	34.94	44.88	45.93	27.51	25.03	22.36	30.40	30.74	30.41

Table 7: Comparison of dBLEU results for WMT24 across different setups and language pairs for TowerInstruct-7B-v0.2, Llama-3.1-8B-Instruct, Qwen-2.5-7B-Instruct and GPT-4o-mini.

Model	Setup	ja-zh	cs-uk	en-de	en-zh	en-es	en-hi	en-ru	en-cs	en-is	en-ja	en-uk
TowerInstruct-7B-v0.2	Segment-level	67.04	81.11	82.67	48.04	48.46	74.42	80.84	77.74	82.67	75.70	77.62
	Multi-turn	64.76	80.38	82.19	45.17	42.19	70.72	78.93	73.44	81.43	75.90	76.79
	Segment-level + ICL	66.02	80.77	82.36	46.55	43.09	76.30	80.32	72.57	82.53	75.54	78.07
	Multi-turn + ICL	64.01	75.43	77.51	42.45	37.80	73.11	79.61	68.00	81.29	73.52	76.19
Llama-3.1-8B-Instruct	Segment-level	75.82	80.73	78.10	79.41	80.31	71.39	77.36	78.45	59.48	81.66	77.71
	Multi-turn	79.84	83.32	79.58	82.03	81.52	72.94	78.64	79.82	61.92	83.61	79.12
	Multi-turn sp	80.84	84.32	80.07	82.84	82.05	73.47	79.77	80.46	61.79	83.55	79.28
	Segment-level + ICL	79.70	82.08	78.95	82.39	81.56	73.04	79.28	78.93	62.68	83.51	80.25
	Multi-turn + ICL	81.55	83.70	79.70	82.94	82.23	73.79	79.46	79.90	62.10	84.26	79.61
	Multi-turn sp + ICL	81.42	84.28	80.26	83.03	82.31	73.55	79.69	80.51	62.79	84.27	80.13
Qwen-2.5-7B-Instruct	Segment-level	80.84	68.21	75.58	82.25	77.79	55.79	73.66	70.65	45.58	76.83	66.76
	Multi-turn	81.93	72.15	77.75	83.57	80.39	58.84	74.27	72.93	47.55	78.44	68.22
	Multi-turn sp	83.00	75.07	78.96	84.32	81.42	61.35	76.45	74.78	48.25	82.59	70.33
	Segment-level + ICL	82.39	72.39	76.53	82.99	79.26	58.08	73.52	71.36	46.91	80.41	67.89
	Multi-turn + ICL	83.07	73.69	77.89	83.82	80.61	58.99	73.61	72.89	47.77	80.63	69.61
	Multi-turn sp + ICL	83.16	74.30	78.64	84.10	81.18	61.42	75.74	74.15	48.10	81.85	69.82
GPT-4o-mini	Segment-level	83.60	88.83	82.59	84.58	82.99	76.78	81.65	84.18	78.10	86.00	84.13
	Multi-turn	84.74	89.24	83.11	85.23	84.59	78.44	83.36	86.05	79.63	87.20	85.75
	Multi-turn sp	84.80	89.24	83.06	85.35	84.52	78.40	83.38	86.10	79.23	87.51	85.63
	Segment-level + ICL	84.29	89.00	83.00	84.84	84.33	78.49	83.00	85.78	79.46	86.92	85.73
	Multi-turn + ICL	84.77	89.23	83.14	85.32	84.57	78.46	83.23	86.03	79.54	87.22	85.78
	Multi-turn sp + ICL	84.73	89.26	83.24	85.36	84.45	78.50	83.47	86.32	79.79	87.72	85.80

Table 8: Comparison of COMET score for WMT-24 across different setups and language pairs for Tower-7B-Instruct, Llama-3.1-8B-Instruct, Qwen-2.5-7B-Instruct, and GPT-4o-mini.