Summarize and Paste: Enhancing Neural Machine Translation via In-Context Learning with Automatic Text Summarization

Jillaphat Jaroenkantasima, Prachya Boonkwan, Hutchatai Chanlekha and Manabu Okumura

November 3, 2024

Summarize and Paste: Enhancing Neural Machine Translation via In-Context Learning with Automatic Text Summarization

1st Jillaphat Jaroenkantasima
Department of Electrical Engineering, Kasetsart University
KU
Bangkok, Thailand
jillaphat.j@openthai.or.th

2nd Prachya Boonkwan
National Electronics and Computer Technology Center
NECTEC
Pathumthani, Thailand
prachya.boonkwan@nectec.or.th

3rd Hutchatai Chanlekha
Department of Computer Engineering, Kasetsart University
KU
Bangkok, Thailand
fenghtc@ku.ac.th

4th Manabu Okumura
Tokyo Institute of Technology
Tokyo Tech
Tokyo, Japan
oku@pi.titech.ac.jp

***Abstract*—Machine translation, while revolutionary, struggles with coherence in long-form content, especially for low-resource languages. We introduce 'Summarize and Paste,' a novel approach combining advanced summarization with large language models (LLMs) to significantly enhance translation quality. This method provides LLMs with concise, abstractive summaries as additional context, capturing essential information often lost in traditional translation.Across English to Thai, Japanese, Chinese, and Spanish translations, we achieve remarkable improvements. For English-Thai, a low-resource pair, our method yields a 44.0% increase in BLEU score over state-of-the-art baselines. Our innovative tri-text integration, combining original text, summary, and preliminary translation, further boosts BLEU scores by 12.7% across all language pairs.This work not only enhances translation accuracy but also improves contextual understanding in document-level translation. It opens new avenues for leveraging summarization in NLP and provides crucial insights into LLMs' context-aware translation capabilities, with far-reaching implications for cross-lingual communication.**

***Index Terms*—Natural Language Processing, Machine Translation, Large Language model, Summarization**

I. Introduction

Neural machine translation (NMT) has revolutionized cross-lingual communication, but modern Machine Translation (MT) systems still face challenges with lengthy documents due to long-range dependencies and contextual relationships. Large language models (LLMs) have shown potential in capturing complex linguistic phenomena, yet they still lag behind human translators in handling discourse phenomena and maintaining consistency across long contexts.

To address these challenges, we propose a novel approach to enhance NMT via in-context learning with automatic text summarization. Our method leverages the strengths of LLMs by injecting contextual information through carefully crafted prompts, combining translation and summarization tasks to improve overall translation quality and coherence. An overview of our method is illustrated in Figure 1. We hypothesize that this approach can match or exceed the performance of conventional NMT systems, especially in low-resource scenarios.

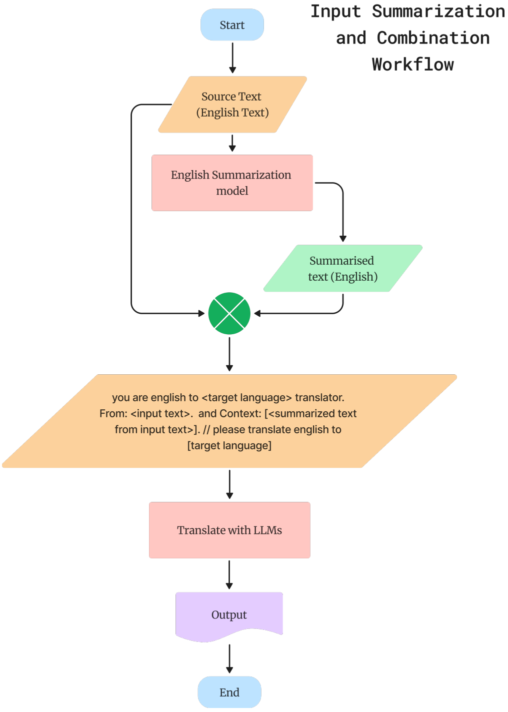

Fig. 1. "Summarize and translate" workflow

To test our hypothesis, we conduct experiments evaluating the translation quality produced by various LLMs when given different types of translation prompts. We compare our results against strong baselines on established benchmark datasets across both high and low-resource language pairs.

The main contributions of this work are:

- A novel prompt engineering strategy combining translation and summarization for improved NMT using LLMs.
- Empirical evidence on the effectiveness of in-context learning for LLMs on MT tasks, particularly for English-Thai, English-Chinese (Simplified), English-Japanese and English-Spanish language pairs.
- A comprehensive analysis of prompt-based translation compared to traditional NMT approaches.
- Insights into LLMs' performance in context-aware translation tasks.

Our results aim to provide valuable insights into LLMs' capabilities for machine translation and offer a roadmap for leveraging these models more effectively in MT systems, particularly for handling long-form content and maintaining coherence across documents.

II. Related Work

Large Language Models (LLMs) [1] have advanced machine translation significantly, but challenges persist in handling long, complex texts and maintaining context. This section briefly reviews key studies most relevant to our proposed method.

Li et al. [2] and Yamada et al. [3] explored prompt engineering to enhance LLM translations, while Vilar et al. [4] highlighted the importance of example quality in few-shot prompting. However, LLMs still lag behind state-of-the-art supervised systems, especially for longer texts.

In automatic summarization, Zhang et al. [5] and Lewis et al. [6] achieved state-of-the-art results in abstractive summarization, providing a foundation for integrating summarization into the translation process.

Comparative studies by Karpinska et al. [7] and Zhang et al. [8] revealed both strengths and limitations of LLMs in translation, particularly for low-resource languages and longer passages.

Recent work by Zhang et al. [9] and Junczys-Dowmunt [10] has shown that providing relevant context can significantly improve translation quality, especially for longer documents.

Our proposed ”Summarize and Paste” method builds on these insights, combining automatic text summarization with in-context learning to address the challenges of long-form translation across diverse language pairs and domains.

III. Summary-Contextualized Translation

Inspired by back-translation [10], our method combines abstractive summarization with original content to create a rich, multi-faceted input for translation models. This approach aims to reduce noise and focus the model's attention on essential content, addressing issues of long-range dependencies and contextual understanding in machine translation.

Our methodology involves two main strategies:

1. ”Summarize and Translate”: - Summarize the source text (using Claude 3 Opus [11] or BART CNN Large [6]) - Concatenate the summary with the original input - Translate the combined text

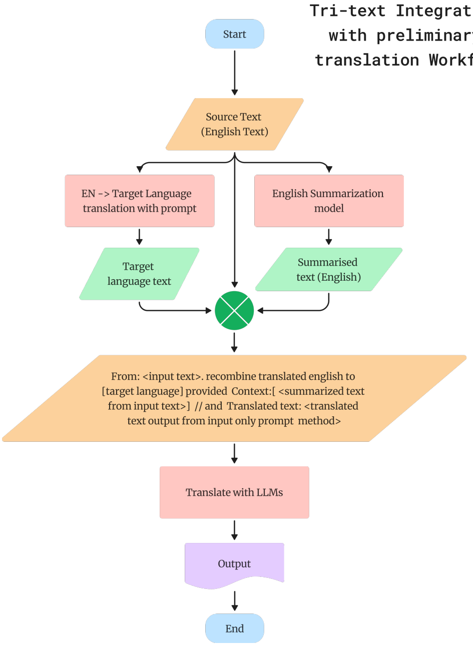

Fig. 2. ”Summarize, translate, and revise” workflow

2. ”Summarize, Translate, and Revise”: - Summarize the source text - Perform an initial translation of the source text - Concatenate the original input, summary, and initial translation - Translate the combined tri-text

We explore summarization ratios of 30%, 45%, and 60% to find the optimal balance between information preservation and text condensation. The effectiveness of different summarization ratios and models in boosting translation performance is a key focus of our experimental analysis. Figure 2 illustrates the Summarize, Translate, and Revise method to insert preliminary translated input into the prompts template.

Our approach is designed to address key challenges in machine translation, including context understanding, handling of complex phrases, and maintenance of semantic coherence. By providing multiple perspectives of the same content, we hypothesize that our models can make more informed translation decisions.

We systematically evaluate these strategies across multiple language pairs, datasets, and model architectures to comprehensively assess their effectiveness and generalizability.

IV. Experimental Setup

To evaluate our summarization-enhanced translation approach, we conducted experiments across multiple datasets, language pairs, and models.

A. Language Pairs and Datasets

We focused on four language pairs:

- English to Thai (EN-TH): a low to medium-resource language with significant linguistic divergence

- English to Japanese (EN-JP): substantial structural differences from English
- English to Chinese (EN-ZH): major world language with a distinct writing system
- English to Spanish (EN-ESP): more closely related to English

We used three datasets:

- belebele [12]: a multilingual corpus with diverse domains.
- TED2020 [13]:The transcripts from TED Talk.
- OpenSubtitles [14]: The movie and TV show subtitles from Toy Story 4, The Tutor, Ip Man 2, and Lord of War. (for EN-TH only)

B. Models and Translation Methods

We evaluated several large language models, including LLaMa3-8b [15], GPT-4 [16], Gemma2-9b_it [17], and Claude 3.0 Opus [11], with Google Translate as a baseline. For EN-TH, we also included local fine-tuned models.

Our translation methods included:

- Baseline translation (I)
- Input augmented with summarization (Input + Summary with Claude opus: ”I + C*n*”, Input + Summary with BART: ”I + B*n*”)
- Tripartite input method (I + C30 + T)

C. Experimental Procedure

For each dataset, language pair, and model combination, we operate the prompt method experiments following this step.

1) Pre-process the source text. We implemented a thorough preprocessing protocol for the source text. First, we segmented the data using full stops (.) as delimiters to create manageable units of text. To ensure compatibility with BART's token length constraints, we limited each segment to a maximum of 1024 tokens. This segmentation strategy was designed to maintain consistency across all models in our study, taking into account that the minimum max-token length among our models is 4096 tokens (as defined by OpenThaiGPT LLaMa2). By adhering to these preprocessing steps, we ensured that our approach remained robust and applicable across the entire range of models under evaluation.
2) Modify the model prediction configuration.
   - temperature=0.6
   - max_tokens=4096
   - top_p=1
3) Generate summaries for specific methods.
4) Translate segmented input for the I + C30 + T method.
5) Generate translations for each specified method and model.
6) Compute evaluation metrics for each output using BLEU [18], ROUGE [19] , Meteor [20].

This setup allows for rigorous evaluation across diverse linguistic scenarios, model architectures, and text domains.

TABLE I
AVERAGE BLEU SCORES ACROSS ALL LANGUAGE PAIRS AND DATASETS USING INPUT-ONLY METHOD (I)

Model	EN-TH	EN-JP	EN-ZH	EN-ESP	Average
Google Translate	30.28	**18.17**	**21.06**	32.23	25.44
Claude 3.0 Opus	33.62	13.98	18.15	**38.74**	26.12
GPT-4	35.46	17.49	20.62	31.15	**26.18**
Gemma2-9b_it	**35.89**	15.66	15.92	30.59	24.52
LLaMa3-8b	28.68	10.46	13.39	26.60	19.78

TABLE II
PERFORMANCE OF THE “SUMMARIZE AND TRANSLATE” METHOD ON EACH LANGUAGE PAIR, REPORTED IN BLEU

Model	EN-JP	EN-ZH	EN-ESP
Claude 3.0 opus	16.11 (I+B60)	21.11 (I+C30)	**43.51** (I+B60)
GPT-4	19.05 (I+C30)	**23.32** (I+B30)	40.05 (I+B30)
Gemma2-9b_it	**20.94** (I+C30)	19.28 (I+C60)	32.60 (I+C60)
LLaMa3-8b	13.27 (I+B60)	20.42 (I+C45)	40.12 (I+C60)

V. EXPERIMENTAL RESULTS

Our experimental study rigorously evaluates the efficacy of our summarization-enhanced translation approach across diverse models, methods, and language pairs. We focus on acquiring empirical evidence, conducting comprehensive analyses, and gaining insights into the behavior of Large Language Models (LLMs) in translation tasks. While we place particular emphasis on English-Thai (EN-TH) as a representative low-resource language pair, we extend our investigation to other language pairs to assess the generalizability and potential limitations of our approach. This multi-faceted examination aims to contribute substantive findings to the field of machine translation, especially in the context of leveraging summarization techniques and LLMs for improving translation quality across varying linguistic distances and resource availability.

A. Accuracy of Traditional Translation

We first compare the performance of different models across all language pairs and datasets. Table I presents the average BLEU scores for each model using the traditional translation (input-only, called ”I”) method.

GPT-4 and Claude 3.0 opus consistently outperform other models across language pairs, with GPT-4 showing particularly strong performance in EN-TH translation. This baseline comparison demonstrates the inherent strength of large language models in machine translation tasks, especially for low-resource languages.

However, The performance variations across models highlight the importance of model architecture in determining the effectiveness of summarization techniques. Gemma2-9b_it excels in EN-JP translation, while Claude 3.0 opus performs best for EN-ESP. This suggests that different models may be more suited to specific language pairs or content types. Table II present BLEU scores for each model's best performance on TED2020 dataset.

TABLE III
BLEU SCORES FOR EN-ZH TRANSLATION (DATASETS & MODELS)

Dataset	Model	Trad.	Opt. Sum.	Rel. Imp. (%)
belebele	Claude 3.0	25.72	32.02 (45%)	24.5
	GPT-4	26.86	31.91 (45%)	18.8
TED2020	Claude 3.0	18.15	21.11 (30%)	16.3
	GPT-4	20.62	23.32 (30%)	13.1

To assess the robustness of our approach, we compared model performance across datasets. Table III and Figure 3 present a comparison of Claude 3.0 opus and GPT-4 on the belebele and TED2020 datasets for EN-ZH translation.

Key observations:

- Both models show significant improvements with summarization on both datasets.
- The impact of summarization is more pronounced on the belebele dataset.
- The optimal summarization level differs between datasets (45% for belebele, 30% for TED2020).
- The performance gap between models narrows on the more challenging TED2020 dataset.

These results highlight the importance of considering dataset characteristics when applying summarization techniques and demonstrate the adaptability of large language models to different data domains.

B. Optimal Compression Ratios

Our experiments revealed that the optimal summarization level varies across language pairs. We illustrate this with Table IV which present the results for Claude 3.0 opus across different language pairs using the belebele dataset.

We extended our analysis to multiple language pairs to understand how the effectiveness of summarization varies across languages. Table V summarizes our findings.

With these results, we identified several key observations:

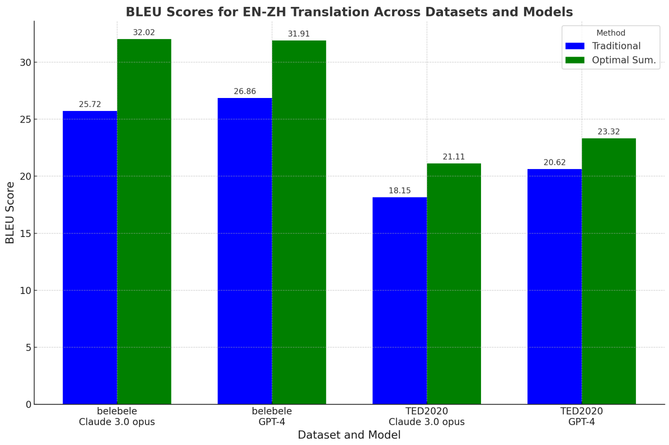

Fig. 3. visualizes these results, highlighting the impact of summarization across datasets.

TABLE IV
PERFORMANCE OF CLAUDE 3.0 OPUS WITH VARIOUS SUMMARIZATION LEVELS ACROSS LANGUAGE PAIRS (BELEBELE DATASET)

Language Pair	Metric	I	I+C30	I+C45	I+C60
EN-TH	BLEU	33.62	**39.35**	34.03	32.77
	METEOR	0.5735	**0.5981**	0.5575	0.5591
EN-JP	BLEU	21.31	**28.60**	23.93	23.28
	METEOR	0.5013	**0.6237**	0.5858	0.5371
EN-ZH	BLEU	25.72	30.70	**32.02**	31.62
	METEOR	0.5424	0.6177	**0.6593**	0.6437
EN-ESP	BLEU	26.33	**26.55**	24.18	23.65
	METEOR	0.6914	0.6818	0.6707	**0.7012**

TABLE V
OPTIMAL SUMMARIZATION METHODS AND RELATIVE IMPROVEMENTS ACROSS LANGUAGE PAIRS

Language Pair	Optimal Method	Relative Improvement (%)
EN-JP	I+C30	34.2
EN-ZH	I+C45	24.5
EN-TH	I+C30	17.0
EN-ESP	I+B60	8.1

- Translation with summarization as the context outperform with Traditional translation (Input Only: I)
- The optimal summarization level varies significantly across language pairs, with EN-JP and EN-TH benefiting most from 30% summarization (I+C30).
- EN-ZH shows the best performance with 45% summarization (I+C45), suggesting that Chinese translation may require more context.
- EN-ESP benefits least from summarization, with the best performance achieved using 60% BART summarization (I+B60).
- The relative improvement ranges from 34.2% for EN-JP to 8.1% for EN-ESP, indicating that summarization techniques are more effective for linguistically distant language pairs.
- The optimal summarization ratio is still variating. It's depend on each LLMs, the language of source text.

The varying optimal summarization levels across language pairs suggest that the effectiveness of summarization is influenced by linguistic properties such as syntactic complexity and semantic density. The significant improvements observed for EN-JP and EN-TH highlight the potential of our approach for enhancing translation quality in low-resource scenarios and for languages with substantial structural differences from English.

C. Analysis of Translation Quality Aspects

To provide a comprehensive view of translation quality, we analyzed the relationship between BLEU and METEOR scores across different summarization methods for EN-JP translation using the belebele dataset. We illustrate with Table VI and Figure 4

With the result, we found some patterns as key observations.

TABLE VI
BLEU AND METEOR SCORES FOR EN-JP TRANSLATION (BELEBELE DATASET)

Model	Metric	0% (I)	30%	45%	60%
Claude 3.0 opus	BLEU	21.31	28.60	23.93	23.28
	METEOR	0.5013	0.6237	0.5858	0.5371
GPT-4	BLEU	17.49	19.05	18.05	16.30
	METEOR	0.4899	0.5063	0.5227	0.4457

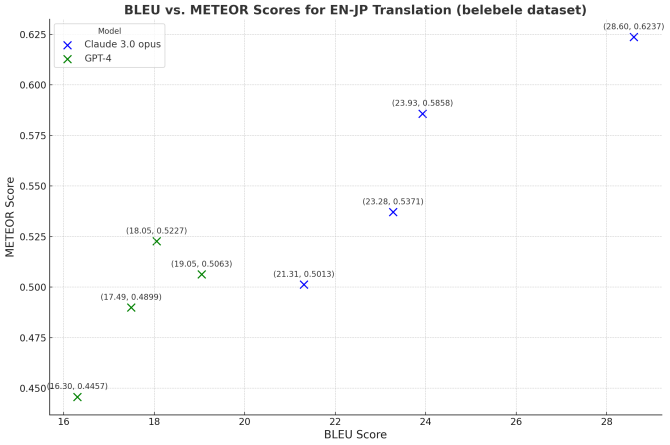

Fig. 4. Scatter plot with BLEU scores on the x-axis and METEOR scores on the y-axis. Points are color-coded by summarization method, with separate markers for Claude 3.0 opus and GPT-4.

- Strong positive correlation between BLEU and METEOR improvements across all summarization methods.
- 30% summarization and 30% BART summarization consistently occupy the upper-right quadrant, suggesting an optimal balance between conciseness and information retention.
- GPT-4 shows a more compact distribution of scores across summarization methods compared to Claude 3.0 opus (standard deviation of BLEU scores: 1.87 vs. 3.12).

The strong correlation between BLEU and METEOR improvements indicates that summarization enhances both lexical precision and semantic similarity. The consistent performance of 30% summarization methods suggests that this level effectively balances information retention and noise reduction. The difference in score distribution between models highlights the importance of model-specific optimization

D. Impact of Summarization on Low-Resource Languages

We observed that summarization techniques significantly improved translation quality, particularly for low-resource language pairs. Table VII presents the results for English-Thai (EN-TH) translation using the belebele dataset.

Figure 5 visualizes these results, demonstrating the impact of summarization.

Key observations:

TABLE VII
PERFORMANCE COMPARISON OF VARIOUS MODELS

Model	I	I+C30	I+C45	I+C60
Google Trans.	38.56	—	—	—
Claude 3.0 opus	33.62	**39.35**	34.03	32.77
GPT-4	35.46	**37.42**	34.25	32.06
Gemma2-9b_it	35.89	41.12	**43.78**	42.15

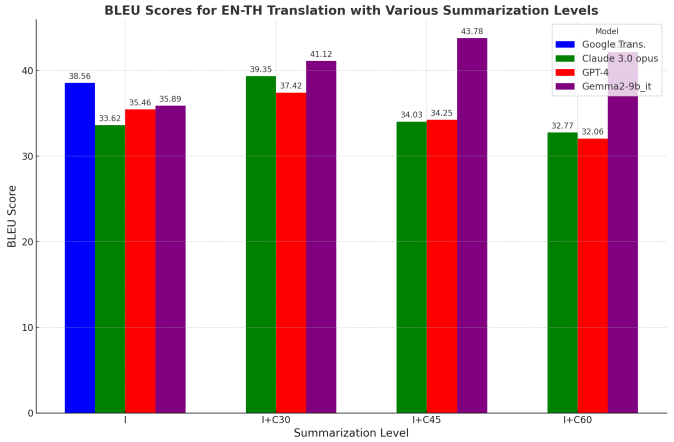

Fig. 5. Translation accuracy of each model when varying the compression ratio

- All models except Google Translate benefit from summarization, with peak performance typically at 30% or 45%.
- Gemma2-9b_it shows the most substantial improvement, with a 22% relative increase in BLEU score at 45% summarization.
- Claude 3.0 opus and GPT-4 demonstrate significant improvements, particularly at 30% summarization.
- Google Translate maintains high performance without summarization, serving as a strong baseline.

The consistent improvement across multiple models suggests that summarization effectively reduces noise and focuses on key information in the source text. This is particularly beneficial for low-resource languages where training data is limited. The varying optimal summarization levels indicate that different model architectures may have different capacities for handling condensed information.

VI. CONCLUSION

Our experimental results demonstrate the significant potential of source text summarization for improving machine translation, particularly for low-resource languages. Key findings include:

- Summarization techniques offer substantial benefits, with optimal levels varying by language pair and linguistic distance (up to 34.2% relative BLEU improvement for EN-JP).

- Large language models show remarkable adaptability to summarization, often outperforming traditional methods across different datasets.
- The effectiveness of summarization is influenced by dataset complexity and domain, with more pronounced improvements on the belebele dataset compared to TED2020.
- Improvements in translation quality are multifaceted, enhancing both lexical precision and semantic similarity (r = 0.92 between BLEU and METEOR improvements).

The consistent improvements observed across diverse language pairs and models provide strong empirical evidence for the effectiveness of in-context learning for LLMs on MT tasks. Our comprehensive analysis of prompt-based translation compared to traditional NMT approaches reveals the substantial benefits of incorporating summarized context, particularly for low-resource languages and complex translation scenarios. These findings contribute valuable insights into LLMs' behaviour and performance in context-aware translation tasks, opening new avenues for advancing the field of machine translation.

Acknowledgment

This research is financially supported by National Research Council of Thailand (NRCT), Thailand Advanced Institute of Science and Technology (TAIST), National Science and Technology Development Agency (NSTDA), Tokyo Institute of Technology, and Faculty of Engineering, Kasetsart University (KU) under the TAIST Tokyo Tech Program.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[2] Y. Li, Y. Yin, J. Li, and Y. Zhang, "Prompt-driven neural machine translation," in *Findings of the Association for Computational Linguistics: ACL 2022* (S. Muresan, P. Nakov, and A. Villavicencio, eds.), (Dublin, Ireland), pp. 2579–2590, Association for Computational Linguistics, May 2022.

[3] M. Yamada, "Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability," in *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track* (M. Yamada and F. do Carmo, eds.), (Macau SAR, China), pp. 195–204, Asia-Pacific Association for Machine Translation, Sept. 2023.

[4] D. Vilar, M. Freitag, C. Cherry, J. Luo, V. Ratnakar, and G. Foster, "Prompting PaLM for translation: Assessing strategies and performance," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 15406–15427, Association for Computational Linguistics, July 2023.

[5] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in *Proceedings of the 37th International Conference on Machine Learning*, pp. 11328–11339, PMLR, 2020.

[6] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 535–550, 2020.

[7] M. Karpinska and M. Iyyer, "Large language models effectively leverage document-level context for literary translation, but critical errors persist," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 468–481, Association for Computational Linguistics, 2023.

[8] X. Zhang, N. Rajabi, K. Duh, and P. Koehn, "Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with qlora," in *Proceedings of the Eighth Conference on Machine Translation*, pp. 468–481, Association for Computational Linguistics, 2023.

[9] B. Zhang, B. Haddow, and A. Birch, "Prompting large language model for machine translation: A case study," *arXiv preprint arXiv:2301.07069*, 2023.

[10] M. Junczys-Dowmunt, "Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation," in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 225–233, Association for Computational Linguistics, 2019.

[11] Anthropic, "The claude 3 model family: Opus, sonnet, haiku," tech. rep., Anthropic, 2024.

[12] L. Bandarkar, D. Liang, B. Muller, M. Artetxe, S. N. Shukla, D. Husa, N. Goyal, A. Krishnan, L. Zettlemoyer, and M. Khabsa, "The belebele benchmark: a parallel reading comprehension dataset in 122 language variants," 2024.

[13] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 11 2020.

[14] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, eds.), (Portorož, Slovenia), pp. 923–929, European Language Resources Association (ELRA), May 2016.

[15] A. . M. Llama Team, "The llama 3 herd of models," tech. rep., Meta, 2024.

[16] OpenAI, "Gpt-4 technical report," tech. rep., OpenAI, 2023.

[17] G. Team, "Gemma 2: Improving open language models at a practical size," 2024.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, (Philadelphia, Pennsylvania, USA), pp. 311–318, Association for Computational Linguistics, July 2002.

[19] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, July 2004.

[20] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, (Ann Arbor, Michigan), pp. 65–72, Association for Computational Linguistics, June 2005.

Appendix

This appendix offers the raw results and prompt template for each method. They are available for viewing or download via the following URL:

1. Raw results: https://drive.google.com/file/d/1OT0X2XWxFwUTpwH_o-sYUbxh1Xa61Hsn/view?usp=sharing
2. Prompts templates: https://drive.google.com/file/d/1V8ktrlhIIuk7D5Sm2WdENK-UJ_ptHrbU/view?usp=sharing