STAMP: STYLE TRANSFER THROUGH ADAPTER-BASED MEMORY AND PROMPT ENGINEERING FOR CLASSICAL CHINESE TRANSLATION

Chenxiao Yang (2023010573)
Institute for Interdisciplinary Information Sciences
Tsinghua University
cx-yang23@mails.tsinghua.edu.cn

Yingxi Lu (2023011435)
Institute for Interdisciplinary Information Sciences
Tsinghua University
lu-yx23@mails.tsinghua.edu.cn

Zhuo Lin (2023011443)
Institute for Interdisciplinary Information Sciences
Tsinghua University
lin-z23@mails.tsinghua.edu.cn

ABSTRACT

The rapid advancement of large pre-trained models has significantly enhanced classical Chinese translation task. However, translating from modern to Classical Chinese remains challenging due to the unique syntactic and stylistic features of Classical Chinese. This paper addresses these challenges by introducing a dual-component framework combining prompt engineering and memory-augmented adapters. Our approach leverages the generalization capabilities of pre-trained models while enabling precise stylistic control. We employ prompt-based techniques to guide the model and a memory-augmented adapter to maintain stylistic fidelity. Extensive experiments demonstrate significant improvements in translation quality and stylistic consistency, validating the effectiveness of our method.

1 INTRODUCTION

The rapid development of large pre-trained models like GPT-4 (Achiam et al. (2023)) and GLM (GLM et al. (2024b)) has significantly advanced NLP tasks by learning rich linguistic representations from vast and diverse data. This enables strong performance in tasks like text generation, machine translation, and sentiment analysis. However, the broad range of training data also results in models inheriting a complex mix of styles and linguistic features, complicating tasks that require precise stylistic control—such as translating from modern to Classical Chinese, where consistency and fidelity are crucial.

Classical Chinese, with its unique syntax, literary devices, and historical layers of meaning, poses distinct challenges for modern NLP models, which are typically trained on contemporary language data. Recent studies on Classical Chinese have focused on specialized tasks such as poetry generation (Yi et al. (2018)) and couplet creation (Song (2022)), often relying on models tailored to specific genres through techniques like template-based design and rule-based post-processing. However, these approaches are narrow in scope, excelling only within defined contexts. Their ability to generalize to broader tasks, such as cross-temporal translation or stylistic adaptation, remains limited, hindering their application to more diverse, cross-disciplinary tasks requiring flexible style control.

Given this challenge, a key question arises: how can we leverage the ability of pre-trained models to handle Classical Chinese translation tasks? Central to this challenge is the concept of pluggable components—adapting a pre-trained model to new tasks without altering its core structure. Techniques like fine-tuning, in-context learning (ICL) (Dong et al. (2022)), and LoRA (Hu et al. (2021)) enable efficient adaptation with minimal disruption to the model's original knowledge base. However, the primary bottleneck lies in striking a balance between preserving the model's general knowledge and ensuring robust performance on specialized tasks. The goal is to minimize interfer-

ence with foundational capabilities while enabling the model to excel in nuanced tasks like stylistic control in Classical Chinese translation.

To achieve this, we propose a dual-component structure. First, to preserve the integrity of the pre-trained model, we employ prompt-based techniques. By providing task-specific instructions through carefully designed prompts, we guide the model toward the desired output while minimizing disruption to its original knowledge. Additionally, a memory-augmented adapter is applied exclusively to the decoder, ensuring precise control over stylistic elements while maintaining the model's generalization ability.

Second, to effectively adapt the model to the new task, we use advanced prompt engineering. This includes strategies to handle semantically diverse sentence structures and isolate modern words, such as neologisms or anachronisms, for separate processing, preventing them from affecting the classical style. Lastly, we implement a full-granularity memory-augmented mechanism, enabling the model to retain and apply stylistic cues consistently throughout the translation process, ensuring stylistic fidelity to Classical Chinese without compromising broader model capabilities.

We conducted extensive experiments to validate our approach. First, we fine-tuned a large pre-trained models, ChatGLM, which showed significant improvements in Classical Chinese translation, confirming the effectiveness of our method. An ablation study on the prompt engineering component further demonstrated that each technique contributed to the overall performance. Additionally, incorporating a memory-augmented adapter in the decoder enhanced stylistic fidelity, highlighting the importance of this mechanism for style control.

To sum up, the main contribution of this work is the introduction of a structured framework that combines prompt engineering and memory augmentation, enabling pre-trained models to achieve specialized task adaptation—such as Classical Chinese translation—while preserving their ability to generalize across diverse tasks.

2 Related Work

2.1 Classical Chinese Translation

Previous studies on Classical Chinese have primarily focused on ancient-modern Chinese translation (Guo et al. (2023)), Classical Chinese poetry generation (Yu et al. (2024)), and couplets (Wang et al. (2021)). However, there has been limited research on generating broader Classical Chinese texts, which presents unique challenges such as the lack of a clear definition, time-invariant language habits, and absence of standardized evaluation metrics. Given the significant stylistic variation across historical periods, time-aware methods have been proposed to improve the understanding and generation of Classical Chinese (Chang et al. (2021), Ren et al. (2023)).

2.2 Prompt Engineering for Style and Domain Adaptation

Recent studies on style and domain adaptation in machine translation (MT) have focused on controlling stylistic elements such as formality and tone. Sennrich et al. (2016) introduced side constraints to control politeness levels in translation, while Niu et al. (2017) proposed a Formality-Sensitive Machine Translation (FSMT) approach, using lexical formality models to adjust the formality of outputs. However, most of these methods require task-specific models, making them less flexible for diverse stylistic needs. In contrast, recent advancements in prompt engineering have shown promise in adapting large language models (LMs) to various tasks with minimal modifications. Techniques such as In-Context Learning (Wang et al. (2022)) and domain-specific keyword selection (Ben-David et al. (2022)) have demonstrated the potential for achieving style adaptation through carefully designed prompts, without altering the underlying model architecture.

2.3 Memory Augmentation for Style Control

Memory-augmented approaches have gained attention for their ability to enhance the flexibility of pre-trained models without requiring extensive fine-tuning. For example, kNN-MT (Khandelwal et al. (2020)) incorporates external memory to support domain adaptation, while recent works (Borgeaud et al. (2022); Chen et al. (2022)) have leveraged memory mechanisms to integrate non-

parametric information into language models. These methods allow for dynamic adjustments based on the input context, providing a more adaptable solution for tasks like machine translation.

3 METHOD

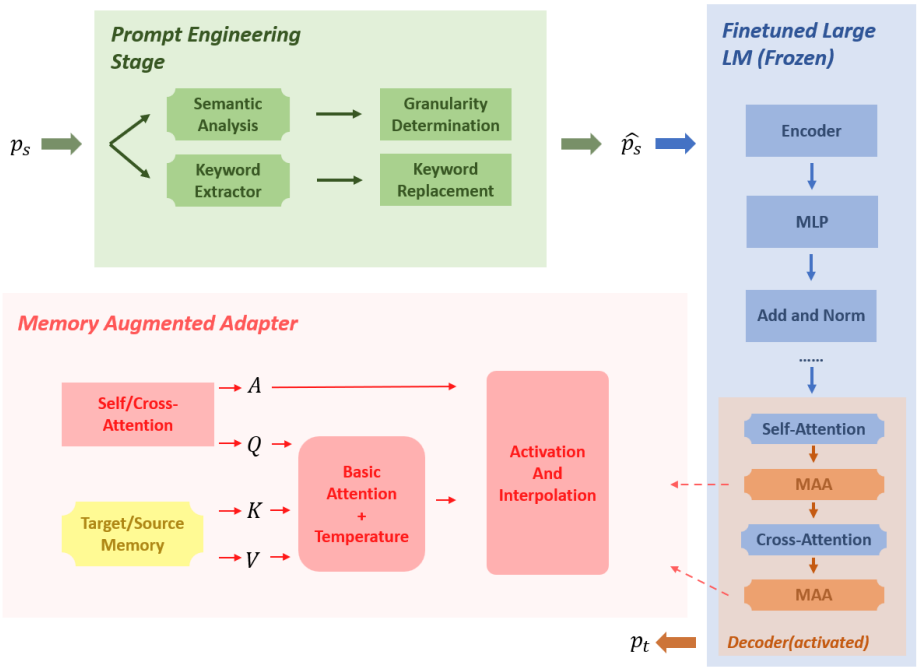


Figure 1: Main Pipeline

Our method tackles the challenge of translating modern texts into Classical Chinese by focusing on two main strategies: prompt engineering and memory-augmented adapters. First, we use prompt engineering to address the differences between modern and classical concepts. By identifying modern terms that would be incomprehensible to ancient readers and replacing them with appropriate Classical Chinese equivalents, we ensure semantic consistency. Additionally, we introduce a multi-granularity translation approach, where narrative sentences are processed in chunks to maintain context, and evaluative statements are translated sentence by sentence to capture stylistic nuances. For memory augmentation, we integrate a memory adapter into the model, allowing it to adapt stylistically while preserving core model capabilities. This approach is modular, enabling the addition of new styles without retraining the entire model.

3.1 PROMPT ENGINEERING

3.1.1 KEYWORD REPLACEMENT

A significant challenge in translating modern texts into Classical Chinese lies in the presence of concepts and expressions that were absent in ancient times. To address this issue, we propose a prompt engineering strategy. Initially, we direct the foundation model to identify ”modern concepts that would be incomprehensible to an ancient audience” within the source text. Subsequently, we instruct the model to ”replace these modern concepts with their Classical Chinese equivalents, ensuring semantic consistency or contextual appropriateness.” Finally, we prompt the model to ”generate the fully modified passage,” incorporating these adjustments while maintaining the overall coherence of the translation.

3.1.2 MULTI-GRANULARITY PROCESSING

Another key challenge in translating modern texts, such as biographical works, lies in the structural differences between modern and Classical Chinese. Modern texts often feature lengthy, laudatory

passages or analytical statements, whereas Classical Chinese favors concise, formal expressions, as seen in historical texts like the "Records of the Grand Historian." To address this, we propose a multi-granularity approach: for narrative sentences, we translate in chunks of four sentences to maintain contextual cohesion, while for laudatory statements, we translate sentence by sentence to capture stylistic nuances. Additionally, we introduce a penalty for non-translation to prevent low-quality outputs and ensure stylistic fidelity.

3.2 Pluggable Memory-augmented Adapter

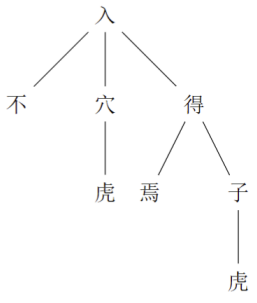
To enable the translator to generate output in a specific style that meets the user's requirements, similar to the approach proposed by Xu et al. (2023), we implemented a methodology that achieves the desired outcome by integrating a memory-augmented adapter into the original model in a pluggable manner. During the testing phase, users can simply provide a few examples in the target style, which will allow the adapter to mimic the style and generate output in the desired style while preserving the original model's performance.
Specifically, we first design a multi-granular memory-augmented adapter that encodes the input examples provided by the user into memory slots. This adapter is then integrated into the original model, combining the external memory information with the model's output to generate a style-specific translation result.

3.2.1 Memory Construction

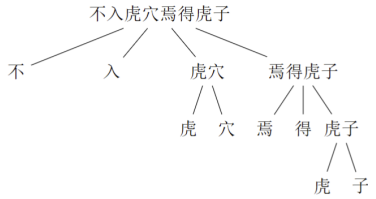
To design the memory augmentation for the adapter, we focus on two main objectives. First, it should capture as much style information as possible from the input examples. Second, it must be easily accessible and usable by the model. Since the amount of style information in a text often correlates with its length, and longer memory units increase the cost of access, we adopt a multi-granular memory structure, as suggested by Xu et al. (2023), to strike a balance between capturing stylistic information and retrieval efficiency.
Multi-granularity Data Structure
To preserve as much semantic information as possible, we segment the input text according to classical Chinese grammar rules and utilize the parse tree to construct a multi-granular memory structure. Specifically, we employ the SuPar-Kanbun dependency parser Yasuoka et al. (2022) to extract the syntactic dependencies between words in each sentence. Using the dependency information provided by SuPar-Kanbun (as shown in Figure 2(a)), we can then construct the parse tree for each sentence (Figure 2(b)). The multi-granular memory is built by associating each layer of the parse tree with a corresponding level of granularity. Each node in the tree is assigned a text that consists of the words in the subtree rooted at that node (Figure 2(c)). The memory at a specific granularity level is then defined as the set of texts corresponding to the nodes at that level of the parse tree. The pipeline for this step is shown in Figure 2. To address the lack of bilingual examples during training,

Word Id	Word	Head Id
1	不	2
2	入	0
3	虎	4
4	穴	2
5	焉	6
6	得	2
7	虎	8
8	子	6

(a) SuPar-Kanbun output



(b) Depandency tree



(c) Multi-granularity

Figure 2: Multi-granular data construction pipeline

we utilize a reliable classical-to-modern Chinese translator to convert each text in the memory into modern Chinese, thereby creating a bilingual memory.
Memory Encoding
To minimize the retrieval cost, we employ the same model for both encoding and accessing the memory. To further bridge the gap between encoding and utilization, we construct the source-side

memory from the encoder's output and the target-side memory from the self-attention outputs of each decoder layer. The key idea is to ensure that both memory encoding and utilization occur within the same layer of the original model. The pipeline for this step is shown in Figure 3.

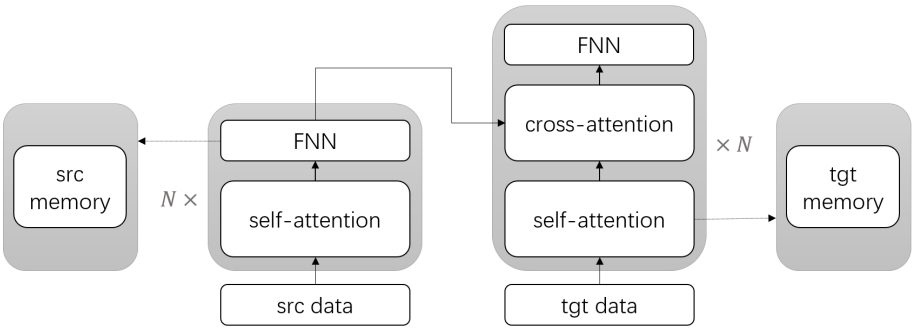


Figure 3: Memory encoding pipeline

3.2.2 MEMORY-AUGMENTED ADAPTER

Adapter Architecture
The memory adapter has 4 inputs: anchor A, query Q, key K, and value V. A and Q are from the original model, while K and V are from the memory. The retrieval output is Xu et al. (2023):

$$R = \text{softmax}(QW_qW_k^TK^T/T)VW_v,$$

where W_q, W_k, W_v are learnable parameters, and T is the temperature parameter.
We use anchors to prevent the translation results from deviating significantly from the original model, as such deviations could lead to serious errors in certain cases.

$$\lambda = \text{sigmoid}(\text{relu}([A; R]W_1)W_2),$$
$$O = \lambda A + (1 - \lambda)R,$$

where O is the final output of the adapter.Xu et al. (2023)

Training Strategy
As suggested in Xu et al. (2023), we used a memory dropout strategy to avoiding over dependency to the memory. Let M be the full memory, we randomly dropout a subset of M to get $\hat{M}$. Then, the overall loss can be expressed as:

$$L = L_{NLL}(P(y|x, \theta, M)) + \alpha L_{NLL}(P(y|x, \theta, \hat{M})) + \beta L_{dist}(P(y|x, \theta, M), P(y|x, \theta, \hat{M})),$$

where L_{NLL} is the negative log-likelihood loss, L_{dist} is the distance loss, and α, β are both learnable parameters.

4 EXPEIEMENTS

4.1 PROMPT ENGINEERING RELATED EXPERIMENTS

4.1.1 EXPERIMENTAL DETAILS

We finetune the glm-4-9b model GLM et al. (2024a) on subsets of the Classical Chinese dataset DBU NiuTrans (2024) with modern-ancient sentence pairs. We selected classical Chinese texts from the Tang Dynasty and earlier, focusing primarily on historical works. There are around 120k training bilingual sentence pairs in total. The training took 3 hours for 5 epochs on $4\times$ RTX 4090 24GB GPUs.

4.1.2 QUANTITATIVE EVALUATION

Dataset. We evaluated the model using two test datasets. The first was a reserved test set from before training, not included in the training process and within the same textual category. The second was derived from the "Song History," which is outside the "pre-Tang" category of the training data, thereby assessing the model's generalization ability. Each dataset comprised 2,400 sentence pairs.

Metrics. We evaluate the translation quality using BLEU Papineni et al. (2002) and ROUGE Lin (2004) scores. BLEU measures the overlap of n-grams between the machine translation and reference translation. We select precision of ROUGE-1 and ROUGE-l to evaluate the quality, where ROUGE-1 calculates the unigram overlap, measuring the precision of individual words between the generated and reference translations and ROUGE-L evaluates the longest common subsequence (LCS) between the machine-generated and reference translations.

Baseline Models. We use the glm-4-9b-chat model as a baseline, prompting it with the sentence "Please translate this sentence from modern Chinese to classical Chinese".

Table 1: Translation performance comparison.

MODEL	TESTSET			SONG SHI		
	BLEU	ROUGE-1	ROUGE-L	BLEU	ROUGE-1	ROUGE-L
glm-4-9b-chat	12.78	36.36	42.09	18.54	42.23	47.22
Our	29.00	51.17	54.52	24.38	52.14	57.40

Results. The results in Table 1 indicate that our model demonstrates a substantial improvement over glm-4-9b-chat across both test datasets. On the reserved test set, our model shows significant enhancements in BLEU, ROUGE-1, and ROUGE-L scores. Similarly, when evaluated on the "Song History" dataset, which tests the model's ability to generalize beyond the training category, our model again outperforms glm-4-9b-chat by a notable margin in all metrics. These improvements highlight the model's strong translation capabilities and effective generalization.

4.1.3 QUALITITATIVE EXAMPLES

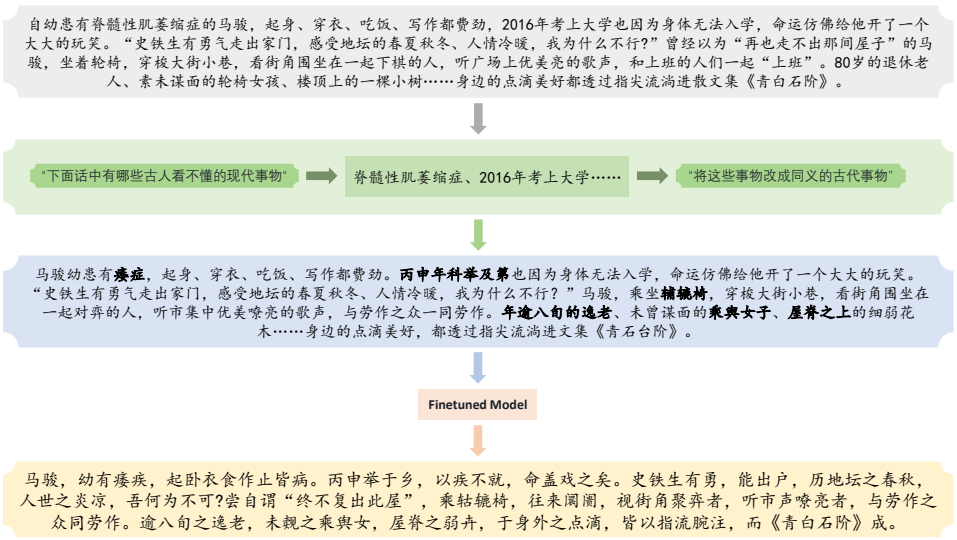


Figure 4: Case Study on Prompt Engineering

As illustrated in Figure 4.1.3, the source modern sentence undergoes processing through prompt engineering, resulting in a high-quality output in the classical form.

4.1.4 ABLATION STUDY

To evaluate the effectiveness of our prompt engineering approach, we compare the model's performance before and after applying prompt engineering across four key metrics: fluency (grammatical accuracy and linguistic smoothness), coherence (logical consistency and contextual relevance), aesthetics (overall elegance), and diversity (variation in expression and lexical richness).

Since expert human evaluation is unavailable, we use multiple large language models (LLMs) to evaluate our model's performance. To reduce the risk of memorization or bias, each evaluation focuses on a single dimension, with a new chat session initiated for each dimension to ensure independence. To address potential biases from input order (before or after prompt engineering), we randomize the sequence of dataset presentation. Additionally, we use two different LLMs, ChatGPT and ChatGLM, to mitigate reliance on a single model. This approach ensures a more reliable and unbiased evaluation of the translation model.
We evaluate the model's performance before and after prompt engineering by scoring each version across multiple metric dimensions. The final score is a weighted sum of these dimension scores, normalized to a percentage. Instead of assigning absolute scores, LLMs compare the two versions for each dimension, increasing the score of the preferred version by 1. The formula for the final score is:

$$User(O) = \frac{s_{after}}{s_{before} + s_{after}},$$

while $s_i = \sum_j w_j s_{i,j}$ is the overall score of version i, $s_{i,j}$ is the score of version i at dimension j, and w_j is the weight for dimension j.
The results for key word replacement:

Table 2: Translation performance comparison for Key Word Replacement.

Model	**Fluency**	**Coherence**	**Aesthetics**	**Diversity**	**Weighted Average**
Before Prompt Engineering	8	6	8	7	7.7
After Prompt Engineering	2	4	2	3	2.3

The final User(O) score is:

$$User(O)_{keyword} = 77\%.$$

The results for granularity determination:

Table 3: Translation performance comparison for Granularity Determination.

Model	**Fluency**	**Coherence**	**Aesthetics**	**Diversity**	**Weighted Average**
Before Prompt Engineering	7	7	10	10	7.4
After Prompt Engineering	5	5	3	3	2.6

The final User(O) score is:

$$User(O)_{granularity} = 74\%.$$

As a result of the evaluation conducted using the LLM, we found that after applying keyword replacement, 77% of evaluators preferred our approach, and after implementing granularity determination, 74% of evaluators expressed a preference for our method.

4.2 MEMORY ADAPTERS BASED EXPERIMENTS

4.2.1 EXPERIMENTAL DETAILS

We integrate the memory-augmented adapter into a model fine-tuned from the GLM-4-9B architecture GLM et al. (2024a), using a subset of bilingual data derived from "The Records of the Grand

Historian". The goal is to adapt the model to generate translations in a style reminiscent of Sima Qian's writings. The training corpus comprises approximately 10k bilingual sentence pairs. The fine-tuning process was conducted over 10 epochs, requiring approximately 1 hour on a single RTX 4090 24GB GPU.

4.2.2 QUANTITATIVE EVALUATION

Dataset. We evaluated the model using the test dataset generated from "The Records of the Grand Historian" sentence pairs.

Metrics. We assess the translation quality using BLEU Papineni et al. (2002) and ROUGE Lin (2004) scores, with a detailed analysis provided in Section 4.1.2. Additionally, we leverage the capabilities of the foundation model to design a custom evaluation metric, akin to the one introduced in Section 4.1.4. In this context, "User(O)" corresponds exactly to the metric described in that section, while "User(S)" represents a similar setup where the model's In-Context Learning (ICL) ability is utilized to infer the stylistic characteristics of "The Records of the Grand Historian". Higher scores in this metric indicate that the generated output is more likely to align with the writing style of Sima Qian. We compare the results between our method and the finetuned model without style transfer.

Table 4: Translation performance comparison.

MODEL	ROUGE-1	ROUGE-L	BLEU	User(O)	User(S)
Simply Finetuned	45.24	59.02	26.49	49.3%	15.6%
Our Method	60.91	68.21	40.35	50.7%	84.4%

Results. The results in Table 4 demonstrate that our method significantly enhances translation quality, particularly in terms of stylistic control, while maintaining similar overall performance to the simply finetuned model. Our method outperforms the simply finetuned model in ROUGE-1, ROUGE-L, and BLEU scores, indicating better semantic and syntactic alignment with the reference, and higher fidelity to the original content.

In terms of User(O), both models show comparable performance with only a 1.4% difference, suggesting minimal impact on general translation quality. However, the most notable improvement is in User(S), where our method achieves 84.4% compared to 15.6% for the simply finetuned model. This strongly supports the effectiveness of our approach in capturing the stylistic nuances of "The Records of the Grand Historian."

5 CONCLUSION

In this study, we presented a novel framework, STAMP, that enhances the translation of modern texts into Classical Chinese by finetuning large language models with a mass of data, integrating prompt engineering and memory-augmented adapters. The experimental results show substantial improvements in translation accuracy and stylistic fidelity, confirming the utility of our method. Future work could expand the limited high-quality classical Chinese dataset and explore how to utilize it efficiently. There is a significant gap between the fashionable expressions of modern Chinese and Classical Chinese, and there are considerable challenges in translating new words. Future work could focus on developing more effective translation strategies to bridge this gap.

6 KEY INFORMATION

This work isn't under the instructions of a PI.

In this work, Chenxiao Yang designed the fine-tuning framework and developed the prompt engineering methodology. Yingxi Lu proposed the key structure for memory-augmented adaptation and conducted related experiments involving prompt engineering. Zhuo Lin implemented the memory-augmented adapter framework and carried out experiments associated with it. All team members

contributed significantly to the entire process, demonstrating a strong spirit of collaboration and effective teamwork.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Eyal Ben-David, Nadav Oved, and Roi Reichart. Pada: Example-based prompt learning for on-the-fly adaptation to unseen domains. *Transactions of the Association for Computational Linguistics*, 10:414–433, 2022.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2206–2240. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/borgeaud22a.html.

Ernie Chang, Yow-Ting Shiue, Hui-Syuan Yeh, and Vera Demberg. Time-aware ancient chinese text translation and inference. *arXiv preprint arXiv:2107.03179*, 2021.

Wenhu Chen, Pat Verga, Michiel De Jong, John Wieting, and William Cohen. Augmenting pre-trained language models with qa-memory for open-domain question answering. *arXiv preprint arXiv:2204.04581*, 2022.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024a.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024b.

Geyang Guo, Jiarong Yang, Fengyuan Lu, Jiaxin Qin, Tianyi Tang, and Wayne Xin Zhao. Towards effective ancient chinese translation: Dataset, model, and evaluation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 416–427. Springer, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. *arXiv preprint arXiv:2010.00710*, 2020.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Xing Niu, Marianna Martindale, and Marine Carpuat. A study of style in machine translation: Controlling the formality of machine translation output. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2814–2819, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1299. URL https://aclanthology.org/D17-1299.

NiuTrans. Classical-modern. https://github.com/NiuTrans/Classical-Modern, 2024. Accessed: 2024-12-20.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Han Ren, Hai Wang, Yajie Zhao, and Yafeng Ren. Time-aware language modeling for historical text dating. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 13646–13656, 2023.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Controlling politeness in neural machine translation via side constraints. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 35–40, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1005. URL https://aclanthology.org/N16-1005.

Yan Song. Chinese couplet generation with syntactic information. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 6436–6446, 2022.

Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. Training data is more valuable than you think: A simple and effective method by retrieving from training data. *arXiv preprint arXiv:2203.08773*, 2022.

Yufeng Wang, Jiang Zhang, Bo Zhang, and Qun Jin. Research and implementation of chinese couplet generation system with attention-based transformer mechanism. *IEEE Transactions on Computational Social Systems*, 9(4):1020–1028, 2021.

Yuzhuang Xu, Shuo Wang, Peng Li, Xuebo Liu, Xiaolong Wang, Weidong Liu, and Yang Liu. Pluggable neural machine translation models via memory-augmented adapters. *arXiv preprint arXiv:2307.06029*, 2023.

Koichi Yasuoka, Christian Wittern, Tomohiko Morioka, Takumi Ikeda, Naoki Yamazaki, Yoshihiro Nikaido, Shingo Suzuki, Shigeki Moro, and Kazunori Fujita. Designing universal dependencies for classical chinese and its application. 63(2):355–363, feb 2022.

Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Zonghan Yang. Chinese poetry generation with a working memory model. *arXiv preprint arXiv:1809.04306*, 2018.

Chengyue Yu, Lei Zang, Jiaotuan Wang, Chenyi Zhuang, and Jinjie Gu. Charpoet: A chinese classical poetry generation system based on token-free llm. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 315–325, 2024.