NPRIP: Nucleus-to-Periphery Retrieval-Iterative Prompting for Improved Abstractive Summarization in Low-Resource Mongolian

Menghan Li [1] , **Nier Wu**[1*] , **Yang Liu**[2] , **Yatu Ji**[1] and **Shuo Sun**[1]
[1]Inner Mongolia University of Technology, Hohhot, China
[2]Inner Mongolia Autonomous Region Big Data Center, Hohhot, China
{20241800116,wunier04,mljyt,sunshuo07}@imut.edu.cn, 656044725@qq.com

Abstract

Large language models often face challenges in low-resource agglutinative language text summarization tasks due to poorly designed prompts, leading to core information dilution, reduced fidelity, and critical information loss caused by the complex grammatical structures of agglutinative languages. For traditional Mongolian, a typical low-resource agglutinative language, this paper proposes a Nucleus-to-Periphery Retrieval-Iterative Prompting (NPRIP). This method first guides the model to extract highly condensed semantic core information (events, persons, time, etc.) from the original text. Subsequently, through multiple rounds of self-refinement iteration, it progressively expands peripheral details (background, causes, consequences, secondary facts, etc.). The model performs fact consistency checks, redundancy removal, and fidelity correction on the current draft, achieving gradual improvements in information completeness and fidelity. To enhance Mongolian language representation, we perform parameter-efficient fine-tuning on LLaMA3-8B using the CCMT2019 Mongolian-Chinese parallel corpus, and construct a larger abstractive Mongolian news summarization dataset MoSum along with its augmented version. Experiments on traditional Mongolian text summarization tasks demonstrate that our proposed method significantly outperforms multiple baseline models on automatic evaluation metrics including ROUGE-1, ROUGE-2, and ROUGE-L. This validates the effectiveness of core-priority structured iterative prompting in low-resource agglutinative language summarization scenarios.

1 Introduction

In the era of information explosion, automatic text summarization technology has become a significant research direction in the field of natural language processing. However, generating summaries for Mongolian texts still faces considerable challenges: high-quality parallel corpora are scarce, and the complex agglutinative grammar structure and non-linear writing characteristics of traditional Mongolian often lead to semantic loss and reduced fidelity during the summarization process. These issues severely constrain the development of Mongolian text summarization technology. In recent years, large language models (LLMs) have demonstrated formidable capabilities in handling general tasks and deep language understanding. Their extensive world knowledge and zero-shot/few-shot learning properties offer a novel research paradigm for low-resource text summarization tasks. Related studies indicate that LLMs can preliminarily comprehend and summarize text through prompting instructions without relying on task-specific annotated data, thereby mitigating data scarcity to some extent. The application of large language models in automatic text summarization traces back to OpenAI's research, which optimized GPT-3 through reinforcement learning to generate summaries more aligned with human preferences[Stiennon *et al.*, 2020]. However, as models grow larger, the computational and storage costs of full-parameter fine-tuning become prohibitive. This has driven research toward more efficient fine-tuning and adaptation techniques, such as adapter-based few-shot fine-tuning[Bražinskas *et al.*, 2022], prompt tuning, and prefix tuning. Zhang et al. enabled models to generate summaries meeting specific requirements by adding learnable control prompts before input[Zhang *et al.*, 2022]. Chen et al. enhanced model adaptability for few-shot summarization by combining multi-task pre-training with task-specific prefix vectors [Chen *et al.*, 2023]. Prompt engineering effectively unlocks the potential of large language models by designing high-quality input prompts, significantly reducing computational overhead and improving summary generation quality without extensive parameter updates. While template engineering offers an intuitive approach to prompt design, its construction heavily relies on human expertise, making it challenging to ensure optimality for complex tasks[Jiang *et al.*, 2020]. To address these limitations, some studies have proposed automated prompt design and structured summarization methods. For instance, [Narayan *et al.*, 2021] guided summary generation by constructing entity chains within text ; Zhou et al. integrated retrieved information using a "retrieval-summarization" strategy. [Zhou *et al.*, 2023]The Chain-of-Thought (CoT) technique, by intro-

*Corresponding author

ducing intermediate reasoning steps, significantly improved LLM performance in complex reasoning tasks [Wei *et al.*, 2022] and has gradually been applied to text summarization. To address factual hallucinations—a common issue in abstract generation—[Wang *et al.*, 2023] proposed the Chain-of-Thought method, which simulates human stepwise abstract writing to guide models in better integrating source information . Adams et al. further introduced the entity-centric "Density Chain" approach to generate summaries with higher information density. Their experiments demonstrated advantages in abstractness and integration, while exhibiting less prompting bias compared to summaries generated by GPT-4 under conventional prompts [Adams *et al.*, 2023]. Although these approaches have achieved significant progress in multilingual summarization tasks, text summarization for traditional Mongolian—a low-resource agglutinative language—still faces unique challenges. The complex grammatical structures, limited corpus resources, and distinctive cultural semantics of Mongolian cause existing models to underperform in key information extraction, logical coherence, and domain adaptation. This frequently results in information omission, factual hallucinations, or distorted cultural expressions. Therefore, it is imperative to explore more efficient and interpretable summarization strategies that enhance fidelity and overall quality without relying on external resources. In recent years, advanced prompt engineering methods such as Self-Refine [Madaan *et al.*, 2023] and Self-RAG [Asai *et al.*, 2024] have provided significant insights for addressing the aforementioned challenges. These approaches embed few-shot examples within prompts and guide models through multiple rounds of self-feedback and iterative refinement (typically 2–5 iterations) on initial outputs, significantly enhancing generation quality. Unlike traditional Retrieval-Augmented Generation (RAG)[Lewis *et al.*, 2020] that relies on external vector databases, these approaches emphasize self-reflection and iterative simulation at the prompt level. They effectively mitigate hallucination issues and enhance output consistency through carefully designed prompts, making them particularly suitable for low-resource language tasks. However, existing iterative prompting methods (e.g., Chain-of-Density [Adams *et al.*, 2023]) predominantly employ a forward-generation sequence that progressively focuses from contextual details to core information. This approach remains prone to diluting or omitting core semantic information during generation, especially in summarization tasks for traditional agglutinative languages like Mongolian. Based on these observations, this paper proposes a Nucleus-to-Periphery Retrieval-Iterative Prompting (NPRIP) method. This method introduces a structured reasoning sequence of "core-first, periphery-expansion," guiding the model to first focus on the most critical semantic core information in the original text (e.g., events, persons, and time). Subsequently, through multiple rounds of self-reflection, it progressively supplements secondary details and background descriptions. This approach better adapts to the summarization needs of low-resource agglutinative languages like Traditional Mongolian. NPRIP achieves significant improvements in summary fidelity and structural integrity solely through prompt optimization, without requiring external retrieval systems or additional knowledge bases. Experimental results demonstrate that NPRIP outperforms baseline methods on traditional Mongolian text summarization tasks across metrics including ROUGE-1, ROUGE-2, and ROUGE-L.

2 Related Work

The advent of large language models has brought numerous opportunities and challenges to the field of automatic text summarization, simultaneously driving a shift in the technical paradigm from traditional pre-training and fine-tuning approaches toward more flexible prompt-based learning methods. In recent years, extensive research has explored the application of large language models in automatic text summarization tasks [Tang *et al.*, 2023]. [Zhang *et al.*, 2024] found that despite stylistic differences, summaries generated by large language models now rival human-written summaries in overall quality. [Goyal *et al.*, 2022] conducted human evaluations comparing GPT-3 operating in zero-shot mode against pre-trained models fine-tuned on summary datasets. Results showed evaluators preferred GPT-3-generated summaries, which exhibited no significant dataset-specific bias. However, as model sizes continue to grow, the computational and storage costs of full-parameter fine-tuning become prohibitive. Consequently, research focus has shifted toward more cost-effective parameter-efficient approaches, including knowledge distillation, parameter-efficient fine-tuning, and prompt engineering. Regarding parameter-efficient fine-tuning, due to the massive parameter count of large language models, direct full-parameter updates are often impractical. Researchers have proposed several efficient alternatives. [Bražinskas *et al.*, 2022] employed an adapter-based few-shot fine-tuning approach, pre-training the adapter module on a large dataset before fine-tuning it on a small, manually annotated dataset. Additionally, other studies have explored techniques like prompt tuning and prefix tuning. [Zhang *et al.*, 2022] introduced learnable control prompts before input documents, enabling the model to learn generating summaries that meet specific requirements during training. At inference, these control prompts can be flexibly specified by users, allowing fine-grained control over summary outputs. [Chen *et al.*, 2023] injected task-specific prefix vectors into each summarization task within a multi-task pretraining framework, jointly optimizing them with model parameters to enhance adaptability for few-shot summarization tasks.

3 Method

To address the challenges faced by traditional Mongolian—a low-resource agglutinative language—in abstract summarization tasks, this paper proposes a phased optimization framework encompassing dataset construction and augmentation, efficient model parameter adaptation, and structured prompt optimization during inference. Specifically, our approach comprises the following three stages:

- **Dataset Construction and Data Augmentation Phase:** Constructed the large-scale abstract traditional Mongolian abstract dataset MoSum, and performed data augmentation through context rewriting and entity replacement driven by large language models to gener-

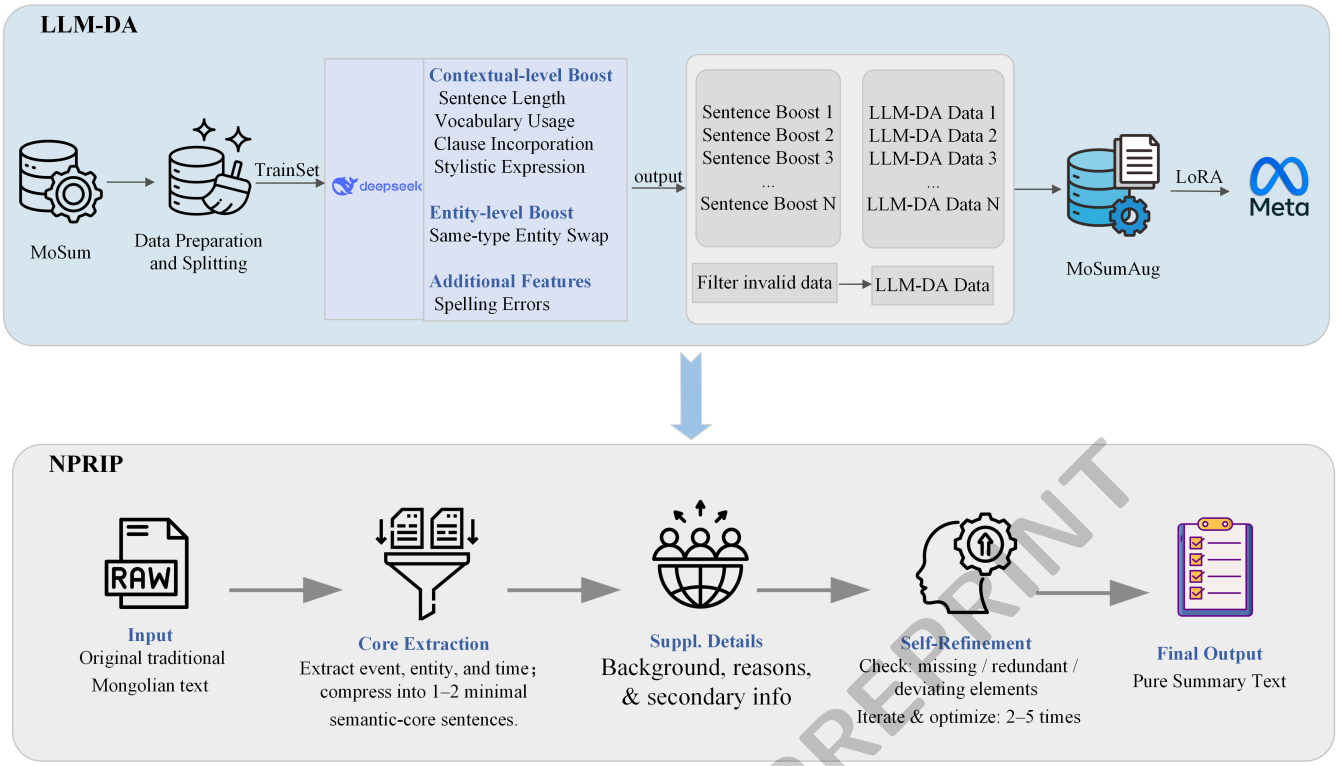

Figure 1: Overall flowchart.Upper part: Data augmentation pipeline using DeepSeek-V3 to extend MoSum into MoSumAug via contextual rewriting and non-core entity substitution,followed by LoRA fine-tuning of the LLaMA. Lower part: NPRIP inference process, consisting of Nucleus Extraction, Periphery Expansion, and Self-Refine (iterative 2–5 times for fact-checking, redundancy removal, and optimization).

ate MoSumAug, thereby mitigating the scarcity of low-resource data.

- **Model-Side Parameter Efficient Summarization Adaptation Phase:** Based on MoSum and its augmented dataset MoSumAug, we perform a low-cost supervised fine-tuning on large language models to enhance their adaptability for traditional abstractive summarization tasks in the Mongolian language.
- **Structured Prompt Optimization Phase for Inference:** Further enhance summary quality through a self-iterative prompt strategy progressing from core to periphery, without introducing additional parameter updates.

3.1 Traditional Mongolian Abstract Dataset and Data Augmentation

Traditional Mongolian, as a typical low-resource agglutinative language, suffers from an extreme scarcity of high-quality publicly available parallel abstracting corpora. Currently, the most representative single-document abstracting dataset for Mongolian is MTESum.[Qi *et al.*, 2025]. It contains approximately 1,000 original news article-summary pairs, primarily constructed using a headline-guided extractive approach. Reference summaries are mostly direct extracts from the source text, making it a typical extractive dataset. Although MTESum holds significant value as the first publicly available Mongolian summary dataset, its small scale, extractive-oriented summary types, and limited domain coverage make it insufficient for fully supporting the training and evaluation of modern abstractive summarization models.

To overcome these limitations, this paper constructs MoSum, a larger-scale traditional Mongolian news summary dataset supporting abstractive generation. The dataset and code are publicly available at: https://github.com/LeeMhh/Mo_Summary. Compared to MTESum, MoSum offers the following significant advantages:

1. Larger Scale: Contains approximately 17,000 high-quality original text-reference summary pairs (versus only 1,000 in MTESum), providing more abundant supervised signals for model training;

2. Broader Domain Coverage: Spans news across multiple domains including politics, economy, culture, society, and technology (2018–2024), enhancing model generalization;

3. Higher Data Quality: Combines original texts from authentic news websites with LCSTS translation versions, undergoing rigorous preprocessing and human proofreading.

This dataset primarily originates from two sources:

(1) The Inner Mongolia Daily, a Mongolian-language news website in Inner Mongolia[1], covering multi-domain news reports from 2018 to 2024;

(2) The high-quality human-translated version of the Chinese abstract dataset LCSTS (Part II)[Hu *et al.*, 2015].

The original corpus underwent a unified preprocessing

[1]https://www.mgyxw.cn

workflow, including symbol normalization, special character filtering, and duplicate content removal, ultimately yielding approximately 17,000 samples. Data is organized in Alpaca instruction format, with each sample comprising a task instruction, the original news article, and a human-reviewed reference summary.

To alleviate the problem of limited training data scale in Mongolian and enhance model robustness, this study selects DeepSeek-V3, which possesses multilingual adaptability and cost-effectiveness advantages, as the foundational model and designs a data augmentation method. Its multilingual capability optimizes for low-resource languages, effectively supporting Mongolian processing, while its high cost-effectiveness meets the expansion needs under resource constraints. Based on this, MoSum is extended into an enhanced training dataset named MoSumAug This method references the LLM-DA framework [Ye *et al.*, 2024], with task-specific adaptations for summarization. Specifically, the data augmentation process includes: (1) Contextual rewriting with semantic preserv ation utilizes the DeepSeek-V3 model to diversify the original text while maintaining factual consistency. (2) non-core entity substitution, replacing entities, dates, or numerical values that do not affect core semantics with similar types, while simultaneously updating the reference summary to ensure sample alignment. These strategies effectively expand training sample diversity without introducing semantic drift.

To ensure evaluation fairness, data augmentation applies only to the training set, while the validation and test sets retain their original distributions. Ultimately, the original approximately 17,000 samples were expanded into roughly 84,000 high-quality augmented samples.Detailed statistics are shown in Table 1.

Dataset	Samples	Avg. Src. Len.	Avg. Summ. Len.
Orig. Train	17295	1178	142
Aug. Train	83474	1118	143
Val.	2250	1301	117
Test	2250	1394	118

Note: Abbreviations used in the table: Orig. = Original, Aug. = Augmented, Val. = Validation

Table 1: Dataset Statistics

3.2 Mongolian Text Summarization Capability Adaptation

To achieve efficient adaptation of large language models for traditional Mongolian text summarization tasks, full-parameter updates incur high computational costs, consume substantial GPU memory, and are prone to catastrophic forgetting of pre-trained knowledge. To address this issue, this paper employs the LoRA (Low-Rank Adaptation) parameter-efficient fine-tuning technique [Hu *et al.*, 2022]. The core idea of LoRA is to represent updates to the pre-trained weight matrix as a low-rank decomposition. Specifically, for the original weight matrix W_0, the final weight after fine-tuning is expressed as:

$$W = W_0 + \Delta W \tag{1}$$

Among these, ΔW is the low-rank update matrix to be learned, which can be decomposed as:

$$\Delta W = A \cdot B \tag{2}$$

Where these $A \in R^{d\times r}$,$B \in R^{r\times k}$ with $r \ll \min(d, k)$ representing the low-rank value. During fine-tuning, only the low-rank matrices and are trained, while the original weights remain frozen. This reduces the number of trainable parameters to a minimal fraction of the original model while preserving its generalization capability.

To enhance the adaptability of large language models to low-resource Mongolian, we employed a domain adaptation approach using the CCMT2019 Mongolian-Chinese parallel corpus to fine-tune the LLaMA3-8B base model. This method effectively improved the model's understanding of Mongolian morphological features and syntactic structures while preserving pre-trained knowledge, laying the groundwork for subsequent Mongolian summarization tasks.

Due to the limited size of the Mongolian abstract dataset, full-parameter fine-tuning is prone to overfitting and incurs high computational costs. Therefore, this paper applies LoRA for prompt fine-tuning based on the LLaMA3-8B model.

3.3 Nucleus-to-Periphery Retrieval-Iterative Prompting Method

Although supervised fine-tuned models have demonstrated basic summarization capabilities in Mongolian, traditional prompting approaches remain prone to diluting or omitting core semantic information during generation in low-resource agglutinative language scenarios. Existing prompting methods typically follow a forward generation sequence from contextual details to core information, a limitation particularly pronounced in traditional Mongolian summarization tasks. Inspired by Self-Refine [Madaan *et al.*, 2023] and Self-RAG [Asai *et al.*, 2024], this paper proposes a Nucleus-to-Periphery Retrieval-Iterative Prompting (NPRIP) method. By explicitly reversing the reasoning sequence, NPRIP guides the model to follow a structured iterative process during the inference phase:

- **Core Extraction:** First, highly condense the most critical semantic core information (such as events, people, and time) from the original text, forming 1–2 sentences of minimalist expression;
- **Peripheral Expansion:** Gradually supplement secondary details, background information, and processes based on core information to generate a complete draft summary.
- **Self-Refinement:** Conduct self-review and revision of the draft, identifying missing information, redundancy, or potential biases, and perform 2–5 iterations of optimization.

The structured iterative process of NPRIP can be expressed in the following form:

Let $d^{(0)}$ be the original document, and $s^{(0)} = \emptyset$ the initial (empty) summary draft. NPRIP proceeds in T iterative steps ($T \approx$ 3–5 in practice), For each iteration $t = 1, \ldots, T$:

Configuration	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore F1
LLaMA3-8B (Zero-shot)	52.28	30.80	22.74	85.07
Mo-LLaMa3-8B (Zero-shot)	77.15	42.29	28.99	96.24
LLaMa3-8B + SFT	87.72	56.61	52.63	96.51
Mo-LLaMa3-8B + SFT	88.18	56.95	53.46	96.89
+Aug	89.12	58.24	55.18	97.28
+NPRIP	**89.35**	**58.87**	**56.42**	**97.89**

Table 2: Experimental results. Experimental results under different configurations. Among them, Mo represents LLaMA3-8B, which has been fine tuned in the Mongolian language domain using the Mongolian part of the CCMT2019 dataset; SFT stands for Supervisory Fine tuning; Aug represents training using MoSumAug enhanced dataset; NPRIP represents the application of the kernel to periphery retrieval iterative prompting method proposed in this paper.

1. **Nucleus Extraction**: Generate the core semantic nucleus

$$n^{(t)} = LLM\Big(d^{(0)},\, s^{(t-1)},\, Prompt_{\text{core}}\Big)$$

2. **Periphery Expansion**: Retrieve and append peripheral details

$$p^{(t)} = LLM\Big(d^{(0)},\, n^{(t)} \cup s^{(t-1)},\, Prompt_{\text{expand}}\Big)$$

3. **Self-Refinement**: Refine the current draft for faithfulness and conciseness

$$s^{(t)} = LLM\Big(d^{(0)},\, s^{(t-1)} \cup n^{(t)} \cup p^{(t)},\, Prompt_{\text{refine}}\Big)$$

The final summary is $s^{(T)}$.

To enhance contextual guidance capabilities, NPRIP embeds fixed few-shot examples within prompts as retrieval-style references. During the inference stage, it employs a bundle search strategy with summary length constraints to improve output consistency and information density. Operating entirely within the inference phase, NPRIP significantly boosts summary fidelity, completeness, and structural coherence without requiring external retrieval systems, additional knowledge bases, or model parameter updates.

The overall flowchart is shown in Figure 1.

4 Experiments

4.1 Model and Training

This paper employs LLaMA3-8B-instruct as the base model and utilizes the parameter-efficient fine-tuning method LoRA to adapt the model for summarization tasks. During the Mongolian adaptation phase, we performed pre-training(PT) fine-tuning using the Mongolian section of CCMT2019. This approach preserved the model's general language capabilities while significantly enhancing its performance on Mongolian text. Without increasing model size, just 0.4 cycles of LoRA fine-tuning improved the model's ROUGE-L score by 6.25 percentage points over the baseline on zero-shot summarization tasks. During the summarization fine-tuning phase, the model was trained using supervised fine-tuning (SFT). LoRA introduced low-rank updates only on the q_proj, k_proj, v_proj, and o_proj layers of the attention mechanism to minimize disruption to the model's generative distribution and enhance generalization. LoRA parameters were set with rank=32, alpha=64, and dropout=0.05. The learning rate was set to 1×10^{-5} with a cosine scheduling strategy and a warmup ratio of 0.06. Training batch size was 2, gradient accumulation steps were 4, and total training epochs were 3. 8-bit quantization and FP16 mixed precision were enabled during training to reduce GPU memory usage and improve training efficiency. During inference, to ensure highly deterministic outputs and strictly adhere to NPRIP's structured inference steps, we employ a beam search decoding strategy (num_beams=3). By setting temperature=0 and do_sample=false to disable sampling, we achieve fully deterministic generation, significantly improving summary consistency and fidelity. All experiments were conducted on NVIDIA V100s (32GB) GPUs.

4.2 Main Experiment Results

We conducted a comprehensive evaluation of the proposed method against various baselines on the MoSum and MoSumAug datasets. As detailed in Table 2, the experiments progressively introduce Mongolian domain adaptation, Supervisory Fine-Tuning (SFT), data augmentation (+Aug), and finally the proposed NPRIP framework.

As shown in Table 2, the zero-shot LLaMA3-8B struggles significantly with Mongolian abstractive summarization (ROUGE-L: 22.74, BERTScore F1: 85.07), highlighting the inherent challenges posed by low-resource agglutinative languages. However, Mongolian domain adaptation (Mo-LLaMA3-8B Zero-shot) yields substantial gains across all metrics, particularly in semantic alignment, where the BERTScore F1 increases by over 11 points (to 96.24). Subsequent Supervised Fine-Tuning on the MoSum dataset further boosts performance, with Mo-LLaMA3-8B + SFT achieving a ROUGE-L of 53.46 and a BERTScore F1 of 96.89. The introduction of the augmented dataset (+Aug) continues this upward trend, improving semantic faithfulness to a BERTScore F1 of 97.28. Critically, applying NPRIP at inference on this strongest baseline delivers the best overall results, pushing ROUGE-L to 56.42 (+1.24) and BERTScore F1 to 97.89 (+0.61). This notable improvement underscores NPRIP's ability to preserve the semantic core while iteratively expanding peripheral details, effectively mitigating the

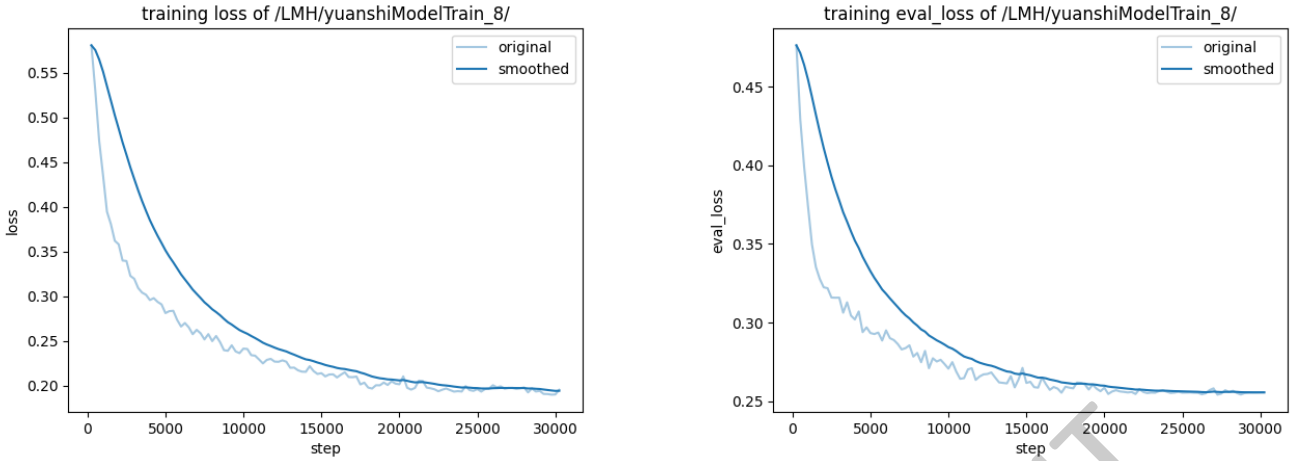

Figure 2: Training and Evaluation Loss Image

information dilution often observed in summarizing agglutinative languages.

Overall, these experimental results validate the effectiveness of a layered strategy that combines parameter-efficient fine-tuning, data augmentation, and structured prompt optimization. As a parameter-free, incremental approach during the inference phase, NPRIP delivers stable and significant performance gains beyond existing high baselines. This demonstrates the unique value of the ”core-to-periphery retrieval self-iteration” strategy in preserving core information and optimizing fidelity for Traditional Mongolian abstract generation.

To further analyze the necessity and optimal configuration of iteration rounds in NPRIP, we examined the trend of ROUGE-L and BERTScore F1 under different iteration rounds T, with results shown in Figure 3. As the iteration rounds increased, both metrics exhibited a trend of rapid improvement followed by convergence. Specifically, ROUGE-L rapidly climbs from 54.76 to 56.42 (+1.66) over the first three rounds, with gains subsequently slowing. This indicates that core extraction and peripheral expansion stages contribute most significantly to long-sequence structural consistency. BERTScore F1 shows substantial improvement in the first two rounds, reaching near saturation after the third round, reflecting the self-refinement step's sustained optimization of semantic fidelity. Overall, T=3 approaches the performance ceiling, with limited marginal gains from further iterations. This result validates NPRIP's multi-round iterative design for its efficiency and robustness in low-resource Mongolian summarization tasks. T=3 is recommended as the default configuration in practical applications. The loss image is shown in Figure 2.

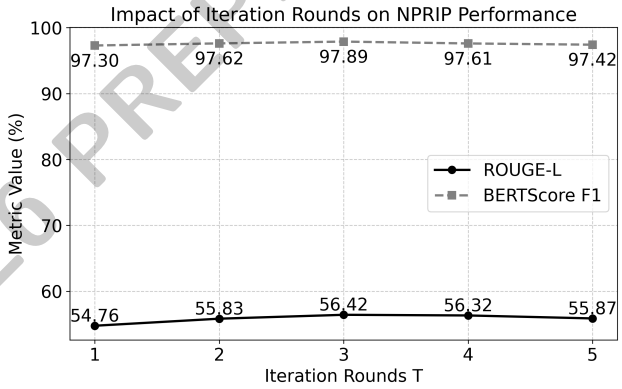

Figure 3: Impact of iteration rounds T on NPRIP performance: ROUGE-L and BERTScore F1 show rapid improvement in the first three rounds, primarily due to core extraction and periphery expansion, indicating the effectiveness of self-refinement. Performance converges after T=3.

4.3 Ablation Experiments

To verify the independent contributions and synergistic effects of the NPRIP framework's components (core extraction, peripheral expansion, and self-refinement), we conducted ablation experiments by sequentially removing each step from the strongest baseline (Mo-LLaMA3-8B + SFT + Aug), as shown in Table 3. The complete NPRIP framework achieves optimal performance across all metrics (ROUGE-L: 56.42, BERTScore F1: 97.89). The ablation results reveal distinct contributions from each module: Core Extraction: Removing this component leads to the most significant drop in ROUGE-L (-0.96). This indicates that the ”core-first” strategy is essential for capturing long-distance semantic dependencies and maintaining structural integrity. Peripheral Expansion: The removal of this module primarily impacts BERTScore F1 (-1.06), demonstrating that the ordered supplementation of contextual details significantly contributes to overall semantic completeness. Self-Refinement: Removing self-refinement leads to a slight ROUGE-L drop (-0.28), confirming its role in fine-tuning generation and factual accuracy. Collectively,

Configuration	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore F1
Mo-LLaMA3-8B + SFT + Aug	89.12	58.24	55.18	97.28
+ NPRIP	**89.35**	**58.87**	**56.42**	**97.89**
w/o Core extraction	88.64	57.82	55.46	97.01
w/o Peripheral Expansion	89.14	57.85	55.21	96.83
w/o Self refinement	89.23	58.12	56.14	97.32

Table 3: Ablation study on NPRIP components. All experiments are conducted on the strongest baseline (Mo-LLaMA3-8B + SFT + Aug). Removing any component leads to performance degradation, with core extraction having the largest impact on ROUGE-L and self-refinement on BERTScore F1.

these results validate the NPRIP design. Specifically, the "core-first" strategy yields the most substantial gains, while the ablation study demonstrates the framework's modularity and interpretability.

5 Conclusion

This paper addresses challenges in traditional Mongolian text summarization tasks, including data scarcity, complex agglutinative structures, and susceptibility to core information loss. We propose a hybrid framework combining parameter-efficient fine-tuning with structured prompt optimization. We construct MoSum, a large-scale Mongolian news summarization dataset, and augment it into MoSumAug via LLM-based semantic retention, alleviating data scarcity in low-resource settings. Built on LLaMA3-8B, we apply LoRA for efficient supervised fine-tuning, reducing computational cost while improving adaptability to Mongolian morphological and syntactic characteristics. Building upon this foundation, the core innovation lies in our proposed Nuclear-to-Peripheral Retrieval Iterative Prompting (NPRIP) method. This approach reverses the traditional generation sequence by first extracting a highly condensed core semantic nucleus, then progressively expanding peripheral details through multi-round self-refinement iterations. It leverages the model's self-refinement mechanism for fact consistency checks, redundancy removal, and logical optimization. This process requires no additional parameter updates or external knowledge, relying solely on structured prompts to significantly enhance summary fidelity, information completeness, and long-sequence consistency—particularly suited for low-resource languages like Mongolian with rich morphology and flexible syntax. Experimental results demonstrate that NPRIP achieves stable performance improvements over zero-shot, SFT, and augmented data baselines on the MoSum test set, fully validating the effectiveness of the "core-first + iterative refinement" strategy. This study demonstrates that combining parameter-efficient fine-tuning with structured prompting during inference offers a computationally efficient and scalable solution for low-resource agglutinative languages. The framework not only provides a practical approach for Mongolian summarization but also establishes a reference paradigm for abstractive generation tasks in other low-resource languages. Future work may explore integration with RAG and multimodal inputs, as well as extension to other Altaic languages.

Acknowledgments

This research was funded by Universities Directly Under the Autonomous Region Funded by the Fundamental Research Fund Project grant number JY20230048 and Autonomous Region Introduced Talent Support Project grant number DC2300001441 and Inner Mongolia Natural Science Foundation Project grant number 2024QN06021, 2024MS09006.

References

[Adams *et al.*, 2023] Griffin Adams, Alex Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. From sparse to dense: Gpt-4 summarization with chain of density prompting. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 68–74, 2023.

[Asai *et al.*, 2024] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. 2024.

[Bražinskas *et al.*, 2022] Arthur Bražinskas, Ramesh Nallapati, Mohit Bansal, and Markus Dreyer. Efficient few-shot fine-tuning for opinion summarization. *arXiv preprint arXiv:2205.02170*, 2022.

[Chen *et al.*, 2023] Yulong Chen, Yang Liu, Ruochen Xu, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Yue Zhang. Unisumm and summzoo: Unified model and diverse benchmark for few-shot summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12833–12855, 2023.

[Goyal *et al.*, 2022] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*, 2022.

[Hu *et al.*, 2015] Baotian Hu, Qingcai Chen, and Fangze Zhu. Lcsts: A large scale chinese short text summarization dataset. *arXiv preprint arXiv:1506.05865*, 2015.

[Hu *et al.*, 2022] 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. In *International Conference on Learning Representations*, 2022.

[Jiang *et al.*, 2020] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language

models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

[Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

[Madaan *et al.*, 2023] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

[Narayan *et al.*, 2021] Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492, 2021.

[Qi *et al.*, 2025] Qirilige Qi, Qintu Si, Siriguleng Wang, and Yongshun Han. Construction and study of a mongolian single-document summarization dataset. *Data Intelligence*, 2025.

[Stiennon *et al.*, 2020] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

[Tang *et al.*, 2023] Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. Evaluating large language models on medical evidence summarization. *NPJ digital medicine*, 6(1):158, 2023.

[Wang *et al.*, 2023] Yiming Wang, Zhuosheng Zhang, and Rui Wang. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. *arXiv preprint arXiv:2305.13412*, 2023.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[Ye *et al.*, 2024] Junjie Ye, Nuo Xu, Yikun Wang, Jie Zhou, Qi Zhang, Tao Gui, and Xuanjing Huang. Llm-da: Data augmentation via large language models for few-shot named entity recognition. *arXiv preprint arXiv:2402.14568*, 2024.

[Zhang *et al.*, 2022] Yubo Zhang, Xingxing Zhang, Xun Wang, Si-qing Chen, and Furu Wei. Latent prompt tuning for text summarization. *arXiv preprint arXiv:2211.01837*, 2022.

[Zhang *et al.*, 2024] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57, 2024.

[Zhou *et al.*, 2023] Yijie Zhou, Kejian Shi, Wencai Zhang, Yixin Liu, Yilun Zhao, and Arman Cohan. Odsum: New benchmarks for open domain multi-document summarization. *arXiv preprint arXiv:2309.08960*, 2023.