Prompt Optimization Methods for Large Language Models with Long Text Input

Yi Ren [1] and Shoubin Li [1]
[1]The Institute of Software, Chinese Academy of Sciences
renyi@iscas.ac.cn

***Abstract*—When faced with long text input, the generated results from large language models sometimes fail to meet user expectations. Due to the length and complexity of the input content, users often do not know how to modify the input to obtain the desired results. To address this dilemma, we propose a Prompt optimization method for large language models with long text input. This method determines the influence weights of different semantic segments on the results, providing guidance for users to generate desired text using large language models. Experimental results show that by evaluating the importance of different semantic segments in military question-answering system text and improving the input content, the quality and usability of the generated military question-answering text can be enhanced.**

***Index Terms*—Long text input, Large language model, Prompt, Question-answering system**

I. Introduction

Large language models, as a product of the combination of "big data + high computing power + strong algorithms," are a collection of implicit knowledge extracted from massive training data. In particular, large language models represented by ChatGPT have demonstrated outstanding performance in the field of text generation. However, when using large language models for text generation, especially when the user input information is lengthy, if the content generated by the large language model does not meet the user's expectations, users usually attempt to modify the input to guide the large language model to generate content that aligns with their expectations. Nevertheless, due to the length of the input text, users find it challenging to grasp the key points when modifying the input. Even after multiple adjustments, the desired output results may still not be obtained which is shown in Figure 11. To solve this problem, this paper proposes a Prompt optimization method for large language models with long text input. This method determines the influence weights of different semantic segments on the results, providing guidance for users to generate desired text using large language models. On this basis, this paper applies the method to the generation of military forum question-answering system text to verify its effectiveness.

Military forum question-answering systems serve as an important entry point for military enthusiasts to quickly learn about past battles and weapons and equipment, playing a crucial role in military education. They assist relevant personnel in understanding and analyzing past battles and equipment, enabling them to quickly and accurately acquire relevant knowledge and experience. This paper combines the Prompt optimization method for large language models with long text input and the generation of military content, aiming to explore the impact of different input semantic segments on the generation of military content text, assisting military enthusiasts in utilizing large language models to obtain military information quickly and accurately.

II. Related Concepts of Military Question-Answering Systems

Question-answering systems for military forums are becoming increasingly complex and intelligent. Wang Xiaoming and Li Xiaohong [1] studied the key technologies involved, such as multi-round dialogue for mining users' deep information needs and semantic matching for precise interaction with knowledge graphs. The construction of knowledge graphs is also gaining attention, with Liu Xiaoming [2] exploring methods suitable for building military domain knowledge graphs. Meanwhile, in the context of big data, traditional matching methods face challenges, and Michael Gray [3] proposed deep semantic matching techniques for better knowledge association. It is evident that research on military forum question-answering systems has begun to take shape, and key technologies are continuously developing. In the future, it will be necessary to build even larger knowledge graphs, achieve dynamic knowledge updates, and enable multi-round dialogue mechanisms to reflect personalized user interest models, making the question-answering services more intelligent.

III. Composition of Military Question-Answering Systems

Military question-answering systems are complex and precise frameworks designed to accurately parse user queries and provide comprehensive answers. The system first uses a question parsing module to understand the user's query intent and key points, identifying the question type and extracting key entities. The subsequent content encoding module is responsible for constructing and continuously enriching the military domain knowledge graph and providing necessary knowledge support by computing entity embedding vectors. This process ensures the system's deep understanding of the military domain and accurate encoding of information.

The matching and retrieval module employs decision tree-based algorithms to achieve deep semantic matching with the knowledge graph, effectively retrieving and linking to the most

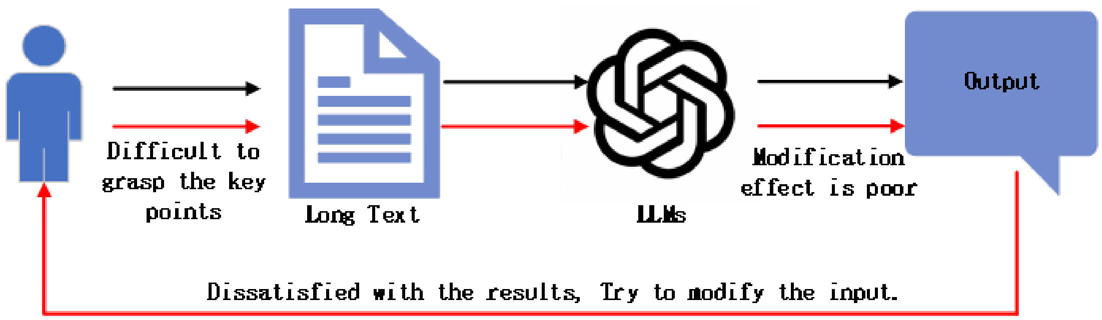

Fig. 1. How to make efficient changes to long text input

relevant information, ensuring that the answers provided to users are highly relevant to their queries. Once the necessary information is retrieved, the reply generation module begins its work, organizing the reply framework and utilizing a rich corpus for training to generate complete and accurate answers.

To provide more personalized services, the user interest modeling module analyzes users' historical topic interest preferences, enabling personalized information provision. Furthermore, the knowledge adjustment module relies on user feedback to promptly update and adjust the knowledge graph, ensuring the timeliness and accuracy of the system's content. The smooth operation of this entire process ensures that the military question-answering system can effectively meet users' queries for military information and provide high-quality personalized services.

IV. Related Work

A. Large Language Models

In 2018, OpenAI proposed the GPT (Generative Pre-Training Transformer) model [4], which uses the Decoder part of the Transformer [5] architecture with certain modifications made to the original Decoder. However, due to the difficulty of the generative direction, its performance was not as good as the BERT model [6] proposed by the Google team in the same year. Subsequently, the OpenAI team proposed GPT2 [7], which expanded the model parameters from 117 million to 1.5 billion and the training data from 5GB to 40GB compared to GPT. The larger model brought better results, and OpenAI shifted its focus to zero-shot learning. GPT2 demonstrated excellent performance in zero-shot learning, but still had gaps compared to traditional models. In GPT3, OpenAI changed zero-shot learning to few-shot learning [8]. GPT3's parameter scale is over a hundred times that of GPT2, reaching an astonishing 175 billion, while the training data expanded a thousandfold to 45TB. From a performance perspective, GPT3 can generate news articles that are difficult for humans to distinguish as being generated by a model. However, from the perspective of safety and other aspects, GPT3 still has numerous issues: GPT3 cannot guarantee the correctness of its output and may generate negative or even harmful information. In 2022, OpenAI combined RLHF (Reinforcement Learning from Human Feedback) [9] with GPT3 and proposed InstructGPT [10]. RLHF technology can help models better understand human instructions and ensure that the generated content is useful and harmless.

B. Prompt Tuning

When fine-tuning for downstream tasks, there may be cases where the gap between the downstream task objective and the pre-training objective is too large, resulting in insignificant training effects. To address this, GPT3 proposed a fine-tuning paradigm called Prompt-Tuning [8].

So far, three Prompt techniques have been proposed and proven to have good effects: In-Context Learning (ICL), Instruction Fine-tuning (IFT), and Chain-of-Thought (CoT). In May 2020, OpenAI first introduced the concept of In-Context Learning in GPT3, which selects a small number of labeled samples from the training set and designs task-relevant instruction templates to guide the generation of corresponding results for test samples. However, this method suffers from high variance and instability. In October 2021, Google released FLAN [11] and proposed Instruction Fine-tuning. The data for IFT is typically a collection of human-written instructions and instruction instances guided by language models. These instruction data consist of three main components: instruction, input, and output. For a given instruction, there can be multiple input and output instances. To enhance the ability of large models to solve mathematical reasoning problems, in 2022, Google released LAMDA (137B) [12] and introduced the Chain-of-Thought mechanism. By providing the model with reasoning step prompts, the model learns to think and reason step by step like humans, enabling it to possess basic reasoning capabilities and ultimately solve simple or even relatively complex mathematical problems.

C. Generating Military Question-Answering Systems Using Knowledge Graphs

Currently, most information generation methods applied in military question-answering systems use intelligent algorithms to realize the mapping from concept models to simulation scenarios. For example, knowledge graph techniques [13] are used to complete the generation of simulation scenarios; by representing concepts of weapons and equipment and combat actions in the combat domain, a domain knowledge base is

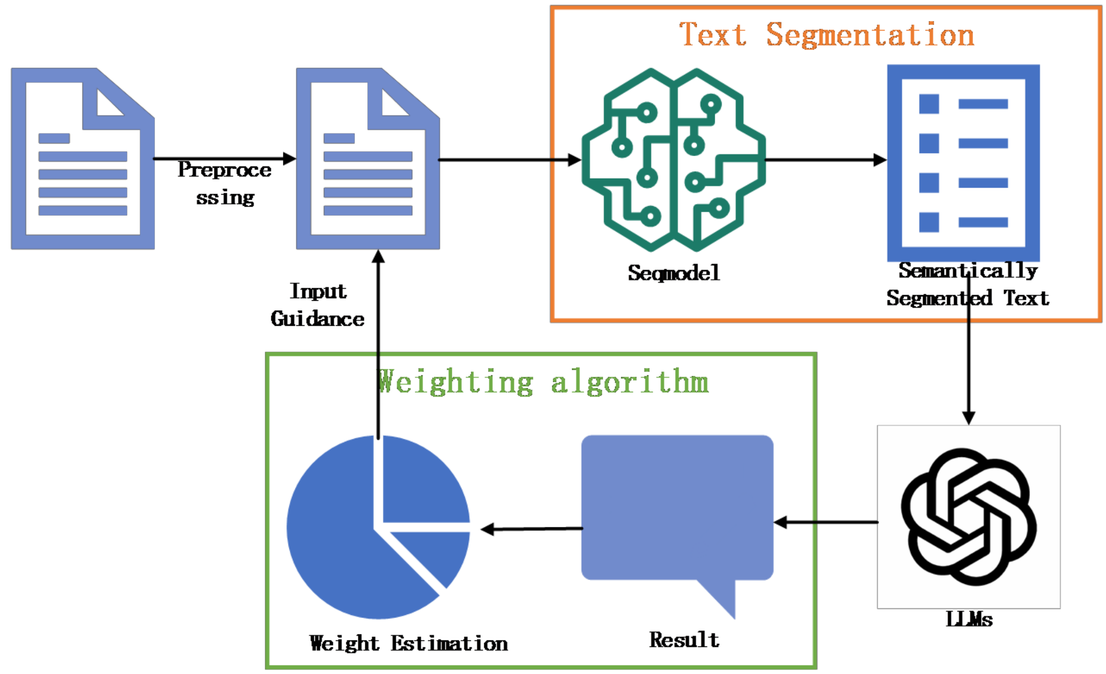

Fig. 2. Flowchart of experimental method

established, and military question-answering text is generated using intelligent mapping algorithms and semantic reasoning methods based on semantic similarity and field similarity [14]; XMLSchema is used to compose the elements of military question-answering deduction models, Backus-Naur Form grammar is used to establish formatted description templates for each component element, and military question-answering text is generated through mapping [15], or programs are used to load electronic nautical charts and military symbol libraries to describe the battlefield situation and combat tasks on a two-dimensional plane, further generating military information text [16].

V. Prompt Optimization Method

This paper designs a Prompt optimization method for large language models with long text input. By detecting the influence weights of each semantic segment in the input long text on the output of the large model, it guides users to precisely modify the input content to obtain the desired output. The flowchart of this method is shown in Figure 2.

A. Semantic Segmentation of Long Text

In this method, the collected text data needs to be preprocessed first to remove parts of the text data that do not meet the input requirements.

The preprocessed text data is input into the Seqmodel [17] for semantic segmentation. Seqmodel is a semantic segmentation model based on the BERT architecture that can effectively utilize contextual information for precise text segmentation. Compared with traditional methods, Seqmodel can simultaneously process more sentences and model longer

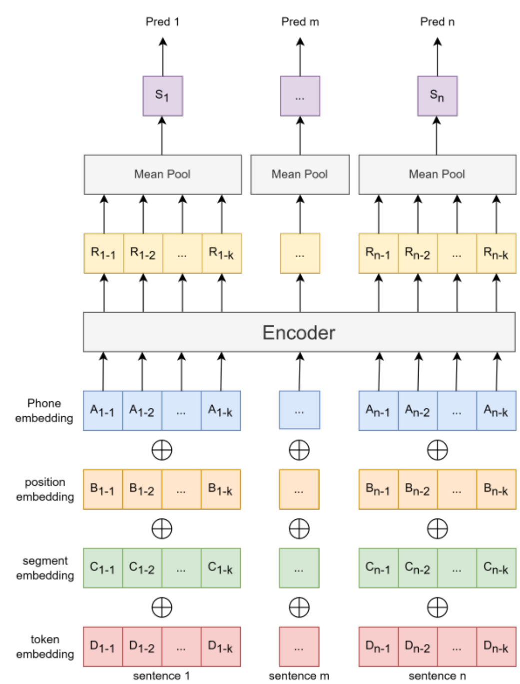

Fig. 3. Seqmodel Model Architecture [17]

context and dependencies between sentences through the self-attention mechanism. The architecture of the Seqmodel is shown in Figure 3.

B. Semantic Segment Influence Weight Algorithm

The long text after semantic segmentation is input into the large language model to obtain the output results. Then, through manual evaluation, the influence weights of different semantic segments on the output are calculated to help users more efficiently modify the input to obtain the desired output.

Specifically, this paper uses a weight calculation method based on keyword hits. The calculation method is as follows: For the i-th semantic segment of the long text input, i_n keywords appearing in it are manually selected, and each occurrence of these keywords in the output is recorded as a hit. Define the total number of hits of the keywords in the i-th input semantic segment in the output text as $Shot_i$. In a long text input composed of n semantic segments, the output influence weight of the i-th input semantic segment is $\frac{Shoti}{\sum i=1^n Shot_i}$. By sorting the obtained influence weights, the input semantic segments with the greatest influence on the output can be determined.

VI. Experiment

Using the battle mentioned in the composition of the military question-answering system in Chapter 2 as an example, it involves parts such as the combat background, objectives, force composition, combat preparations, basic tactics of each party, combat plans, combat actions, etc. The content involved in these components overlaps. To clarify the format of the military information text in the experiment, this paper defines its format before using the large language model to generate the military information text. This paper summarizes the above parts into the following six parts: combat background, force deployment, combat objectives, combat plan, combat process, and combat results. The summarized mapping relationship is shown in Figure 4.

A. Experimental Data and Preprocessing

In terms of experimental data, this paper obtained 30 battle-related texts from forums. Since the combat plan, combat process, and combat results are the generated content of the large language model and cannot be used as input, these contents are removed during the preprocessing stage.

B. Experimental Process

The preprocessed text data is input into the Seqmodel for semantic segmentation.

The segmented text by Seqmodel is input into the large language model to obtain the output results of the large language model.

Finally, the influence weights of different input semantic segments on the generation of battle text are calculated according to the output influence weight algorithm defined in this paper.

C. Algorithm Effectiveness Verification

This paper verifies the effectiveness of the proposed algorithm by modifying the input semantic segments with higher influence weights. Specifically, new content is added to the input semantic segments with larger weights, and the changes in the model-generated results are observed.

If the generated results of the large language model do not contain content related to the added input information, it cannot be determined whether the added information in the input semantic segment is utilized by the large language model. This paper defines the modification of the input semantic segment in this case as an invalid modification.

If the generated results of the large language model only contain content related to the added input information but do not generate new content related to the battle but unrelated to the added information, this paper defines the modification of the input semantic segment in this case as a non-important modification.

If the generated results of the model not only contain content related to the added input information but also generate new content related to the battle but unrelated to the added information, this paper defines the modification of the input semantic segment in this case as an important modification.

If the modifications made to the input semantic segments with higher influence weights during the experiment are always important modifications, while the modifications made to the input semantic segments with lower influence weights are always non-important modifications or even invalid modifications, it can be considered that the algorithm proposed in this paper is effective.

VII. Experimental Results

Through experiments on 30 battle texts, the experimental results are obtained as shown in Table I.

TABLE I
EXPERIMENTAL RESULTS

Semantic Segment Name	Influence Weight
Combat Objectives	47%
Force Deployment	38%
Combat Background	15%

The experimental results show that the two semantic segments of combat objectives and force deployment are relatively important for using large models to generate battle text. Scenario designers can improve the quality of the generated battle text by focusing on describing the combat objectives and force deployment parts.

This paper provides a comparison example of combat objectives. Figure 5 shows the generated results without a detailed description of the combat objectives semantic segment.

Only the combat objectives semantic segment is modified, while the other input semantic segments remain unchanged. The generation effect is shown in Figure 6.

The generated results in Figure 6 not only contain content related to the added input information but also generate new content related to the battle but unrelated to the added information, such as electronic warfare, which aligns with the definition of important modification in Chapter 5 of this paper.

VIII. Conclusion

This paper proposes a Prompt optimization method for large language models with long text input, aiming to help users better utilize large models to generate desired text by detecting the influence weights of different semantic segments in the input on the output of the large model. At the same time,

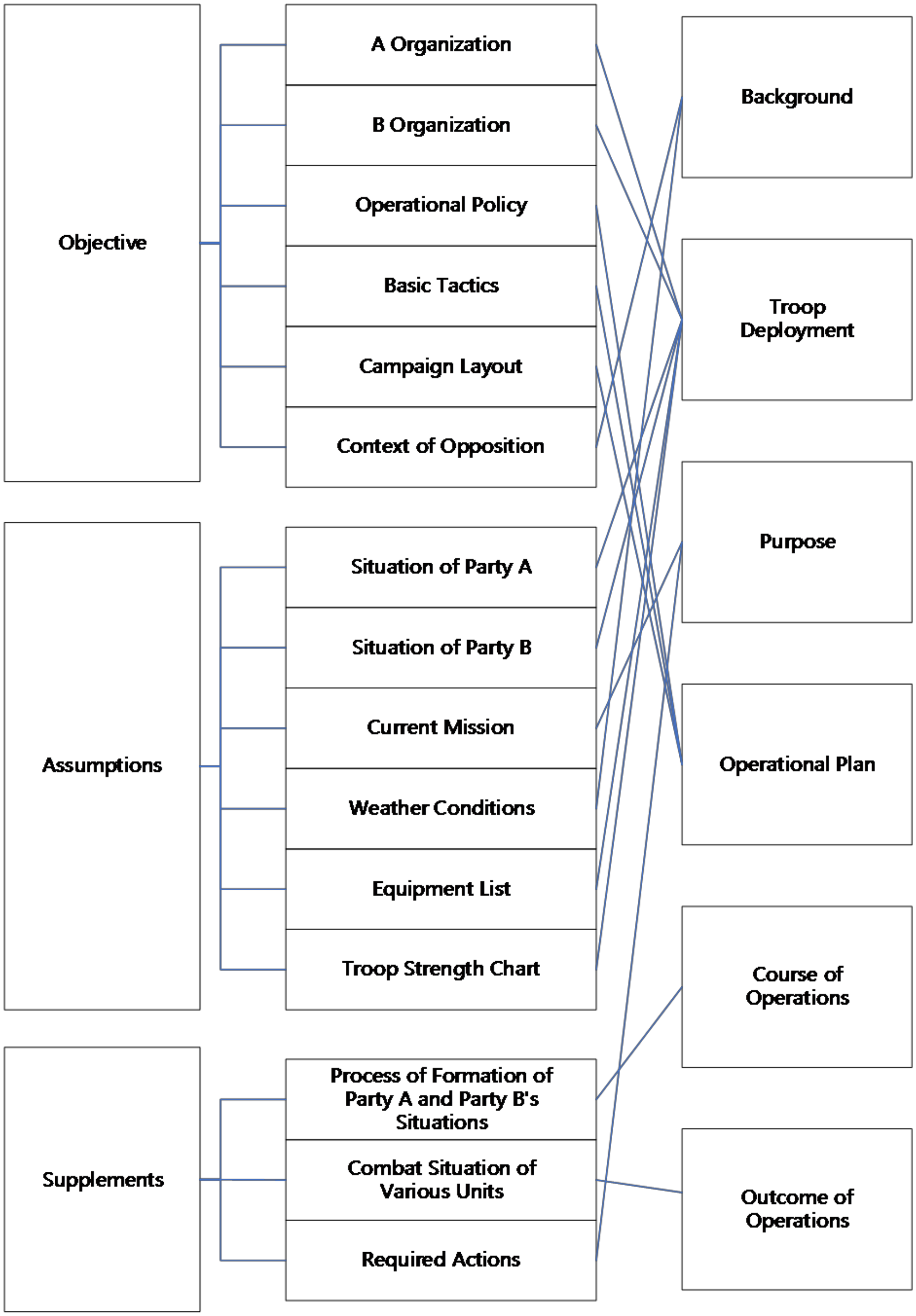

Fig. 4. Mapping maps for the definition of military ideograms

Input:
Please generate a military scenario based on the following context:
Scenario Context:
In 2030, a conflict erupted between the United States and Russia, leading to Russia launching a comprehensive ballistic missile assault on the United States and its Pacific allies. This offensive caused substantial damage to almost all U.S. military installations, resulting in the destruction of hundreds of aircraft and numerous naval vessels at both ground facilities and harbors. In response, the U.S. mobilized its available aircraft, deploying long-range air force capabilities to the newly restored Misawa Air Base. Before hostilities commenced, three U.S. Marine Corps F-35Bs were pre-positioned at a dispersed site in Japan, surviving the initial ballistic missile onslaught. Additionally, Australia has reallocated its refueling and Intelligence, Surveillance, and Reconnaissance (ISR) assets to the Royal Australian Air Force Base in Darwin to support its allies.

Objective:
As the commander of the United States Air Force's long-range capabilities, the operational objective is to employ our aerial forces to detect, locate, and annihilate the Kuznetsov aircraft carrier strike group traversing the Sea of Japan.

U.S. Forces:

Long-range strike capabilities situated at Misawa Air Base, Japan, comprising:
The 90th Fighter Squadron (6 x F-22A)
The 355th Fighter Squadron (6 x F-35A)
The 23rd Bomber Squadron (2 x B-52H)
The 37th Bomber Squadron (2 x B-1B)
The 517th Transport Squadron (2 x C-17 [Globemaster III])
The 168th Air Refueling Squadron (6 x KC-135R)
U.S. Marine Corps fighter detachment:
Marine Fighter Attack Squadron 121 (3 x F-35B)
Marine Aerial Refueler Transport Squadron 152 (1 x KC-130J)
Royal Australian Air Force (RAAF), positioned at Darwin Royal Air Force Base:

The 11th Squadron (2 x P-8A)
The 33rd Squadron (3 x KC-30A)
The 92nd Wing (2 x MQ-4C)

Output:
Operational Plan:
Intelligence Analysis:
Intelligence has identified that the Kuznetsov aircraft carrier strike group is navigating through the Sea of Japan. Our mission is to leverage the United States Air Force's long-range capabilities to discover, pinpoint, and destroy this enemy strike group, thereby diminishing Russia's maritime threat and bolstering our allies' defense lines.

Operational Objectives:

To utilize our aerial forces for intelligence, reconnaissance, and strike operations to obliterate the Kuznetsov aircraft carrier strike group.
To maximize the protection of our air assets against the enemy's air defense systems.
To collaborate with allies for intelligence and resource sharing, implementing a comprehensive operational strategy.
Operational Phases:

Intelligence, Surveillance, and Target Acquisition:

The RAAF's 11th Squadron (2 x P-8A) will conduct maritime reconnaissance and intelligence gathering in the Sea of Japan to ascertain the precise location and trajectory of the enemy strike group.
The RAAF's 92nd Wing (2 x MQ-4C) will provide high-altitude reconnaissance and target acquisition support.
Air Strike Preparation and Deployment:

The USAF's 168th Air Refueling Squadron (6 x KC-135R) will ensure the refueling and operational continuity of the long-range strike forces.
The 517th Transport Squadron (2 x C-17 [Globemaster III]) will transport critical supplies and equipment.
Air Strike Operations:

The 90th Fighter Squadron (6 x F-22A), the 355th Fighter Squadron (6 x F-35A), and Marine Fighter Attack Squadron 121 (3 x F-35B) will form the vanguard air strike team, exploiting air superiority and stealth capabilities for precision strikes against the enemy strike group.
The 23rd and 37th Bomber Squadrons (2 x B-52H and 2 x B-1B, respectively) will carry out strategic bombing missions, targeting the Kuznetsov aircraft carrier and its escort vessels for extensive bombardment.

Fig. 5. Military Identified Text Generation Results I

Input:
Objective:
As the commander of the United States Air Force's long-range forces, the operational objective is to employ our aerial capabilities to detect, locate, and destroy the Kuznetsov aircraft carrier strike group navigating the Sea of Japan. Prior to the commencement of hostilities, the Kuznetsov aircraft carrier strike group was already deployed to the Sea of Japan, supported by a Surface Action Group (SAG) to bolster the Kuznetsov carrier battle group. The primary target is the Kuznetsov aircraft carrier itself, with secondary targets being its escort vessels and the Surface Action Group. Given the likelihood that the forces deployed at Misawa Air Base have not gone unnoticed by Russia, it is imperative to launch your strike within six hours to preempt a potential preemptive ballistic missile strike by Russia.

Output:
Military Scenario: Operation Plan for the Destruction of the Kuznetsov Aircraft Carrier Strike Group

1.Intelligence Collection and Analysis
Initially, it is crucial to reestablish reconnaissance and intelligence gathering capabilities to compensate for the satellite reconnaissance systems disabled by cyber attacks. The Royal Air Force's 11th Squadron, equipped with two P-8A Anti-Submarine Warfare Patrol aircraft, and the 92nd Wing, operating two MQ-4C Unmanned Reconnaissance Drones, will undertake reconnaissance and intelligence gathering missions to search for and locate the Russian Navy fleet. They will utilize radar, electronic reconnaissance systems, and other sensors to monitor the Sea of Japan.

2. Air Combat Support
The 517th Transport Squadron, with its two C-17 transport aircraft, will be responsible for delivering ammunition, fuel, and other essential supplies to Misawa Air Base. Concurrently, the 168th Air Refueling Squadron's six KC-135R tankers will provide aerial refueling support to ensure the range and endurance of all our combat aircraft.
3. Aerial Combat Operations
Upon reconnaissance and location of the Kuznetsov aircraft carrier strike group, aerial strike operations will commence. This combat operation will be jointly executed by the United States Air Force's long-range strike capabilities and the Marine Corps fighter detachment.
The 90th Fighter Squadron's six F-22A stealth fighters will be tasked with securing air superiority, ensuring our combat aircraft operate beyond the enemy's reach.
The 355th Fighter Squadron's six F-35A fighters will carry out precision strikes against the Kuznetsov aircraft carrier, leveraging their advanced stealth and missile systems.
The 23rd Bomber Squadron's two B-52H strategic bombers and the 37th Bomber Squadron's two B-1B bombers will conduct long-range strikes against the Kuznetsov aircraft carrier and its escort vessels, utilizing their extensive strike range and payload capacity.
The Marine Fighter Attack Squadron 121's three F-35B fighters will be armed with anti-ship missiles for air-to-ground and anti-ship strike missions. They will exploit their vertical/short takeoff and landing capabilities for rapid deployment and return.
4. Electronic Warfare and Air Defense
While executing aerial strike missions, electronic warfare measures will be employed to disrupt the Russian fleet's communication and radar systems, weakening the enemy's countermeasures. Additionally, the United States Air Force will establish air defense cover, using the stealth capabilities and advanced missile systems of the F-22A and F-35A to protect our aircraft from enemy air defense threats.

Fig. 6. Military Intended Text Generation Results II

this paper verifies the effectiveness of the method through experiments on generating battle text.

The method proposed in this paper also has potential application value in other fields besides battle text generation. By determining the influence weights of different semantic segments in the input, users can better understand the importance of various parts of the input text, thereby improving the effect of generated text. In future work, the semantic segmentation method can be improved to achieve finer-grained segmentation. Furthermore, more efficient algorithms can be proposed to calculate the influence weights of different semantic segments. In summary, this research provides new ideas and methods for the application of large language models in the field of text generation and promotes further research and development in related fields.

REFERENCES

[1] Wang X.M. and Li X.L. Research on Key Technologies of Forum-Oriented Question-Answering Systems. Computer Systems & Applications, 2023, 32(2): 12-15.

[2] Liu X.M. Exploration of Military Knowledge Graph Construction Methods. Journal of Intelligence, 2023, 35(1): 5-10.

[3] Gray M. Deep Semantic Matching Technology in Big Data Environment. Journal of Software, 2023, 34(3): 405-412.

[4] Radford A., Narasimhan K., Salimans T., et al. Improving language understanding by generative pre-training. 2018.

[5] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In Advances in neural information processing systems 2017, 30.

[6] Devlin J., Chang M.W., Lee K., et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[7] Radford A., Wu J., Child R., et al. Language models are unsupervised multitask learners. OpenAI blog, 2019, 1(8): 9.

[8] Brown T., Mann B., Ryder N., et al. Language models are few-shot learners. In Advances in neural information processing systems 2020, 33: 1877-1901.

[9] Christiano P.F., Leike J., Brown T., et al. Deep reinforcement learning from human preferences. In Advances in neural information processing systems 2017, 30.

[10] Ouyang L., Wu J., Jiang X., et al. Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems 2022, 35: 27730-27744.

[11] Wei J., Bosma M., Zhao V.Y., et al. Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652, 2021.

[12] Thoppilan R., De Freitas D., Hall J., et al. Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239, 2022.

[13] Ge B., Tan Z., Zhang Y., et al. Research on Military Knowledge Graph Construction Technology. Journal of Command and Control, 2016, 2(4): 302-308.

[14] Tian X.Y., Zeng G.X., Gao Y.B., et al. Research on Model Reuse Technology Based on Semantic Matching and Combination. Journal of System Simulation, 2021, 33(12): 1-10.

[15] Xiao B., Wu J.P. A Formalized Description Method for Simulation Deduction Scenarios and an Instantiation Method for Deduction Models. Patent: 202310281676, 2023-09-08.

[16] Hou G.C. and Yang L. Research on Generation and Application Technology of Simulation Deduction Scenarios for Naval Battles. Ship Electronic Engineering, 2019, 39(7): 4.

[17] Zhang Q., Chen Q., Li Y., et al. Sequence Model with Self-Adaptive Sliding Window for Efficient Spoken Document Segmentation. In IEEE ASRU 2021.