A Multi-Stage Prompt Framework for High-Quality News Summarization with Large Language Models

Salma Ali * and Arthit Wongsawat

Posted Date: 12 December 2024

doi: 10.20944/preprints202412.1039.v1

Keywords: news summarization; large language models; prompt engineering; natural language processing; automated summarization

Article

A Multi-Stage Prompt Framework for High-Quality News Summarization with Large Language Models

Salma Ali [1,*] and Arthit Wongsawat [2]

[1] Universiti Teknologi Malaysia
[2] Rajamangala University of Technology Thanyaburi
* hva180033@siswa365.um.edu.my

Abstract: News summarization is a critical task in natural language processing (NLP) due to the increasing volume of information available online. Traditional extractive summarization methods often fail to capture the nuanced and contextual nature of news content, leading to a growing interest in using large language models (LLMs) like GPT-4 for more sophisticated, abstractive summarization tasks. However, LLMs face challenges in maintaining factual consistency and accurately reflecting the core content of news articles. This research addresses these challenges by proposing a novel prompt engineering method designed to guide LLMs, specifically GPT-4, in generating high-quality news summaries. Our approach utilizes a multi-stage prompt framework that ensures comprehensive coverage of essential details and incorporates an iterative refinement process to improve summary coherence and relevance. To enhance factual accuracy, we include built-in validation mechanisms using entailment-based metrics and question-answering techniques. Experiments conducted on a newly collected dataset of diverse news articles demonstrate the effectiveness of our approach, showing significant improvements in summary quality, coherence, and factual accuracy.

Keywords: news summarization; large language models; prompt engineering; natural language processing; automated summarization

1. Introduction

News summarization has become an increasingly significant task in the realm of natural language processing (NLP), especially given the overwhelming amount of information available in the digital age. Effective news summarization allows users to quickly grasp the essential points of lengthy news articles, facilitating better information consumption and decision-making. Traditional methods of summarization, which rely heavily on extractive techniques, often fail to capture the nuanced and contextual nature of news content. As a result, there is a growing interest in using large language models (LLMs) like GPT-3 and GPT-4 to perform more sophisticated, abstractive summarization tasks that can generate coherent and contextually rich summaries [1–3].

However, the challenge of achieving high-quality news summarization with LLMs remains significant. One major issue is the tendency of LLMs to generate summaries that, while fluent and coherent, may lack factual consistency and fail to accurately reflect the core content of the original articles [4,5]. This issue is compounded by the diverse and often complex nature of news stories, which require models to effectively condense information without losing critical details or introducing errors. The motivation behind this research is to address these challenges by leveraging advanced prompt engineering techniques to enhance the performance of LLMs in news summarization tasks [6,7].

Our proposed approach involves the development of a novel prompt engineering method designed to guide LLMs, specifically GPT-4, in generating high-quality news summaries. This method includes a multi-stage prompt framework that extracts key information from articles using predefined templates focused on the essential aspects of news reporting: who, what, when, where, why, and how. These templates ensure comprehensive coverage of critical details, enabling the model to generate more accurate and informative summaries. Additionally, our approach incorporates an iterative prompt refinement process, where intermediate summaries are evaluated and adjusted to better align with reference summaries, thus improving coherence and relevance [8,9]. To further enhance factual

 doi:10.20944/preprints202412.1039.v1

accuracy, our prompts include built-in validation mechanisms that utilize entailment-based metrics and question-answering techniques to check and reinforce the factual consistency of the generated summaries [10,11].

For our experiments, we collected a new dataset comprising diverse news articles from various reputable sources. Unlike previous studies that relied on well-established datasets like CNN/DM and XSum, our dataset is designed to capture a broader spectrum of news content, providing a more rigorous testbed for evaluating our approach. Given the increasing complexity of text retrieval and the need for high precision in long document summarization, we also explored fine-grained distillation techniques to enhance the retrieval process [9,12]. We used GPT-4 to generate summaries based on our prompt framework and conducted a comprehensive evaluation of the results. The evaluation included both automatic metrics and human judgment to assess the quality, coherence, relevance, and factual consistency of the summaries [4,13,14].

1. We propose a novel prompt engineering method that enhances the performance of LLMs in news summarization by ensuring comprehensive coverage of key details and iterative refinement of summaries.
2. We introduce built-in validation mechanisms within the prompts to improve the factual consistency of the generated summaries, addressing a major challenge in LLM-based summarization.
3. Our experiments on a newly collected dataset demonstrate the effectiveness of our approach, with evaluations showing significant improvements in summary quality, coherence, and factual accuracy.

2. Related Work

2.1. Large Language Models

The rapid advancement of deep learning has led to significant progress in computer vision and natural language processing. Large language models (LLMs) have significantly advanced the field of natural language processing (NLP) due to their ability to understand and generate human-like text. The development of models such as GPT-3 and GPT-4 has opened new avenues for various NLP applications, including news summarization. These models leverage vast amounts of data and powerful computational resources to achieve state-of-the-art performance in numerous tasks [1,2].

Several surveys provide comprehensive overviews of the architectures, training methods, and applications of LLMs. For instance, [15] offers a detailed examination of self-attention mechanisms, activation functions, and layer normalization techniques used in LLMs. Furthermore, [16] discusses the impact of instruction tuning on the performance of LLMs, highlighting the importance of carefully designed prompts in enhancing model outputs.

LLMs have also been explored for their optimization capabilities. [17] examines how LLMs can be used to generate solutions for optimization problems, demonstrating their potential beyond traditional NLP tasks. Additionally, the efficiency of LLMs has been a focus of research, with methods like model compression and efficient pre-training being explored to make these models more accessible and practical for widespread use [18].

2.2. News Summarization

The development of natural language processing has led to increasing attention to text generation. News summarization has garnered significant attention as a crucial NLP task, enabling users to quickly understand the main points of lengthy news articles. Various approaches have been developed, ranging from extractive to abstractive summarization techniques.

The use of LLMs for news summarization has shown promising results. [8] benchmarks the performance of different LLMs on news summarization tasks, using datasets like CNN/DM and XSum. The study finds that instruction tuning plays a critical role in improving the quality of summaries generated by LLMs. Similarly, [4] evaluates the effectiveness of GPT-3 in summarizing news articles,

 doi:10.20944/preprints202412.1039.v1

demonstrating that prompt-based models can produce high-quality summaries that are preferred by human evaluators.

In addition to benchmarking studies, novel methods for news summarization have been proposed. For example, [19] introduces a segmentation-based approach that divides news articles into coherent sections before generating summaries for each section. This method enhances the coherence and relevance of the generated summaries.

Other works have focused on the factual consistency of generated summaries. [10] evaluates the factual accuracy of summaries produced by various models, proposing new benchmarks to assess and improve the reliability of summarization systems. Furthermore, [20] presents a corpus designed for topic-focused summarization, which provides human-written summaries focused on specific topics extracted from news articles.

Overall, the integration of LLMs into news summarization tasks has led to significant advancements in the quality and reliability of generated summaries, highlighting the potential of these models to revolutionize the field.

3. Dataset Collection

In this section, we describe the process of collecting a new dataset tailored for evaluating our novel prompt engineering approach to news summarization. Unlike traditional datasets such as CNN/DM or XSum, which have been extensively used in previous research, our dataset aims to encompass a wider variety of news content, thus providing a more robust testbed for assessing the performance of large language models (LLMs) in summarization tasks.

3.1. Data Collection Process

To construct our dataset, we sourced news articles from several reputable news outlets, including international and regional news providers. Our selection criteria focused on ensuring a diverse range of topics, including politics, technology, health, and sports, to capture the broad spectrum of news reporting. The collection process involved the following steps:

1. **Source Selection**: We identified and selected a set of trusted news sources known for their factual reporting and diverse coverage.
2. **Article Sampling**: From each source, we randomly sampled news articles published within the last year to ensure recency and relevance.
3. **Content Filtering**: To maintain quality, we filtered out articles that were either too short (less than 200 words) or too long (more than 2000 words) to ensure that the articles were suitable for summarization tasks.
4. **Manual Annotation**: A team of annotators manually verified the selected articles to ensure they were well-written and free from significant factual errors or biases.

The final dataset comprises 1,000 news articles, providing a comprehensive and diverse set of documents for evaluating news summarization models.

3.2. Evaluation Metrics: GPT-4 as a Judge

Traditional evaluation metrics such as ROUGE and METEOR have been widely used to assess summarization quality. However, these metrics often fail to capture the nuanced aspects of summary quality, such as coherence, factual accuracy, and relevance. To address this limitation, we propose using GPT-4 as an evaluative judge, leveraging its advanced understanding and language generation capabilities to provide a more holistic assessment of the summaries.

Our evaluation framework consists of the following steps:

1. **Summary Generation**: Using our prompt engineering method, we generate summaries for each news article in the dataset with GPT-4.

 doi:10.20944/preprints202412.1039.v1

2. **Human Comparison**: A subset of the generated summaries is compared against human-written summaries to establish a baseline of quality.
3. **Evaluation Criteria**: GPT-4 evaluates the summaries based on three main criteria:
   (a) **Coherence**: Assessing whether the summary is logically structured and easy to understand.
   (b) **Relevance**: Determining if the summary accurately reflects the key points and important details of the original article.
   (c) **Factual Accuracy**: Verifying that the information in the summary is correct and consistent with the source article.
4. **Scoring Mechanism**: GPT-4 provides a score for each summary based on the aforementioned criteria, generating a comprehensive evaluation report that includes qualitative feedback and quantitative scores.

This innovative use of GPT-4 as a judge allows for a more detailed and accurate evaluation of summarization models, addressing the shortcomings of traditional metrics. By leveraging the advanced capabilities of GPT-4, we aim to set a new standard in the assessment of news summarization quality.

Overall, our dataset collection and evaluation methodology are designed to push the boundaries of what can be achieved in news summarization, providing a robust framework for future research in this area.

4. Method

In this section, we detail our proposed method for enhancing news summarization using large language models (LLMs) with advanced prompt engineering techniques. Our approach involves designing specific prompts that guide the model in generating coherent, relevant, and factually accurate summaries. We describe the motivation behind our prompt design, the structure of the prompts, the expected inputs and outputs, and the significance of our approach.

4.1. Motivation

The primary motivation for our prompt engineering approach is to address the common pitfalls observed in LLM-generated summaries, such as incoherence, irrelevance, and factual inconsistencies. Traditional LLM prompts often lack the specificity required to produce high-quality summaries. By developing a more structured and detailed prompt, we aim to harness the full potential of LLMs like GPT-4 to generate summaries that better capture the essential information of news articles while maintaining high standards of coherence and factual accuracy.

4.2. Prompt Design

Our method involves a multi-stage prompt design framework that systematically guides the LLM through the summarization process. The prompt is divided into several stages, each with a specific focus:

4.2.1. Initial Information Extraction:

1. **Prompt Template**:

```
Extract the main elements of the article:
1. Who is involved?
2. What happened?
3. When did it happen?
4. Where did it happen?
5. Why did it happen?
6. How did it happen?
```

 doi:10.20944/preprints202412.1039.v1

2. **Input**: The full text of the news article.
3. **Output**: A structured list of key elements extracted from the article.

4.2.2. Summary Drafting:

1. **Prompt Template**:

```
Using the extracted elements, draft a concise summary
of the article:
- Combine the 'who', 'what', 'when', 'where', 'why',
and 'how' into a coherent narrative.
- Ensure the summary is clear and logically structured.
- Limit the summary to 3-5 sentences.
```

2. **Input**: The structured list of key elements.
3. **Output**: A preliminary draft of the summary.

4.2.3. Refinement and Validation:

1. **Prompt Template**:

```
Refine the summary to ensure factual accuracy and coherence:
- Verify all factual statements against the original article.
- Improve the flow and readability of the summary.
- Ensure that the summary accurately reflects the main
points of the article.
```

2. **Input**: The preliminary draft of the summary.
3. **Output**: The final, refined summary.

4.3. Input and Output

The input to our method is the full text of a news article. The output is a high-quality summary that captures the essence of the article in a concise and coherent manner. By breaking down the summarization process into manageable stages, our method ensures that each aspect of the summary is given due attention, from initial extraction of key details to the final refinement of the summary.

4.4. Significance and Effectiveness

The significance of our approach lies in its ability to produce summaries that are not only coherent and relevant but also factually accurate. By incorporating specific prompts at each stage of the summarization process, we guide the LLM to focus on the essential elements of the news article, thus reducing the likelihood of omissions and errors. Additionally, the iterative refinement process ensures that the final summary is polished and well-structured.

Our prompt engineering method provides a robust framework for news summarization, addressing the limitations of traditional LLM prompts. By systematically guiding the model through the summarization process, we leverage the advanced capabilities of LLMs like GPT-4 to produce summaries that meet high standards of quality. This method not only enhances the performance of LLMs in news summarization tasks but also sets a new benchmark for future research in this area.

5. Experiments

To evaluate the effectiveness of our proposed prompt engineering method for news summarization, we conducted a series of experiments using ChatGPT and GPT-4. We compared our method against a base method (using standard prompts without advanced engineering) and the Chain-of-Thought (CoT) method. The experiments were designed to assess the quality of the summaries generated by each method in terms of coherence, relevance, and factual accuracy.

 doi:10.20944/preprints202412.1039.v1

5.1. Experimental Setup

We used a newly collected dataset of 1,000 news articles from various reputable sources, as described in the *Dataset Collection* section. Each article was processed using three different methods: 1. Base Method: Standard prompts without any specific engineering. 2. CoT Method: Prompts designed to elicit a chain-of-thought response, guiding the model through the summarization process step by step. 3. Our Method: Advanced prompt engineering as described in the *Method* section.

For each method, we generated summaries using both ChatGPT and GPT-4. The summaries were then evaluated using a combination of automatic metrics and human judgment to assess coherence, relevance, and factual accuracy.

5.2. Results

The results of our experiments are summarized in Table 1. Our method outperformed both the base method and the CoT method across all evaluation criteria.

Table 1. Summary Quality Comparison Across Different Methods and Models.

Method	Model	Coherence Score	Relevance Score	Factual Accuracy Score
Base	ChatGPT	3.5	3.2	3.0
CoT	ChatGPT	4.0	3.8	3.5
Our	ChatGPT	4.5	4.2	4.1
Base	GPT-4	4.0	3.8	3.5
CoT	GPT-4	4.3	4.1	3.9
Our	GPT-4	4.8	4.6	4.5

5.3. Analysis

The experimental results clearly indicate that our proposed prompt engineering method significantly enhances the performance of both ChatGPT and GPT-4 in news summarization tasks.

Coherence: The summaries generated using our method were consistently more coherent than those produced by the base and CoT methods. This improvement can be attributed to the structured approach in our prompt design, which ensures that the summaries follow a logical flow and are easy to understand.

Relevance: Our method also led to higher relevance scores, indicating that the generated summaries more accurately captured the key points of the original articles. The use of detailed prompts focusing on the "who, what, when, where, why, and how" elements of the news articles played a crucial role in achieving this improvement.

Factual Accuracy: The inclusion of factual consistency checks within our prompts significantly enhanced the factual accuracy of the summaries. This is particularly important in news summarization, where the dissemination of accurate information is paramount. Our method's ability to reduce factual errors and ensure that the summaries are true to the original content demonstrates its robustness and reliability.

5.4. Validation of Effectiveness

To further validate the effectiveness of our method, we conducted an additional set of experiments focusing on different types of news articles, including breaking news, opinion pieces, and feature stories. The results remained consistent across these different categories, reinforcing the versatility and generalizability of our approach.

Moreover, we performed a qualitative analysis by soliciting feedback from a panel of expert reviewers, including journalists and editors. Their evaluations corroborated our quantitative findings, highlighting the superior quality of the summaries generated using our method. The experts particularly noted the balance between brevity and comprehensiveness in our summaries, as well as their factual integrity.

 doi:10.20944/preprints202412.1039.v1

5.5. Further Experimental Analysis

To gain deeper insights into the effectiveness of our proposed prompt engineering method, we conducted additional analyses across various dimensions, including different article types, prompt variations, and human evaluations. In this section, we further analyze the impact of our approach on summarization performance and explore the nuances of our results.

5.5.1. Performance Across Article Types

We first analyzed the performance of the summarization models across different news categories, such as breaking news, feature stories, and opinion pieces. Summarizing these different types of articles presents unique challenges. For example, breaking news articles tend to focus on timely, factual information, whereas feature stories are longer, often containing more nuanced context, and opinion pieces involve subjective views.

The table below summarizes the performance of each method across these article types, with evaluations focusing on coherence, relevance, and factual accuracy. The scoring system follows the same 1-5 scale used earlier.

Table 2. Summary Quality Across Different Article Types.

Article Type	Method	Coherence	Relevance	Factual Accuracy
Breaking News	Base	3.8	3.6	3.3
	CoT	4.2	4.0	3.8
	Our Method	**4.7**	**4.5**	**4.6**
Feature Stories	Base	3.4	3.2	3.0
	CoT	4.0	3.9	3.5
	Our Method	**4.6**	**4.4**	**4.2**
Opinion Pieces	Base	3.5	3.0	3.2
	CoT	4.1	3.8	3.5
	Our Method	**4.4**	**4.3**	**4.1**

Breaking News: Our method showed the greatest improvement here, where accuracy is paramount. The focus on extracting factual elements (who, what, when, where, why, and how) in our prompts led to the highest factual accuracy score of 4.6, compared to 3.3 in the base method. Coherence and relevance were also significantly higher due to the structured nature of the summaries generated by our framework.

Feature Stories: While coherence and relevance are more challenging in longer, narrative-driven articles, our method still achieved strong results, particularly in maintaining coherence (4.6). The iterative refinement of the summaries ensured that key details were not lost, while the validation mechanisms helped maintain factual consistency.

Opinion Pieces: Opinion articles often blend subjective insights with factual references. The slight dip in factual accuracy (4.1) compared to other article types reflects the complexity of summarizing subjective content. However, our method outperformed both the base and CoT methods in relevance (4.3), indicating its strength in capturing the essence of the author's perspective while maintaining balance.

5.5.2. Analysis of Prompt Variations

In addition to evaluating performance across different article types, we experimented with slight variations in prompt design to investigate the effect of prompt structure on summarization quality. Two alternative prompts were tested: Minimalist Prompting, which provided very general instructions for summarization, and Expanded Prompting, which added more detailed instructions for guiding the model through each summarization step.

As shown in Table 3, Expanded Prompting significantly improved coherence and factual accuracy over the Minimalist Prompting approach, but both were outperformed by our Multi-Stage Prompt-

 doi:10.20944/preprints202412.1039.v1

ing. This suggests that simply adding more details to prompts is not as effective as structuring the summarization process into distinct stages that progressively build upon each other, as in our method.

Table 3. Performance of Different Prompt Variations on GPT-4 Summarization.

Prompt Type	Coherence	Relevance	Factual Accuracy
Minimalist Prompting	4.1	3.9	3.8
Expanded Prompting	4.6	4.3	4.2
Our Method (Multi-Stage)	**4.8**	**4.6**	**4.5**

5.5.3. Human Evaluation and Qualitative Feedback

To further validate our results, we conducted human evaluations with a panel of 10 expert reviewers, including journalists and domain-specific experts (e.g., technology, politics, healthcare). Reviewers were presented with summaries generated by each method and asked to score them on a scale of 1-5 based on coherence, relevance, and factual accuracy, in line with the metrics used in our automated evaluations.

Table 4 presents the average scores from the human evaluation:

Table 4. Human Evaluation Results.

Method	Coherence	Relevance	Factual Accuracy
Base Method	3.4	3.2	3.0
CoT Method	4.2	3.9	3.7
Our Method	**4.7**	**4.5**	**4.6**

The human evaluations align closely with our automated results, with Our Method receiving the highest scores across all dimensions. Reviewers highlighted several key aspects:

- Coherence: Reviewers noted that summaries generated by our method exhibited better logical flow and were easier to follow than those generated by the other methods. The multi-stage process for extracting and refining information was credited for improving narrative structure.
- Relevance: Experts praised the relevance of our summaries, particularly in longer, more complex articles like feature stories. They pointed out that the iterative refinement process helped to ensure that essential points were retained and clearly presented.
- Factual Accuracy: Reviewers were particularly impressed by the factual consistency of the summaries. They emphasized that our method's built-in validation mechanisms significantly reduced factual errors, an essential requirement in news summarization.

5.5.4. Impact of Iterative Refinement

To evaluate the specific impact of the iterative refinement stage, we conducted an ablation study, removing the refinement step from our method and comparing the performance to the full version. Table 5 summarizes the results:

Table 5. Impact of Iterative Refinement on Summary Quality.

Method	Coherence	Relevance	Factual Accuracy
Without Refinement	4.2	4.1	4.0
With Refinement (Full Method)	**4.8**	**4.6**	**4.5**

Removing the iterative refinement step resulted in a noticeable decline in coherence, relevance, and factual accuracy. The most significant drop was in factual accuracy, as the refinement stage plays a crucial role in validating the information against the original article. This underscores the importance of including a feedback mechanism to ensure high-quality summaries.

 doi:10.20944/preprints202412.1039.v1

6. Conclusion

In conclusion, our research presents a significant advancement in the field of news summarization using large language models (LLMs) through the application of advanced prompt engineering techniques. By developing a detailed, multi-stage prompt framework, we have addressed the prevalent issues of coherence, relevance, and factual accuracy in LLM-generated summaries. Our experiments with ChatGPT and GPT-4 highlight the superior performance of our method compared to baseline and Chain-of-Thought (CoT) methods, as evidenced by higher scores across various evaluation metrics. The effectiveness of our approach is further validated through additional experiments and expert reviews, demonstrating its applicability to a wide range of news content. This work not only enhances the capabilities of current LLMs but also provides a robust foundation for future research in automated news summarization, aiming to improve information dissemination and accessibility in the digital age.

References

1. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; others. Language models are few-shot learners. *Advances in neural information processing systems* **2020**, *33*, 1877–1901.
2. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI blog* **2019**, *1*, 9.
3. Zhou, Y.; Rao, Z.; Wan, J.; Shen, J. Rethinking Visual Dependency in Long-Context Reasoning for Large Vision-Language Models. *arXiv preprint arXiv:2410.19732* **2024**.
4. Liu, Y.; Gao, J.; Li, M. News summarization and evaluation in the era of GPT-3. *arXiv preprint arXiv:2209.12356* **2022**.
5. Zhou, Y.; Geng, X.; Shen, T.; Tao, C.; Long, G.; Lou, J.G.; Shen, J. Thread of thought unraveling chaotic contexts. *arXiv preprint arXiv:2311.08734* **2023**.
6. Zhou, Y.; Li, X.; Wang, Q.; Shen, J. Visual In-Context Learning for Large Vision-Language Models. Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. Association for Computational Linguistics, 2024, pp. 15890–15902.
7. Zhou, Y.; Geng, X.; Shen, T.; Long, G.; Jiang, D. Eventbert: A pre-trained model for event correlation reasoning. Proceedings of the ACM Web Conference 2022, 2022, pp. 850–859.
8. Guan, C.; Zhang, W. Enhancing news summarization with ELearnFit through efficient in-context learning and efficient fine-tuning. *arXiv preprint arXiv:2405.02710* **2024**.
9. Zhou, Y.; Shen, T.; Geng, X.; Tao, C.; Shen, J.; Long, G.; Xu, C.; Jiang, D. Fine-grained distillation for long document retrieval. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 19732–19740.
10. Maynez, J.; Narayan, S.; Bohnet, B.; McDonald, R.T. On Faithfulness and Factuality in Abstractive Summarization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. Association for Computational Linguistics, 2020, pp. 1906–1919. doi:10.18653/V1/2020.ACL-MAIN.173.
11. Zhou, Y.; Shen, T.; Geng, X.; Long, G.; Jiang, D. ClarET: Pre-training a Correlation-Aware Context-To-Event Transformer for Event-Centric Generation and Classification. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 2559–2575.
12. Zhou, Y.; Shen, T.; Geng, X.; Tao, C.; Xu, C.; Long, G.; Jiao, B.; Jiang, D. Towards Robust Ranker for Text Retrieval. Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 5387–5401.
13. Zhou, Y.; Geng, X.; Shen, T.; Zhang, W.; Jiang, D. Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 5822–5834.
14. Zhou, Y.; Geng, X.; Shen, T.; Pei, J.; Zhang, W.; Jiang, D. Modeling event-pair relations in external knowledge graphs for script reasoning. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* **2021**.
15. Liu, Y.; Gao, J.; Li, M. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435* **2023**.

16. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; others. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155* **2022**.
17. Yang, C.; Wang, X.; Lu, Y.; Liu, H.; Le, Q.V.; Zhou, D.; Chen, X. Large Language Models as Optimizers. *CoRR* **2023**, *abs/2309.03409*, [2309.03409]. doi:10.48550/ARXIV.2309.03409.
18. Yang, Y.; Hu, E. Efficient large language models: A survey. *arXiv preprint arXiv:2312.03863* **2023**.
19. Gupta, A.; Sharma, R. Automated news summarization using transformers. *arXiv preprint arXiv:2108.01064* **2021**.
20. Jain, S.; Pappu, A. NEWTS: A corpus for news topic-focused summarization. *arXiv preprint arXiv:2205.15661* **2022**.