Context-Aware and Task-Specific Prompting with Iterative Refinement for Historical Texts

Jingjing Zhang *

Posted Date: 8 October 2024

doi: 10.20944/preprints202410.0470.v1

Keywords: Iterative Refinement; Natural Language Processing

 doi:10.20944/preprints202410.0470.v1

Article

Context-Aware and Task-Specific Prompting with Iterative Refinement for Historical Texts

Iterative Refinement for Historical Texts

Jingjing Zhang * and Yangshu Lin

Zhoukou Normal University

* Correspondence: 20190121@stu.sqxy.edu.cn

Abstract: The advent of Large Language Models (LLMs) has significantly advanced natural language processing (NLP), yet their application to historical texts remains challenging due to archaic language, distinct terminologies, and varied contextual backgrounds. This study introduces Historical Domain Large Language Models, designed to bridge this gap by adapting LLMs for better comprehension and processing of historical data. Our approach leverages context-aware and task-specific prompts to enhance model performance in tasks such as named entity recognition (NER), sentiment analysis, and information extraction within historical contexts. We propose an iterative refinement process to improve prompt quality and model outputs continuously. Instruction tuning on newly collected evaluation data ensures our methods' efficacy, avoiding biases from previously used datasets. Evaluations using GPT-4 demonstrate significant improvements in handling historical texts, underscoring the potential of our approach to unlock profound insights from historical data. This work highlights the importance of tailored LLM adaptations for specialized domains, offering a robust framework for future research in historical NLP.

Keywords: iterative refinement; natural language processing

1. Introduction

In recent years, the development of Large Language Models (LLMs [1,2]) has revolutionized natural language processing (NLP [3,4]), enabling significant advancements across various domains. However, applying these models to historical texts presents unique challenges due to the archaic language, distinct terminologies, and varied contextual backgrounds inherent in such documents. Historical Domain Large Language Models aim to bridge this gap by adapting LLMs to better understand and process historical data, providing tools for tasks such as named entity recognition (NER), sentiment analysis, and information extraction in historical contexts [5].

The significance of this adaptation lies in the vast amount of historical data available, which, if properly analyzed, can offer profound insights into past events, social structures, and linguistic evolution. Traditional LLMs, however, often fall short when dealing with historical texts, primarily due to their training on contemporary language corpora that lack the specific nuances of historical language [6]. This leads to issues such as misinterpretation of context, failure to recognize outdated terminologies, and poor performance on specialized tasks [7].

Our motivation stems from the need to enhance the capabilities of LLMs in handling historical texts. Previous works, such as hmBERT for multilingual historical NER tasks [8], and efforts to extract information from historical well records using LLMs [9], have highlighted both the potential and the limitations of current approaches. By addressing these limitations, we aim to develop a more robust framework for historical domain language modeling.

To tackle these challenges, we propose a novel approach that combines LLMs with instruction tuning tailored for historical domains. The core of our method involves crafting detailed, context-aware prompts that guide the model in understanding the specific characteristics of historical data. This includes:

 doi:10.20944/preprints202410.0470.v1

1. **Contextual Prompts**: Providing background information relevant to the historical period and events, such as "Identify key entities in the following text from 18th-century France, considering the historical context of the French Revolution."
2. **Task-Specific Prompts**: Tailored for specific tasks like NER or sentiment analysis, for example, "Extract all named entities related to geographic locations from this 19th-century British text."
3. **Iterative Refinement**: Implementing a feedback loop where initial outputs are analyzed to refine the prompts further, ensuring continuous improvement in the model's performance.

We further enhance this approach by conducting instruction tuning on the data obtained using these prompts. This process involves fine-tuning the LLMs on the specially crafted prompts to better adapt to the intricacies of historical texts. Our experiments collect a new set of evaluation data specifically for this purpose, avoiding reliance on previously used datasets to ensure unbiased assessment. The results are then evaluated using GPT-4, providing a comprehensive analysis of the model's performance and the effectiveness of our prompting strategy.

1. We introduce a novel method for crafting context-aware and task-specific prompts to improve LLM performance on historical texts.
2. We propose an iterative refinement process to continuously enhance the quality of prompts and model outputs.
3. We demonstrate the effectiveness of instruction tuning using newly collected evaluation data, assessed with state-of-the-art LLMs, specifically GPT-4.

2. Related Work

In this section, we review the existing literature relevant to our research, focusing on two main areas: Large Language Models (LLMs) and Instruction Tuning. These areas form the foundation of our approach and provide the necessary background to understand the advancements and challenges in this domain.

2.1. Large Language Models

Large Language Models (LLMs) have revolutionized natural language processing [10,11] and computer vision [12,13] with their ability to understand and generate human-like text. The development of models such as GPT-3 and GPT-4 by OpenAI has demonstrated unprecedented capabilities in a wide range of tasks, from language generation to complex reasoning [1,6,14–17]. These models utilize the transformer architecture, which allows for efficient handling of long-range dependencies in text [18,19].

Several surveys and studies have explored the architecture and applications of LLMs. For instance, [5] provides a comprehensive overview of the opportunities and risks associated with foundation models, including LLMs. Another survey by [7] discusses the advancements in LLMs, highlighting their impact on various AI applications and the ongoing research to enhance their efficiency and performance.

Recent advancements have focused on improving the efficiency of LLMs through techniques like model compression, efficient pre-training, and fine-tuning [20]. Additionally, [21] introduced the T5 model, which explores the use of transfer learning and pre-training on diverse NLP tasks, setting a new benchmark for LLM performance.

2.2. Instruction Tuning

Instruction tuning is a technique used to improve the performance of LLMs by training them to follow specific instructions or prompts. This method has shown significant improvements in the ability of LLMs to perform zero-shot and few-shot learning tasks [22,23]. Instruction tuning involves fine-tuning models on datasets that contain instruction-like prompts, enabling them to better understand and execute complex tasks based on user instructions.

 doi:10.20944/preprints202410.0470.v1

A notable survey by [24] provides an extensive review of instruction tuning techniques, datasets, and their impact on LLM performance. The survey highlights key datasets such as FLAN [25] and Alpaca [26], which are designed specifically for instruction tuning. These datasets provide a diverse set of tasks and prompts, helping models to generalize better across various applications.

Recent studies have also explored the application of instruction tuning in multi-modal contexts. For example, [9] discuss visual instruction tuning, where models are trained to follow instructions related to visual data, improving their performance on tasks that require understanding both text and images. Another study by [27] investigates the use of instruction tuning for LLMs in scientific domains, showing how it can enhance the models' ability to understand and generate scientific text.

In summary, these works on LLMs and instruction tuning provides a robust foundation for our research. The advancements in these areas have paved the way for developing models that are more efficient, versatile, and capable of understanding complex instructions, which is crucial for tasks involving historical texts and other specialized domains.

3. Dataset

The success of our proposed approach hinges on the quality and relevance of the datasets used for instruction tuning and evaluation. In this section, we outline the process of dataset collection, detailing the sources, preprocessing steps, and the rationale behind the selection of data. Furthermore, we introduce a novel evaluation framework using GPT-4 as the judge, moving away from traditional metrics to a more holistic and context-aware assessment.

3.1. Instruction Tuning Dataset

For instruction tuning, we curated a dataset that captures the linguistic diversity and historical specificity required for training our LLMs. The dataset comprises historical texts from multiple periods and regions, ensuring a broad representation of linguistic styles and terminologies. Our primary sources include:

1. **Historical Archives**: Digitized archives from libraries and museums, including letters, diaries, and official documents from different historical periods.
2. **Newspaper Archives**: Collections of historical newspapers that provide rich contextual information and reflect the vernacular of their time.
3. **Literary Works**: Classical literature that captures the narrative styles and linguistic nuances of different eras.

Each document undergoes preprocessing, including OCR (for scanned texts), tokenization, and normalization to standardize spelling and grammatical conventions without losing historical authenticity. This preprocessing ensures that the model receives clean and consistent input for effective training.

3.2. Evaluation Dataset

For evaluating the performance of our instruction-tuned LLMs, we assembled a separate dataset that mirrors the diversity of the training data but remains unseen during the instruction tuning phase. This dataset includes:

1. **Unpublished Historical Documents**: Manuscripts and letters from private collections and less accessible archives.
2. **Historical News Reports**: Specific events and reports that were not included in the training data, providing a testbed for the model's ability to generalize.
3. **Annotated Corpora**: Historical texts annotated by historians for NER and sentiment analysis, offering a benchmark for comparison.

 doi:10.20944/preprints202410.0470.v1

3.3. GPT-4 as Judge: Novel Evaluation Metrics

Traditional evaluation metrics, such as BLEU or F1 scores, often fall short in capturing the nuanced understanding required for historical texts. Therefore, we employ GPT-4 as a judge to provide a more context-aware evaluation of the model's performance. This method leverages the advanced reasoning and comprehension capabilities of GPT-4 to assess outputs based on several novel criteria:

1. **Contextual Accuracy**: GPT-4 evaluates whether the model's output accurately reflects the historical context and language of the input text.
2. **Entity Recognition and Relevance**: The evaluation focuses on the correct identification of historical entities and their relevance to the context provided.
3. **Linguistic Fidelity**: The assessment considers the fidelity of the language used, ensuring that the output maintains the stylistic and grammatical conventions of the historical period.
4. **Holistic Understanding**: GPT-4 judges the overall coherence and understanding of the text, beyond mere token-level accuracy.

By incorporating these criteria, we aim to provide a comprehensive evaluation framework that better reflects the real-world applicability of LLMs in historical research. This approach not only highlights the strengths and weaknesses of the model but also provides actionable insights for further refinement.

In summary, our dataset collection strategy and novel evaluation framework lay a solid foundation for enhancing the performance of LLMs in historical domains. The careful selection and preprocessing of training and evaluation data, combined with the innovative use of GPT-4 as a judge, ensure that our approach addresses the unique challenges posed by historical texts effectively.

4. Method

In this section, we detail our proposed method for enhancing the performance of large language models (LLMs) in the historical domain through instruction tuning with specially crafted prompts. We begin by discussing the motivation behind our approach, followed by a description of the specific prompts used and the instruction tuning process. Finally, we explain the significance and benefits of our method.

4.1. Motivation

The motivation for our method arises from the inherent complexities of historical texts, which often feature archaic language, unique terminologies, and varied contextual backgrounds. Standard LLMs, typically trained on contemporary text corpora, struggle to accurately interpret and process these texts. Our goal is to develop a robust framework that allows LLMs to handle historical data more effectively. By using context-aware and task-specific prompts, we aim to guide the models in understanding the nuances of historical language and context, thereby improving their performance in tasks such as named entity recognition (NER), information extraction, and sentiment analysis.

4.2. Crafting the Prompts

The core of our method involves creating detailed prompts that provide context and specific instructions to the LLMs. We categorized our prompts into two types: contextual prompts and task-specific prompts.

4.2.1. Contextual Prompts

These prompts provide the necessary background information about the historical period and events relevant to the text. They are designed to immerse the model in the historical context, enabling it to understand and process the text more accurately. An example of a contextual prompt is:

> "Identify the key entities in the following text from 18th-century France, considering the historical context of the French Revolution."

 doi:10.20944/preprints202410.0470.v1

4.2.2. Task-Specific Prompts

These prompts are tailored to specific NLP tasks such as NER or sentiment analysis. They guide the model on what to focus on in the text, ensuring that the outputs are relevant to the task at hand. An example of a task-specific prompt is:

> "Extract all named entities related to geographic locations from this 19th-century British text."

4.3. Instruction Tuning with Prompts

The instruction tuning process involves fine-tuning the LLMs on data obtained using these crafted prompts. The process is as follows:

1. **Data Collection**: Historical texts are collected and preprocessed as described in the previous section.
2. **Prompting**: Each text is paired with a relevant contextual or task-specific prompt.
3. **Fine-Tuning**: The LLM is fine-tuned on the prompted data. This involves adjusting the model's weights to optimize its performance on the given tasks.
4. **Iterative Refinement**: The outputs are analyzed and used to refine the prompts further, creating a feedback loop that continually enhances the model's understanding and performance.

4.4. Significance and Benefits

The significance of our method lies in its ability to address the specific challenges posed by historical texts. By using context-aware and task-specific prompts, we ensure that the LLMs are better equipped to handle the unique characteristics of historical language. This approach offers several benefits:

1. **Improved Accuracy**: The detailed prompts guide the model to understand and process historical context and terminologies accurately, leading to better performance in NLP tasks.
2. **Contextual Understanding**: By immersing the model in the historical context, our method enhances its ability to interpret texts from different periods accurately.
3. **Flexibility and Adaptability**: The iterative refinement process ensures that the prompts and the model continually improve, allowing for adaptability to various historical domains and tasks.

In conclusion, our method of combining LLMs with instruction tuning through carefully crafted prompts offers a robust solution for processing and analyzing historical texts. It leverages the strengths of LLMs while addressing their limitations in handling specialized domains, ultimately contributing to more accurate and insightful historical research.

5. Experiments

To validate the effectiveness of our proposed method, we conducted a series of experiments comparing our instruction-tuned LLM with several baseline LLMs, including LLaMA 7B, LLaMA-2 7B, and Qwen 7B. The experiments aimed to evaluate the performance of these models on tasks involving historical texts, such as named entity recognition (NER) and information extraction.

5.1. Experimental Setup

We used the datasets described in the *Dataset Collection* section for both instruction tuning and evaluation. The baseline models (LLaMA 7B, LLaMA-2 7B, and Qwen 7B) were fine-tuned on the same datasets without the use of our specialized prompts. Our model, however, was fine-tuned using the context-aware and task-specific prompts as detailed in the *Method* section.

The evaluation metrics included contextual accuracy, entity recognition, linguistic fidelity, and holistic understanding, assessed using GPT-4 as described earlier. Additionally, we conducted human evaluations to further substantiate the performance differences between the models.

 doi:10.20944/preprints202410.0470.v1

5.2. Experimental Design

The experimental design included the following steps:

1. **Data Preparation**: Historical texts were collected, preprocessed, and annotated for NER and information extraction tasks.
2. **Baseline Model Fine-Tuning**: The baseline models (LLaMA 7B, LLaMA-2 7B, and Qwen 7B) were fine-tuned on the historical datasets without using our specialized prompts. This involved standard fine-tuning procedures using learning rate scheduling, early stopping, and regularization techniques to optimize performance.
3. **Instruction-Tuned Model Fine-Tuning**: Our instruction-tuned model was fine-tuned using the same datasets but with the addition of context-aware and task-specific prompts. This fine-tuning process was iterative, where initial model outputs were analyzed and used to refine the prompts further.
4. **Evaluation**: Both the baseline and instruction-tuned models were evaluated using predefined metrics on a separate test set of historical texts. The evaluation involved both automated metrics (contextual accuracy, entity recognition, linguistic fidelity, and holistic understanding) and human evaluations by experts in historical linguistics and digital humanities.

5.3. Results

The results of our experiments are summarized in Table 1. As seen, our instruction-tuned model outperformed the baseline models across all evaluation metrics.

Table 1. Performance Comparison of Models on Historical Text Tasks.

Model	Contextual Accuracy	Entity Recognition	Linguistic Fidelity	Holistic Understanding
LLaMA 7B	76.4%	72.1%	74.3%	71.2%
LLaMA-2 7B	78.5%	75.4%	76.9%	73.5%
Qwen 7B	77.2%	73.8%	75.6%	72.8%
Our Method	**85.3%**	**83.7%**	**84.5%**	**82.9%**

5.4. Analysis of Results

The superior performance of our model can be attributed to the effectiveness of the tailored prompts in guiding the LLM to better understand the historical context and language. The context-aware prompts provided the necessary background, enhancing the model's ability to interpret historical texts accurately. The task-specific prompts ensured that the model focused on relevant details, improving its performance in specific tasks such as NER.

5.5. Ablation Studies

To further understand the contribution of each component in our method, we conducted ablation studies. These studies involved removing or modifying certain aspects of the instruction tuning process to observe their impact on model performance. The following variations were tested:

1. **No Contextual Prompts**: The model was fine-tuned without the context-aware prompts to evaluate their impact on contextual understanding.
2. **No Task-Specific Prompts**: The model was fine-tuned without the task-specific prompts to assess their importance in improving task performance.
3. **No Iterative Refinement**: The model was fine-tuned using the initial set of prompts without iterative refinement to determine the value of the feedback loop.

The results of these ablation studies are presented in Table 2. The findings highlight the significance of each component, with the full model consistently outperforming the variations.

 doi:10.20944/preprints202410.0470.v1

Table 2. Ablation Study Results on Historical Text Tasks.

Model Variation	Contextual ACC	Ent. Recognition	Linguistic Fidelity	Holistic Understanding
No Contextual Prompts	80.1%	78.4%	79.0%	77.3%
No Task-Specific Prompts	82.5%	80.6%	81.2%	79.8%
No Iterative Refinement	83.2%	81.4%	82.0%	80.5%
Full Model (Ours)	**85.3%**	**83.7%**	**84.5%**	**82.9%**

5.6. Human Evaluation

To complement the quantitative analysis, we also conducted human evaluations. Experts in historical linguistics and digital humanities were asked to rate the outputs of each model on a scale of 1 to 5 across the same evaluation criteria. The results are presented in Table 3.

Table 3. Human Evaluation of Model Outputs.

Model	Contextual Accuracy	Entity Recognition	Linguistic Fidelity	Holistic Understanding
LLaMA 7B	3.8	3.6	3.7	3.5
LLaMA-2 7B	4.0	3.9	4.0	3.8
Qwen 7B	3.9	3.7	3.8	3.7
Our Method	**4.5**	**4.4**	**4.5**	**4.4**

5.7. Validation of Effectiveness

The human evaluation results corroborate our quantitative findings, confirming that our instruction-tuned model provides more accurate and contextually relevant outputs compared to the baseline models. The feedback from experts highlighted the model's superior understanding of historical contexts and its ability to maintain linguistic fidelity, further validating the effectiveness of our approach.

5.8. Additional Insights

We also explored additional insights by analyzing the error patterns in model outputs. Common errors included misidentification of historical entities due to similar names across different periods, and difficulty in understanding nuanced historical contexts. By examining these errors, we refined our prompts and instruction tuning process, leading to iterative improvements in model performance.

In summary, our experiments demonstrate that instruction tuning with tailored prompts significantly enhances the performance of LLMs in the historical domain. The combined use of contextual and task-specific prompts, along with iterative refinement, ensures that the model can accurately interpret and process historical texts, making it a valuable tool for researchers in the field.

6. Conclusions

In this study, we have addressed the challenges of applying Large Language Models (LLMs) to historical texts by proposing an innovative approach that combines instruction tuning with specially crafted prompts. Our method focuses on creating detailed, context-aware, and task-specific prompts that guide the model in understanding the unique characteristics of historical language and context. Through a series of experiments, we have demonstrated that our instruction-tuned model significantly outperforms several baseline models, including LLaMA 7B, LLaMA-2 7B, and Qwen 7B, across various evaluation metrics such as contextual accuracy, entity recognition, and linguistic fidelity.

Moreover, human evaluations further validate the effectiveness of our approach, indicating that our model provides more accurate and contextually relevant outputs compared to the baselines. The feedback from experts underscores the model's enhanced ability to interpret historical contexts and maintain linguistic fidelity, which are critical for historical research tasks.

In conclusion, our research not only highlights the limitations of current LLMs in handling historical texts but also offers a robust solution through instruction tuning with tailored prompts. This approach enhances the applicability of LLMs in the historical domain, providing researchers with powerful tools to analyze and interpret historical data. Future work will focus on expanding the range

 doi:10.20944/preprints202410.0470.v1

of historical periods and languages covered, as well as exploring additional applications of our method in other specialized domains.

References

1. Zhao, J.; Wang, T.; Abid, W.; Angus, G.; Garg, A.; Kinnison, J.; Sherstinsky, A.; Molino, P.; Addair, T.; Rishi, D. LoRA Land: 310 Fine-tuned LLMs that Rival GPT-4, A Technical Report. *CoRR* **2024**, *abs/2405.00732*, [2405.00732]. doi:10.48550/ARXIV.2405.00732.
2. Zhou, Y.; Li, X.; Wang, Q.; Shen, J. Visual In-Context Learning for Large Vision-Language Models. *arXiv preprint arXiv:2402.11574* **2024**.
3. 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.
4. 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**.
5. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.B.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N.S.; Chen, A.S.; Creel, K.; Davis, J.Q.; Demszky, D.; Donahue, C.; Doumbouya, M.; Durmus, E.; Ermon, S.; Etchemendy, J.; Ethayarajh, K.; Fei-Fei, L.; Finn, C.; Gale, T.; Gillespie, L.; Goel, K.; Goodman, N.D.; Grossman, S.; Guha, N.; Hashimoto, T.; Henderson, P.; Hewitt, J.; Ho, D.E.; Hong, J.; Hsu, K.; Huang, J.; Icard, T.; Jain, S.; Jurafsky, D.; Kalluri, P.; Karamcheti, S.; Keeling, G.; Khani, F.; Khattab, O.; Koh, P.W.; Krass, M.S.; Krishna, R.; Kuditipudi, R.; et al.. On the Opportunities and Risks of Foundation Models. *CoRR* **2021**, *abs/2108.07258*, [2108.07258].
6. Wang, Z.; Li, M.; Xu, R.; Zhou, L.; Lei, J.; Lin, X.; Wang, S.; Yang, Z.; Zhu, C.; Hoiem, D.; Chang, S.; Bansal, M.; Ji, H. Language Models with Image Descriptors are Strong Few-Shot Video-Language Learners. Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022; Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; Oh, A., Eds., 2022.
7. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Barnes, N.; Mian, A. A Comprehensive Overview of Large Language Models. *CoRR* **2023**, *abs/2307.06435*, [2307.06435]. doi:10.48550/ARXIV.2307.06435.
8. Schweter, S.; März, L.; Schmid, K.; Çano, E. hmBERT: Historical Multilingual Language Models for Named Entity Recognition. Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022; Faggioli, G.; Ferro, N.; Hanbury, A.; Potthast, M., Eds. CEUR-WS.org, 2022, Vol. 3180, *CEUR Workshop Proceedings*, pp. 1109–1129.
9. Du, Y.; Guo, H.; Zhou, K.; Zhao, W.X.; Wang, J.; Wang, C.; Cai, M.; Song, R.; Wen, J. What Makes for Good Visual Instructions? Synthesizing Complex Visual Reasoning Instructions for Visual Instruction Tuning. *CoRR* **2023**, *abs/2311.01487*, [2311.01487]. doi:10.48550/ARXIV.2311.01487.
10. 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. *arXiv preprint arXiv:2203.02225* **2022**.
11. 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.
12. Zhou, Y.; Long, G. Improving Cross-modal Alignment for Text-Guided Image Inpainting. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 3445–3456.
13. Zhou, Y.; Long, G. Multimodal Event Transformer for Image-guided Story Ending Generation. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 3434–3444.
14. Zhou, Y.; Tao, W.; Zhang, W. Triple sequence generative adversarial nets for unsupervised image captioning. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 7598–7602.
15. Zhou, Y. Sketch storytelling. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 4748–4752.

16. 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.
17. 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.
18. Zhang, X.; Yang, H.; Young, E.F.Y. Attentional Transfer is All You Need: Technology-aware Layout Pattern Generation. 58th ACM/IEEE Design Automation Conference, DAC 2021, San Francisco, CA, USA, December 5-9, 2021. IEEE, 2021, pp. 169–174. doi:10.1109/DAC18074.2021.9586227.
19. Zhou, Y.; Long, G. Style-Aware Contrastive Learning for Multi-Style Image Captioning. Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 2257–2267.
20. Elouargui, Y.; Zyate, M.; Sassioui, A.; Chergui, M.; El-Kamili, M.; Ouzzif, M. A Comprehensive Survey On Efficient Transformers. 10th International Conference on Wireless Networks and Mobile Communications, WINCOM 2023, Istanbul, Turkey, October 26-28, 2023; Ibrahimi, K.; El-Kamili, M.; Kobbane, A.; Shayea, I., Eds. IEEE, 2023, pp. 1–6. doi:10.1109/WINCOM59760.2023.10322921.
21. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 140:1–140:67.
22. Wei, J.; Bosma, M.; Zhao, V.Y.; Guu, K.; Yu, A.W.; Lester, B.; Du, N.; Dai, A.M.; Le, Q.V. Finetuned Language Models are Zero-Shot Learners. The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022.
23. 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**.
24. Mishra, S.; Khashabi, D.; Baral, C.; Hajishirzi, H. Cross-Task Generalization via Natural Language Crowdsourcing Instructions. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022; Muresan, S.; Nakov, P.; Villavicencio, A., Eds. Association for Computational Linguistics, 2022, pp. 3470–3487. doi:10.18653/V1/2022.ACL-LONG.244.
25. Longpre, S.; Hou, L.; Vu, T.; Webson, A.; Chung, H.W.; Tay, Y.; Zhou, D.; Le, Q.V.; Zoph, B.; Wei, J.; Roberts, A. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning. International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA; Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; Scarlett, J., Eds. PMLR, 2023, Vol. 202, *Proceedings of Machine Learning Research*, pp. 22631–22648.
26. Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N.A.; Khashabi, D.; Hajishirzi, H. Self-Instruct: Aligning Language Models with Self-Generated Instructions. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023; Rogers, A.; Boyd-Graber, J.L.; Okazaki, N., Eds. Association for Computational Linguistics, 2023, pp. 13484–13508. doi:10.18653/V1/2023.ACL-LONG.754.
27. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.F.; Leike, J.; Lowe, R. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022; Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; Oh, A., Eds., 2022.