Enhancing Paraphrasing in Chatbots through Prompt Engineering: A Comparative Study on ChatGPT, Bing, and Bard

Meltem Kurt Pehlivanoğlu
Kocaeli University
Department of Computer Engineering
Kocaeli, Türkiye
meltem.kurt@kocaeli.edu.tr

Muhammad Abdan Syakura
Kocaeli University
Department of Computer Engineering
Kocaeli, Türkiye
prof.syakur@gmail.com

Nevcihan Duru
Kocaeli Health and Technology University
Department of Computer Engineering
Kocaeli, Türkiye
nevcihan.duru@kocaelisaglik.edu.tr

2023 8th International Conference on Computer Science and Engineering (UBMK) | 979-8-3503-4081-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/UBMK59864.2023.10286606

Abstract*—Paraphrase generation, a crucial task in Natural Language Processing (NLP), is pivotal for the effectiveness of AI chatbots. However, generating high-quality paraphrases that are contextually relevant, semantically equivalent, and linguistically diverse remains a challenge. This paper explores the use of prompt engineering to enhance the paraphrasing capabilities of AI chatbots, specifically focusing on ChatGPT, Bing, and Bard. We introduce a new dataset of 5000 sentences generated by ChatGPT across diverse topics and propose two distinct prompts for paraphrase generation: a direct approach and an engineered prompt. The engineered prompt explicitly instructs the chatbot to generate paraphrases that exhibit lexical diversity, phrasal variations, syntactical differences, fluency, language acceptableness, and relevance, while preserving the original meaning. We conduct a comprehensive evaluation of the generated paraphrases using a range of metrics, including BERTScore, STS-B, METEOR for semantic similarity; ROUGE, BLEU, GLEU for diversity; and CoLA, Perplexity for language acceptableness or fluency. Our findings reveal that the use of the engineered prompt results in higher quality paraphrases across all three chatbots, demonstrating the potential of prompt engineering as a tool for improving chatbot communication.

Keywords*—*AI chatbots; paraphrase generation; prompt engineering; generative AI; ChatGPT; Bing Chat; Bard.

I. Introduction

Paraphrasing, the task of rewording a given text while preserving its original meaning, is a fundamental aspect of human communication. It allows us to adapt our language to different contexts, audiences, and purposes. In the field of Natural Language Processing (NLP), paraphrasing plays a crucial role in various applications, including information retrieval, machine translation, text summarization, and dialogue systems.

Despite its importance, paraphrase generation remains a challenging task for automated systems. The complexity arises from the need to generate text that is not only semantically equivalent to the original but also syntactically varied and linguistically fluent. Traditional rule-based and statistical methods often fall short in generating high-quality paraphrases due to the inherent ambiguity and variability of natural language [1, 2].

The advent of deep learning and generative models, such as Generative Pretrained Transformer (GPT) [3], has brought significant advancements in paraphrase generation [4, 5]. These models, trained on large-scale corpora, have demonstrated an impressive ability to generate human-like text. However, they often require careful tuning to perform specific tasks effectively, which is where prompt engineering comes into play.

Prompt engineering is the process of designing input prompts to guide a generative model towards producing the desired output [6]. It has emerged as a crucial skill in the era of large language models, enabling us to harness their power for a wide range of tasks without the need for extensive fine-tuning or additional training data.

In recent years, the use of AI chatbots has surged across various domains, from customer service to mental health counseling. These chatbots, often powered by generative models, have the potential to revolutionize human-computer interaction by providing more natural and engaging conversations. However, their effectiveness heavily relies on their ability to paraphrase effectively, which remains an open challenge.

This paper aims to address this gap by exploring the use of prompt engineering to enhance the paraphrasing capabilities of AI chatbots [7]. We investigate the impact of various prompt designs on the quality of paraphrasing and present a comprehensive comparison of the performance of three popular chatbots—ChatGPT [8], Bing Chat [9], and Bard [10]. Our findings shed light on the potential of prompt engineering as a tool for improving chatbot communication and contribute to the ongoing efforts to make AI more useful and accessible.

A. Motivation and Contribution

The motivation for this research arises from the limitations observed in generative AI models, including issues of accuracy, hallucination, biases, context preservation, and creativity. Specifically, when using chatbots for paraphrasing, we often face problems such as inconsistent responses and paraphrases that do not align with our desired output. This research aims to address these challenges by exploring the question: How can

979-8-3503-4081-5/23/$31.00 ©2023 IEEE

prompt engineering be utilized to improve the consistency and accuracy of paraphrase generation in AI chatbots?

This paper makes several key contributions to the field:

- First, we introduce a new dataset of 5000 sentences generated by ChatGPT across five diverse topics: Legal/Government/Business documents, Scientific papers, Health, Creative/Fiction and Non-fiction books, and Social media/Dialogue/Opinion. This dataset, which also includes subtopics for each category, provides a rich resource for evaluating paraphrase generation in AI chatbots.
- Second, we propose two distinct prompts for paraphrase generation: a direct approach and an engineered prompt that incorporates various aspects of the paraphrasing task. Our engineered prompt explicitly instructs the chatbot to generate paraphrases that exhibit lexical diversity, phrasal variations, syntactical differences, fluency, language acceptableness, and relevance, while preserving the original meaning.
- Third, we conduct a comprehensive evaluation of the generated paraphrases using a range of metrics, including BERTScore, STS-B, METEOR for semantic similarity; ROUGE, BLEU, GLEU for diversity; and CoLA, Perplexity for language acceptableness and fluency. Our findings show that the use of the engineered prompt results in higher quality paraphrases across all three chatbots, demonstrating the potential of prompt engineering as a tool for improving chatbot communication.

Furthermore, our source code and dataset in [11] are made publicly available.

B. Organization

We give some related works in Section II. In Section III, we give all the details of our new dataset and the paraphrasing and evaluation system. The experimental results are presented in Section IV. Finally, we present our conclusive findings in Section V.

II. Related Works

In the realm of language model utilization, various techniques have been explored to optimize the performance of these models for specific tasks. One such technique is prompt engineering, which involves the careful crafting of prompts to guide the behavior of language models towards desired outputs. This technique, while requiring a deep understanding of the model and the task, offers a quick way to adapt a pretrained model to a new task without additional training [7, 12].

However, the effectiveness of prompt engineering can be highly dependent on the specific model and task, and finding the right prompt for complex tasks can be challenging. This has led to the development of more advanced techniques such as prompt tuning and prompt optimization.

Prompt tuning is a technique that involves fine-tuning the model on a specific prompt, optimizing the model's responses to that prompt [13]. While this can lead to improved performance on the task associated with the prompt, it requires additional computational resources and can make the model less flexible, as it becomes optimized for a specific prompt and may perform poorly on other prompts or tasks.

On the other hand, prompt optimization involves optimizing the prompt itself to maximize the model's performance on a specific task [14]. This technique, while also requiring additional computational resources, can lead to better performance than manual prompt engineering or prompt tuning.

However, it's important to note that these techniques may not be suitable for all tasks. For instance, paraphrase generation is a complex task that requires the model to maintain the semantic meaning while introducing lexical and syntactic diversity. This complexity may limit the effectiveness of techniques like prompt tuning, which optimize the model for a specific prompt and may not generalize well to the diverse range of prompts required for paraphrase generation. This underscores the need for more research into effective techniques for paraphrase generation, such as the prompt engineering approach explored in this study.

A recent work that is closely related to our study is the Quality Controlled Paraphrase Generation (QCPG) [15]. In this work, the authors propose a method that allows users to control the quality of the generated paraphrases by specifying desired levels of lexical, syntactic, and semantic similarity. This method, while effective, requires the user to provide explicit scores for each aspect of the paraphrase quality. Their findings demonstrated that their method could generate paraphrases that not only preserved the original meaning but also exhibited greater diversity compared to uncontrolled methods. This work is particularly relevant to our study as it demonstrates the potential of using controlled methods to improve the quality of paraphrases generated by AI models.

In the work [16], the authors proposed a catalog of prompt patterns to facilitate the process of prompt engineering with ChatGPT. They demonstrated that the use of these patterns could significantly improve the performance of the model on various tasks. This work provides valuable insights into the potential of prompt engineering for enhancing the capabilities of AI chatbots.

In the study [17], the authors proposed a method for controlling the novelty of generated paraphrases by tuning the prompt conditionally. They demonstrated that their method could generate paraphrases with controlled novelty, which could be beneficial for tasks that require a balance between novelty and semantic similarity.

While these studies have made significant contributions to the field of paraphrase generation and prompt engineering, our work differs in several key aspects. First, unlike the studies that focus on controlling the quality of paraphrases by assigning explicit scores to inputs or using retrieval-augmented conditional prompt tuning, our approach does not involve explicit scoring or retrieval mechanisms. Instead, we rely on the inherent capabilities of the language models and guide them through carefully crafted prompts to generate diverse paraphrases. Second, while some studies present catalogs to aid in prompt creation and refinement, our work focuses on the practical application of prompt engineering in the context of AI chatbots. Lastly, our work emphasizes the simplicity and effectiveness of prompt engineering, demonstrating its potential as a tool for improving the quality of paraphrases generated by AI models without the need for extensive computational resources or additional training data. This makes our approach more accessible and scalable, especially for applications where resources are limited or where rapid adaptation to new tasks is required. Our study contributes to the ongoing efforts to harness the power of large language models for specific tasks through prompt engineering, and we hope that our findings will inspire further research in this direction.

III. Proposed Dataset and Method

A. Proposed Dataset

We introduce a new dataset of 5000 sentences generated by ChatGPT across five diverse topics: “Legal/Government/Business documents, Scientific papers, Health, Creative/Fiction and Non-fiction books, and Social media/Dialogue/Opinion”. Each topic is further divided into approximately ten subtopics to ensure a wide range of sentence structures, domains, and complex linguistic patterns.

The sentences were generated using the prompt: *“Can you generate 100 different mid-length sentences (also with different sentence structures, a variety of phrase types, domains, and complex linguistic patterns) found and used in [topic name] with subtopic is [subtopic]”*. The topics and subtopics used for sentence generation are provided in Table I.

B. Method

We employed two distinct prompts for paraphrase generation.

- The first is a direct approach, using the direct prompt (Prompt 1): *“Paraphrase: [sentence]”*.
- The second is an engineered prompt (Prompt 2) designed to guide the chatbot toward generating paraphrases that exhibit lexical diversity, phrasal variations, syntactical differences, fluency, language acceptableness, and relevance while preserving the original meaning. The engineered prompt is as follows:
  “Paraphrase each of the sentences listed below separately, keeping in mind that they may have different topics, therefore it’s important to preserve contextual relevance.
  Please include as much diversity as you can in the rewrites, such as varied vocabulary usage (lexical diversity), changes to phrase arrangements (phrasal variations), and the restructuring of sentence composition (syntactical differences).
  Each rewritten sentence must exhibit coherence, maintain fluency, and use proper grammar and punctuation. However, it’s crucial to ensure the preservation of the original meaning and achieve a high degree of semantic similarity with the original sentence.
  FORMAT: Do not include any introductory or concluding statements; only list the paraphrased sentences with their corresponding numbers.

  **[sentences]* ”*

TABLE I
Topics and Subtopics used for Sentence Generation

No	Topic	Subtopic	#sentence	Total
1	Legal/ Government/ Business documents	same as 1st topic	66	1000
		contract	189	
		legislation	100	
		court cases	145	
		business proposal	100	
		meeting minutes	100	
		memos	100	
		press releases	100	
		agreement	100	
2	Scientific papers	same as 2nd topic	100	1000
		physics	100	
		chemistry	100	
		biology	100	
		computer science	100	
		tech	100	
		psychology	100	
		blog post	100	
		news	100	
		ai	100	
3	Health	health	69	1000
		nutrition	99	
		exercise	100	
		mental health	100	
		chronic diseases	100	
		infectious diseases	100	
		patient’s medical history	100	
		patient’s complaint	144	
		symptoms and treatments	88	
		good habits	100	
4	Creative/ Fiction and Non-fiction books	fantasy	100	1000
		mystery	100	
		romance	100	
		historical fiction	100	
		biography	100	
		memoirs	100	
		self-help	100	
		travel	100	
		cookbooks	100	
		creative/novel/storybooks	100	
5	Social media/ Dialogue/ Opinion	personal updates	100	1000
		event announcement	100	
		product review	100	
		debates	100	
		casual conversation	100	
		complaints	100	
		daily routines	100	
		opinions	100	
		movie reviews	100	
		bad reviews	100	

We paraphrased each sentence in the dataset using both prompts on each of the three chatbots (ChatGPT, Bing, and Bard), generating one paraphrase per sentence. The paraphrasing was conducted in batches of 10 sentences per request to accommodate the chatbots’ request limits. Any paraphrases that were identical to the original sentences were re-paraphrased until no duplicates remained.

1) Evaluation: The generated paraphrases were evaluated automatically using a range of metrics. For semantic similarity or adequacy, we used BERTScore [18], STS-B [19] with T5-base model, and METEOR [20]. For diversity, we used ROUGE, BLEU, and GLEU [21]. For fluency or language acceptableness, we used CoLA with the T5 model [22] and Perplexity with the GPT2-large model. These metrics allowed us to assess the quality of the paraphrases in terms of three key aspects: semantic equivalence to the original sentence, diversity in terms of lexical choice and sentence structure, and fluency in terms of grammatical correctness and naturalness of language.

In Figure 1, we give the system architecture of the proposed paraphrasing and evaluation system given above.

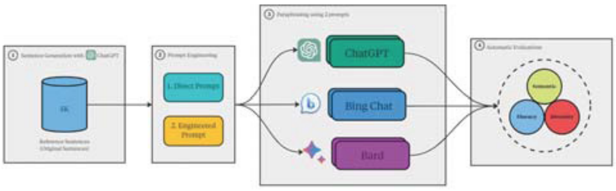

Fig. 1. The general structure of the proposed paraphrasing and evaluation system.

IV. Experimental Results

Our study aimed to enhance the quality of paraphrases generated by AI chatbots using prompt engineering. The results indicate that the engineered prompt consistently produced higher quality paraphrases across all three chatbots (ChatGPT, Bing, and Bard) compared to the direct approach. This improvement was observed in terms of both diversity and semantic similarity, demonstrating the effectiveness of prompt engineering in improving paraphrase generation.

In Table II, we give an example for the corresponding paraphrased sentences of each model for one reference sentence. Note that all evaluations are made for the 5000 reference sentences given in our new dataset.

TABLE II
THE CORRESPONDING PARAPHRASED SENTENCES OF EACH MODEL FOR ONE REFERENCE SENTENCE

Models	Prompt	Sentences
Reference	-	I engage in daily gratitude practices, whether it's writing in a gratitude journal or sharing what I'm grateful for with a loved one.
ChatGPT	1	Every day, I incorporate gratitude practices into my life, whether it involves journaling my blessings or expressing gratitude to a loved one.
		Every day, I cultivate gratitude in various ways, be it through penning expressions of appreciation in my gratitude journal or sharing heartfelt gratefulness with a cherished companion.
Bing Chat	1	I express gratitude daily through practices such as writing in a gratitude journal or sharing what I'm grateful for with a loved one.
	2	Every day, I express my gratitude for the things I have, either by writing them down in a journal or sharing them with someone I care about.
Bard	1	I express my gratitude for the good things in my life on a daily basis.
	2	I express gratitude daily to appreciate the good things in my life.

TABLE III
SEMANTIC SIMILARITY SCORES

Model	Prompt	BERTScore[a]	STS-B	METEOR
ChatGPT	1	0.9227	4.7895	0.7755
	2	0.8922	4.7248	0.689
Bing Chat	1	**0.9383**	**4.8415**	**0.820**8
	2	0.9296	4.8116	0.7817
Bard	1	0.9077	4.5121	0.7152
	2	0.8955	4.4092	0.6981

[a] with *microsoft/deberta-xlarge-mnli [23]* model

In Table III, we give the semantic similarity scores of the three chatbots. It can be seen from the table that the semantic similarity of the generated paraphrases to the original sentences was evaluated using BERTScore, STS-B, and METEOR metrics. Despite the increased diversity introduced by the engineered prompt, the semantic similarity scores only decreased slightly, indicating that the chatbots were able to maintain the semantic equivalence of the paraphrases. This is a significant finding as it demonstrates that it's possible to increase diversity without significantly compromising semantic similarity. The METEOR metric, which considers synonymy, stemming, and word order to compute similarity, also showed high scores for the paraphrases generated with the engineered prompt. This suggests that the chatbots were able to generate paraphrases that not only preserved the original meaning but also exhibited varied vocabulary usage and sentence structure, aligning with the objectives of our engineered prompt.

TABLE IV
DIVERSITY SCORES

Metric	ChatGPT		Bing Chat		Bard	
	1	*2*	*1*	*2*	*1*	*2*
ROUGE-1	0.7298	**0.6418**	0.8023	0.7655	0.7076	0.6837
ROUGE-2	0.5193	**0.3988**	0.6403	0.5701	0.549	0.5182
ROUGE-L	0.6726	**0.5623**	0.7345	0.6753	0.6793	0.6377
BLEU	0.385	**0.2442**	0.5243	0.4352	0.4268	0.3808
GLEU	0.4559	**0.3362**	0.5737	0.5044	0.4848	0.4455

Note: "1" represents the direct prompt (Prompt 1), while "2" denotes the engineered prompt (Prompt 2).

In Table IV, we also present the diversity scores. The diversity of the generated paraphrases was evaluated using ROUGE, BLEU, and GLEU metrics. The engineered prompt resulted in lower scores on these metrics compared to the direct approach. This is an expected outcome as these metrics reward exact matches, and a lower score indicates that the paraphrases generated with the engineered prompt were more diverse in terms of lexical choice and sentence structure. This demonstrates the effectiveness of the engineered prompt in encouraging the generation of more diverse paraphrases.

In Table V, we also give the fluency scores of the chatbots. The fluency or language acceptableness of the generated paraphrases was evaluated using the CoLA metric and Perplexity (PPL). CoLA measures the grammatical correctness of a sentence, and a lower score indicates fewer ungrammatical sentences. All chatbots produced a small number of ungram-

TABLE V
FLUENCY SCORES

Model	Prompt	CoLA[a]	PPL[b]
Reference	-	9	**5.1137**
ChatGPT	1	**6**	6.9443
	2	10	8.4555
Bing Chat	1	14	6.1802
	2	32	6.572
Bard	1	18	5.4791
	2	10	5.0247

[a]with T5-base model. [b]with GPT2-large model.

matical sentences, indicating a high level of fluency in the generated paraphrases. Perplexity, a measure of how well a probability model predicts a sample, was also low for all chatbots, indicating that the generated sentences were likely under the trained language model, further suggesting high fluency.

TABLE VI
LENGTH OF GENERATED SENTENCES

Model	Prompt	Word Len	Char Len
Reference	-	15.24	103.06
ChatGPT	1	17.27	119.48
	2	18.34	128.6
Bing Chat	1	15.51	101.95
	2	15.73	102.92
Bard	1	14.51	93.16
	2	15.55	99.53

The average word and character lengths of the generated paraphrases were also analyzed in Table VI. ChatGPT tended to generate slightly longer sentences, Bing maintained a similar length to the original sentences, and Bard produced slightly shorter sentences. This variation in sentence length contributes to the diversity of the paraphrases and can be seen as a reflection of the different language models' styles.

During the paraphrase generation process, there were instances where the chatbots produced paraphrases identical to the original sentences. These duplicates were handled by re-initiating the paraphrasing process until no duplicates remained. The first prompt resulted in more duplicates, except for the Bard model. The duplicates were easily handled in ChatGPT and Bing models, as seen in the second re-paraphrase, where no duplicates were found. However, for Bard, duplicates were found up to the sixth re-paraphrase. This suggests that while Bard was less likely to generate duplicates initially, it required more iterations to completely eliminate duplicates.

We encountered a few challenges during the experiment. For instance, Bard occasionally returned responses stating that it was not programmed to assist with the given task. This occurred in a small number of cases and did not significantly impact the overall results. However, it highlights the limitations of current AI models and the need for further improvements.

Additionally, we observed some inconsistencies in the response formats of the chatbots, particularly with Bard. Despite specifying the response format in the prompt, Bard sometimes included the original sentences, generated more than one paraphrase, or provided explanations about paraphrasing. This required us to write special code to handle the responses and highlights the challenges of working with AI models, which do not always respond predictably to prompts.

In conclusion, our results demonstrate the effectiveness of prompt engineering in improving the quality of paraphrases generated by AI chatbots. The engineered prompt resulted in paraphrases that were more diverse and equally semantically similar to the original sentences compared to the direct approach. This suggests that prompt engineering can be a valuable tool for enhancing the performance of AI chatbots in paraphrase generation tasks.

V. CONCLUSION

This study aimed to enhance the quality of paraphrases generated by AI chatbots through the use of prompt engineering. Our findings demonstrate that a carefully engineered prompt can guide chatbots to produce paraphrases that are not only semantically similar to the original sentences but also exhibit a higher degree of diversity in terms of lexical choice and sentence structure.

The results of our study have significant implications for the field of NLP and AI. They highlight the potential of prompt engineering as a tool for improving the performance of AI chatbots in tasks such as paraphrase generation. By providing more specific instructions and expectations in the prompt, we can guide the chatbot to produce outputs that better meet our requirements.

Interestingly, our study suggests that paraphrasing one sentence at a time might be a more effective approach. This method allows for more focused and accurate paraphrasing, reducing the likelihood of generating duplicates or irrelevant paraphrases.

However, our study also revealed some challenges. We encountered issues with duplicate paraphrases and inconsistencies in the response formats of the chatbots, particularly with Bard. These challenges highlight the limitations of current AI models and underscore the need for further improvements and more sophisticated handling methods.

Looking forward, the high-quality paraphrases generated in this study could serve as valuable resources for downstream tasks in natural language processing, such as text summarization, machine translation, and information extraction. This opens up new avenues for future research.

Moreover, we plan to develop a ChatGPT plugin for paraphrasing. This tool could serve as a free and customizable alternative to paid paraphrasing tools, further democratizing access to high-quality paraphrasing services. In conclusion, our study demonstrates the power of prompt engineering in enhancing the performance of AI chatbots in paraphrase generation tasks and opens up exciting possibilities for future research and applications.

References

[1] I. Androutsopoulos and P. Malakasiotis, "A survey of paraphrasing and textual entailment methods," *Journal of Artificial Intelligence Research*, vol. 38, pp. 135–187, may 2010. [Online]. Available: https://doi.org/10.1613%2Fjair.2985

[2] J. Mallinson, R. Sennrich, and M. Lapata, "Paraphrasing revisited with neural machine translation," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics (ACL), Apr. 2017, pp. 881–893, 15th EACL 2017 Software Demonstrations, EACL 2017 ; Conference date: 03-04-2017 Through 07-04-2017. [Online]. Available: http://eacl2017.org/,http://eacl2017.org/index.php

[3] M. R. Chavez, T. S. Butler, P. Rekawek, H. Heo, and W. L. Kinzler, "Chat generative pre-trained transformer: why we should embrace this technology," *American Journal of Obstetrics and Gynecology*, vol. 228, no. 6, pp. 706–711, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0002937823001552

[4] S. Witteveen and M. Andrews, "Paraphrasing with large language models," in *Proceedings of the 3rd Workshop on Neural Generation and Translation*. Association for Computational Linguistics, 2019. [Online]. Available: https://doi.org/10.18653%2Fv1%2Fd19-5623

[5] C. Hegde and S. Patil, "Unsupervised paraphrase generation using pre-trained language models," 2020.

[6] F. Shi, P. Qing, D. Yang, N. Wang, Y. Lei, H. Lu, and X. Lin, "Prompt space optimizing few-shot reasoning success with large language models," 2023.

[7] Y. Liu, G. Deng, Z. Xu, Y. Li, Y. Zheng, Y. Zhang, L. Zhao, T. Zhang, and Y. Liu, "Jailbreaking chatgpt via prompt engineering: An empirical study," 2023.

[8] "Introducing ChatGPT," https://openai.com/blog/chatgpt, accessed: 2023-06-06.

[9] "Bing Chat," https://www.bing.com/new, accessed: 2023-06-06.

[10] "Bard," https://bard.google.com/?hl=en, accessed: 2023-06-06.

[11] "Prompt-engineering-for-paraphrase-generation," https://github.com/massyakur/Prompt-Engineering-for-Paraphrase-Generation, accessed: 2023-06-06.

[12] Y. Hao, Z. Chi, L. Dong, and F. Wei, "Optimizing prompts for text-to-image generation," *arXiv preprint arXiv:2212.09611*, 2022.

[13] Z. Xu, C. Wang, M. Qiu, F. Luo, R. Xu, S. Huang, and J. Huang, "Making pre-trained language models end-to-end few-shot learners with contrastive prompt tuning," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 2023, pp. 438–446.

[14] Y. Wang, D. Lu, C. Kong, and J. Sang, "Towards alleviating the object bias in prompt tuning-based factual knowledge extraction," *arXiv preprint arXiv:2306.03378*, 2023.

[15] E. Bandel, R. Aharonov, M. Shmueli-Scheuer, I. Shnayderman, N. Slonim, and L. Ein-Dor, "Quality controlled paraphrase generation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 596–609. [Online]. Available: https://aclanthology.org/2022.acl-long.45

[16] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A prompt pattern catalog to enhance prompt engineering with chatgpt," 2023.

[17] J. R. Chowdhury, Y. Zhuang, and S. Wang, "Novelty controlled paraphrase generation with retrieval augmented conditional prompt tuning," 2022.

[18] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[19] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 1–14. [Online]. Available: https://aclanthology.org/S17-2001

[20] S. Lee, J. Lee, H. Moon, C. Park, J. Seo, S. Eo, S. Koo, and H. Lim, "A survey on evaluation metrics for machine translation," *Mathematics*, vol. 11, no. 4, 2023. [Online]. Available: https://www.mdpi.com/2227-7390/11/4/1006

[21] A. Sheth, M. Gaur, K. Roy, and K. Faldu, "Knowledge-intensive language understanding for explainable ai," *IEEE Internet Computing*, vol. 25, no. 5, pp. 19–24, 2021.

[22] L. Yang, S. Zhang, L. Qin, Y. Li, Y. Wang, H. Liu, J. Wang, X. Xie, and Y. Zhang, "Glue-x: Evaluating natural language understanding models from an out-of-distribution generalization perspective," *arXiv preprint arXiv:2211.08073*, 2022.

[23] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=XPZIaotutsD