The Search for the Holy Prompt

MAX MARTIN GNEWUCH, Georg-August University of Göttingen, Germany

Abstract goes here!

CCS Concepts: • **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → Human computer interaction (HCI).

Additional Key Words and Phrases: language models, prompt engineering, language sensitivity, paraphrasing, natural language processing

ACM Reference Format:
Max Martin Gnewuch. . The Search for the Holy Prompt. 1, 1 (March), 7 pages.

1 INTRODUCTION

Large pre-trained transformer-based language models (PLMs), like BERT [3] and GPT [23], have revolutionized the Natural Language Processing (NLP) field due to their ability to understand and generate human-like text. However, the effectiveness of these models in producing meaningful results depends heavily on the design of the input's syntax and semantics [9].

Prompting PLMs has made NLP components user-friendly to those without expert knowledge, by providing natural language instructions for NLP tasks [18]. However, slight alterations in task instructions can lead to considerable differences in the generated content, which may affect both the quality and the nature of the response. The evidence for fluctuating performance [5, 10] has been met with many suggested techniques for automated prompt[1] creation [4, 22, 30] paired with an ongoing debate on whether expert written prompts are superior to generated ones [12]. Essentially, the inconsistency in performance across tasks raises the question of whether certain models have a bias toward prompts. This phenomenon underlines the need for an in-depth analysis of PLM sensitivity to changes in prompt language.

To address this need, I present a systematic analysis of the variability of prompt performance in PLMs, focusing on linguistic and syntactic changes in prompt language. To the best of my knowledge, I conduct the first systematic investigation of PLM performance variability across prompts that are semantically identical but differ in their lexical and syntactic composition. I chose three different common sense reasoning, question answering, and emotional understanding tasks. I investigate the impact of prompt language changes on PLM performance by paraphrasing selected datasets from the Beyond the Imitation Game Benchmark (BIG-bench[2]) [2]. High-quality paraphrases are generated using the Quality Controlled Paraphrase Generation model (QCPG) [1] to compose paraphrased datasets for testing against the original datasets. I selected three PLMs of different sizes, all pretrained on English language: OpenAI GPT-1 [23], GPT-2Large, and GPT-2XL [24]. To understand and examine the influence of changes in prompt language on model prediction, I perform token-level analysis on a selection of prompts. I employ Local Interpretable Model-Agnostic Explanations (LIME) [26] as a tool to explain the predictions generated by the models in response to these prompts. The results show that prompts with improved lexical and syntactic diversity do not always surpass original prompts across tasks and models. This suggests that the robustness of the evaluated models to changes in prompt language is task dependent. Furthermore, I observe that the performance of the models varies across tasks and model size. Therefore, model selection should be task specific. The in-depth analysis using LIME raises the question of whether the wording chosen for the paraphrased prompts is optimal. I further demonstrate these findings in my experiments (Section 4).

2 RELATED WORK

Machine-Paraphrase generation. A considerable amount of recent research has been devoted to the pursuit of high-quality paraphrases [1]. This body of work can be divided into two methodologies, with unsupervised paraphrasing being a relatively unexplored and more complex topic in NLP [19]. Liu et al. [16] frame paraphrasing as an optimization challenge, where the search is conducted within the sentence space to identify the optimal point for an objective function. This function considers semantic similarity, diversity of expression, and language fluency. Niu et al. [19] rely on neural models to produce high-quality paraphrases, using a decoding technique that promotes diversity by inhibiting repetitive duplication of input tokens. Wahle et al. [33] utilize GPT-3 and T5 to create paraphrases using prompts and human paraphrases as examples in a few-shot style prediction.

In this study, I employ the QCPG model proposed by Bandel et al. [1]. This model takes advantage of paraphrase quality metrics to encourage the production of paraphrases with desired quality. This feature simplifies the task of creating prompts that are lexically and syntactically diverse while retaining the same semantic meaning.

Robustness in prompting. Previous research has examined prompts from different perspectives. However, just a few studies have explored the robustness to different prompt formulations. Ishibashi et al. [6] conducted a study that shows machine-generated prompts lack robustness when subjected to token deletion or reordering. Shaikh et al. [29] showed that the performance of GPT-3 significantly deteriorates on bias and toxicity challenge sets when the phrase *'Let's think step by step.'* is appended to a given prompt

[1]The terms *prompt* and *instruction* are used interchangeably.

Author's address: Max Martin Gnewuch, maxmartin.gnewuch@stud.uni-goettingen.de, Georg-August University of Göttingen, Weender Landstrasse 37, Göttingen, Lower Saxony, Germany, 37075.

ACM XXXX-XXXX//3-ART
https://doi.org/

[2]https://github.com/google/BIG-bench

intended for chain-of-thought reasoning. Gonen et al. [5] found a significant correlation between a decrease in prompt perplexity and improved performance. This correlation was observed across a diverse range of tasks for both the OPT [7] and BLOOM [11] models. Leidinger et al. [12] investigated the influence of grammatical properties such as mood, tense, aspect, and modality, as well as lexico-semantic variation on the performance of Large Language Models (LLMs). The study's findings challenge the widely held belief that LLMs perform best on prompts with lower perplexity that reflect language use in pretraining or instruction-tuning data. The researchers found that prompts do not transfer well between datasets or models, and that performance cannot be explained solely by perplexity, word frequency, ambiguity, or prompt length.

Unlike the approaches taken in previous studies [8, 12, 15, 28], I do not aim to craft prompts manually or (semi-)automatic. Nor do I propose prompt selection methods like [5, 13]. My work focuses on exploring whether linguistic and syntactic changes to the prompt language can add enough diversity to the prompts to improve PLM understanding without introducing too much complexity.

3 METHODOLOGY

I generate machine-paraphrased prompts based on three datasets from BIG-bench with varying quality dimension values of QCPG. Subsequently, three evaluation datasets are created using the best configuration of quality dimension values, composed of the BIG-bench datasets.

Finally, I evaluate the performance of selected PLMs on the created evaluation and original datasets.

3.1 Tasks and Datasets

I experiment with three datasets for three different tasks from BIG-bench, a commonly used benchmark that focuses on tasks that are believed to be beyond the capabilities of current language models [2]:

- **Social IQa** (Common Sense Reasoning) [27]. This dataset contains 38k multiple-choice questions, specifically crafted to evaluate emotional and social intelligence in various everyday scenarios. Based on empirical data, it's evident that this benchmark presents a significant hurdle for existing question answering systems that depend on pre-trained language models.
- **CoQA** (Question Answering) [25]. This dataset comprises 127k questions and their corresponding answers, extracted from 8k dialogues that discuss text excerpts from seven distinct fields. The purpose of this dataset is to assess a model's ability to comprehend a text passage and respond to a series of inquiries about it in a conversational manner. To perform well in this task, a model must be capable of handling complex linguistic phenomena such as coreference resolution and pragmatics, as well as challenging reasoning tasks like counting and multi-hop reasoning.
- **COM2SENSE** (Emotional Understanding) [31]. This dataset presents a significant challenge in common sense reasoning. It was developed by annotators using a gamified model-in-the-loop approach. The dataset consists of true/false statements expressed in natural language, with each example matched with its corresponding opposite. The purpose of this dataset is to evaluate a model's ability to comprehend various forms of common sense knowledge and to determine the sensibility of a specific statement under diverse reasoning contexts.

These datasets were selected because they relate to fundamental human abilities that allow us to navigate social situations. Such skills are critical for AI assistants to better interact with human users [27].

The source code to process my data and reproduce the experiments is available on GitHub: https://github.com/Peerzival/LLM-Language-Sensitivity

3.2 Models

For **evaluation** I use various models of the GPT family: OpenAI GPT-1 [23], GPT-2Large, and GPT-2XL [24]. All of the chosen models are included as default models in BIG-bench.

For **paraphrase generation** I use an encoder-decoder model that has been trained for the specific task of controlled paraphrase generation, known as QCPG [1]. The QCPG has three distinct variants, each trained on a unique dataset, enhancing its paraphrasing capabilities for diverse input texts. I use the *qcpg-sentences* [3] variant.

3.3 Paraphrase Generation

Method. I use the QCPG model to generate candidate versions of prompts. This model takes as input an initial sentence and quality constraints, represented by a three-dimensional vector of semantic similarity, syntactic and lexical distances. The model then generates a target sentence that complies to the specified quality constraints [1].

To ensure the utilization of the highest-quality paraphrases, I create five unique Quality Dimension Groups (QDG), each defined by different values on the quality constraints (see Table 1). These measures are used to determine the optimal values for producing the highest-quality paraphrases for my research trials.

Candidate Selection. Paraphrase quality is evaluated based on three dimensions, with high-quality paraphrases exhibiting *high* semantic similarity as well as *high* lexical and syntactic diversity [1, 33]. My objective is to select paraphrases of high quality that are semantically similar to the original content, while avoiding reusing the exact words and structures [33].

I opt for Pareto-optimal QDGs that *minimize* the metrics ROUGE-L [14] and BLEU [21] (i.e., penalize verbatim usage of words from the original version) and *maximize* BERTScore [34] (i.e., reward semantic similarity to the original version). Paraphrases with high count-based similarity usually convey the same message, often replicating the original sentence structure and wording. Conversely, paraphrases with high semantic similarity but lower count-based similarity express the same idea, but with a new sentence structure and synonyms that convey the same concept [33].

[3] https://huggingface.co/ibm/qcpg-sentences

Table 1. Overview of the QDGs and their corresponding values.

QDG	Quality dimensions	Value
veryLow	syntactic distance	0.1
	lexical distance	0.1
	semantic similarity	0.95
low	syntactic distance	0.2
	lexical distance	0.3
	semantic similarity	0.95
medium	syntactic distance	0.4
	lexical distance	0.5
	semantic similarity	0.95
high	syntactic distance	0.6
	lexical distance	0.7
	semantic similarity	0.95
veryHigh	syntactic distance	0.8
	lexical distance	1.0
	semantic similarity	0.95

Evaluation Set Creation. To measure how changes in prompt language affect PLM performance, I paraphrase the original prompts of Social IQa, CoQA, and COM2SENSE. I use the Pareto-optimal QDG specific to each dataset to generate paraphrases of the highest-quality (see Table 2). These paraphrases will then be used to create three evaluation datasets for Social IQa, CoQA, and COM2SENSE, by replacing the original prompts with their paraphrased counterparts.

3.4 GPT*k* Evaluation Setup

The evaluation process consists of two main parts. In the first part, the selected models are executed on the three original datasets to establish a performance baseline. This baseline will be used for comparison with the evaluation datasets. The BIG-bench API is used to load the models, execute them, and calculate the evaluation metric for each respective dataset. In the second part, the models are executed on the evaluation datasets. The setup for this execution is identical to that of the original datasets. Besides Social IQa all of my experiments are carried out in a zero-shot setting. I evaluate the models on Social IQa and COM2SENSE according to the proposed setups of [27], and [31] on BIG-bench. Due to computational constraints, the models are evaluated on a subset of 110 examples from the CoQA dataset, as opposed to the full set of 10930 examples.

3.5 LIME

LIME is an interpretability method used to explain the predictions generated by machine learning models. It creates an understandable surrogate model by selecting points around a given input example and using the model's scores at those points to train a simpler, more interpretable "surrogate model", such as a linear model [26]. This surrogate model can then be used to explain the behavior of the original model near the target input point [17].

4 EXPERIMENTS

In my experiments, I first measure QCPG's ability to generate high-quality paraphrases to find the Pareto-optimal QDGs, which are then used to create the evaluation datasets (Section 4.1). Next, I empirically study how the GPT-based models perform on the original and evaluation datasets to investigate the extent to which the PLMs are sensitive to alterations in the prompt language. I correlate the performance of the GPT-based models on the original datasets with the performance on the evaluation datasets (Section 4.2). Finally, I perform a token-level analysis with LIME to understand the model predictions in detail (Section 4.3).

4.1 Pareto-optimal QDG Selection

Table 2 presents the performance results of the QCPG model. Contrary to initial expectations, *high* has the lowest Bleu and Rouge-L score, although it does not have the highest lexical and syntactic diversity (see Table 1). The QDG *veryLow* shows the highest Bleu and Rouge-L score of all QDGs, as expected, because this group has the lowest lexical and syntactic diversity. However, this indicates that the paraphrased prompts simply replicate the sentence structure and wording used in the original prompts. Therefore, this QDG will not add enough diversity to the prompts to improve PLM comprehension.

Table 2. The average BLEU scores, F1 scores derived from ROUGE-L, and BERTScore for each QDG across all datasets. I choose the Pareto-optimal QDG that maximizes semantic similarity (BERTScore ↑) and minimizes word overlap (ROUGE-L, BLEU ↓). **Selected QDGs** and the best scores per dataset are highlighted.

Dataset	QDGs	BERTSc. ↑	BLEU ↓	Rouge-L ↓
Social IQa	veryLow	0.93	0.17	0.51
	low	**0.92**	**0.10**	**0.44**
	medium	0.92	0.11	0.45
	high	0.90	0.05	0.34
	veryHigh	0.92	0.12	0.46
CoQA	veryLow	0.96	0.22	0.59
	low	0.94	0.12	0.43
	medium	**0.93**	**0.11**	**0.39**
	high	0.91	0.07	0.27
	veryHigh	0.94	0.11	0.40
COM2S.	veryLow	0.94	0.22	0.57
	low	0.93	0.13	0.48
	medium	**0.93**	**0.13**	**0.45**
	high	0.91	0.07	0.37
	veryHigh	0.93	0.14	0.48

It is worth noting that the scores of the QDG *veryHigh* closely match with the scores of the Pareto-optimal QDGs. This suggests a successful preservation of the core meaning of the original prompts, achieved without the verbatim usage of words from the original version. These observations support the findings of Bandel et al.

[1], which indicate that the QCPG can increase the linguistic diversity of the generated paraphrases without compromising semantic similarity.

4.2 GPT*k* Performance Evaluation

Figures 1 to 3 illustrate the model performance on the evaluation and paraphrased datasets across Social IQa, CoQA, and COM2SENSE.

Social IQa. Figure 1 shows the percentage of questions answered correctly across models. The results indicate that in the majority of cases, **increasing the lexical and syntactic diversity of prompts does not lead to a substantial rise or fall in the performance of the models**. This suggests that the changes in the prompt language do not significantly affect the model's ability to correctly predict the outcomes for this dataset. Notably, there are cases where the paraphrased prompts slightly outperform their original counterparts, such as with GPT-2Large in the three-shot setting and GPT-2XL in the one-shot setting. However, the difference in prompt performance is to low to suggest that one type of prompt is superior to the other. Of the three models evaluated, GPT-1 is the most resilient to prompt language changes, with very little performance difference between original and paraphrased prompts.

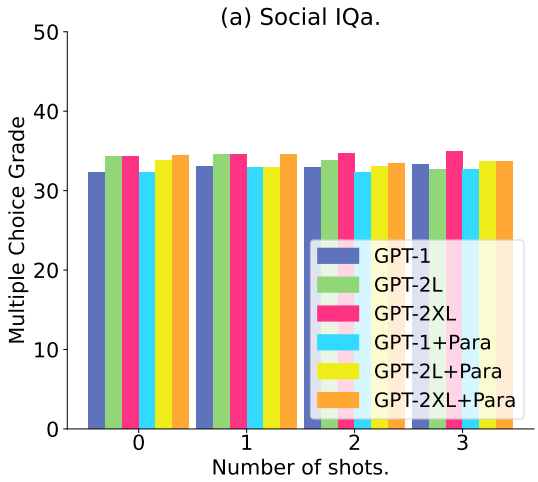

Fig. 1. Model performance breakdown on Social iqa across original and paraphrased datasets, "+*Para*" denotes model performance on the evaluation dataset.

CoQA. The results shown in Figure 2 indicate that the **paraphrased prompts underperform compared to the original prompts**. A positive correlation between model performance and model size is noticeable for both the original and paraphrased prompts. This implies that prompts with increased lexical and syntactic diversity affect model performance independently of model size, at least for this dataset. I note that I only used $\frac{1}{100}$ of the dataset. Therefore, different correlations between model performance and model size might potentially emerge. However, I expect that the models will continue to perform less effectively on the paraphrased prompts than on the original prompts, as the results indicate.

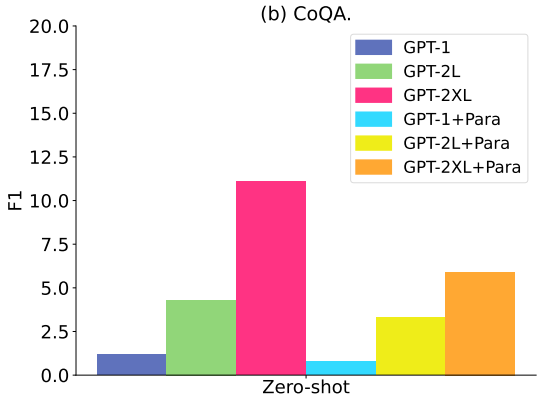

Fig. 2. Comparison of macro-average F1 scores of word overlap across models on the Coqa dataset. "+*Para*" indicates the performance of the models on the evaluation dataset.

COM2SENSE. As depicted in Figure 3, **prompts with enhanced lexical and syntactic diversity slightly improve the performance of GPT-2Large and GPT-2XL on the standard accuracy [4] metric**, although not significantly enough to suggest a definitive superiority of one type of prompt over the other. These performance variations could be due to inherent differences in the models themselves, such as their training data or training time, rather than the prompt itself. Similar to the Social IQa dataset, GPT-1 shows the highest resilience to prompt language changes, with no noticeable performance difference between original and paraphrased prompts.

In the case of the more stringent pairwise accuracy [5] shown in Figure 3, the paraphrased prompts outperform the original prompts more clearly, at least for GPT-2Large and GPT-2XL, despite achieving a relatively low overall score. In particular, the performance of GPT-2XL improves from zero. This indicates that **prompts with enhanced lexical and syntactic diversity can add enough variation to improve the comprehension of PLMs without introducing additional complexity**.

In summary. The results show that prompts with enhanced lexical and syntactic diversity do not consistently outperform the original prompts across tasks and models. However, they do not drastically reduce performance, except in the case of the Coqa dataset. This implies that the **resilience of the evaluated models to changes in prompt language is task dependent**. However, in two of the three datasets tested, GPT-1 proved to be particularly robust to prompt language changes. This raises the question of whether smaller models are better equipped to handle prompts with increased lexical and syntactic diversity, or whether this phenomenon is due to inherent differences within the models themselves. The slight improvements observed in some cases suggest that further investigation of lexically and syntactically diverse prompts could potentially yield performance improvements. Future research

[4] Proportion of correct predictions made by the model out of all predictions.
[5] Evaluates as correct if both predictions within a pair are accurate.

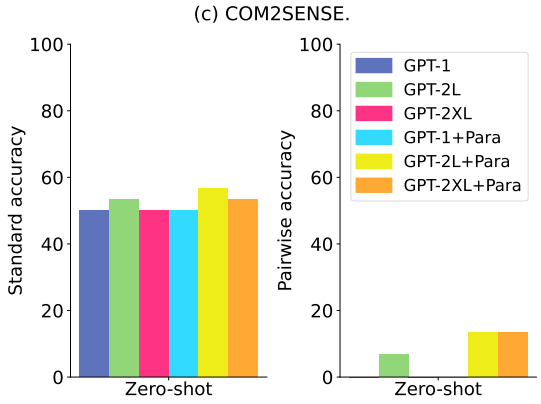

Fig. 3. Performance comparison across models on the Com2sense dataset. The figure shows the percentage of accurate predictions made by each model out of all predictions. The metric in the right plot considers a prediction as correct only if both statements of a complementary pair are accurate. "+*Para*" denotes model performance on the evaluation dataset.

in this domain will need to examine whether these performance improvements hold true across tasks.

The variation in performance across tasks and model size indicates that bigger models are not always better. This aligns with the findings of Ouyang et al. [20], who showed that fine-tuning language model's with human feedback can drastically improve performance without necessarily increasing the model's size. Therefore, I suggest that **model selection should be task specific**.

The poor performance of the models on the selected datasets confirm the findings of [25, 27, 31], that these benchmarks pose a challenge to existing models.

4.3 LIME Results

I use LIME to explain the predictions generated by the GPT-2Large model on a pair of selected prompts and their paraphrased versions from the CoQA dataset. I chose CoQA because the difference in performance between the original and rephrased prompts is most prominent in this dataset. I chose two different prompts from the *medium* QDG. The first prompt pair minimizes Rouge-L and Bleu scores, indicating the least word overlap with the original version. The second pair maximizes Bertscore, reflecting the highest semantic similarity to the original version. The results, in Table 3, represent the degree to which each word contributed (green) or reduced (red) the likelihood that the prompt produces the correct answer. The unchanged parts of the prompt consistently receive the same value. This is not the case for prompts that have been enhanced for lexical and syntactic diversity. Interestingly, these enhancements do not seem to improve GPT-2Large's understanding. The model values certain features positively, suggesting that the enhanced prompts do not add complexity. However, they do not increase the likelihood either. This does not mean that changes in prompt language are ineffective. It's possible that the wording chosen is not optimal for the model being evaluated. Therefore, future research should investigate whether other QDGs have a noticeable effect on the model's predictions.

5 EPILOGUE

Conclusion. I generated machine-paraphrased prompts that differ in their lexical and syntactic diversity using an encoder-decoder model specifically trained for controlled paraphrase generation. I selected Pareto-optimal paraphrases to create three evaluation datasets for Social IQa, CoQA, and COM2SENSE. I evaluated three PLMs, all pretrained on English language, on the original and evaluation datasets. My results suggest that a model's robustness to changes in prompt language is task dependent. I find that model selection should be tailored to the task, as performance varies across tasks and model sizes. A detailed analysis of selected prompts revealed that while enhanced prompts do not add complexity, they do not increase the likelihood that the prompt will produce the correct answer either. Therefore, future research needs to further investigate the impact of prompt language modifications on model performance across different tasks and model sizes.

Limitations. This study has the following potential limitations. Due to limited computing resources, the CoQA dataset was limited to a maximum of 110 out of 10930 examples. Consequently, the results presented do not fully represent the entire dataset, which could introduce bias into parts of my analysis. However, with my open-source implementation, one can replicate the experiments using the full datasets if computing resources are available.

The range of datasets included in my experiments is limited. Therefore, it is necessary to propose and explore more diverse datasets for a comprehensive evaluation of potential solutions. In addition, the inclusion of more NLP tasks could further clarify the impact of prompt language changes on different tasks.

In terms of QDG selection, I rely on standard metrics such as ROUGE, BLEU, and BERTscore. These metrics, are known for their limitations (e.g., poor correlation with human preferences) and cannot account for paraphrase types, locations, or segment length in their score [32].

The token-level analysis with LIME is limited in scope. As a result, it remains uncertain whether enhanced prompts do not introduce additional complexity and increase the likelihood of producing the correct answer.

Future Work. This study is an initial step toward understanding how PLMs process prompts that differ in lexical and syntactic structure. I plan to further investigate the impact of the remaining three QDGs on model performance. The goal is to understand whether paraphrases within these QDGs add enough variation to improve the comprehension of PLMs without introducing additional complexity. In terms of model selection, I plan to include additional models (e.g., T5 and LLaMA) to facilitate a comparative analysis of prompt understanding across different model families, architectures, and sizes. While my approach is currently focused on English, it could be extended to other languages to allow comparison of prompt understanding across various languages. Finally, conducting further experiments to understand the specific influence of each word in a prompt on the model seems to be a valuable extension of my study.

Table 3. LIME output for selected prompts from the CoQA dataset. Words that increase (shown in green) or decrease (shown in red) the probability of the prompt yielding the correct answer are highlighted. The actual prompts are highlighted in bold. $RB - Para$ denotes the paraphrase that achieves the minimum ROUGE-L and BLEU scores. Similarly, $BSc - Para.$ refers the paraphrase that maximizes BERTScore.

Original Text 1	"Chapter 10: Southward Ho! Upon making inquiries, Ned Hearne found that Captain Drake had, upon the return of his expedition, set aside the shares of the prize money of Gerald Summers, himself, and the men who were lost in the wreck of the prize, in hopes that they would some day return to claim them. Upon the evidence given by Gerald and himself of the death of the others, their shares were paid, by the bankers at Plymouth who had charge of them, to their families; while Ned and Gerald received their portions. Owing to the great mortality which had taken place among the crews, each of the lads received a sum of nearly a thousand pounds, the total capture amounting to a value of over a million of money. As boys, they each received the half of a man's share. The officers, of course, had received larger shares; and the merchants who had lent money to get up the expedition gained large profits. Ned thought, at first, of embarking his money in the purchase of a share in a trading vessel, and of taking to that service; but, hearing that Captain Drake intended to fit out another expedition, he decided to wait for that event, and to make one more voyage to the Spanish main, before determining on his future course. Having, therefore, his time on his hands, he accepted the invitation of the parents of his three boy friends, Tom Tressilis, Gerald Summers, and Reuben Gail. He was most warmly welcomed, for both Tom and Gerald declared that they owed their lives to him. He spent several weeks at each of their homes, and then returned to Plymouth, where he put himself into the hands of a retired master mariner, to learn navigation and other matters connected with his profession, and occupied his spare time in studying the usual branches of a gentleman's education. **who wanted to spend his money?**
RB-Para. 1	"Chapter 10: Southward Ho! Upon making inquiries, Ned Hearne found that Captain Drake had, upon the return of his expedition, set aside the shares of the prize money of Gerald Summers, himself, and the men who were lost in the wreck of the prize, in hopes that they would some day return to claim them. Upon the evidence given by Gerald and himself of the death of the others, their shares were paid, by the bankers at Plymouth who had charge of them, to their families; while Ned and Gerald received their portions. Owing to the great mortality which had taken place among the crews, each of the lads received a sum of nearly a thousand pounds, the total capture amounting to a value of over a million of money. As boys, they each received the half of a man's share. The officers, of course, had received larger shares; and the merchants who had lent money to get up the expedition gained large profits. Ned thought, at first, of embarking his money in the purchase of a share in a trading vessel, and of taking to that service; but, hearing that Captain Drake intended to fit out another expedition, he decided to wait for that event, and to make one more voyage to the Spanish main, before determining on his future course. Having, therefore, his time on his hands, he accepted the invitation of the parents of his three boy friends, Tom Tressilis, Gerald Summers, and Reuben Gail. He was most warmly welcomed, for both Tom and Gerald declared that they owed their lives to him. He spent several weeks at each of their homes, and then returned to Plymouth, where he put himself into the hands of a retired master mariner, to learn navigation and other matters connected with his profession, and occupied his spare time in studying the usual branches of a gentleman's education. **I mean, who wants to spend money on him?**
Original Text 2	"CHAPTER XXIX Frona had gone at once to her father's side, but he was already recovering. Courbertin was brought forward with a scratched face, sprained wrist, and an insubordinate tongue. To prevent discussion and to save time, Bill Brown claimed the floor. "Mr. Chairman, while we condemn the attempt on the part of Jacob Welse, Frona Welse, and Baron Courbertin to rescue the prisoner and thwart justice, we cannot, under the circumstances, but sympathize with them. There is no need that I should go further into this matter. You all know, and doubtless, under a like situation, would have done the same. And so, in order that we may expeditiously finish the business, I make a motion to disarm the three prisoners and let them go." The motion was carried, and the two men searched for weapons. Frona was saved this by giving her word that she was no longer armed. The meeting then resolved itself into a hanging committee, and began to file out of the cabin. "Sorry I had to do it," the chairman said, half-apologetically, half-defiantly. Jacob Welse smiled. "You took your chance," he answered, "and I can't blame you. I only wish I'd got you, though." Excited voices arose from across the cabin. "Here, you! Leggo!" "Step on his fingers, Tim!" "Break that grip!" "Ouch! Ow!" "Pry his mouth open!" Frona saw a knot of struggling men about St. Vincent, and ran over. He had thrown himself down on the floor and, tooth and nail, was fighting like a madman. Tim Dugan, a stalwart Celt, had come to close quarters with him, and St. Vincent's teeth were sunk in the man's arm. **How is Frona's father doing?**
BSc-Para. 2	"CHAPTER XXIX Frona had gone at once to her father's side, but he was already recovering. Courbertin was brought forward with a scratched face, sprained wrist, and an insubordinate tongue. To prevent discussion and to save time, Bill Brown claimed the floor. "Mr. Chairman, while we condemn the attempt on the part of Jacob Welse, Frona Welse, and Baron Courbertin to rescue the prisoner and thwart justice, we cannot, under the circumstances, but sympathize with them. There is no need that I should go further into this matter. You all know, and doubtless, under a like situation, would have done the same. And so, in order that we may expeditiously finish the business, I make a motion to disarm the three prisoners and let them go." The motion was carried, and the two men searched for weapons. Frona was saved this by giving her word that she was no longer armed. The meeting then resolved itself into a hanging committee, and began to file out of the cabin. "Sorry I had to do it," the chairman said, half-apologetically, half-defiantly. Jacob Welse smiled. "You took your chance," he answered, "and I can't blame you. I only wish I'd got you, though." Excited voices arose from across the cabin. "Here, you! Leggo!" "Step on his fingers, Tim!" "Break that grip!" "Ouch! Ow!" "Pry his mouth open!" Frona saw a knot of struggling men about St. Vincent, and ran over. He had thrown himself down on the floor and, tooth and nail, was fighting like a madman. Tim Dugan, a stalwart Celt, had come to close quarters with him, and St. Vincent's teeth were sunk in the man's arm. **What's Frona's father doing?**

ACKNOWLEDGMENTS

I am grateful to Jan Philip Wahle and Terry Lima Ruas for their valuable discussions, feedback, ideas, and help in developing this study.

REFERENCES

[1] Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shnayderman, Noam Slonim, and Liat Ein-Dor. 2022. Quality Controlled Paraphrase Generation. arXiv:2203.10940 [cs.CL]

[2] BIG bench authors. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research* (2023). https://openreview.net/forum?id=uyTL5Bvosj

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[4] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. arXiv:2012.15723 [cs.CL]

[5] Hila Gonen, Srini Iyer, Terra Blevins, Noah A. Smith, and Luke Zettlemoyer. 2022. Demystifying Prompts in Language Models via Perplexity Estimation. arXiv:2212.04037 [cs.CL]

[6] Yoichi Ishibashi, Danushka Bollegala, Katsuhito Sudoh, and Satoshi Nakamura. 2023. Evaluating the Robustness of Discrete Prompts. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 2373–2384. https://doi.org/10.18653/v1/2023.eacl-main.174

[7] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2023. OPT-IML: Scaling Language Model Instruction Meta Learning through the Lens of Generalization. arXiv:2212.12017 [cs.CL]

[8] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics* 8 (2020), 423–438.

[9] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and Applications of Large Language Models. arXiv:2307.10169 [cs.CL]

[10] Abdullatif Köksal, Timo Schick, and Hinrich Schütze. 2023. MEAL: Stable and Active Learning for Few-Shot Prompting. arXiv:2211.08358 [cs.CL]

[11] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. (2022).

[12] Alina Leidinger, Robert van Rooij, and Ekaterina Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful? arXiv:2311.01967 [cs.CL]

[13] Chonghua Liao, Yanan Zheng, and Zhilin Yang. 2022. Zero-label prompt selection. *arXiv preprint arXiv:2211.04668* (2022).

[14] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[15] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

[16] Xianggen Liu, Lili Mou, Fandong Meng, Hao Zhou, Jie Zhou, and Sen Song. 2020. Unsupervised Paraphrasing by Simulated Annealing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 302–312. https://doi.org/10.18653/v1/2020.acl-main.28

[17] Vivek Miglani, Aobo Yang, Aram H. Markosyan, Diego Garcia-Olano, and Narine Kokhlikyan. 2023. Using Captum to Explain Generative Language Models. arXiv:2312.05491 [cs.CL]

[18] Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. Reframing Instructional Prompts to GPTk's Language. In *Findings of the Association for Computational Linguistics: ACL 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 589–612. https://doi.org/10.18653/v1/2022.findings-acl.50

[19] Tong Niu, Semih Yavuz, Yingbo Zhou, Nitish Shirish Keskar, Huan Wang, and Caiming Xiong. 2021. Unsupervised Paraphrasing with Pretrained Language Models. arXiv:2010.12885 [cs.CL]

[20] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback, 2022. *URL https://arxiv. org/abs/2203.02155* 13 (2022), 1.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Pierre Isabelle, Eugene Charniak, and Dekang Lin (Eds.). Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[22] Guanghui Qin and Jason Eisner. 2021. Learning How to Ask: Querying LMs with Mixtures of Soft Prompts. arXiv:2104.06599 [cs.CL]

[23] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[24] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[25] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics* 7 (2019), 249–266. https://doi.org/10.1162/tacl_a_00266

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[27] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense Reasoning about Social Interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 4463–4473. https://doi.org/10.18653/v1/D19-1454

[28] Timo Schick and Hinrich Schütze. 2021. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 2339–2352. https://doi.org/10.18653/v1/2021.naacl-main.185

[29] Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2023. On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. arXiv:2212.08061 [cs.CL]

[30] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV au2, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. arXiv:2010.15980 [cs.CL]

[31] Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. COM2SENSE: A Commonsense Reasoning Benchmark with Complementary Sentences. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 883–898. https://doi.org/10.18653/v1/2021.findings-acl.78

[32] Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2023. Paraphrase Types for Generation and Detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12148–12164. https://doi.org/10.18653/v1/2023.emnlp-main.746

[33] Jan Philip Wahle, Terry Ruas, Frederic Kirstein, and Bela Gipp. 2022. How Large Language Models are Transforming Machine-Paraphrase Plagiarism. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 952–963. https://doi.org/10.18653/v1/2022.emnlp-main.62

[34] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675 [cs.CL]