Prompt Engineering in LLMs: Optimizing Performance

Aldi Halili
Final course report, Profilierungsmodul CL 1 2024-2025
{aldi.halili@campus.lmu.de}

Abstract

Prompts play a crucial role in the performance of large language models (LLMs). This study examines how different prompt types influence LLM performance in comparison to simpler techniques such as Zero-Shot Prompting (current paper). In addition to automatic evaluation, human evaluation is incorporated, and GPT-4 is employed as a model-to-model evaluator. The results indicate that complex prompting techniques, such as Chain-of-Thought (CoT) prompting, are not particularly well-suited for summarization tasks. Furthermore, different models produce varying results depending on the specific prompting techniques used.

1 Introduction

Large language Models (LLMs) have made significant advancements in sequence-to-sequence tasks within Natural Language Processing (NLP). Tasks such as text summarization and simplification have become well-established in the field. However, the same cannot be said about the quality of automatic evaluation metrics for these tasks, as developing a robust evaluation strategy remains an unsolved challenge. Although LLMs themselves can conduct evaluations, the quality of these assessments can depend on effective prompting and prompt-engineering techniques (Song et al., 2025; Mendonça et al., 2023; Chu et al., 2024). As discussed in the course "Trustworthy Data-Centric AI," the quality of prompting and appropiate prompt-technical is crucial for guiding LLMs toward high-quality outputs. Despite this, human evaluation is still often viewed as the most reliable method, as highlighted in the same course. Comparing these three evaluation methodologies - automatic metrics, human evaluation, and model-to-model assessment - across multiple generation tasks and LLMs would provide a more comprehensive outlook and a deeper understanding of current model performance. Significant progress has been made in this area. A hybrid evaluation is employed across multiple datasets, models, and tasks, utilizing automatic metrics, human assessments, and model-to-model evaluation with GPT-4. By analyzing both open-source and commercial LLMs, the study provides a broad overview of currently available models (Sottana et al., 2023). Their approach assesses models on various sequence-to-sequence tasks reframed as text-generation tasks without the need for task-specific fine-tuning including text summarization, text simplification, and grammatical error correction.

2 Summary of the Paper

Sottana et al. (2023) has helped advance our understanding of current model performance. To achieve this, the authors conducted a hybrid evaluation of both open-source and commercial generative LLMs across three NLP benchmarks text summarization, text simplification, and grammatical error correction (GEC) using both automated and human assessments. They also explored the published approach of employing GPT-4 as an evaluator. Their findings indicate that, according to human evaluations, ChatGPT consistently outperforms many other established models on most metrics, while performing significantly worse under traditional automatic metrics. Moreover, human reviewers rated the gold reference as notably inferior compared to the outputs of the top-performing models, highlighting how data quality has become a central bottleneck in evaluation research. Finally, the authors observed that GPT-4's assessments of model outputs largely align with human judgments.

For text simplification, the SARI score is used. For text summarization, the ROUGE score is applied, and for GEC, the F0.5 score computed with the ERRANT toolkit is utilized. For human evaluation of text summarisation, is follow the evaluation criteria and their definitions: Relevance, Fluency, Coherence, and Consistency, on a 5-point Likert

scale from 1 to 5 (Sottana et al., 2023).

For text simplification, the evaluation criteria and definitions from Grabar and Saggion (2022) are used: Semantics, Fluency, and Simplicity, each rated on a 5-point Likert scale. For GEC, the Over-correction criterion from Fang et al. (2023) is adopted, and two new criteria are introduced: Semantics and Grammaticality.

All experiments were conducted in a zero-shot, unsupervised manner, without any additional fine-tuning or in-context learning. The following **Hugging Face open-source models** were implemented and evaluated:

- **Flan-T5** (`google/flan-t5-xxl`)
- **T0pp** (`bigscience/T0pp`)
- **OPT-IML** (`facebook/opt-iml-max-30b`)
- **Flan-UL2** (`google/flan-ul2`)

Additionally, the following **OpenAI models** were used:

- **GPT-3** (`text-davinci-003`)
- **InstructGPT** (`davinci-instruct-beta`)
- **ChatGPT** (`gpt-3.5-turbo-0301`)

(Sottana et al., 2023)

3 Critical Analysis

Despite the significant results, the authors primarily highlight the limitations of the prompts used. In the study, only a limited number of prompts (2 to 5 different ones) were employed, relying on a zero-shot, unsupervised approach. The authors also suggest that more advanced prompting techniques, such as in-context learning or chain-of-thought prompting, which have been shown to significantly improve the performance of generative models, could further enhance the quality of model outputs. However, the quality of the gold references will remain unchanged until new datasets become available. The study of prompts in LLMs is becoming increasingly important, as they significantly influence the output quality and overall performance of the models. Serving as the interface between the user and the model, prompts determine the relevance, precision, and content quality of the generated text. Various types of prompts have been examined across multiple NLP tasks, highlighting their significant impact on model performance. For example, research in machine translation has shown that the structure of a prompt plays a crucial role in determining translation quality (García and Firat, 2022; Zhang et al., 2023). Likewise, studies on question-answering systems have demonstrated that variations in prompt design can lead to substantial differences in performance (Fisch et al., 2019; Roberts et al., 2020). The significance of specificity in prompts for text summarization tasks was highlighted by Kryściński et al. (2018). They argued that prompts with more explicit instructions generally produced more accurate and relevant summaries compared to open-ended, generic prompts. However, a study by Shaib et al. (2023) found that GPT-3.5 was capable of generating high-quality summaries for general-domain articles even in few-shot and zero-shot settings.

In addition to identifying the limitations of various prompt types, this project aims to explore more advanced prompting techniques, such as Chain-of-Thought (CoT) prompting (Kojima et al., 2023), self-consistent CoT (Wang et al., 2023), and Self-Refine: Iterative Refinement with Self-Feedback (Madaan et al., 2023). Furthermore, it explores instruction-based zero-shot prompting (Instructing Prompt-to-Prompt Generation for Zero-Shot Learning) and different prompt formulations.

4 Implementation

This project experimented with five different prompt techniques for only two tasks: the **Summarization Task** and the **Simplification Task**. Due to financial constraints, **Grammatical Error Correction (GEC)** was omitted. Another limitation concerns the data: only **100 samples** were considered. The data used by Sotanna et al., (2023) was utilized for human evaluation. In this study, the 100 samples were used for both human and automatic evaluation. Full access to all three datasets for the three tasks was not feasible. As for the models, all those mentioned in Section 2 were used. For the Hugging Face open-source models, Google Colab with an A100 GPU was used, taking 10 hours, while the OpenAI models were run on a standard GPU, requiring 20 hours for both tasks.

Due to the difficulty of finding annotators, only the Summarization Task was manually evaluated. The evaluation was conducted by two non-native English-speaking annotators with a background in computer science. The template for human annotation was finalized in the form of a JSON file. The JSON file was modified so that both reviewers could enter their scores as values under predefined keys (summary metrics). The reviewers evaluated the output of the three best-performing automatically rated models, as well as the gold score. GPT

evaluated the same model outputs. The code is available on GitHub.

5 Results

This section presents the evaluation results of the models based on different prompting techniques. We analyze both automatic evaluation metrics and human assessments to compare model performance. The first subsection provides insights into automatic evaluation scores, while the second examines human and GPT-4 evaluations, highlighting key differences and correlations.

5.1 Automatic Evaluation Results

As shown in Table 1 and Table 3 (both in the Appendix), using various prompting techniques for the summarization task does not significantly enhance the models' outputs. Only the ChatGPT model achieves a slight improvement in automatic evaluation, while the other two models perform marginally worse. The overall ranking remains unchanged: ChatGPT remains in last place, while T0pp is ranked first.

In contrast, for the simplification task (see Table 1 and Table 3 in the Appendix), all models demonstrate noticeable improvements. Interestingly, the open-source models outperform OpenAI's commercial model in this regard. The top-performing models are facebook/opt-iml-max-30b with a score of 49.50 and google/flan-t5-xxl with 49.52, whereas ChatGPT achieves the lowest score. However, there has been no human evaluation for this particular task.

This absence of human judgment introduces uncertainty as to whether reviewers would concur with the automatic scores. For instance, in the few-shot prompting approach for the summarization task, ChatGPT performed the worst according to the automatic metrics—but was ranked highest in the human evaluation (Sottana et al., 2023). This study has also shown similar results (see Table 1 and Table 2 in the Appendix), with the difference that the average of two reviewers was used instead of three, and the reviewers were not native speakers. This could be another limitation of the study.

5.2 Human and GPT-4 Evaluation Results

Human reviewers rated each model's output based on the metrics in Section 2. Their scores were converted into rankings from best (1) to worst (4) for each model and reviewer, and then averaged. Table 2 (Appendix) presents the rankings from human evaluation and GPT-4 (in brackets), along with the interval Krippendorff's α coefficient.

The T0pp model showed a significant decline in relevance according to the reviewers. GPT-4 shares the same opinion. In contrast, GPT-3 demonstrated a notable improvement in both relevance and consistency. However, the scores from human reviewers and GPT-4 differ. Reviewers believe that ChatGPT performed slightly better on relevance and fluency metrics, whereas GPT-3 considers its relevance performance to have declined. In contrast to the automatic evaluation, the reviewers ranked ChatGPT together with GPT-3 as the best models. T0pp was rated significantly lower. Compared to the Zero-Shot Prompting results (Sottana et al., 2023), these rankings remained consistent. However, ChatGPT was not consistently rated as the best model by all reviewers for summarization. This suggests that consistency remains an issue when using the new prompting technique.

6 Reflection & Conclusion

Prompt engineering is currently a major focus within LLMs research. Different studies have yielded varied results, and our findings reinforce that diverse prompting techniques affect different aspects of a task—and different models—in distinct ways. For the summarization task, introducing new prompting techniques led to mixed outcomes across four dimensions of summary quality: while one dimension improved significantly, another worsened. This suggests that crafting prompts that explicitly address all four dimensions might yield more balanced improvements.

Moreover, our results indicate that chain-of-thought prompting may not be particularly effective for summarization (see the repository results for details). One limitation of this study is the absence of human evaluations for tasks beyond summarization, which constrains the generalizability of our conclusions to text generation tasks more broadly.

Additionally, text summarization can be divided into extractive and abstractive methods (Liu and Lapata, 2019). It might be advantageous to tailor the prompt design to a single category of summarization for clearer, more targeted outcomes. Once again, GPT-4 proved to be an effective evaluator, showing a strong—though not perfect—correlation with the reviewers' assessments and again all human reviewers rated the gold reference summaries

as the worst across every metric.

References

KuanChao Chu, Yi-Pei Chen, and Hideki Nakayama. 2024. A better llm evaluator for text generation: The impact of prompt output sequencing and optimization. *arXiv preprint arXiv:2406.09972*.

Tao Fang, Shu Yang, Kaixin Lan, Derek F Wong, Jinpeng Hu, Lidia S Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *arXiv preprint arXiv:2304.01746*.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China. Association for Computational Linguistics.

Xavier García and Orhan Firat. 2022. Using natural language prompts for machine translation. *ArXiv*, abs/2202.11822.

Natalia Grabar and Horacio Saggion. 2022. Evaluation of automatic text simplification: Where are we now, where should we go from here. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 453–463, Avignon, France. ATALA.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.

John Mendonça, Patrícia Pereira, Helena Moniz, João Paulo Carvalho, Alon Lavie, and Isabel Trancoso. 2023. Simple llm prompting is state-of-the-art for robust and multilingual dialogue evaluation.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model?

Chantal Shaib, Millicent Li, Sebastian Joseph, Iain Marshall, Junyi Jessy Li, and Byron Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using GPT-3 (with varying success). pages 1387–1407.

Mingyang Song, Mao Zheng, Xuan Luo, and Yue Pan. 2025. Can many-shot in-context learning help llms as evaluators? a preliminary empirical study.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of GPT-4: Reliably evaluating large language models on sequence to sequence tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8776–8788, Singapore. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models.

Geng Zhang, Qi Chen, Zhifang Zhao, Xinle Zhang, Chao Jiangqin, Dingyi Zhou, Wang Chai, Haiying Yang, Zhibin Lai, and Yangyidan He. 2023. Zhang et al 2023. *Remote Sensing*, 15.

A Prompt Engineering

Below are the formulations of advanced prompt engineering techniques for tasks such as Text Summarization, Text Simplification, and Grammatical Error Correction.

A.1 Text Summarization

```
(a)
( "Provide a structured summary of the
    following text. Follow these steps
    internally but output only the final
    summary:\n\n"
"Step 1: Identify a suitable title based on the
     text.\n"
"Step 2: Extract the three most important key
    points.\n"
"Step 3: Generate a concise summary in 1-2
    sentences using the key points.\n\n"
"**Important:** Do not include the title or key
     points in the output. Only return the
    final summary.\n\n"
"Text: [...]" ),

(b)
( "Summarize the following text in 2-3
    sentences while keeping the key
    information intact.\n"
"**Important:** Output only the final summary,
    without explanations or bullet points.\n"
"Text: [...]" ),

(c)
( "Generate a highly concise summary of the
    following text in a **single sentence**
    while preserving its main idea.\n"
"**Return only the final summary sentence,
    without any additional text.**\n\n" "Text:
     [...]" ),

(d)
( """Summarisation the following Text: [...],
    ensuring to capture key points, highlight
    crucial arguments and provide a concise
    yet comprehensive understanding of the
    content""" ),

(e)
( """Provide a summary of the following Text:
    [...] highlighting the key themes and
    arguments""")
```

A.2 Text simplification

```
(a)
( "Generate TWO different simplified versions
    of the text below. "
"Then pick the one that is simpler and clearer.
     "
"**Important:** Output only your final chosen
    version.\n\n"
"Original text: [...]" ),

(b)
( "Step 1: Identify any difficult words or
    phrases in the text.\n"
"Step 2: Replace them with simpler alternatives
     and rewrite the text in short sentences."
"**Important:** Provide only the simplified
    text.\n\n"
"Text to simplify: [...]" ),

(c)
( "Step 1: Rewrite the text below in a simpler
    and clearer way, while preserving its
    meaning.\n"
"Step 2: Analyze your simplified text and
    identify any remaining complex words or
    unclear structures.\n"
"Step 3: Improve the simplification by making
    it even more accessible. Ensure sentences
    are short, direct, and free of jargon.\n"
"**Important:** Output only the final refined
    version.\n\n"
"Text to simplify: [...]" ),

(d)
( "Rewrite the text below so that it is easily
    understandable for a 8-year-old children.\
    n"
"Ensure that:\n"
"- Sentences are short and direct.\n"
"- Difficult words are replaced with simple
    alternatives.\n"
"- The overall meaning remains the same.\n\n"
"**Important:** Output only the final refined
    version.\n\n"
"Text to simplify: [...]" ),

(e)
( "Step 1: Identify the challenging words,
    phrases, or sentence structures in the
    text and explain why they might be
    difficult to understand.\n"
"Step 2: Rewrite the text in simpler language
    while ensuring that the original meaning
    is preserved.\n"
"Step 3: Verify that the rewritten version is
    clear and easy to understand.\n"
"**Important:** Provide only the final,
    simplified version.**\n\n"
"Text to simplify: [...]" )
```

A.3 Grammatical Error Correction

```
(a)
( "Perform a two-step grammar correction:\n"
"1) Correct all obvious grammar/spelling
    mistakes.\n"
"2) Re-check the sentence and refine if needed
    .\n"
"**Important:** Only return the final corrected
     version..\n\n"
"Input sentence: [...]" ),

(b)
( "First pass: fix grammatical errors.\n"
"Second pass: ensure the sentence is fluent and
     natural in standard English.\n"
"**Important:** Return the final corrected
    version only.\n\n"
"Sentence: [...]"),
(c)
( "Step 1: Identify and list any grammar,
    spelling, or fluency errors in the
```

```
    sentence.\n"
"Step 2: Explain why each mistake is incorrect
    and how it can be improved.\n"
"Step 3: Rewrite the corrected sentence in
    standard English with natural fluency.\n"
"**Important:** Return only the final corrected
     version.\n\n, without any additional text
    .\n\n"
"Sentence: [...]" ),

(e)
( "Perform a multi-level correction of the
    following sentence:\n"
"1) Fix all grammatical, spelling, and
    punctuation mistakes.\n"
"2) Improve clarity by simplifying complex
    structures.\n"
"3) Enhance precision by removing unnecessary
    words or ambiguity.\n"
"**Important:** Provide only the final improved
     version.\n\n"
"Sentence: [...]" )
```

B Advanced Prompting Techniques Result

Table 1: Automatic evaluation of the best open-source model and two commercial models from OpenAI using new prompting techniques.

Task	Model	Open source	Temperature	Score (human eval. subset)
Summarisation(ROUGE score)	T0pp	Yes	0.01	**27.30**
	GPT-3	No	0.7	26.39
	ChatGPT	No	0	26.716
Simplification(SARI score)	Flan-T5	Yes	0.01	**48.71**
	InstructGPT	No	0.01	37.73
	ChatGPT	No	0	36.82

Table 2: Average human annotator rankings (GPT-4 rankings in brackets)

SUMMARISATION	RELEVANCE ($\alpha_1 = 0.95$, $\alpha_2 = 0.90$)	FLUENCY ($\alpha_1 = 0.90$, $\alpha_2 = 0.84$)	COHERENCE ($\alpha_1 = 0.81$, $\alpha_2 = 0.94$)	CONSISTENCY ($\alpha_1 = 1.00$, $\alpha_2 = 0.84$)
Gold reference	3.00 (3.00)	4.00 (3.00)	4.00 (3.00)	4.00 (3.00)
T0pp	4.00 (4.00)	3.00 (4.00)	3.00 (4.00)	2.75 (4.00)
GPT-3	1.25 (1.00)	1.75 (2.00)	2.00 (1.50)	1.25 (2.50)
ChatGPT	1.25 (2.00)	1.25 (1.00)	1.00 (1.50)	2.00 (2.50)

C Zero-Shot-Prompting Results

This section presents the evaluation results reported in the paper (Sottana et al., 2023).

Table 3: Automatic evaluation of the best open-source model and two commercial models from OpenAI. Results are shown both on the main subset and the small subset used for human evaluation. †Due to the specifics of HuggingFace implementation, a temperature of 0.0 cannot be used; we therefore used a value of 0.01 for such cases.

Task	Model	Open source	Temperature	Score (main subset)	Score (human eval. subset)
Summarisation(ROUGE score)	T0pp	Yes	0.01†	**28.82**	**31.62**
	GPT-3	No	0	24.22	27.19
	ChatGPT	No	0	23.76	25.72
Simplification(SARI score)	Flan-T5	Yes	0.01†	**44.98**	**44.61**
	InstructGPT	No	0	44.79	43.25
	ChatGPT	No	0	37.55	35.01

Table adapted from the paper (Sottana et al., 2023).

Table 4: Average human annotator rankings (GPT-4 rankings in brackets)

SUMMARISATION	RELEVANCE ($\alpha_1 = 0.88$, $\alpha_2 = 0.81$)	FLUENCY ($\alpha_1 = 0.88$, $\alpha_2 = 0.82$)	COHERENCE ($\alpha_1 = 1.00$, $\alpha_2 = 0.91$)	CONSISTENCY ($\alpha_1 = 0.97$, $\alpha_2 = 0.86$)
Gold reference	4.00 (3.00)	4.00 (3.00)	4.00 (3.00)	4.00 (3.00)
T0pp	3.00 (4.00)	3.00 (4.00)	3.00 (4.00)	2.83 (4.00)
GPT-3	1.67 (2.00)	1.67 (1.50)	2.00 (2.00)	2.17 (2.00)
ChatGPT	1.33 (1.00)	1.33 (1.50)	1.00 (1.00)	1.00 (1.00)

Table adapted from the paper (Sottana et al., 2023).