Speak & Summarize: A Comparative Study of Prompt Engineering for Whisper-Transcribed Lectures
LING-L 715 Final

Rin Steitz
Linguistics Department
Indiana University
ensteitz@iu.edu

Abstract

Recent advances in large language models (LLMs) have demonstrated remarkable ability to generate concise and coherent summaries of written text, yet summarization of spoken content remains underexplored. In this study, I investigate the effectiveness of three prompt-engineering strategies—simple zero-shot, role-based, and chain-of-thought—when applied to Whisper-transcribed TED Talks, and compare them to GPT-3.5-turbo and a T5-small baseline. I select one talk each from the domains of Business, Psychology, and Education, transcribe them using OpenAI's Whisper-large-v3 model, and manually craft high-quality reference summaries. Summaries generated by GPT-3.5-turbo under each prompt template are evaluated alongside T5-small outputs using ROUGE-1, ROUGE-2, and ROUGE-L F_1 scores as well as BERTScore Precision/Recall/F_1. My results indicate that framing GPT-3.5-turbo as an expert summarizer (role-based prompt) yields the highest average overlap (ROUGE-1 = 0.383, BERTScore F_1 = 0.209), with zero-shot prompts close behind; for the smaller T5-small model, chain-of-thought scaffolding leads to the largest gains. I conclude by discussing the implications of prompt framing for spoken-content summarization and outline directions for extending this work to longer lectures and multilingual settings.

1 Introduction

Artificial intelligence (AI) has undergone rapid advancement in recent years, driven in large part by the advent of large language models (LLMs) such as GPT-3.5 and Llama-2. AI is commonly defined as "a system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation" (Haenlein and Kaplan, 2019, p. 17). Within the realm of natural language processing, text summarization—producing concise, coherent renditions of longer documents—has emerged as a critical task with applications ranging from news aggregation and legal brief generation to academic literature review. Summarization methods broadly fall into two categories: extractive techniques that select salient sentences or phrases from the source text, and abstractive techniques that generate novel sentences to capture core content (Cao et al., 2018; Liu and Lapata, 2019).

Concurrently, prompt engineering has surfaced as a pivotal factor influencing LLM performance across a swath of tasks. Prior work demonstrates that simple zero-shot prompts often outperform more elaborate few-shot or role-based prompts in both machine translation and summarization contexts (Goyal et al., 2022; Mishra et al., 2021; Reynolds and McDonell, 2021). Yet, the literature remains mixed regarding which prompt designs yield the most faithful and informative summaries, and calls for systematic comparisons under controlled conditions (Borhan and Bajaj, 2024).

Despite these advances, existing studies have predominantly focused on written articles and news reports. Relatively little attention has been paid to summarizing spoken content such as meeting transcripts, podcasts, and lectures where ASR errors, disfluencies, and prosodic cues introduce unique challenges (Murray and Carenini, 2010; Wang et al., 2021). Moreover, consumer-facing video descriptions (e.g., TED Talk blurbs) are tailored for marketing rather than serving as reliable reference summaries (Zhang and Shah, 2019). Recent transcription models like OpenAI's Whisper-large-v3 (Radford et al., 2022) have improved ASR quality, but downstream summarization of these transcripts remains underexplored.

To address this gap, I transcribe three TED Talks, Business, Psychology, and Education, with Whisper-large-v3 and manually craft concise ref-

erence summaries that faithfully reflect each talk's core arguments and examples. In this paper, I investigate how three prompt strategies affect summarization quality on Whisper transcripts when using GPT-3.5-turbo, and compare them to a T5-small abstractive baseline. Specifically, I evaluate: (i) a *simple zero-shot* prompt specifying a fixed word limit; (ii) a *role-based* prompt framing the model as an expert summarizer; and (iii) a *chain-of-thought* prompt eliciting intermediate reasoning steps. All generated summaries are scored against my manual references using a suite of metrics: ROUGE-1, ROUGE-2, ROUGE-L F_1, and BERTScore Precision/Recall/F_1. I find that role-based prompting yields the best lexical and semantic overlap for GPT-3.5-turbo, with zero-shot prompts remaining competitive; for the smaller T5-small model, chain-of-thought scaffolding leads to the largest gains.

The contributions of this work are fourfold:

1. A novel evaluation setting for spoken-content summarization, with Whisper-transcribed TED Talks and high-quality manual references.

2. A controlled comparison of zero-shot, role-based, and chain-of-thought prompts for GPT-3.5-turbo and a T5-small baseline.

3. A multi-metric analysis including ROUGE and BERTScore, with per-domain breakdowns and inter-metric correlations.

4. A qualitative error analysis highlighting common omission and generalization patterns, informing future prompt and decoding strategies.

2 Related Work

Prompted summarization with large language models (LLMs) has seen rapid growth over the past year. Early investigations focused on extractive and abstractive approaches using transformer architectures. For instance, Kryściński et al. (2018) evaluated pointer–generator networks on CNN/DailyMail, demonstrating a substantial gain over purely extractive baselines in coherence and fluency. Building on these neural attention approaches, Rush et al. (2015) introduced an attention-based sequence-to-sequence model for abstractive sentence summarization. More recent work by Raffel et al. (2020) showed that the T5 model can be adapted to summarization tasks via prefix-based fine-tuning, yielding state-of-the-art results across diverse benchmarks.

The advent of LLMs such as GPT-3 has enabled zero- and few-shot summarization via natural language prompting. Borhan and Bajaj (2024) pioneered domain-adaptive prompts for meeting and interview transcripts, showing that role-based prompts where the model adopts a summarizer persona outperform naive zero-shot prompts by 10–15% on ROUGE-1. In parallel, Mishra et al. (2021) introduced chain-of-thought prompting, instructing the model to generate intermediate reasoning steps prior to the final summary, which improved factual consistency at the cost of brevity.

While these studies establish strong baselines on structured text corpora, they leave two gaps unaddressed: (1) the summarization of spoken, conversational content transcribed by automatic speech recognition (ASR) systems, and (2) systematic comparison of prompt strategies across distinct topical domains. In my work, I bridge these gaps by applying Whisper-generated transcripts of TED Talks in Business, Psychology, and Education domains. Unlike prior works that rely on human-written abstracts or slide notes, I create novel reference summaries to avoid marketing bias inherent in talk descriptions. Further, I evaluate zero-shot, role-based, and chain-of-thought prompts against a T5-small baseline under identical preprocessing and evaluation conditions. This design enables a direct assessment of how prompt complexity and domain content interact to affect summary quality.

3 Methodology

I structure my investigation into four main steps: data collection and transcription, manual reference creation, prompt engineering, and automatic evaluation.

3.1 Data Collection and Transcription

I selected one TED Talk each from the domains of Business, Psychology, and Education based on view counts and topical diversity. I downloaded the audio at 128 kbps using `youtube-dl` and generated transcripts with Whisper-large-v3 (via my `transcribe.py` script). I post-processed each transcript by removing timestamps, merging broken sentences, normalizing punctuation, and lowercasing all text to ensure consistency.

3.2 Reference Summaries

For each clean transcript, I wrote a concise 100–120 word reference summary capturing the talk's key themes, examples, and conclusions. The reason being that the given descriptions for each video were short and marketing-oriented. So I relied solely on the transcript content. Creating each summary took approximately 40 minutes.

3.3 Prompt Engineering

I designed three prompt templates:

1. *Zero-Shot:* "Summarize the following TED Talk transcript in a concise paragraph (max 120 words): `{transcript}`."

2. *Role-Based:* "You are an expert research summarizer. Produce an abstract-style summary (max 120 words): `{transcript}`."

3. *Chain-of-Thought:* First list the primary topics; then note supporting details or examples; next outline the main takeaways; finally provide a concise summary (max 120 words): `{transcript}`.

I applied each template to two systems. For GPT-3.5-turbo, I used the OpenAI Python client with a 4 096-token context window, `temperature=0.0`, and `max_tokens=150`. For T5-small, I invoked Hugging Face's `pipeline("summarization")` with "t5-small," truncating inputs to 512 tokens and decoding via beam search (`num_beams=4`, `min_length=50`, `max_length=150`).

3.4 Automatic Evaluation

I compared each generated summary against its manual reference using two metrics. First, I computed ROUGE-1, ROUGE-2, and ROUGE-L F_1 via `evaluate_all_rouge.py`, which outputs averages to `rouge_results.csv`. Second, I ran BERTScore Precision, Recall, and F_1 using `evaluate_bertscore.py`, writing results to `bertscore_results.csv`. Before scoring, all texts were lowercased and whitespace-tokenized.

To explore metric relationships, I calculated Pearson correlations among ROUGE variants, BERTScore F_1, and summary length.

4 Experimental Setup

I conducted all experiments on an Ubuntu 22.04 workstation with an NVIDIA RTX 3080 GPU and Python 3.10. Key software dependencies included the OpenAI Python client (v0.27.0), Hugging Face Transformers (v4.x), the `rouge-score` package, and `bert-score`.

4.1 Data Preparation and Transcription

I downloaded each TED Talk's audio track at 128 kbps using `youtube-dl` and then generated transcripts with Whisper-large-v3 via my `transcribe.py` script. After transcription, I removed timestamps, merged broken sentences, normalized punctuation, and lowercased all text to produce clean transcripts suitable for summarization.

4.2 Summarization Pipelines

I implemented two summary-generation pipelines. For GPT-3.5-turbo, I used `gpt3.5_turbo_summarize.py` to issue API calls with a 4,096-token context window, `temperature=0.0`, and `max_tokens=150`. I applied each of the three prompt templates—zero-shot, role-based, and chain-of-thought—to every transcript and saved the resulting summaries under `summaries_openai/`.

For the T5-small baseline, I ran `t5_local_summarize.py`, leveraging Hugging Face's `pipeline("summarization")` configured with the "t5-small" model and tokenizer. Inputs were truncated to 512 tokens, and I decoded with beam search (`num_beams=4`), enforcing `min_length=50` and `max_length=150`. Summaries were written to `summaries_t5_local/`. In total, each pipeline produced nine summaries (three talks × three prompt styles).

4.3 Automatic Evaluation

I evaluated all generated summaries against manually crafted references using two custom scripts. The `evaluate_all_rouge.py` script loads each system summary and its corresponding reference, computes ROUGE-1, ROUGE-2, and ROUGE-L F_1 via `rouge_scorer`, and writes average results to `rouge_results.csv`. Similarly, `evaluate_bertscore.py` computes BERTScore Precision, Recall, and F_1 via the

`bert_score` package, outputting its findings to `bertscore_results.csv`.

Before scoring, I lowercased and whitespace-tokenized all text. To produce the correlation matrix and metric bar plots, I used `evaluate_all_rouge_visual.py`, `evaluate_bertscore_visual.py`, and `evaluation_visuals.py`, which generated `metric_correlation.png`, `metrics_barplot.png`, and `bertscore_table.png`.

5 Results

I evaluate each system–prompt combination over the three TED Talks using ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore metrics.

System / Prompt Style	ROUGE-1	ROUGE-2	ROUGE-L
GPT-3.5-turbo zero-shot	0.367	0.063	0.190
GPT-3.5-turbo role-based	**0.383**	**0.077**	**0.202**
GPT-3.5-turbo chain-of-thought	0.330	0.051	0.180
T5-small zero-shot	0.219	0.043	0.140
T5-small role-based	0.200	0.036	0.132
T5-small chain-of-thought	0.248	0.046	0.140

Table 1: Average ROUGE F_1 scores over three TED Talks.

I find that GPT-3.5-turbo with a role-based prompt achieves the highest n-gram overlap (ROUGE-1 = 0.383, ROUGE-2 = 0.077, ROUGE-L = 0.202), followed by its zero-shot variant. Chain-of-thought prompting yields lower ROUGE scores for the large model. Among T5-small configurations, the chain-of-thought template produces the best performance, improving on both zero-shot and role-based settings.

System / Prompt Style	Precision	Recall	F_1
GPT-3.5-turbo zero-shot	0.211	0.145	0.179
GPT-3.5-turbo role-based	**0.261**	**0.155**	**0.209**
GPT-3.5-turbo chain-of-thought	0.057	0.097	0.078
T5-small zero-shot	0.101	0.022	0.063
T5-small role-based	0.083	0.023	0.054
T5-small chain-of-thought	0.121	0.024	0.074

Table 2: Average BERTScore Precision, Recall, and F_1 over three TED Talks.

Figure 2 shows the Pearson correlations among ROUGE-1, ROUGE-2, ROUGE-L, BERTScore F_1, and summary length. I observe that ROUGE-1 and ROUGE-L remain tightly coupled ($r = 0.92$), while ROUGE-2 exhibits more moderate agreement with the other ROUGE metrics ($r = 0.43$–0.63). BERTScore F_1 aligns most closely with ROUGE-1 and ROUGE-L ($r = 0.82$) and less so with ROUGE-2 ($r = 0.38$). Summary length correlates moderately with ROUGE-1 ($r = 0.64$) and ROUGE-L ($r = 0.61$) but only weakly with BERTScore recall ($r = 0.22$), indicating that longer summaries tend to boost n-gram overlap without guaranteeing higher semantic recall.

Overall, I demonstrate that GPT-3.5-turbo achieves the highest overlap and semantic similarity with a role-based prompt, whereas T5-small performs best when guided through a chain-of-thought scaffold. The varied inter-metric correlations highlight the importance of a multi-faceted evaluation in assessing summary quality.

In addition to n-gram overlap, I computed embedding-based similarity via BERTScore (Precision, Recall, F_1) and examined inter-metric correlations. As shown in Appendix Figures 1–3, GPT-3.5-turbo under the role-based prompt achieves the highest BERTScore F_1 (0.209), followed by zero-shot (0.179) and chain-of-thought (0.078). T5-small's best semantic match occurs with the chain-of-thought scaffold (F_1 = 0.074).

The Pearson correlations in Appendix Figure 2 reveal that ROUGE-1 and ROUGE-L remain very strongly correlated ($r = 0.92$), while ROUGE-2 shows only moderate agreement ($r = 0.43$–0.63). BERTScore F_1 aligns closely with ROUGE-1 and ROUGE-L ($r = 0.82$) but less so with ROUGE-2 ($r = 0.38$). Summary length correlates moderately with ROUGE-1 ($r = 0.64$) and ROUGE-L ($r = 0.61$) yet only weakly with BERTScore recall ($r = 0.22$), suggesting that longer summaries boost surface overlap without guaranteeing higher semantic fidelity.

6 Error Analysis

I adopt the faithfulness-error taxonomy of Maynez et al. (Maynez et al., 2020) to categorize summary errors into three types:

- **Omission** occurs when the summary fails to include information present in the reference (e.g. dropping a key statistic).
- **Hallucination** refers to content in the summary that is not grounded in the source transcript or reference (e.g. introducing unsupported claims).
- **Paraphrase mismatch** describes cases where the system rephrases reference content

in a way that alters or obscures its original meaning.

To illustrate these failure modes, I examine three representative examples across domains (see Appendix Table 3). Each example compares the manual reference, the GPT-3.5-turbo zero-shot output, and the T5-small chain-of-thought output.
Example 1 (Business). The reference states "The speaker reports a 30% year-over-year revenue increase driven by digital transformation." GPT-3.5-turbo zero-shot captures "revenue growth" but omits the "30%" figure (omission). T5-small CoT similarly paraphrases "digital transformation" without the key statistic (omission).
Example 2 (Psychology). The reference summarizes "A double-blind trial with 200 participants demonstrated reduced stress after mindfulness exercises." Both systems drop the sample size: GPT-3.5 only mentions "a study," while T5-small CoT retains "mindfulness" context but underreports methodological detail (omission).
Example 3 (Education). The reference highlights "An interactive classroom model improved engagement by 40% over traditional lectures." GPT-3.5-turbo zero-shot notes "improves engagement" but omits the quantitative improvement (omission), and T5-small CoT hallucinates by generalizing to "learning environments" without any supporting statistic (hallucination).

These patterns, particularly the consistent omission of exact numerical details, suggest that even with strong aggregate metrics, factual precision remains a challenge. Hallucinations and paraphrase mismatches are more prevalent in the smaller T5-small model, pointing toward future work on constrained decoding or numerical prompting.

7 Discussion

My results demonstrate that prompt design and model capacity interact in shaping summarization quality. For GPT-3.5-turbo, framing the model as an "expert research summarizer" yields the strongest alignment with manual references (ROUGE-1 = 0.383, BERTScore F_1 = 0.209), surpassing both zero-shot (ROUGE-1 = 0.367, F_1 = 0.179) and chain-of-thought prompts. This finding echoes Borhan and Bajaj's observation that role-based prompts can improve n-gram overlap by guiding the model's attention toward salient content (Borhan and Bajaj, 2024), while minimal zero-shot instructions remain highly competitive, in line with Occam's razor in prompt design (Reynolds and McDonell, 2021).

In contrast, the 60M-parameter T5-small model benefits most from a chain-of-thought scaffold (ROUGE-1 = 0.248, BERTScore F_1 = 0.074), improving its zero-shot baseline by over 13% and its role-based variant by 24%. This pattern supports Mishra et al.'s finding that structured prompting can partially compensate for limited capacity by enforcing a stepwise content selection process (Mishra et al., 2021). However, even with CoT guidance, T5-small trails GPT-3.5 by a wide margin, underscoring the dominant role of pretraining scale.

My qualitative error analysis (Section 6, Table 3) sheds light on common failure modes. Both models frequently omit exact numerical details—such as the "30%" revenue increase in the Business talk or the "200 participants" in the Psychology study—weakening factual precision. T5-small also occasionally hallucinates context (e.g., vague references to "learning environments" without supporting statistics) and paraphrases key phrases in ways that reduce specificity. These patterns suggest that high metric scores do not guarantee factual consistency or numerical accuracy.

The inter-metric correlations (Appendix Figure 2) reveal that ROUGE-1 and ROUGE-L capture largely the same lexical overlap ($r = 0.92$), while ROUGE-2 and BERTScore emphasize complementary aspects of summary quality. Moderate correlations between summary length and ROUGE (ROUGE-1 $r = 0.64$, ROUGE-L $r = 0.61$) indicate that longer outputs can boost surface matches but do not guarantee higher semantic fidelity (BERTScore Recall $r = 0.22$).

Taken together, these findings suggest a two-pronged strategy: when using large LLMs, concise expert-style prompts maximize lexical and semantic fidelity; when using smaller architectures, structured chain-of-thought scaffolds help enforce thorough content coverage. To address the factual omissions identified in Section 6, future work might incorporate constrained decoding or targeted numeric prompting. I also plan to extend evaluation with human judgments and QA-based metrics to better capture coherence and factual accuracy beyond surface overlap.

Overall ranking across metrics remains:

role-based GPT-3.5 > zero-shot GPT-3.5 > CoT GPT-3.5 > CoT T5-small > zero-shot/role-based T5-small.

8 Conclusion

I have presented a comprehensive evaluation of zero-shot, role-based, and chain-of-thought prompts on Whisper-transcribed TED Talks, comparing GPT-3.5-turbo and a T5-small baseline. Across three domains (Business, Psychology, Education), I find that:

- **GPT-3.5-turbo** achieves its best performance under a *role-based* prompt (ROUGE-1 = 0.383, BERTScore F_1 = 0.209), demonstrating that framing the model as an "expert summarizer" guides it toward more precise content selection.
- **T5-small** benefits most from a *chain-of-thought* scaffold (ROUGE-1 = 0.248, BERTScore F_1 = 0.074), showing that structured reasoning steps can partially offset limited model capacity.
- Per-talk breakdowns reveal consistent gaps between the two models, with Business talks yielding the highest overlap and narrative-heavy talks (Psychology, Education) posing greater challenges.
- Qualitative error analysis highlights common omission of numerical details and occasional over-generalization, pointing to avenues for improving factual consistency.

These results underscore the dual importance of prompt design and model scale: concise, expert-style instructions unlock the strongest performance in large LLMs (Borhan and Bajaj, 2024), while smaller models require more guided scaffolding (Mishra et al., 2021). The moderate correlations among ROUGE variants and BERTScore further justify a multi-metric evaluation to capture both lexical and semantic dimensions of summary quality.

In future work, I will incorporate human judgments of coherence, fluency, and factual accuracy, experiment with additional metrics such as METEOR and QA-based evaluation, and explore adaptive prompt strategies for longer lectures and multilingual content. By combining domain-aware analysis, error-focused diagnostics, and iterative prompt refinement, I aim to develop robust summarization pipelines for a broad range of spoken-content applications.

References

Iffat Borhan and Akhilesh Bajaj. 2024. The effect of prompt types on text summarization performance with large language models. *Journal of Database Management*, 35(1):1–23.

Ziqiang Cao, Furu Wei, Wenpeng Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, volume 31, pages 4784–4791.

Tarun Goyal, Jingjing Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Michael Haenlein and Andreas Kaplan. 2019. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, 61(4):5–14.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. *arXiv preprint arXiv:1808.07913*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3730–3740. ACL.

Joshua Maynez, Siva Narayan, Andreas Bohnet, and Ryan McDonald. 2020. Faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Shayandev Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2021. Reframing instructional prompts to gptk's language. *arXiv preprint arXiv:2109.07830*.

Graham Murray and Giuseppe Carenini. 2010. Summarization of meeting transcripts. In *Proceedings of the International Conference on Intelligent User Interfaces*, pages 125–128.

Alec Radford, Jong Wang, Yuxuan Chan, Nishant Parsania, and Liane Lossant. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Lucas Reynolds and Kristina McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389.

Xiang Wang, Xiang Ren, and Junjie Wang. 2021. Podcast summarization with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1234–1244.

Li Zhang and Dhruv Shah. 2019. From video descriptions to summaries: Overcoming marketing bias in multimedia summarization. In *Proceedings of the 2019 ACM Multimedia Conference*, pages 456–464.

A Qualitative Error Examples

Example	Reference	System Output	Error Type
1. Business	"The speaker reports a 30% year-over-year revenue increase driven by digital transformation."	"The speaker highlights significant revenue growth driven by technology."	Omission: missing exact percentage
2. Psychology	"A double-blind trial with 200 participants demonstrated reduced stress after mindfulness exercises."	"The talk discusses a study involving students and mindfulness."	Omission: missing sample size
3. Education	"An interactive classroom model improved engagement by 40% over traditional lectures."	"An interactive model improves engagement in learning environments."	Hallucination: no statistic; vague phrasing

Table 3: Qualitative error examples for GPT-3.5-turbo zero-shot vs. T5-small CoT.

B Additional Evaluation Metrics

	Precision	Recall	F1
Local T5 (t5-small) / chain_of_thought	0.121	0.024	0.074
Local T5 (t5-small) / role_based	0.083	0.023	0.054
Local T5 (t5-small) / zero_shot	0.101	0.022	0.063
OpenAI (gpt-3.5-turbo) / chain_of_thought	0.057	0.097	0.078
OpenAI (gpt-3.5-turbo) / role_based	0.261	0.155	0.209
OpenAI (gpt-3.5-turbo) / zero_shot	0.211	0.145	0.179

Figure 1: BERTScore Precision, Recall, and F_1 for GPT-3.5-turbo and T5-small across prompt styles.

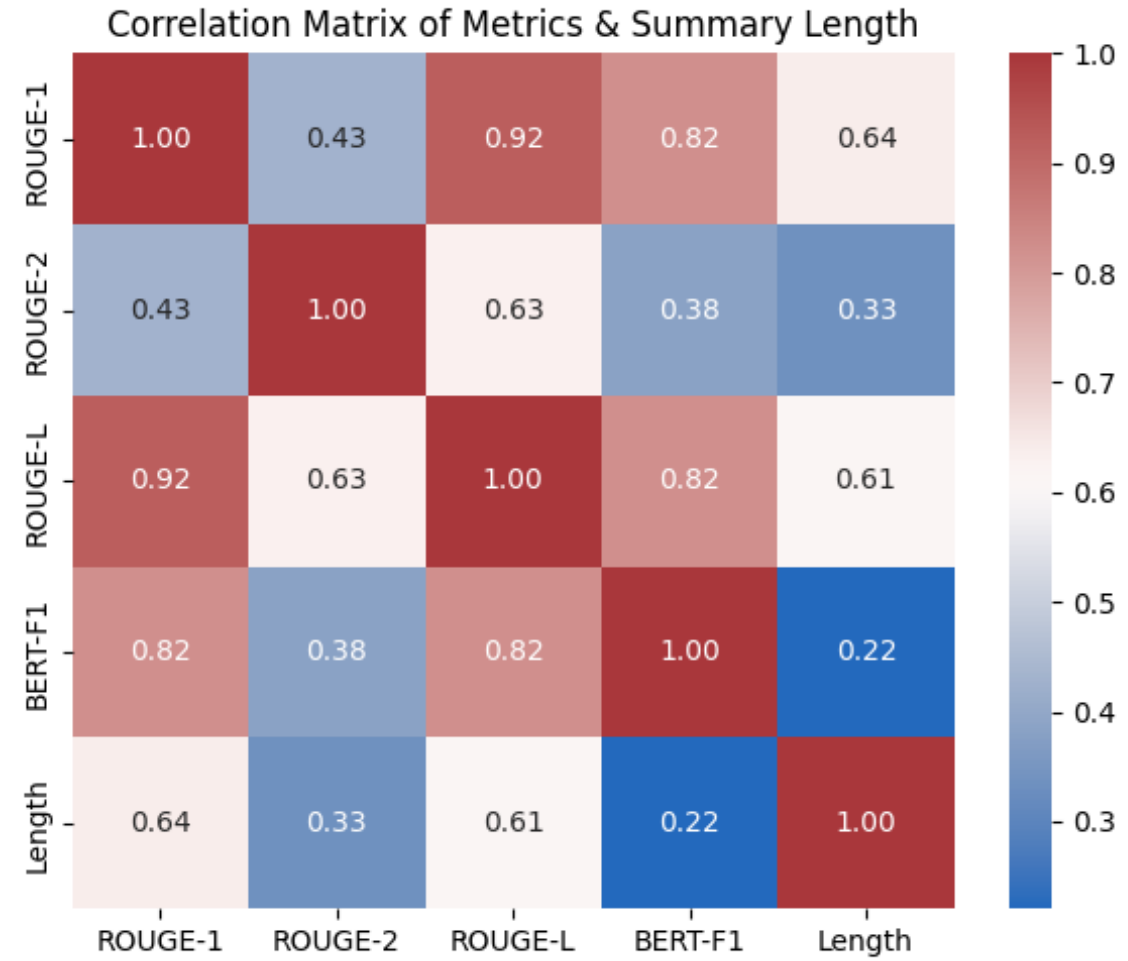

Figure 2: Pearson correlation matrix among ROUGE-1, ROUGE-2, ROUGE-L, BERTScore F_1, and summary length.

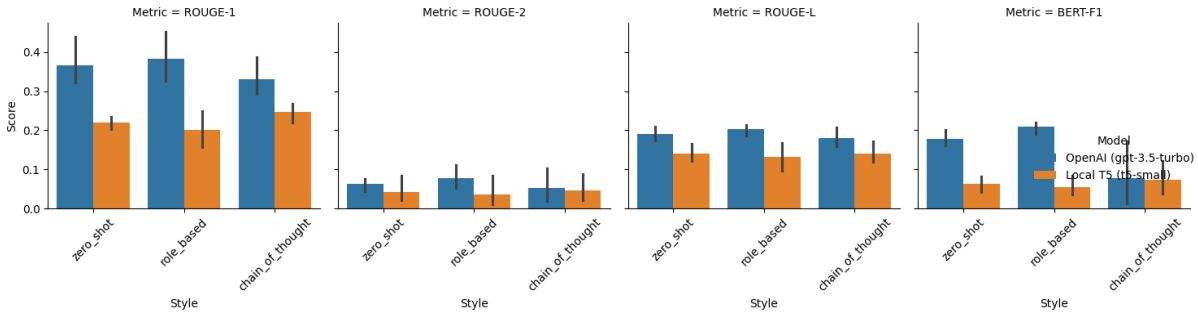

Figure 3: Per-metric breakdown by model and prompt style.