Prompt Evaluation for Summarization Review

To assess the varying techniques for summarizing product reviews with OpenAI's GPT-3.5-turbo, and determine the best approach to producing concise, respectful summaries.

Overall Process

1. Real customer reviews were collected for Sony WH-CH520 headphones from BestBuy using the Unwrangle API.
2. Five prompt variants using Jinja2 templates were implemented:
    - Zero-shot
    - Few-shot
    - Chain-of-Thought
    - Tree-of-Thoughts
    - ReAct
3. Each of the reviews was dynamically inserted within these prompts in Python.
4. The OpenAI GPT-3.5-turbo model was used to generate summaries.
5. Outputs were scored on word count, TextBlob sentiment polarity, and rubric ratings by hand.
6. Visual analysis and average scores helped identify the best-performing prompting strategy.

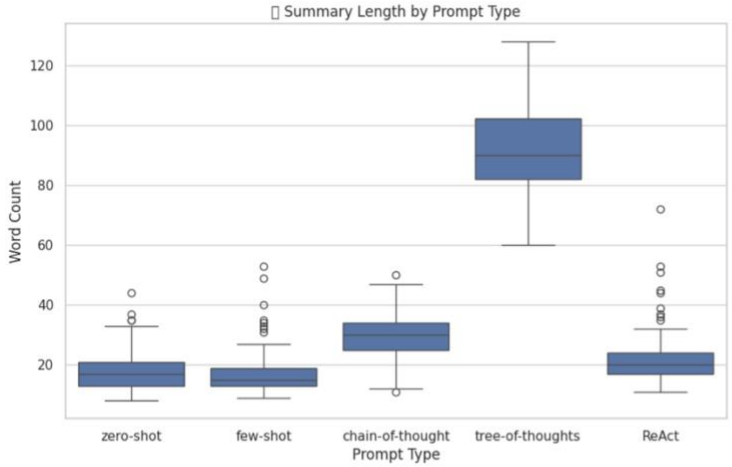
Prompting Strategies for Review Summarization

Prompt Type	Description	Goal
Zero-shot	Directly asks the model to summarize without any examples	Test LLM's raw understanding
Few-shot	Provides 1-2 examples of review → summary before asking it to summarize	Guide the model with examples
Chain-of-Thought	Breaks down reasoning in steps (sentiment → main point → summary)	Encourage stepwise/structured thinking
Tree-of-Thoughts	Generates multiple summary options and selects the best with reasoning/justification	Explore diverse reasoning paths
ReAct	Combines thinking + acting in one flow: Think sentiment → Act as summarizer	Emulate real-time analysis

---

Key Findings

- Tree-of-Thoughts created extensive, in-depth outputs but occasionally stretched beyond optimal length.
- Few-shot and ReAct proved to be steady in creating polite, coherent, and accurate summarizations.
- Small wording changes had considerable impacts on tone, clarity, and usability.
- Prompt structure impacts the model's behavior and the length of its output directly.



Longer summaries often reflect deeper reasoning steps (e.g., Tree-of-Thoughts), while shorter ones result from direct instructions.

- **Tree-of-Thoughts** prompts led to the longest, most detailed summaries
- **Few-shot** and **Zero-shot** kept the summaries short and to the point
- **Chain-of-Thought** produced medium-length outputs with step-by-step reasoning
- **ReAct** struck a balance between reasoning and brevity

Note - The way we design our prompt directly shapes how much the model says

---

Final Optimized Prompt (Refinement of Few-shot)

Examples:
Review: *Battery drains too quickly. Not practical for travel.*
Summary: *User was disappointed in the battery's limited life while on the move.*
Review: *Very comfortable with great clarity! Would purchase again.*
Summary: *User appreciated the comfort and clarity, and was willing to buy again.*
Now do an eloquent but brief summary in 1–2 sentences
"{{ review }}"
Summary:

Before & After: Refining the Prompt for Better Output

```
from jinja2 import Template

refined_few_shot = Template('''
Examples:
Review: "Battery dies too fast. Not usable for travel."
Summary: "User was unhappy with the short battery life and travel performance."

Review: "Very comfortable and amazing clarity! Would buy again."
Summary: "User praised the comfort and clarity, and expressed willingness to repurchase."

Now write a polite and concise summary (1–2 sentences):
"{{ review }}"
Summary:
''')

review = "Great Head phones. I use these when doing Yard work."
prompt = refined_few_shot.render(review=review)

response = openai.chat.completions.create(
    model="gpt-3.5-turbo",
    messages=[{"role": "user", "content": prompt}],
    temperature=0.7,
    max_tokens=60
)
```

Type	Summary Output
Original Summary	The reviewer found the headphones great for yard work.
Refined Summary	User enjoys these headphones for yard work due to their great quality.

- Original summary was correct but sounded robotic
- Refined prompt made the output more natural and specific
- Small prompt tweaks improved tone, clarity, and usefulness
- Few-shot examples helped guide the model effectively

Conclusion

The improved few-shot prompt achieved the optimal balance of clarity, tone, and completion of tasks. It was chosen as the final optimized prompt based on its uniform performance in evaluations.