

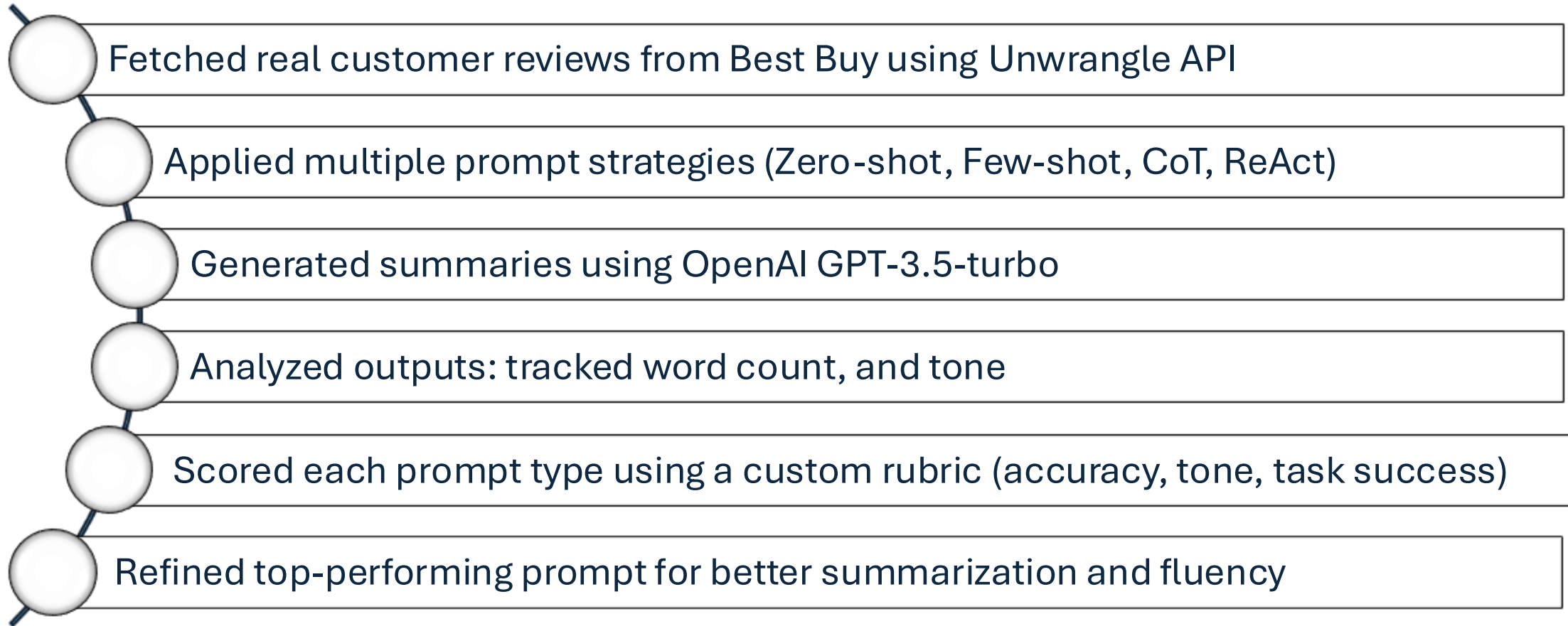
# *Evaluating Prompt Engineering Strategies for Product Review Summarization*

Student Name – Darshika Verma

Assignment A03-D: "Prompt Flow & Evaluation" – Optimizing the Dialogue

CISC 692-51- B-2025/Summer - ST: Next-Gen AI System Des Patterns

# Project Workflow: Prompt Evaluation Pipeline



# Input Data: Customer Reviews



Sony

### Sony - WH-CH520 Wireless Headphone with Microphone - Black

Model: WHCH520/B SKU: 6533161

★★★★★ 4.6 (1,606 reviews)

**\$39.99**

Save \$20

Comp. Value: \$59.99 ⓘ

4 payments starting at

**\$10.00**

with Zip

[Learn more >](#)

ⓘ **Finance Options** View your offers >


🛒 **Special Offers** On related products >

Trade in and save

🔄 **Check your trade-in value.**

Save when you trade in a similar device.

Color: **Black**

<      >

Sony WH-CH520 Wireless Headphones (BestBuy)

### Reviews

★ 4.6

1,607 reviews

5 ★ 1,202

4 ★ 264

3 ★ 76

2 ★ 25

1 ★ 40

✓ 93% would recommend to a friend

#### Top Mentions

● Sound Quality (738)

● Battery Life (316)

● Comfort (229)

● Price (63)

● Weight (60)

● Volume (9)









● Compatibility (7)

#### Customers are saying

Customers are pleased with the sound quality, battery life, price, and comfort of the WH-CH520 Wireless Headphones with Microphone. They appreciate the clear and rich sound, long battery life, affordable price, and comfortable fit. Customers also find the wireless connectivity seamless and easy to pair with their devices.

🔗 This summary was generated by AI based on customer reviews.

#### Customer Images



The vast majority of our reviews come from verified purchases. Reviews from customers may include My Best Buy members, employees, and Tech Insider Network members (as tagged). Select reviewers may receive discounted products, promotional considerations or entries into drawings for honest, helpful reviews.

★★★★★

**A great headphones for everyone!**


Incentivized

Verified Purchaser

My Best Buy® Member

Owned for 3 weeks

These budget headphones are shining in terms of light weight and sound quality considering its price. Especially in sound, it is much better than I thought. The overall tonal balance and resolution far exceeds my expectation at this price range. I have been an audiophile and experienced a lot of premium headphones and loud-speaker system. But with this one, I can even listen to classical music without any major annoying points. It can most of the Pop, RnB, soul, Ja...[See More](#)



Posted 7 months ago by Injae

Average Rating: 4.6 stars from 1,607 reviews

## Sample Reviews:

- "Great headphones. I use these when doing yard work."
- "These wireless headphones are comfortable and the sound comes through clearly! My friends admired them for their beauty — pretty long usage too."
- "My son loves these headphones and uses them every day. He can listen to what he wants without making too much noise."
- "I bought it mainly to use it in the office to take calls, and so far it’s been great! Wireless allows me to pace as I talk."
- "Perfect! Sound is excellent. Would purchase again for gifts!"

# Prompting Strategies for Review Summarization

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

# System Components for Prompt Evaluation

- ◆ Model: OpenAI GPT-3.5-turbo
- ◆ Type: Large Language Model (LLM)
- ◆ Provider: OpenAI
- ◆ API Used:  
`openai.chat.completions.create()`

◆ Purpose:

Generate concise and polite summaries of real customer reviews using advanced natural language understanding.

◆ Key Advantages:

- Multi-turn chat support
- Optimized for cost-effective experimentation
- Robust with various prompting styles (zero-shot, few-shot, CoT, ReAct, etc.)

Tool/Library	Purpose
<b>Jinja2</b>	Used to fill review text into different prompt formats automatically
<b>OpenAI Python SDK</b>	Helps connect with the GPT-3.5-turbo model and send prompts via API
<b>Pandas</b>	For data manipulation
<b>TextBlob</b>	Used to measure the tone of summaries by calculating their sentiment scores
<b>Matplotlib + Seaborn</b>	Used to create charts for comparing summary length and sentiment across prompt types
<b>Requests</b>	Makes API calls to get review data from external sources like Unwrangle
<b>Unwrangle API</b>	The service used to collect real customer reviews from BestBuy.com
(Optional) <code>os.environ</code>	To securely set API keys for OpenAI

# Comparison of Prompt Styles Using One Review

**Review Used for All Prompts:**  
*“Great Headphones. I use these when doing yard work.”*

## Zero-shot

**Prompt:**

Summarize this product review into one polite sentence:  
*"Great Head phones. I use these when doing Yard work."*

**Output:**

*These headphones are great for use during yard work.*

## Few-shot

**Prompt:**

Here are examples of product review summaries:  
Review: "The sound quality is terrible and it broke after a week."  
Summary: "The user reported poor sound quality and a broken device."  
Review: "Great fit and comfort, with amazing bass response."  
Summary: "The reviewer praised the headphones for comfort and bass."  
Now summarize this review:  
*"Great Head phones. I use these when doing Yard work."*  
Summary:

**Output:**

*"The reviewer found the headphones great for yard work."*

## Chain-of-Thought

**Prompt:**

Here is a product review:  
*"Great Head phones. I use these when doing Yard work."*  
Step 1: Identify whether the feedback is positive, negative, or neutral.  
Step 2: Find the main point of praise or complaint.  
Step 3: Write a professional summary in one line.  
Answer:

**Output:**

*Positive feedback; main point of praise is the headphones being great for yard work.*  
*Summary: The headphones are great for yard work, according to the review.*

## Tree-of-Thoughts

**Prompt:**

You're analyzing a product review. First, write two different summaries based on different viewpoints. Then choose the best one and explain why.  
Review: *"Great Head phones. I use these when doing Yard work."*  
Option 1:  
Option 2:  
Chosen Summary:  
Reason:  
**Output:**  
*Option 1: The headphones are great for outdoor use during yard work.*  
*Option 2: The headphones are perfect for listening to music while doing chores outside.*  
*Chosen Summary: The headphones are great for outdoor use during yard work.*  
*Reason: This summary effectively captures the main point of the review, which is that the headphones are specifically used for yard work. It is more specific and directly relates to the context in which the headphones are being used.*

## ReAct

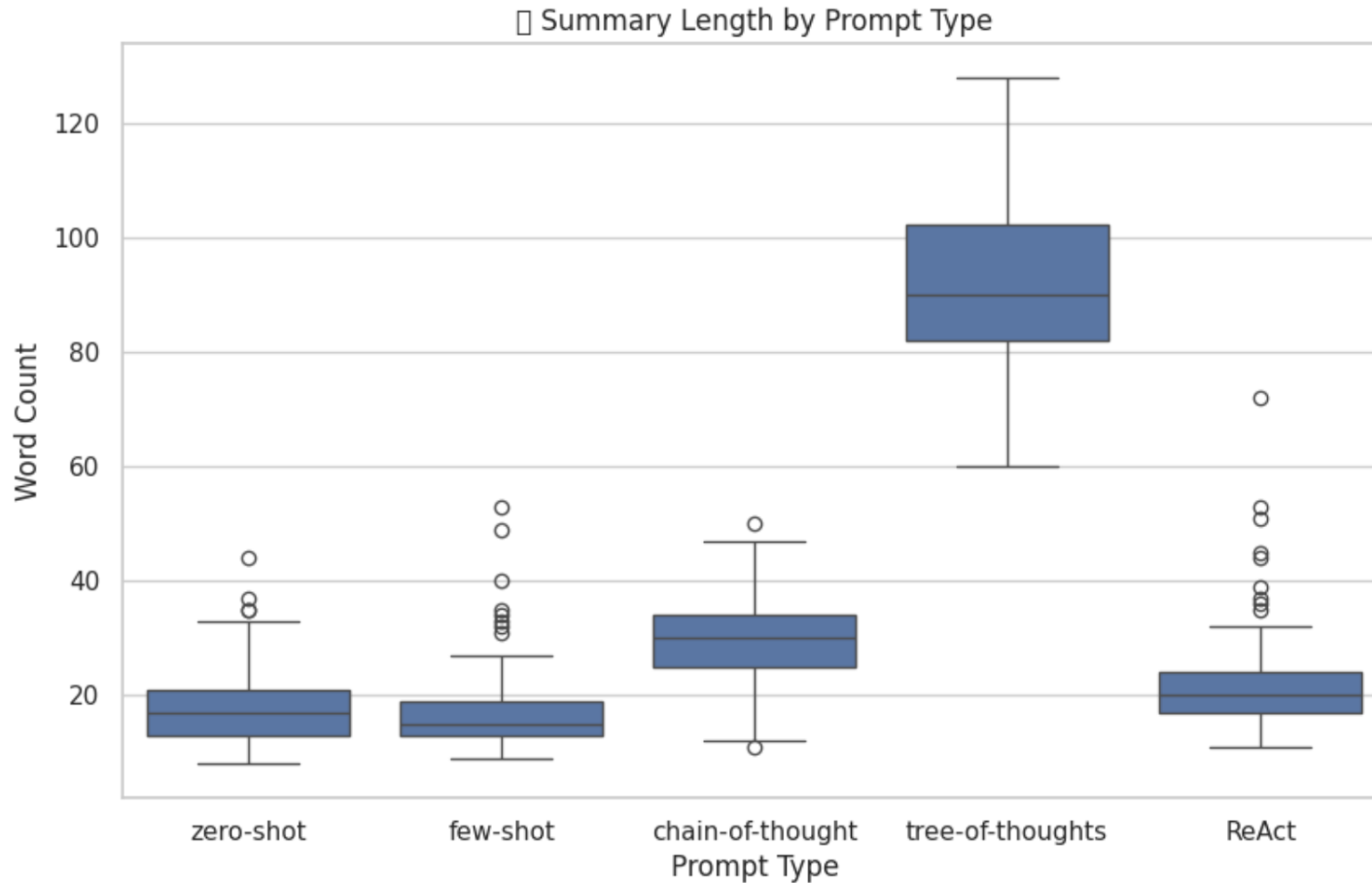
**Prompt:**

You are a helpful AI reviewer. Your task is to:  
– Think through the user feedback  
– Generate a short, polite summary  
Review: *"Great Head phones. I use these when doing Yard work."*  
Think: What is the core sentiment and message?  
Act: Write a summary in 1–2 lines.

**Output:**

*Summary: The user finds the headphones great for use during yard work.*

# Prompt Strategies vs. Summary Word Count



- **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*

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

# Prompt Output Evaluation: Accuracy, Tone & Task Fit

## accuracy\_score (0–2)

### What it checks:

How well the summary matches the original review content.

Score	Meaning
0	Summary is wrong or misleading
1	Partially correct (somewhat vague)
2	Fully accurate and on point

## tone\_score (0–2)

### What it checks:

Whether the summary sounds polite, professional, and well-structured.

Score	Meaning
0	Unclear or robotic tone
1	Somewhat readable but not polished
2	Polite, fluent, and human-like

## task\_success (0–2)

### What it checks:

Whether the model completed the task as asked (e.g., summary in 1–2 sentences).

Score	Meaning
0	Didn't follow instructions (e.g., too long, irrelevant)
1	Task partially met (too verbose or too short)
2	Task fully completed (clear, correct summary)

	accuracy_score	tone_score	task_success	total_score
prompt_type				
ReAct	2.0	2.0	2.0	6.0
few-shot	2.0	2.0	2.0	6.0
zero-shot	2.0	2.0	2.0	6.0
chain-of-thought	2.0	1.8	1.6	5.4
tree-of-thoughts	2.0	2.0	0.6	4.6

- **Accuracy** – Did the summary reflect the review correctly?

- **Tone** – Did it sound polite and professional?

- **Task Success** – Was it concise and within 1–2 sentences?

Each score ranges from 0 (poor) to 2 (excellent)

**Top performers:** ReAct, Few-shot, Zero-shot — all scored a perfect 6

**Tree-of-Thoughts** was accurate, but often too long

**Chain-of-Thought** had great logic but slightly missed the tone and task format



# 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

*Appreciate Your  
Time!*