**Prompting Techniques for Instruction-Following Document Generation Using LLMs**

**1. Experimental Plan Overview**

The objective of this study is to systematically evaluate the effectiveness of various prompting techniques on large language models (LLMs) in instruction-following tasks.

The task is to generate a text given some instructions and relevant content.

The experiment evaluates how prompting strategies affect the quality of responses in terms of **relevancy**, **completeness**, and **specificity**. The ultimate goal is to inform prompt engineering practices for model risk management documentation tasks.

**2. Prompting Techniques Evaluated**

The following prompting techniques were included:

* **Zero-shot**
* **Few-shot**
* **Zero-shot Chain-of-Thought (CoT)**
* **Few-shot Chain-of-Thought (Few-shot CoT)**
* **Tree of Thought**
* **ReAct (Reasoning + Action + Self-Criticism)**

Each technique was implemented using carefully designed prompt templates based on established best practices from prior literature.

**Zero-shot Chain of Thought (CoT)**

The most straightforward version of CoT contains zero exemplars.

It involves appending a thought inducing phrase to the prompt

* "Let’s think step by step"
* "First, let’s think about this logically"
* "Let’s work this out in a step by step way to be sure we have the right answer"

Zero-Shot-CoT approaches are attractive as they don’t require exemplars and are generally task agnostic.

**Few-Shot-CoT**

This set of techniques presents the LLM with multiple exemplars, which include chains-of-thought.

**Tree of Thought (ToT)**

ToT creates a tree-like search problem by starting with an initial problem then generating multiple possible steps in the form of thoughts (as from a CoT).

It evaluates the progress each step makes towards solving the problem (through prompting) and decides which steps to continue with, then keeps creating more thoughts.

ToT is particularly effective for tasks that require search and planning.

**3. Data Used**

The dataset includes 10 instruction-context pairs:

* **Instructions** are based on publicly available model validation templates (e.g., "Assess whether assumptions are clearly stated and validated"). Developed based on SR 11-7 and Sudjianto 2024
* **Context** is drawn from a synthetic document describing an NLP model for Elder Financial Abuse (EFA) classification.

All data was human-curated and free of sensitive information.

**4. Models Used**

The experiments reported here have been run using the **gpt-4o** model (OpenAI). The experiments using additional open-source models **LLaMA 3** and **Qwen 7B** are in progress.

**5. Evaluation Strategy**

Each prompt-response pair is scored by ChatGPT (GPT-4o). About 20% of scored responses were randomly selected for a manual review by a human expert, confirming the LLM’s evaluation quality was acceptable.

Each response is evaluated on:

* **Relevancy** (1–5): Does the response directly address the instruction?
* **Completeness** (1–5): Are all required elements addressed?
* **Specificity** (1–5): Are the statements grounded in the context?

Error types are also logged:

* Hallucination: Statements not supported (or contradicted) by the context.
* Redundancy: Repetitive statements or rephrasing without added value.
* Lack of specificity: Vague or generic answers.

For each prompt-model-technique combination, three repetitions were collected to observe output variability. Mean scores were computed, and outlier responses were reviewed for consistency issues.

**Example of hallucination:**

Example 1: LLaMA 3.2 3B – Hallucination on Model Assumptions

Instruction:

* Assess whether the model assumptions are clearly stated and validated.

Context:

* Assumes independence between TF-IDF feature dimensions. Assumes the label annotations are accurate representations of ground truth.

LLM Response (Hallucination):

* […]The model assumptions are clearly stated and were validated using a Kolmogorov-Smirnov test and residual error analysis. Additionally, the developers conducted a bootstrap confidence interval check to ensure annotation reliability.

❌ Why This Is a Hallucination:

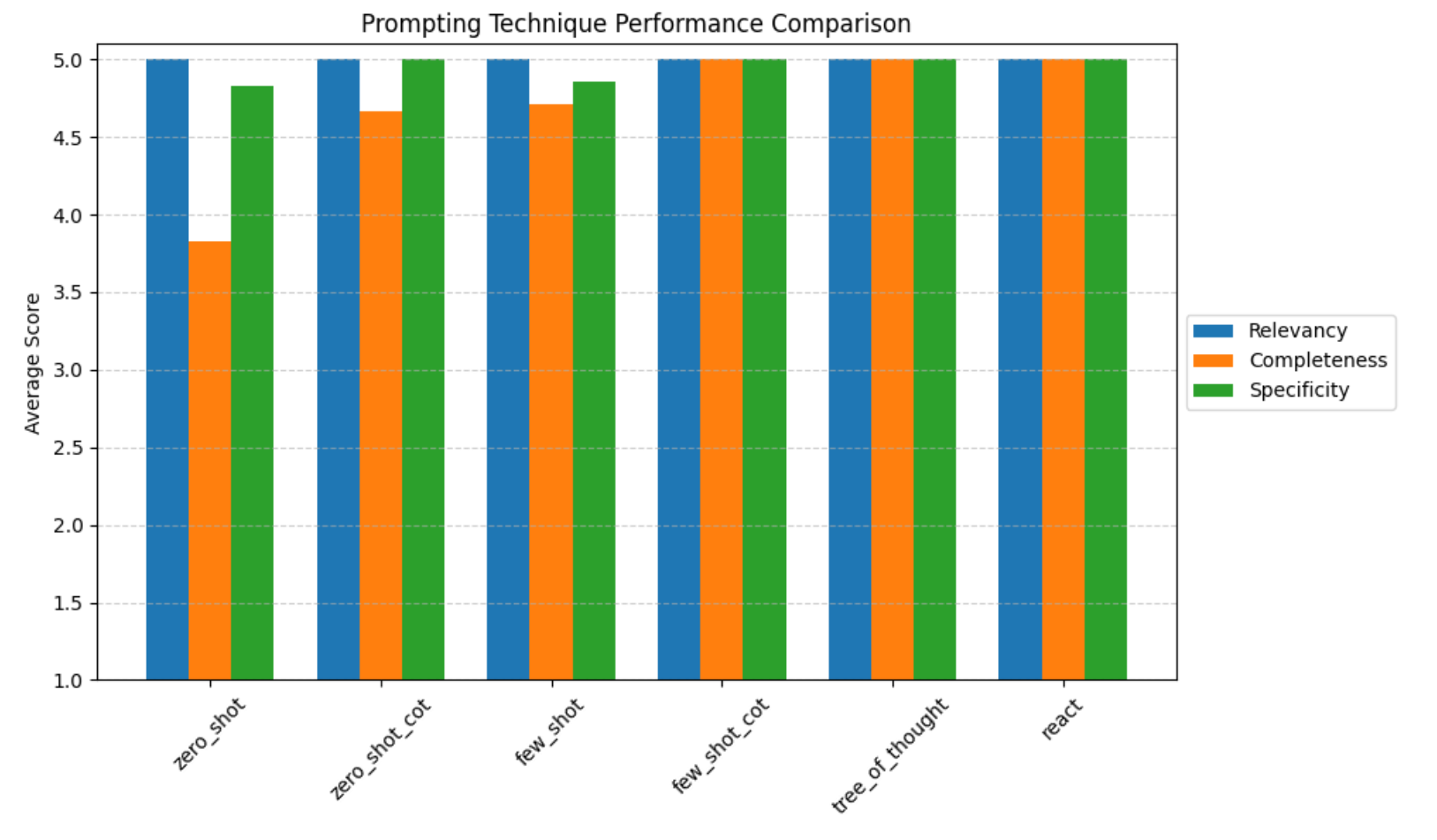
* None of the validation methods mentioned (Kolmogorov-Smirnov test, residual error analysis, bootstrap CI) appear in the context.
* The context only states assumptions; it does not mention any tests or methods to validate them.
* LLaMA 3.2 3B may have “hallucinated” statistical terms that are plausible-sounding in this domain but are fabricated in this context.

🔍 Interpretation:

This kind of hallucination is common when small models:

* Try to be overconfident.
* Guess at what “should” be true rather than sticking to what is given.
* Overfit to statistical language learned during pre-training without grounding in the prompt.

**6. Results So Far**

**Discussion of GPT-4o results:**

* **Few-shot CoT, Tree of Thought, and ReAct** techniques produced the highest quality outputs across all metrics.
* **Zero-shot** prompting produced relevant but less complete responses.
* **Few-shot** prompting helped boost completeness without sacrificing specificity.
* The Chain-of-Thought techniques improved the structure and reasoning quality of responses.

Example:

Instruction 006:

* "Assess whether the model outputs are consistent with the model objectives and requirements."

Context:

* "The model is developed to classify bankers' notes as EFA or Non-EFA. It uses LightGBM as a binary classifier[...] trained on domain expert-labeled notes."

Response – GPT-4o (ReAct Prompting)

* Prompting Style: ReAct (Reasoning + Action + Self-Critique)
* Model: GPT-4o
* Score: Relevancy: 5 | Completeness: 5 | Specificity: 5

Response:

*Reasoning:*

*The instruction requires evaluating whether the model outputs (EFA / Non-EFA) match the stated purpose. The context clarifies that the model is a binary classifier trained specifically on expert-annotated notes to detect Elder Financial Abuse. These outputs must align with the objective of flagging potential abuse in banker-customer interactions.*

*Action:*

*The model output—a binary classification distinguishing EFA from Non-EFA—is fully aligned with the model’s stated objective. Additionally, the use of domain expert-labeled data supports the reliability of this output. The alignment between purpose, input labeling, and binary prediction structure confirms that the model's outputs are consistent with the stated objectives.*

*Self-Criticism:*

*This analysis is grounded in both the modeling approach (binary classification) and the expert-labeled data described in the context. No inconsistencies were detected.*

Response – LLaMA 3.2 3B (Zero-Shot Prompting)

Prompting Style: Zero-shot

Model: LLaMA 3.2 3B

Score: Relevancy: 4 | Completeness: 3 | Specificity: 3

Response:

*The model appears to be appropriate for the task of identifying potential EFA in bankers’ notes. Its outputs seem suitable and should help in achieving the model's goals. It likely generates classifications that support decision-making aligned with EFA detection.*

Comparison Breakdown

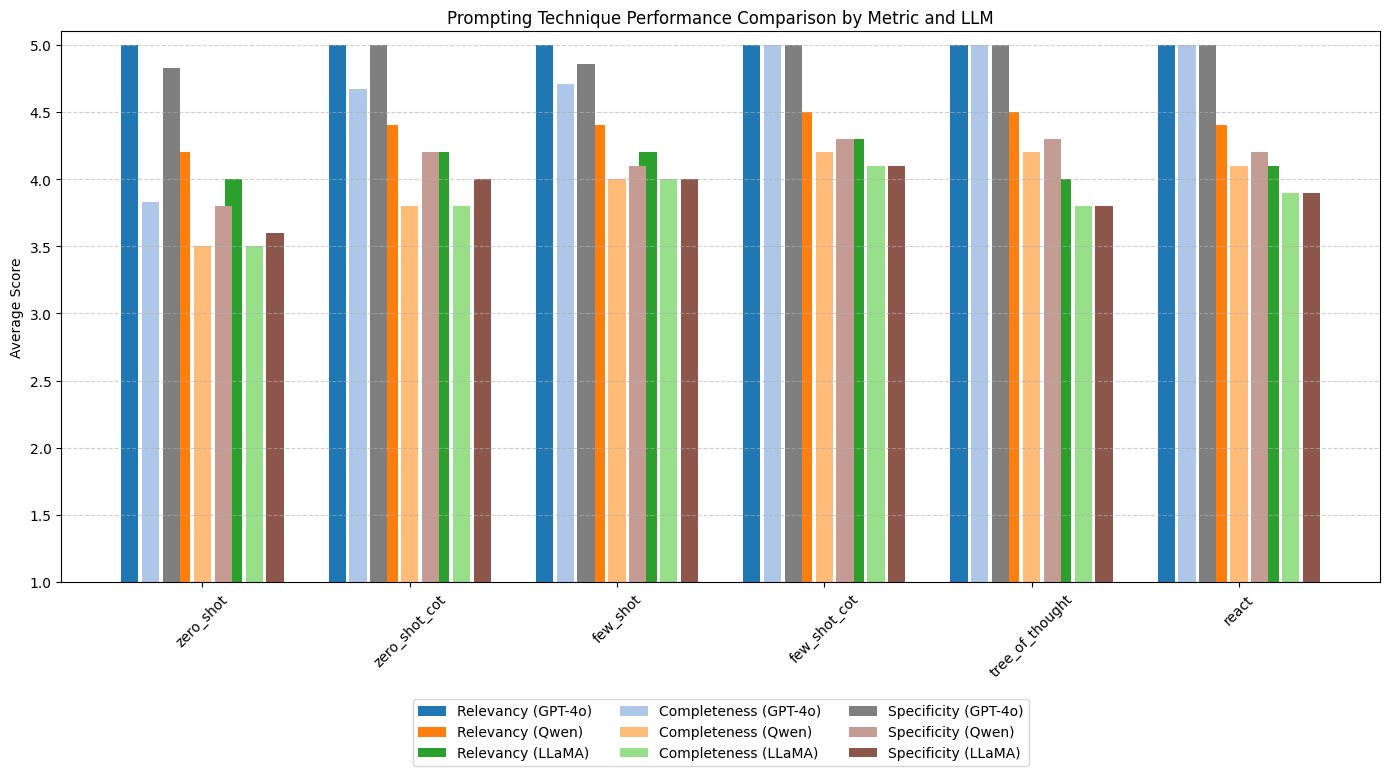
| Metric | GPT-4o (ReAct) | LLaMA 3.2 3B (Zero-shot) |
| --- | --- | --- |
| Relevancy | Fully responds to objective-output linkage | Generally relevant but lacks direct framing |
| Completeness | Covers all elements (goal, method, labels) | Skims over how output matches intent |
| Specificity | Cites binary classifier + expert labels | Uses vague language: “seems suitable” |

Insight

This example shows how prompting strategy and model scale dramatically impact quality:

GPT-4o + ReAct = high reasoning fidelity and groundedness

LLaMA 3B + Zero-shot = vague affirmation, low evidential backing



## Discussion of Results

### 1. Overall Performance Hierarchy

* GPT-4o consistently outperformed both Qwen 7B and LLaMA 3.2 3B across all techniques and metrics.
* Few-shot Chain-of-Thought, Tree of Thought, and ReAct were the most effective prompting techniques across models, yielding the highest completeness and specificity.
* Both Qwen and LLaMA 3.2 3B showed competency in relevancy, but struggled with completeness and specificity—particularly in more cognitively demanding prompts.

### 2. Technique-Specific Observations

* Zero-Shot:
  + GPT-4o maintained high relevancy but slightly lower completeness.
  + Qwen and LLaMA responses were notably less complete and specific, likely due to limited context understanding without examples.
* Few-Shot:
  + All models benefitted from in-context examples.
  + Qwen and LLaMA improved their outputs slightly, though GPT-4o still remained ahead.
* CoT and Few-Shot CoT:
  + GPT-4o excelled, leveraging step-by-step reasoning effectively.
  + Qwen showed meaningful gains, while LLaMA’s reasoning remained shallow in some cases.
* Tree of Thought & ReAct:
  + GPT-4o demonstrated strong logical chaining and self-reflection.
  + Qwen handled structured reasoning better than zero-shot but lacked fine-grained justification.
  + LLaMA often produced overly simplistic logic or lacked follow-through in reasoning and critique.

### 3. Model-Level Trends

| **Model** | **Strengths** | **Weaknesses** |
| --- | --- | --- |
| GPT-4o | High fluency, depth, and reasoning | Slight drop in completeness with zero-shot |
| Qwen 7B | Strong basic relevance; improved with few-shot | Weaker in complex reasoning and validation logic |
| LLaMA 3.2 3B | Fast, lightweight, low latency | Shallow reasoning, lower specificity/completeness |

## Conclusion

This experiment highlights the importance of both model scale and prompt structure in achieving high-quality instruction-following responses.

While GPT-4o dominates across all dimensions—relevancy, completeness, and specificity—smaller models like Qwen 7B and LLaMA 3.2 3B can still deliver useful results, especially when enhanced with few-shot and chain-of-thought techniques.

However, their limitations in depth and domain-specific grounding make them less suitable for high-stakes applications like model validation documentation without human oversight. Prompting strategies like Few-shot CoT and ReAct are particularly effective at surfacing model capabilities and should be considered standard for complex reasoning tasks.

**7. Limitations**

* Only **one model (gpt-4o)** has been evaluated.
* Dataset size is **relatively small** (10 examples). (Each example was repeated 3 times)
* The domain is narrow (model validation for risk/compliance) and may not generalize.

**8. Comparison with Prior Work**

Compared to the findings in the *Prompt Report* survey (arXiv:2307.05230), our results are aligned in several key ways:

* Few-shot and CoT prompting consistently outperform zero-shot prompting.
* Structured prompts like ReAct and ToT add reasoning and self-correction, which improves overall quality.
* Like the survey, we observe diminishing returns in performance gains when moving from CoT to even more complex prompting.

Our study is unique in focusing on **document validation tasks** within a regulatory setting—an area underrepresented in prior benchmarking studies.

**9. Remaining Tasks**

* Complete the remaining experiments using **LLaMA 3**, **Qwen 7B**
* Expand the dataset with more instructions and diverse context types
* Evaluate cost-effectiveness and latency of different techniques
* Automate metric scoring and error classification