**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. These tasks are grounded in a model validation use case, where the LLM is asked to produce professional-quality assessments based on instructions and excerpts from model development documents (MDDs).

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.

**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 Model Development Document (MDD) 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 Metrics**

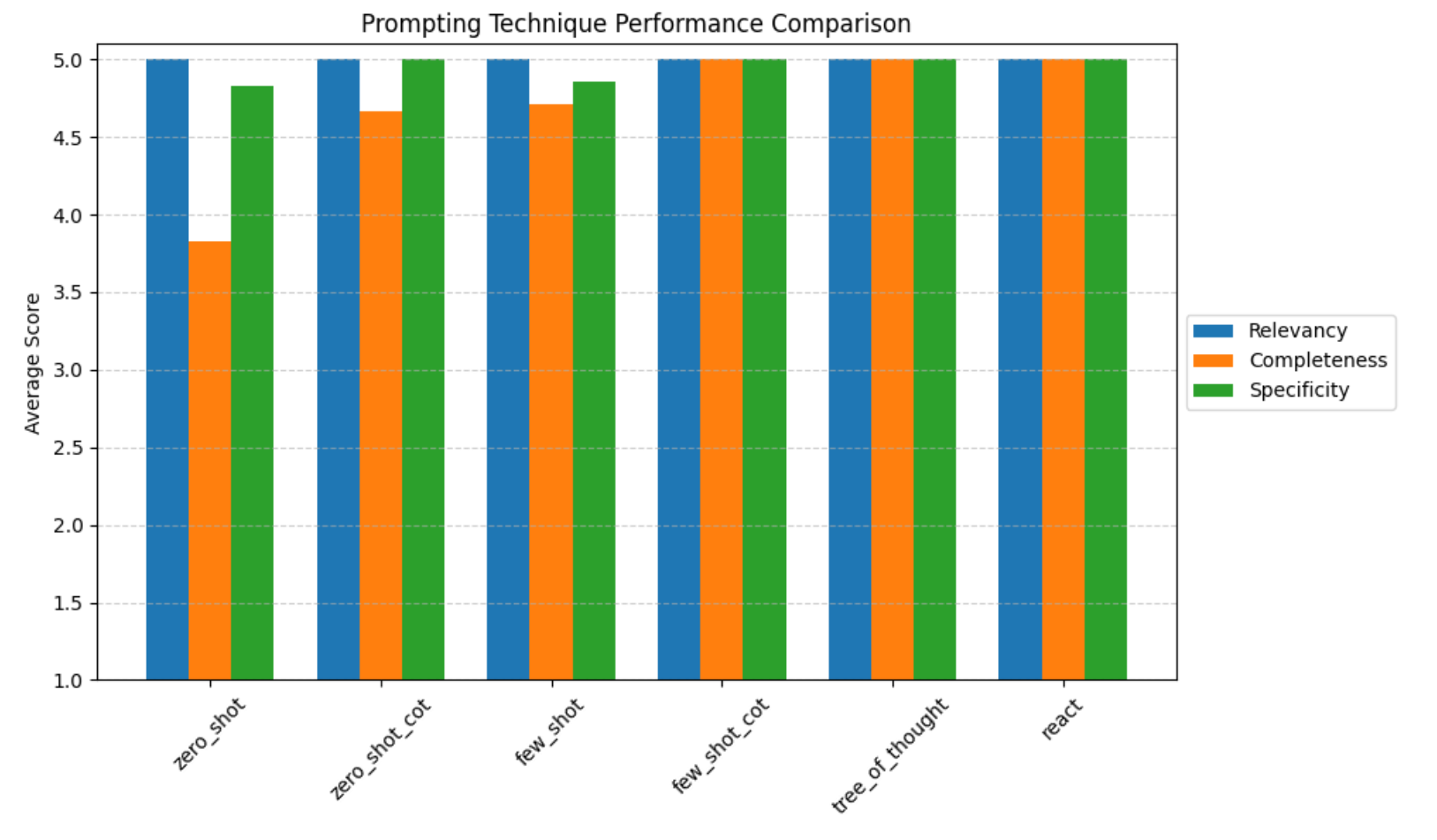
Each response is evaluated on:

* **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
* Redundancy
* Lack of specificity

**6. Results So Far**

****

| **Technique** | **Relevancy** | **Completeness** | **Specificity** |
| --- | --- | --- | --- |
| Zero-shot | 5.00 | 3.83 | 4.83 |
| Zero-shot CoT | 5.00 | 4.67 | 5.00 |
| Few-shot | 5.00 | 4.71 | 4.86 |
| Few-shot CoT | 5.00 | 5.00 | 5.00 |
| Tree of Thought | 5.00 | 5.00 | 5.00 |
| ReAct | 5.00 | 5.00 | 5.00 |

**Discussion:**

* **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.

**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**

* Run the same experiments using open-source models (e.g., **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.