# Are Large Language Models Ready for Healthcare? A Comparative Study on Clinical Language Understanding

Yuqing Wang[1], Yun Zhao[2], Linda Petzold[1]
[1]University of California, Santa Barbara, [2]Meta Platforms, Inc.

## Motivation and Goals

- **Large Language Models (LLMs)** have substantial untapped potential for healthcare revolution - a topic yet to be comprehensively evaluated and fully appreciated.
- There is a need to explore the **efficacy of diverse prompting techniques**, such as the proposed self-questioning prompting, in clinical tasks and healthcare settings.
- Assessing **GPT-3.5, GPT-4, and Bard** in diverse clinical language tasks emphasizes the evolving role of LLMs in healthcare.

## Tasks

Overview of six biomedical and clinical language understanding tasks, encompassing eight datasets for experimental evaluation.

| | |
|---|---|
| **Named Entity Recognition** | **Document Classification** |
| **Relation Extraction** | **Question Answering** |
| **Semantic Textual Similarity** | **Natural Language Inference** |

## Self-questioning Prompting

**Construction process of SQP:**

1. Extract key details from text;
2. Create targeted questions for understanding;
3. Enrich task context via Q&A;
4. Customize strategy for task-specific outputs.

**Prompting Methods Comparison:**

Table: Comparison between standard, chain-of-thought, and self-questioning prompting.

| Prompting Strategy | Guideline | Purpose |
|---|---|---|
| Standard | Use a direct, concise prompt for the desired task. | To obtain a direct response from the model. |
| Chain-of-Thought | Create interconnected prompts guiding the model through logical reasoning. | To engage the model's reasoning by breaking down complex tasks. |
| Self-Questioning | Generate targeted questions and use answers to guide the task response. | To deepen the model's understanding and enhance performance. |

## Performance Comparison

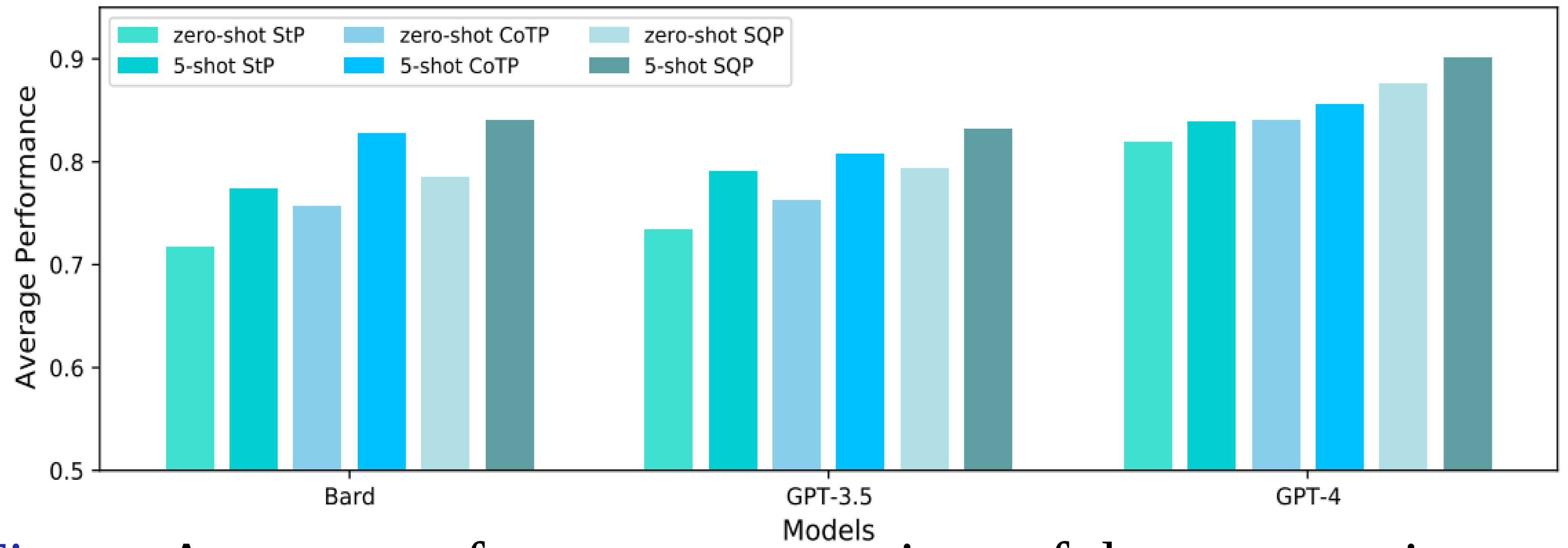


Figure: Average performance comparison of three prompting methods in zero and 5-shot learning settings across three models.

## Error Analysis

Table: Average error type distribution for DDI (relation extraction) across Bard, GPT-3.5, and GPT-4. Error types are identified manually.

| Error Type | Description | Error Proportion (%) Bard | GPT-3.5 | GPT-4 |
|---|---|---|---|---|
| Wording Ambiguity | unclear wording | **32** | 23 | 24 |
| Lack of Context | incomplete context usage | 25 | **31** | 19 |
| Complex Interactions | multiple drug interactions | 19 | 12 | 14 |
| Negation and Qualification | Misinterpreting negation/qualification | 8 | 27 | **25** |
| Co-reference Resolution | Misidentifying co-references | 16 | 7 | 18 |

## Conclusion

- LLMs exhibit potential in various clinical language tasks.
- Task-specific prompts, like SQP, enhance LLMs' understanding and response generation.
- LLMs support, not replace, human expertise in existing workflows.

## Further Questions?

Please don't hesitate to reach out via **email**: wang603@ucsb.edu
**Code** is available at: https://github.com/EternityYW/LLM_healthcare