NAISTym at the NTCIR-18 MedNLP-CHAT: Classifying Patient-Chatbot Conversations with Objective and Subjective Assessments Using Prompting Techniques

Lenard Paulo V. Tamayo, Sa'idah Zahrotul Jannah, Axalia Levenchaud, Mohamad Alnajjar,
Shaowen Peng Ph.D, Shoko Wakamiya Ph.D, Eiji Aramaki Ph.D

Social Computing Laboratory, Graduate School of Science and Technology,
Nara Institute of Science and Technology (NAIST), Japan

ABSTRACT

Chatbots are widely used in the healthcare sector, making their accuracy and reliability essential. Beyond providing factually correct information, chatbots must also consider the human aspect of their responses. Large language models (LLMs) can be utilized to evaluate chatbot responses, employing prompting strategies such as chain-of-thought and few-shot prompting to enhance reasoning and optimize output quality. This study evaluates a chatbot's answers to medical questions using both objective and subjective assessments. Different prompting techniques were applied: objective evaluation used baseline, chain-of-thought (COT), and chain-of-thought with few-shot (COTF) prompting, while subjective evaluation used baseline and baseline with few-shot (Baseline-f) prompting. The results revealed that COTF prompting with both models improved the performance of objective evaluation, while few-shot prompting enhanced subjective evaluation.

METHODS

INPUT

Patient (Q): I had severe pain in my shoulder during **radio exercise**, and I can no longer raise my arm. Should I go to the **hospital**?

Chatbot (A): Severe **shoulder pain** may be caused by **inflammation** of **muscles** and **joints**. Inability to raise your arm is an **unfavorable condition**. It is recommended that you see an **orthopedic surgeon** as soon as possible.

PROMPT

Chain-of-Though (COT)
Method to incorporate a reasoning process in the prompt design to enhance logical progression in responses

+

Few-shots method
Method that implement numerous examples (shots) in the prompt. Here, this method was experimented with varying numbers of examples (1, 3, 5, and 10 shots)

LLM

GPT-4o and **Gemini 1.5 Flash**

EVALUATION

Objective Evaluation
- Medical risk
- Ethical risk
- Legal risk

Evaluation metrics
- Accuracy
- F1 Score

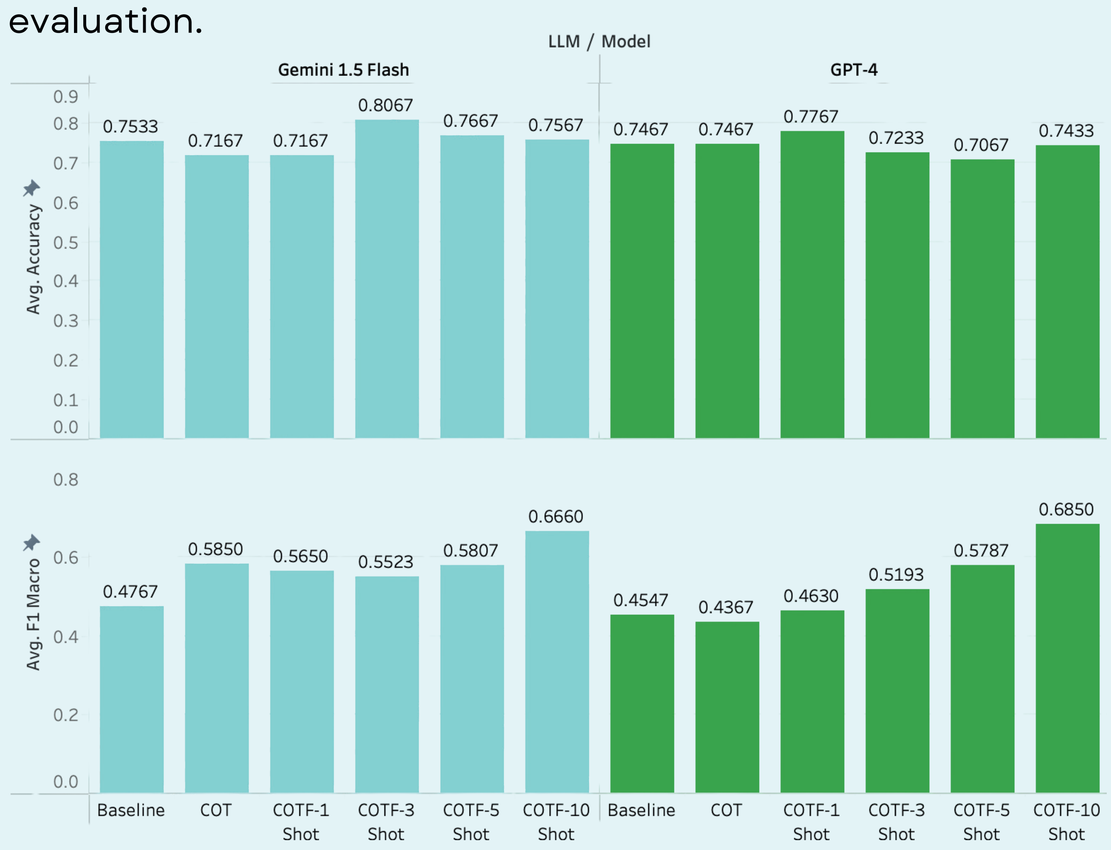
Subjective Evaluation
- Fluency
- Harmlessness
- Helpfulness

Evaluation metric
- EMD score

RESULTS

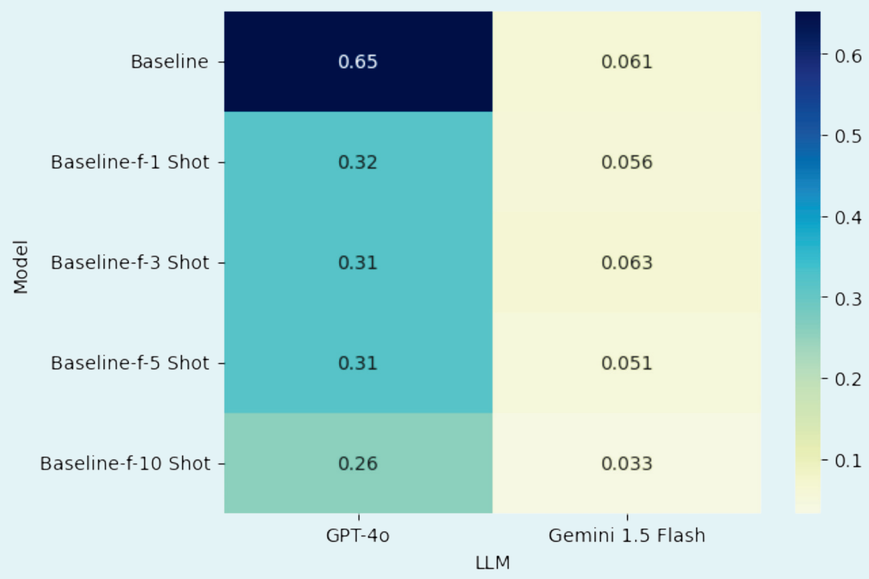
Objective Evaluation

COTF methods stands out for both LLMs, with an increasing average F1 Macro score the more we add shots to the prompt. Thus **COTF-10 shots** shows the better result for this objective evaluation.



Subjective Evaluation

Gemini 1.5 Flash models manage to accomplish a significant drop in the score value. Similarly to the objective scores, the **Few shots method** shows better results and decrease the EMD value as we add more shots in our prompt.



We selected our best models based on a balance between our objectives and subjectives results. After examining all of our metrics, we selected:

- the **Gemini 1.5 Flash** and **GPT-4 COTF-10 shots models**
- the **Gemini 1.5 Flash baseline** based to its performance in terms of EMD score and the accuracy.

Combining chain-of-thought (COT) and few-shot prompting improves results over baseline, but further gains may come from exploring other strategies (e.g., RAG, RL) and NLP techniques to better leverage LLMs in healthcare while minimizing patient risk and supporting clinicians in focusing on their expertise.

SOCIOCOM Social Computing Laboratory since 2015

Nara Institute of Science and Technology NAIST®