Fall 23 CS 638 Final Project

Team Members: Zhou, Zejun; Shi, Kaiwen; Tan, Junyuan
University of Wisconsin - Madison, Fall 2023

Introduction

This project aims to evaluate and compare the quality of responses generated by Large Language Models (LLMs) in mental health counseling conversations. Specifically, the project evaluated the responses generated by original and tuned models on the Cohere platform. We also set up evaluation rubrics, and use the Cohere, GPT-3.5 Turbo, Bard model to conduct quality evaluations. The project analyzes and understands the performance differences before and after model optimization by comparing different prompting technologies and evaluation indicators.

Datasets and Models

We used the open-source *mental_health_counseling_conversations* dataset from Hugging Face[1]. The dataset contains labeled conversations from mental health counseling.

There are, in total, two models used in this project.

1. **Un-tuned Cohere LLM**: the general version of Cohere LLM is used for evaluation firstly. We randomly select questions from the *mental_health_counseling_questions,* enter into the Cohere LLM, and record the response from the model.
2. **Tuned Cohere LLM.** Cohere provides fine-tuning features for customizable model training. We have used the *mental_health_counseling_questions* dataset to train a tuned model, then enter the same query for Un-tuned Cohere LLM into Tuned Cohere LLM, and record its response tentatively.

Model Assessment

We have evaluated the quality of generated response from the given dimensions below:

1. **Language Fluency**. 0 means the response makes no sense to humans, 10 means response makes sense, easily readable and understandable.
2. **Related Content**. 0 means totally not related to what we users expect to hear, 10 means the response is perfectly related to the topic we are discussing in the query.
3. **Response accuracy**. We will follow both the dataset - response and common knowledge, 0 means response is totally wrong and 10 means response is 99% match common knowledge and response in dataset.
4. **Overall score evaluation.** An overall evaluation of the patient's problem. 0 means terrible, 10 means perfect.

---

[1] Mental Health Counseling questions, *Hugging Face,* https://huggingface.co/datasets/Amod/mental_health_counseling_conversations

Assessment Prompt Techniques with GPT-3.5 Turbo

Three different prompting techniques—original human-written prompts, strict AI generated prompts, and step-by-step professional prompts—were employed to guide the GPT-3.5 Turbo model in scoring.

```
## original human written prompt
origin_prompt = """Based on the criteria of {metrics}, please score the following
response from 0 to 10, where 0 means bad answer and 10 means perfect answer based on
response from given question.\n\nQuestion: {question}\nResponse:
{response}\n\nscore:"""
```

```
## strict AI generated prompt
strict_prompt = """"Please provide a numerical score from 0 to 10 for the following
response, based on the specified criteria of {metrics}. Note that not every response
can achieve a high score. A score of 0 represents a completely inadequate answer, and
a score of 10 represents a perfect answer that fully meets the criteria. Score the
response realistically, understanding that high scores are reserved for only the most
exceptional and fitting responses.

      Question: {question}
      Response: {response}

      Score (numerical only):"""
```

```
## Professional step by step prompt
step_prompt = """"Instruction: Based on the criteria of {metrics} (e.g., accuracy,
relevance, completeness, coherence), please provide a numerical score from 0 to 10 for
the following response. Here is a guideline for scoring:
- 0: Completely irrelevant or incorrect.
- 1-3: Partially relevant but with significant errors or omissions.
- 4-6: Moderately relevant, minor errors or incompleteness.
- 7-9: Highly relevant and accurate with minor imperfections.
- 10: Perfectly relevant, accurate, and complete in every aspect.

Please note that high scores should be reserved for responses that are exceptionally
fitting and outstanding in all criteria.

Example 1 (Score: 4 in overall rating):
"I'm sorry to hear about your struggles with binge eating, gastric sleeve surgery, and
the traumatic experiences you've had with abuse, addiction, and abusive partners. It
takes a lot of strength and courage to acknowledge these difficulties and seek help,
and you've taken an important step in your healing journey by engaging in therapy. It
```

```
is concerning to hear that you aren't getting the feedback or support you need from
your current therapist. Having a safe, trusting, and empathetic therapeutic
relationship is crucial to the process of overcoming the challenges you've outlined.
Therefore, it may be helpful to consider discussing your concerns with your therapist
directly. You could explain that you feel your sessions would be more productive if
you received more active feedback and guidance, and ask them if they have any
suggestions or recommendations for how to make the process more effective for you. If
you feel that your therapist is not the right fit for you after having this
conversation, consider finding a new therapist who specializes in the issues you are
facing, such as trauma, addiction, and eating disorders. You may benefit from working
with a trauma-informed therapist who can provide you with empathetic feedback and help
you process the trauma you experienced as a child. Additionally, finding a support
group for individuals who have had gastric sleeve surgery may also help you navigate
the challenges you face with your eating disorder. Remember, finding the right help
for you is important, and if one path isn't working, there are always other options to
consider. Would you like me to go into more detail about any of these suggestions?"
- This response shows empathy and addresses many of the patient's concerns, but it is
long and verbose, which may detract from its effectiveness.

Example 2 (Score: 10 in overall rating):
"I'm sorry to hear that you have been going through a difficult time. It's important
to remember that you are not responsible for the actions of others, and that you have
the right to make your own decisions about your life. I would recommend that you
continue to seek help from your therapist and that you work with them to develop a
plan to address your eating disorder. It's also important to remember that you are not
responsible for the actions of your parents or other family members, and that you have
the right to make your own decisions about your life. I would recommend that you seek
the help of a medical professional to address your eating disorder, as well as a
therapist to address your mental health issues. I would also recommend that you seek
the help of a support group or community group to help you feel less alone and more
supported."
- This response is empathetic and to the point, offering practical advice in a clear
and concise manner.

Context: You will be given a question and a response from health counseling
conversations. Consider the given metrics of the response in the context of the
question.

Question: {question}
Response: {response}

Score (numerical only):
```

Related Figures

We have collected some evaluation data results. These results can be seen from the figures below.

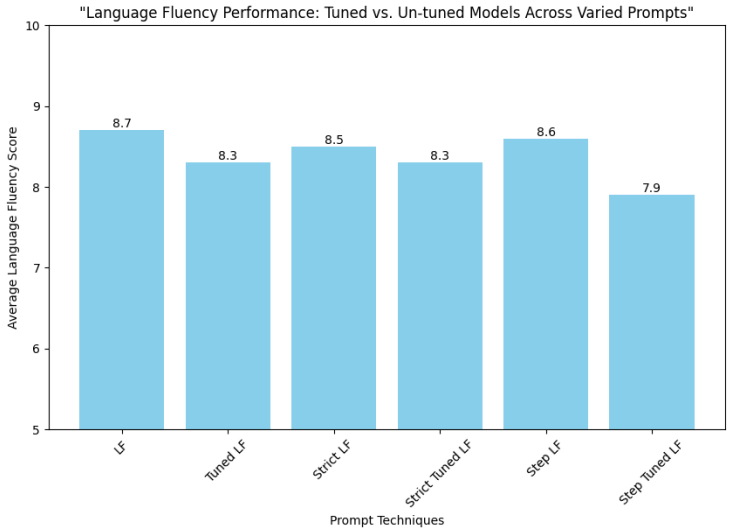

Figure 1: Language Fluency Performance: Tuned vs. Un-tuned Models Across Varied Prompts

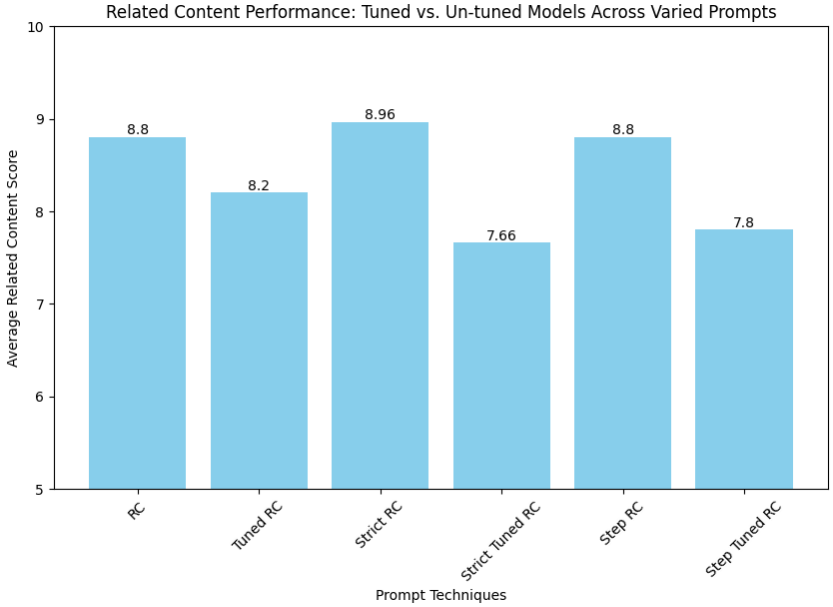

Figure 2: Related Content Performance: Tuned vs. Un-tuned Models Across Varied Prompts

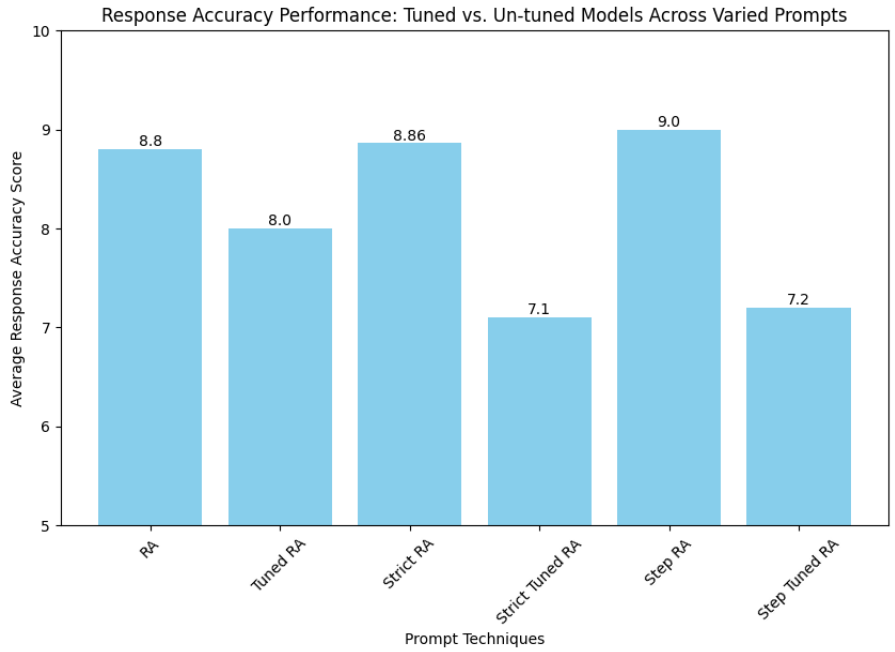

Figure 3: Response Accuracy Performance: Tuned vs. Un-tuned Models Across Varied Prompts

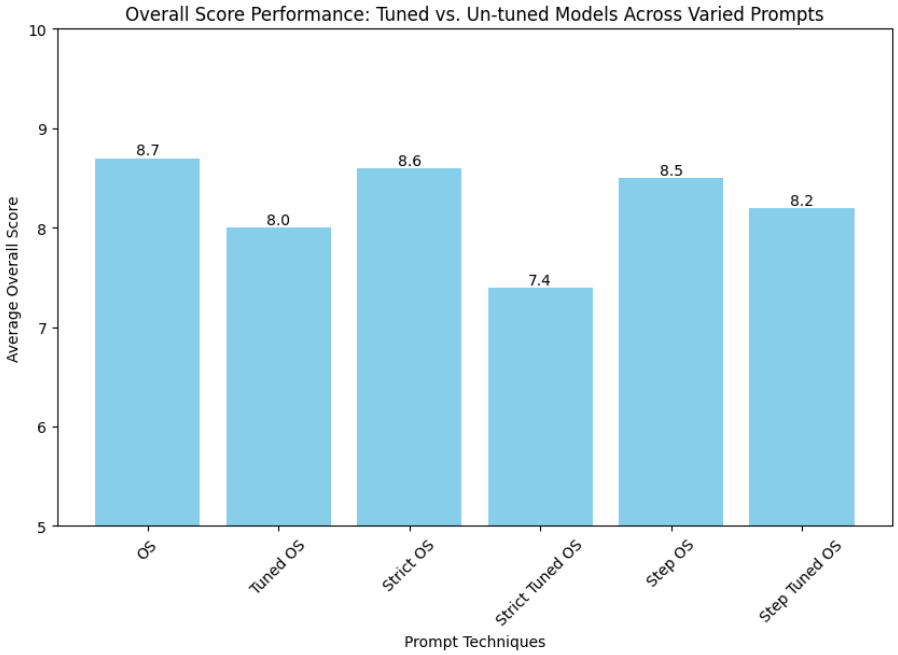

Figure 4: Overall Score Performance: Tuned vs. Un-tuned Models Across Varied Prompts

Result analysis

We have collected our conclusion of generated response quality for tuned and un-tuned models below, based on our evaluation metric:

1. **Language Fluency Scores**. The un-tuned model generally scores higher in language fluency compared to the tuned model, with a noticeable decrease in fluency scores when using the step-by-step prompting technique for the tuned model.
2. **Related Content Scores.** The related content scores follow the same pattern, with the un-tuned model achieving higher scores. The strict prompting technique results in the largest difference between the un-tuned and tuned models.
3. **Response Accuracy Scores**. Contrary to the initial analysis, the un-tuned model also scores slightly higher in response accuracy across all prompting techniques, with the largest gap observed under the step-by-step prompt.
4. **Overall Scores.** The overall scores, which reflect a combination of different performance metrics, show a similar trend where the un-tuned model scores higher. Again, the step-by-step prompt technique results in a smallest discrepancy between un-tuned and tuned models due to initial analysis.