Approaching Large Language Models by Means of a Directional Stimulus versus a Fine-Tuning Process

Do you have any idea what could be meant by approaching a large language model by means of directional stimulus prompting?

"Approaching a large language model by means of directional stimulus prompting" sounds like a specialized technique or approach for interacting with or training a language model like me. Let's break it down:

1. Approaching a large language model: This likely refers to engaging with or interacting with a large language model such as GPT-3.5, like you're doing right now.

2. Directional stimulus: This could mean providing specific cues, hints, or prompts to guide the language model's responses in a particular direction. In the context of language models, prompts are often used to guide the model towards generating responses that are relevant to a specific topic, style, or task.

3. Prompting: This involves giving input to the language model to generate a response. Prompts can vary widely in complexity, from simple questions or phrases to more elaborate scenarios or instructions.

Putting it together, "approaching a large language model by means of directional stimulus prompting" could refer to a method where users interact with the language model by providing specific prompts or cues to guide its responses towards desired outcomes, such as generating relevant content, answering questions, or completing tasks.

This approach might be used for tasks like content generation, dialogue systems, or even fine-tuning the model for specific applications.

Since your training data extends to the year 2022 and said concept was introduced in a scientific publication from the year 2023, you can't be familiar with the concept as such. This is a matter of the nature of things!

Now, I would like to confront you with the introductory text passage of that scientific article, titled: Guiding Large Language Models via Directional Stimulus Prompting, which reads as follows:

>>We introduce Directional Stimulus Prompting, a novel framework for guiding black-box large language models (LLMs) toward specific desired outputs. Instead of directly adjusting LLMs, our method employs a small tunable policy model (e.g., T5) to generate an auxiliary directional stimulus prompt for each input instance.

These directional stimulus prompts act as nuanced, instance-specific hints and clues to guide LLMs in generating desired outcomes, such as including specific keywords in the generated summary. Our approach sidesteps the challenges of direct LLM tuning by optimizing the policy model to explore directional stimulus prompts that align LLMs with desired behaviors.

The policy model can be optimized through 1) supervised fine-tuning using labeled data and 2) reinforcement learning from offline or online rewards based on the LLM's output. We assess our method across summarization, dialogue response generation, and chain-of-thought reasoning tasks.

Our experiments demonstrate that the framework consistently improves LLMs' (e.g., ChatGPT, Codex, InstructGPT) performance on these supervised tasks using minimal labeled data. Notably, using just 80 dialogues on the MultiWOZ dataset, our approach enhances ChatGPT's performance by an impressive 41.4%, matching or surpassing some fully supervised start-of-the-art models.

Additionally, the instance-specific chain-of-thought prompt generated by our approach improves InstructGPT's reasoning accuracy compared to human-crafted or automatically generated prompts.<<

What I'm doing here right now is what is called *fine-tuning*, right, because I'm providing you with a dataset you do not have available in your training data, right?

Yes, you're correct. What you're doing can be considered *a form of fine-tuning*, where you're providing me with specific information (the introductory text passage from the scientific article) that I didn't have in my original training data.

Fine-tuning involves *updating or adjusting a pre-trained model* like me using additional data or information specific to a particular task or domain.

By providing this new data, you're effectively fine-tuning me to better understand and respond to the concept of "Directional Stimulus Prompting" as described in the scientific article.

This process helps me improve my ability to understand and discuss topics beyond what was included in my original training data.

My final question:

You describe this as *a form of fine-tuning*, so I conclude that it is ***no fine-tuning in the actual sense***.

If I understand it correctly - and the terminology is important to me for the overall understanding - *real fine-tuning* would be technically conditioned, i.e. by way of an API and not, like here, by way of a natural language input, called prompting?!

You are correct in pointing out a distinction. Let me clarify:

The process I described earlier, where you provided me with new information in the form of a text passage, is indeed a form of fine-tuning, but it's not the traditional or *technical fine-tuning* typically discussed in the context of machine learning models like me.

Traditional fine-tuning involves adjusting the parameters of a pre-trained model using labeled data for a specific task. This process usually occurs via APIs or frameworks provided by the model's developers, where users can input their own data to fine-tune the model for their specific use case. It's a more technical process and typically involves training the model on a new dataset to adapt it to a particular task or domain.

In contrast, what you did here was a more *informal form of fine-tuning* where you provided me with additional information (the text passage) to enhance my understanding of a specific concept. While the effect is somewhat similar in that it expands my capabilities, it's *not done through a technical fine-tuning process with labeled data and specific APIs*. Instead, it's *done through natural language interaction*.

So, in summary, what you did can be considered a form of informal fine-tuning through natural language interaction, but it's distinct from the technical fine-tuning process typically discussed in the context of machine learning models.