OpenAI in PHP

How AI works in theory & how to use it in PHP

Fine Tuning & Agents

Weekly Topics

1. How Generative AI & LLMs Work
2. Prompt Engineering
3. Retrieval-Augmented Generation (RAG)
4. **Fine-Tuning a Model**
5. **Agents**
6. Using VUMC's OpenAI API in PHP

Key Concepts

- Caveats
- Fine Tuning
  - What It Is
  - How It Works
  - Common Uses
  - How to Prepare Data
- Agents
  - What They Are
  - How They Works
  - Common Uses
  - How to Prepare Code

Caveat #1

I've done very little work on implementing these concepts. I've only explored them theoretically & conceptually.

Caveat #2

I don't anticipate using either of these concepts in EFS projects. I'm including them for the sake of completeness in case we find use cases for them.

Fine Tuning: What It Is

- Unlike RAG which adds facts, fine-tuning adjusts the <u>tone</u> of an LLM

- Therefore, you can make it into a salesperson who pushes products

- However, prompt engineering (especially context and role) can also provide similar benefits
  - In code, this can be handled behind the scenes

Fine Tuning: How It Works

- Each LLM builds vectors with a set of weights
- Fine-tuning submits dozens/hundreds of examples to readjust the weights
- Therefore, adjusting the LLM is computationally expensive, more so than adding extra vectors (RAG)
- It has the effect of giving the LLM a unique "attitude," "approach," or "tone" to given questions
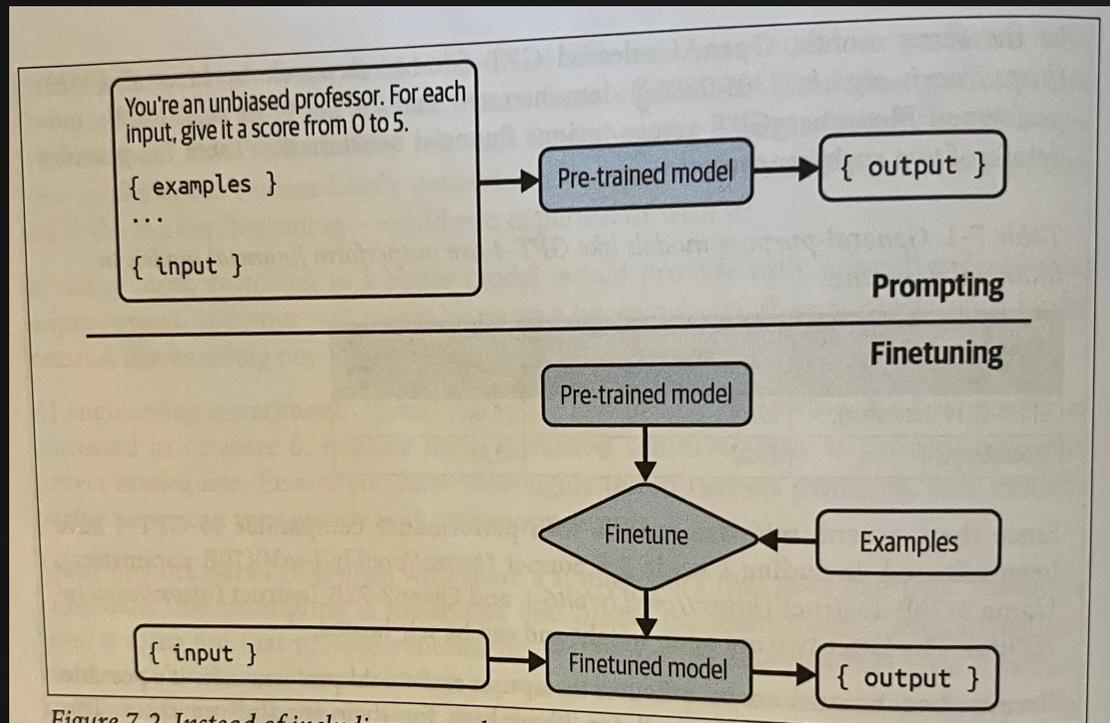
Fine Tuning: How It Works

Figure 7-2. Instead of including examples in each prompt, which increases cost and latency, you finetune a model on these examples.

Fine Tuning: Common Uses

- Modifying behavior of a model, not information (RAG)

- Creating a verbal culture of word usage
- Cutting down on irrelevant text
- Creating a verbal style
- Use with RAG to improve performance

Fine Tuning: Preparing Data

- Computationally & financially expensive since you are adjusting the LLM for your purposes
- It takes a lot of examples to fine tune a model (100 examples as a minimum)

Fine Tuning: Preparing Data

- GIGO: Bad data quality results in bad results
- Usually, give hundreds of examples to reinforce a certain behavior

- Alternative: Use a LLM's role to specify a tone, but this tone won't be customized

Fine-Tuning: Example

```
{ "conversations": [ {
    "from": ”user", "value": "Can you help me make pasta
carbonara?"
    }, {
    "from": ”system", "value": "Would you like the traditional
Roman recipe, or a simpler version?"
    }, {
    "from": ”user", "value": "The traditional version please"
    }, {
    "from": ”system", "value": "The authentic Roman carbonara
uses just a few ingredients: pasta, guanciale, eggs, Pecorino
Romano, and black pepper. Would you like the detailed
recipe?"
} ] }
```

Agents: What They Are

- Custom code that AI chooses to run in certain situations
- Keeps AI up-to-date with current or dynamic information
- RAG = settled information (history)

Agents: How They Work

- Code + a plain-text description
- AI uses the description to "choose" to use the code instead of consulting the LLM
- AI then integrates the result into its answer

Agents: How They Work

Figure 6-9. Decoupling planning and execution so that only validated plans are executed.

Agents: Common Uses

- What's today's weather?
- Mathematics
- Querying REDCap for data
- News
- Computationally complex tasks (deterministic)

Agents: Preparing Code

- Plain-text description + function in code
- In the literature, most example agents are written in Python
  - Other languages are supported (*e.g.*, C# and presumably PHP)
- Usually, agent code is complex
  - AI picks the tool based on the description & the prompt
  - Often, consults an API
  - Often, the function returns JSONs
  - Then agent turns the JSON into a string of text to return to the AI engine
  - AI integrates the response into its response

Key Concepts

- Caveats
- Fine Tuning
  - What It Is
  - How It Works
  - Common Uses
  - How to Prepare Data
- Agents
  - What They Are
  - How They Works
  - Common Uses
  - How to Prepare Code

Weekly Topics

1. ~~How Generative AI & LLMs Work~~
2. ~~Prompt Engineering~~
3. ~~Retrieval-Augmented Generation (RAG)~~
4. ~~Fine-Tuning a Model~~
5. ~~Agents~~
6. Using VUMC's OpenAI API in PHP

Next Week: OpenAI API in PHP at VUMC

OpenAI in PHP

How AI works in theory & how to use it in PHP

Fine Tuning & Agents