

GPT-4o Machine Learning Fine-Tuning: live chat Integration

Overview

I built a Python application that combined the fine-tuned model with retrieval of product and policy information. We then commissioned the developer of the company's live chat software to integrate my application into the existing system, allowing sales staff to use AI suggestions within their usual chat interface.

The problem

The fine-tuned model and vector database needed an application around them before they could be used in customer live chats. Each incoming message had to be combined with the conversation history, relevant product and policy information had to be retrieved, and the resulting answer had to be returned to the company's live chat software. This required coordinating the locally stored vector database with the external embedding and language-model APIs.

I also needed a practical way to test different model versions and response settings.

The application needed a consistent runtime environment, with its Python dependencies, embedding dataset and configuration available wherever it was hosted.

What was achieved?

I built a testing application for trained models to evaluate training and variations on the RAG (vector DB).

Edit Prompt

Clear Chat

Save to File

- ☒ ChatGPT4, 5 epochs
- ☐ ChatGPT4, 3 epochs
- ☐ ChatGPT 3.5 Turbo, 5 epochs

0.1

Hello, my name is <employee>, how can I help you today?

I'm looking to make an

Okay sure, when would you like to come in?

Maybe sometime

We are open 7 days a week, is there a particular day that would work best for you?

could we do anything on th

Yes of course, we are open 7 days a week. Is there a particular date that would work best for you?

Edit Prompt

Clear Chat

Save to File

- ☐ 5 epoch model checkpoint 1
- ☐ 5 epoch model checkpoint 2
- ☒ 5 epoch model
- ☐ 10 epoch model checkpoint 1
- ☐ 10 epoch model checkpoint 2
- ☐ 10 epoch model

☐ Show embedded

0.7 Temperature

I'd like to book an appointment, how do I do this?

Hello, my name is <employee>, thank you for your enquiry. May I take your email address and I can send you a link to our online booking system.

Edit Prompt

Clear Chat

Save to File

5 epoch model gpt4o mini

5 epoch model gpt3.5 turbo

Show embedded

0.7 Temperature

0.9 Top P

generate me links please for the olivia deisgn and also for the page of all solitaire designs

```
[{"query": "generate me links please for the olivia deisgn and also for the page of all solitaire designs", "user_result": "name: [Aurelia] short description: [engagement ring enticing solitaire crossover ring] description: [An enchanting four claw Round Brilliant diamond ring, with 10 extra round diamonds gracefully curving up each side of the setting, adding to the overall sparkle effect when worn. The 20 small diamonds are G colour and VS quality and they total 0.04 carats. The shank is D-shaped.] URL: [ https://www.samarajames.com/engagement-ring-enticing-solitaire-crossover-ring.html ]"}, {"query": "generate me links please for the olivia deisgn and also for the page of all solitaire designs", "user_result": "[Link to all solitaire set with oval shape: https://www.samarajames.com/engagement-rings.html?settingstyle=solitaire&diamondshape=oval-shape]"}, {"query": "generate me links please for the olivia deisgn and also for the page of all solitaire designs", "user_result": "[Link to all solitaire set with princess cut: https://www.samarajames.com/engagement-rings.html?settingstyle=solitaire&diamondshape=princess-cut]"}]
```

Hello, my name is <employee>, I can help with this. Here is the link for the Olivia design: <https://www.samarajames.com/engagement-ring-oval-basket-set.html>. And for all solitaire designs: <https://www.samarajames.com/engagement-rings.html?settingstyle=solitaire>. Is there any other way I can help?

olivia design

```
[{"query": "olivia design", "user_result": "name: [Taryn] short description: [engagement ring three oval design] description: [A lovely, graduated oval ring where the three diamonds are displayed in a pretty, crossover setting. The cushion shaped shank is slightly tapered.] URL: [ https://www.samarajames.com/engagement-ring-three-oval-design.html ]"}, {"query": "olivia design", "user_result": "name: [Priyanka] short description: [engagement ring oval halo setting] description: [A choice ring, with an irresistible oval centre diamond set within a halo of delightful round brilliant diamonds. These total between 0.14 carats and 0.18 carats, depending on the size of the oval diamond.] URL: [ https://www.samarajames.com/engagement-ring-oval-halo-setting.html ]"}, {"query": "olivia design", "user_result": "name: [Viola] short description: [engagement ring halo oval rubover set] description: [A truly alluring design, which enhances the luxurious central oval shaped diamond by surrounding it with a cluster of round brilliant diamonds.] URL: [ https://www.samarajames.com/engagement-ring-halo-oval-rubover-set.html ]"}]
```

Hello, my name is <employee>, please can you kindly confirm what you mean by "Olivia design"?

Send

Sales staff also used the application to evaluate the model and report incorrect responses.

Robb, 12:18

Edit Prompt

Clear Chat

Save to File

gpt4o mini

gpt4o

Show embedded

0.7 Temperature

0.9 Top P

Can I make an appointment on bank holiday monday?

Our showroom is open 7 days a week, but I'm afraid we are closed on Bank Holidays.

This response isn't correct, we're open on BH Mondays

12:18

interesting

I'll update it's information

did you find anything else?

Robb, 12:19

4-o mini got it right by the way

We commissioned the developer of the company's live chat software to integrate my Python application into the existing operator interface. Employees could review AI-generated suggestions, send them directly, edit them before sending, or write their own replies.

I also requested a report button so employees could flag poor responses. The integration recorded the model's input and suggested reply, together with the associated chat and employee, for later review.

The model was trained to use an <employee> placeholder instead of an employee's name. The integration replaced this placeholder with the name of the employee handling the chat, so the same model could generate personalised greetings for different members of staff.