# AI Engineering - LLM challenge

Hi there 👋

Congrats on passing to the assessment stage of our AI Engineer hiring process at saas.group. We're excited to have you in the process and can't wait to see your solutions.

🎯 **Objective**:

Develop a solution that can provide answers to users' questions by matching their queries with semantically similar questions in a predefined FAQ database. This system should further be enhanced to interact with the OpenAI API when it can't find a close enough match within the local FAQ database.

**Resources:**

- https://python.langchain.com/docs/get_started/introduction.html
- https://platform.openai.com/docs/guides/embeddings
- https://platform.openai.com/docs/guides/gpt/chat-completions-api
- https://fastapi.tiangolo.com/

> 💡 The purpose of the challenge is to assess technical skills and coding maturity. Code quality will be thoroughly reviewed.

**Requirements**:

1. **Embedding Computing**

   Use the provided Q&A examples (given below) to compute embeddings that will be used for similarity searches.

2. **Similarity Search**

When a user submits a query, search for the most similar question in your local FAQ database using the computed embeddings.

3. **Interacting with OpenAI API (via LangChain)**

   If the similarity score for the best local match is below a certain threshold, forward the question to the OpenAI API for an answer.

   > ❗ Here's an OpenAI key that you can use. Note that it is limited to a small amount of requests, so use it wisely! The key will be rotated or disabled every week.
   >
   > `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`
   >
   > Here are the permissions:
   >
   > `/v1/chat/completions`
   >
   > `/v1/embeddings`
   >
   > Allowed models:
   >
   > `gpt-4o-mini`
   >
   > `gpt-4o`
   >
   > `gpt-4.1`
   >
   > `gpt-4.1-mini`
   >
   > `chatgpt-4o-latest`
   >
   > `text-embedding-3-small`
   >
   > `text-embedding-3-large`
   >
   > Limits:
   >
   > 50 requests/minute
   >
   > 100k tokens/minute

4. **FastAPI Endpoint**

   Design an API endpoint using FastAPI where users can submit their questions and get answers.

5. **LangChain**,

   Use LangChain for any necessary natural language processing outside the scope of OpenAI.

**Challenge Steps**:

1. Compute embeddings for each question in the FAQ database using the provided text and any necessary tools from LangChain.
2. Create a similarity search function that takes in a user's question, computes its embedding, and finds the most similar question from the FAQ database.
3. Design a decision function to decide whether the found match is close enough or if the question should be sent to OpenAI's API.
4. Implement the API endpoint using FastAPI where users can submit their question and retrieve an answer based on the logic from the previous steps.
5. Make sure that all interactions with the OpenAI API (via LangChain) are correctly handled and errors are gracefully managed.

**Example Input/Output**:

*Input*:

```
POST /ask-question
{
   "user_question": "How do I reset my account?"
}
```

*Output if a local match is found*:

```
{
   "source": "local",
   "matched_question": "How can I restore my account settings?",
   "answer": "Go to settings and click on 'restore default'."
}
```

*Output if forwarded to OpenAI API*:

```
{
   "source": "openai",
   "matched_question": "N/A",
    "answer": "To reset your account, typically, you'd navigate to account settings and look for the 'reset' option. However, specific instructions may vary based on the platform."
}
```

**Use this database to generate embeddings**:

```
faq_database = [
    {
        "question": "How can I update my username?",
        "answer": "Go to your account settings, select 'Manage Username', enter your desired new username, and click 'Save Changes'."
    },
    {
        "question": "What's the process for recovering a forgotten username?",
        "answer": "On the login page, click 'Forgot Username', enter the email associated with your account, and follow the instructions sent to your email."
    },
    {
        "question": "How do I switch to dark mode in the application?",
        "answer": "In the application settings, look for 'Display Options'. Toggle the 'Dark Mode' switch to enable or disable it."
    },
    {
        "question": "Is it possible to link multiple email addresses to my account?",
        "answer": "Yes, in your profile settings, find 'Linked Emails', click 'Add Email', and follow the verification steps for each new email."
    },
    {
        "question": "How can I export my user data from the platform?",
        "answer": "Navigate to 'Privacy and Data' in your account settings, select 'Export Data', choose the data types you want, and click 'Generate Export'."
    },
    {
        "question": "What are the requirements for creating a secure passphrase?",
        "answer": "We recommend using a passphrase of at least 4 random words, totaling 20 characters or more. Avoid common phrases or quotes."
    },
    {
        "question": "How do I enable biometric login for the mobile app?",
        "answer": "Open the mobile app, go to 'Security Settings', and toggle on 'Enable Biometric Login'. Follow the prompts to set it up with your device."
    },
    {
        "question": "What steps should I take to permanently delete my account?",
        "answer": "Go to 'Account Management', select 'Delete Account', read through the implications, enter your password, and confirm deletion. This action cannot be undone."
    },
    {
        "question": "What should I do if I suspect unauthorized access to my account?",
        "answer": "Immediately log out of all sessions, change your password, and contact our security team through the 'Report Security Concern' form in the Help Center."
    },
    {
        "question": "How can I manage the frequency of email digests I receive?",
        "answer": "In your communication preferences, find 'Email Digest Settings', where you can adjust the frequency to daily, weekly, or monthly, or turn them off completely."
    }
]
```

This FAQ database covers various common queries related to user accounts. The challenge would be to identify the most similar question to a user's query and return the corresponding answer. If no sufficiently similar question is found in the database, the query would be forwarded to the OpenAI API for a response.

📋 **Evaluation Criteria**:

1. Quality of embeddings and similarity search results.
2. Correctness of the interaction with OpenAI API.
3. Robustness and error handling of the FastAPI application.
4. Env management.
5. Code structure, quality, and readability.

💡 The bonus points are *optional*. They do however allow us to assess some of the candidate's extended technical skills.
Choose what feels straight forward, easy, and that you would like to display as your best skills.
Keep in mind that by not solving the bonus points you are not in any disadvantage.
**You are still assessed on the main challenge.**

### ➕ Bonus Points:

1. Authentication added to endpoints by using FastAPI's dependency mechanism ( `Depends(get_token)` - a hint)
2. PostgreSQL usage for info and embedding storage ( `pgVector` )
3. Have the `postgres` instance run in `docker compose`
4. Scripts for managing embeddings DB: creating embeddings, updating embeddings (without deleting existing ones and being token-efficient), adding new collections
5. Create an AI Router using LangChain (docs)

   The router will allow us to handle 2 type of questions:
   a. is the question IT related and follows the topics that the system is designed to answer?
   b. if not, we have to route it to a Compliance Agent which will answer with a default output of
   "This is not really what I was trained for, therefore I cannot answer. Try again."
6. `Dockerfile` - run the system using a docker image, perhaps even `docker-compose.yaml` !
7. `Celery` - create a task for async embeddings processing

### ⏳ Timeframe:

Our expectation is that you spend no more than 3 hours on the main challenge (and no more than 6 hours if you choose to complete the extra parts of the challenge).
Please submit your solutions within 3 business days.

### ✈️ Submission:

Please submit your solution via email to vadim@saas.group with either a public link where the code can be downloaded, or an attached .zip file.

### 🤔 What you can expect:

A thorough code review of the solution on the mentioned criteria.

Best of luck! 🤞🏻