Step-by-Step Guide: Running Your AI Code in Google Colab

With Hugging Face Gated Model Access & Google Drive Setup

Step 1: Open Google Colab

1. Visit https://colab.research.google.com

2. Open your project notebook from:

   - 'Recent' if you've used it before

   - 'Google Drive' tab to browse your saved notebooks

   - Or upload it from your computer

Step 2: Mount Google Drive via GUI

1. In the Colab notebook interface, look on the left sidebar (click the small folder icon if it's hidden).

2. Click the 'Mount Drive' button at the top.

3. Follow the authentication prompt and approve access.

4. Your Google Drive will now be accessible at /content/drive/

Step 3: Install Required Python Packages

Use the following command to install required libraries:

!pip install -U transformers accelerate bitsandbytes sentence-transformers chromadb streamlit docx2txt pymupdf mammoth

Step 4: Get Access to a Gated Model on Hugging Face

1. Visit the model page on Hugging Face (e.g., mistralai/Mistral-7B-Instruct-v0.1).

2. Click 'Access Repository' or 'Request Access'.

3. Accept terms and wait for approval.

Step 5: Create Your Hugging Face Access Token

1. Go to https://huggingface.co/settings/tokens

2. Click 'New Token'

3. Name it (e.g., 'colab_token'), set role to 'Read'

4. Click 'Generate' and copy the token

Step 6: Log in to Hugging Face from Colab

Use this code snippet in a cell:

```
from huggingface_hub import login
login(token="your_token_here")
```

Step 7: Set File Paths and Load ChromaDB

Example path for ChromaDB:

CHROMA_PATH = "/content/drive/MyDrive/AI Chatbot Data/Glassdoor Chroma Store/chroma.sqlite3"

Use PersistentClient from ChromaDB to access your collection.

Step 8: Run the Model and App Code

Run your AI assistant logic (loading the LLM, running inference, etc.)

Ensure paths and model identifiers are correctly set.

Optional: Enable GPU

1. Runtime > Change runtime type
2. Set 'Hardware accelerator' to 'GPU'
3. Run: import torch; torch.cuda.is_available()