# Instructions: Extracting Dataset Info from Biomedical Papers

Step 1: Get Your OpenAI API Key  
You’ll need a personal OpenAI (ChatGPT) API key:  
1. Go to: https://platform.openai.com/account/api-keys  
2. Log in with your OpenAI account (create one if you don’t have one).  
3. Click “Create new secret key” → Copy the key (it starts with “sk-...”)

Step 2: Add API Key to the Code  
1. Open the file called `api\_keys.json`  
2. Replace the placeholder value under `"SaraQ"` with your own key:

```json  
{  
 "SaraQ": "sk-xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx"  
}  
```

Step 3: Clean the Paper (.xml) File  
This will turn the raw XML into a readable plain-text version:  
Open Terminal or Command Prompt and run:

```bash  
python preprocess\_xml.py --xml\_file /path/to/PMCpaper.xml --output\_txt cleaned\_PMCpaper.txt  
```

Replace `/path/to/PMCpaper.xml` with the path to your actual file.

Step 4: Run the Dataset Extraction  
Now, use the cleaned `.txt` file to extract information about datasets used in the paper:

```bash  
python extract\_info\_from\_paper.py --txt\_file cleaned\_PMCpaper.txt --output\_dir dataset\_info  
```

This will create a JSON file in the `dataset\_info` folder. The file contains:  
- Dataset names  
- Dataset sources  
- Data types  
- Brain regions studied  
- Tools used (e.g., Seurat, MAGMA)  
- Key findings

That’s it! You’ve extracted dataset knowledge using AI!