💡 Ollama **直接支援**以 **SafeTensors** 格式儲存的模型，允許使用者直接從 Hugging Face 等來源匯入模型，而無需預先將其轉換為 GGUF 格式。

Ollama 是透過其 **Modelfile** 系統來實現這種支援的。

💻 SafeTensors 格式模型的匯入方式

匯入 **SafeTensors** 模型是 Ollama 建立客製化模型工作流程的一部分，主要透過在 `Modelfile` 中使用 `FROM` 指令來完成。

1. 匯入 SafeTensors 模型權重

您可以直接在 `Modelfile` 中指向包含 SafeTensors 權重檔案的目錄，Ollama 會自動處理匯入和模型設定。

- **建立 `Modelfile`**： 在包含您的 SafeTensors 檔案（以及模型的配置檔如 `config.json` 和分詞器檔案）的目錄中，建立一個 `Modelfile`。
- **使用 `FROM` 指令**： 在 `Modelfile` 中，使用 `FROM` 指令指向該目錄。

  > **範例 (在模型權重目錄內建立 Modelfile):**
  >
  > ```
  > FROM .
  > ```

  或者，您也可以指向該目錄的完整路徑：

  > **範例 (指向特定目錄):**
  >
  > ```
  > FROM /path/to/safetensors/directory
  > ```

- **建立模型**： 使用 Ollama CLI 命令建立模型：

  ```
  ollama create my_model -f ./Modelfile
  ```

  Ollama 將讀取 SafeTensors 權重並將其轉換/封裝為可運行的內部格式（通常基於 llama.cpp 的架構）。

2. 匯入 SafeTensors 微調（Adapter）權重

如果您只有微調的權重（例如 LoRA 或其他 Adapter），您可以將它們套用到一個已知的**基礎模型**上。

- **使用 `FROM` 和 `ADAPTER` 指令**： 在 `Modelfile` 中， `FROM` 指令指向您用於微調的**基礎模型**名稱（可以是 Ollama 庫中的模型或您已建立的模型），而 `ADAPTER` 指令指向包含 SafeTensors 微調權重的目錄。

  > **範例:**
  >
  > ```
  > FROM mistral:7b
  > ADAPTER /path/to/safetensors/adapter/directory
  > ```
  >
  > - **注意**：確保 `FROM` 中使用的基礎模型與建立 Adapter 時使用的模型**相同**，以避免結果錯誤。

⚙️ 內部機制與 GGUF 的關係

雖然 Ollama 現在直接支援匯入 SafeTensors 格式，但其底層的推理引擎（基於 **llama.cpp**）仍然主要使用 **GGUF** (GPT-GEneration Universal Format) 格式。

- **自動處理**：當您匯入 SafeTensors 模型時，Ollama 在內部會處理必要的步驟，包括解析模型架構、讀取 SafeTensors 檔案中的張量權重，並將其轉換和/或封裝成其運行時所需的優化格式（通常是 GGUF 格式）。
- **量化 (Quantization)**：Ollama 也支援在模型匯入後進行量化。您可以將 FP16 或 FP32 的 SafeTensors 模型匯入，然後在 `ollama create` 時使用 `-q` 旗標來進行量化，以減少模型大小和提高推理速度。

> **總結**：對於使用者來說，Ollama **提供了簡化的流程**，允許您直接指向 SafeTensors 檔案目錄來匯入模型，而無需手動使用 `llama.cpp` 腳本進行 **SafeTensors 到 GGUF 的手動轉換**。

Would you like me to find a specific example of a Hugging Face SafeTensors model and show you the exact `Modelfile` content needed to import it into Ollama?