Functional UI Elements Inspired by LM Studio

This document outlines the core functional user interface elements and options available in LM Studio. This is intended to serve as a blueprint for creating a program with similar capabilities.

1. Main Dashboard / Home Screen

- **Search Bar:** For discovering and finding new models from repositories like Hugging Face.
  - *Functionality:* Keyword search, filtering.
- **Model Cards:** Displaying individual models found via search.
  - *Functionality:* Shows model name, creator, likes, download count, and file size.
  - **Download Button:** For each available quantization/version of the model.
  - **Download Progress Bar/Indicator:** Shows the status of model downloads.
- **Featured/Trending Models Section:** A curated list of popular or recommended models.

2. Chat Interface

- **Model Selection Dropdown:** Allows the user to load and switch between different downloaded language models for a chat session.
- **Chat History/Pane:** Displays the conversation between the user and the model.
  - *Functionality:* Scrollable history, user and AI messages are visually distinct.
- **Prompt Input Area:** A text box where the user types their messages.
  - *Functionality:* Multi-line input, send button, stop generation button.
- **New Chat Button:** Clears the current session and starts a new conversation.
- **Chat Presets/Personas:** A way to load predefined system prompts or configurations to steer the model's behavior (e.g., "AI Assistant," "Code Helper," "Story Writer").
- **Chat Settings Panel (Side Panel):**
  - **Inference Parameters:**
    - **Temperature:** Slider or input field to control randomness.
    - **Top-p (Nucleus Sampling):** Slider or input field.
    - **Top-k Sampling:** Slider or input field.
    - **Repetition Penalty:** Slider to discourage repetitive output.
    - **Max Tokens (Prediction Length):** Sets the maximum length of the model's response.
  - **Hardware Settings:**
    - **GPU Offloading:** Slider to select how many model layers to offload to the GPU (if available).
    - **CPU Threads:** Input to specify the number of CPU threads to use.
  - **Prompt Formatting:** Options to select the correct prompt format for the loaded model (e.g., ChatML, Alpaca, Vicuna).

    - **System Prompt Editor:** A text area to define the model's role, instructions, and context for the conversation.

3. Local Inference Server

- **Start/Stop Server Button:** To launch or shut down the local HTTP server.
- **Server Log/Output:** A console-like window showing server status, requests, and errors.
- **Model Selection:** Dropdown to choose which downloaded model to serve.
- **Server Configuration Options:**
    - **Host/Port Selection:** Input fields to change the server's network address and port.
    - **CORS (Cross-Origin Resource Sharing) Toggle:** To allow or block requests from other web pages.
    - **API Endpoint Information:** Displays the local server URL and example API calls (e.g., cURL, Python requests).

4. "My Models" Section

- **List of Downloaded Models:** A view showing all models saved on the user's local machine.
- **File Path Display:** Shows the location of the model file on the disk.
- **Delete/Remove Model Button:** To manage local storage.
- **"Reveal in Finder/Explorer" Button:** To open the folder containing the model file.
- **Model Information:** Displays details like file size and quantization version.

5. General Application Settings

- **Hardware Acceleration Toggle:** A primary switch to enable or disable GPU usage.
- **Theme Selection:** Options for light mode, dark mode, or system default.
- **Model Download Path:** Field to specify the default folder for saving downloaded models.
- **Check for Updates Button/Indicator:** To update the application itself.
- **Help/Documentation Links:** Links to external resources, guides, or a Discord community.