Internal server hosted AI assistance

Purpose

1. Embedded software:
    a. Automatic structuring of the code.
    b. Module generations based on the datasheets
    c. Readymade commenting
    d. Compliance against company standards and rules
    e. Automatic reports.
    f. Reliable and capable on the principle of FTR (First Time Right)

2. FE (python utilities and tools):
    a. Support in creating handy tools and small softwares, utilities, in python for speeding design activities.
    b. Ex: 1. MID tool ; 2. Cropee tool.

3. Technical documentation:
    a. Creating, validating documents based on user-friendly UI with Poka-yoke approach for more accuracy and reliability in the document creation, making the error rate lower.

4. Selpro:
    a. Creating POU on the json, xml logic generated by the AI, with simple sentence based inputs/prompts from the user.
    b. For effective and time saving in the ladder creation stage.
    c. Opening opportunities for developing an in-built json, xml convertor, where other similar ladder files can be also imported in Selpro's DSL (Domain Specific Language)
    (P.S.: complex task as handling of each minute factors are important and requires deep dive to understand the convergence logic

5. Document understanding:
    a. Providing summary tables, QnA, over the uploaded documents, in-turn reducing time for manual work.
    b. Smart suggestions over the available data as good as an intelligence for project considerations and initialization

Security

1. Completely air-gapped system.
2. Upgrades, uploads, can be done periodically via manual copy paste method, without indulging in the internet.
3. Strictly bound to the Selec IP.

Reference and best practices

1. Selection of various models, systems, methods, has been done on the basis of what is available in the world and should be open-source, should prevent making things from scratch, and provide edge features to be used directly without any kind of cost.
2. Models and platforms:
    a. The open source platform Ollama provides various models that can be downloaded and used directly. Those models have variants based on parameters trained (in billion)
    b. After that we have the open source web app which eliminates most of the work, providing rich features and security.
    c. This application is Open-WebUI. This is also an open source platform which delivers the exact requirements which is important.
    d. The models currently utilized in POCs are:
        - Qwen2.5-coder:7b
        - hf.co/bartowski/Phi-3.1-mini-4k-instruct-GGUF:Q5_K_M
          phi3:mini
        - qwen3:8b
        - gemma3:4b-it-qat
        - deepseek-r1:1.5b

Requirements

1. As we are working with AI, which is nonetheless an Artificial brain, so as the human brain requires a healthy environment, body, food, rest, etc.. for its proper functioning, similarly the artificial brain requires all the necessary things to work properly.
2. The models which we will be using vary in size because of the parameters, so the heavier the model the more resources it consumes.
3. Hence the processor as well the GPU should be capable enough to handle the model and multiple users at the same time.
4. An ideal system specifications are as follows:

CPU	16–32 cores
RAM	128 GB ECC
GPU	48 GB GDDR6
STORAGE	1 - 2 TB

(P.S.: IT support would be required for taking appropriate decisions and ideal recommendations.)

SELEC
Creating Best Value

Current status

1. Currently the trials on the various models stated above have been taken on desktop trials.
2. Document validation and generation using user-friendly UI and the local model running as the backend has been tested.
3. Code creations, document summarization has been tested.