**Technology Stack (Architecture & Stack)**

Date: 27 June 2025

Team ID: LTVIP2025TMID22707

Project Name: SmartSDLC: AI-Enhanced Software Development Lifecycle

**Components & Technologies**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Layer** | **Component / Tech** | **Function** | | --- | --- | --- | | **UI & Hosting** | Gradio | Build interactive text inputs, language selectors, exercise modules, and charts in Colab. | |  | Google Colab + ngrok | GPU-hosted notebook exposed via ngrok tunnel for external access | | **Language Detection** | langid | Detect input language and provide confidence score, enabling valid multilingual feedback | | **Model Inference** | ibm-granite/granite-3.3-2b-instruct via HF Transformers, PyTorch, Accelerate | Load and serve the 2B instruction-tuned Granite model with extended context capabilities | |  | Quantized models (GGUF 8‑bit, 4‑bit, etc.) | Reduce memory and speed for inference using converters like GGUF | | **Caching** | In-memory dict or functools.lru\_cache | Speed up repeated inference to improve real-time performance | | **Visualization** | Matplotlib / Plotly | Generate bar charts and radar charts for analysis feedback | | **Tunnel / API access** | ngrok, pyngrok, (optional) FastAPI | Securely expose the Gradio or custom API interface from Colab | | **Optional Production** | Docker + FastAPI + Redis | Containerize and productionize the stack with caching and REST endpoint support | | **Logging & Monitoring** | Python logging, optional Prometheus / Grafana | Track inference latencies, errors, and usage analytics | |