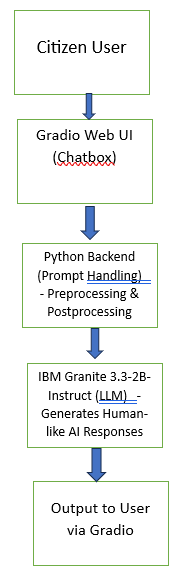
**Project Design Phase-II**

**Technology Stack (Architecture & Stack)**

|  |  |
| --- | --- |
| Date | 26 June 2025 |
| Team ID | LTVIP2025TMID35580 |
| Project Name | Citizen AI – Intelligent Citizen Engagement Platform |
| Maximum Marks | 4 Marks |

**Technical Architecture:**



**Table-1 : Components & Technologies:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Component** | **Description** | **Technology** |
|  | User Interface | Web-based chat interface for citizen querie | Gradio |
|  | Application Logic-1 | Handles prompt input/output flow | Python |
|  | Application Logic-2 | Model interaction logic (sending prompt to LLM and formatting response) | Transformers Library (Hugging Face) |
|  | Application Logic-3 | Not used in current phase | (Future) IBM Watson Assistant / STT |
|  | Database | Not required in current MVP | N/A |
|  | Cloud Database | Not used currently, potential for feedback logging | (Future) IBM Cloudant |
|  | File Storage | Local storage for downloaded model and tokenizer | Google Colab Local FS / /content/granite-model |
|  | External API-1 | Not used currently | N/A |
|  | External API-2 | Not used currently | N/A |
|  | Machine Learning Model | Language model to generate responses | IBM Granite 3.3-2B-Instruct (from Hugging Face) |
|  | Infrastructure (Server / Cloud) | Temporary deployment via Gradio (colab environment) | Google Colab (Local), Gradio Share |

**Table-2: Application Characteristics:**

| **S.No** | **Characteristics** | **Description** | **Technology** |
| --- | --- | --- | --- |
|  | Open-Source Frameworks | Hugging Face Transformers, Gradio | Python, Transformers, Gradio |
|  | Security Implementations | No user authentication in demo; can be extended with access controls | (Future) IAM, OAuth, HTTPS |
|  | Scalable Architecture | Modular design: UI + Prompt Handler + Model Pipeline (can be containerized) | Microservice-friendly, Docker-ready |
|  | Availability | Gradio sharing enables temporary public access; future: cloud deploymentthe availability of application (e.g. use of load balancers, distributed servers etc.) | Hugging Face Spaces / Cloud Foundry (planned) |
|  | Performance | Fast responses with local model loading; no latency from remote APIs | Token caching, device\_map="auto", CPU/GPU usage |