CODEGENIE

AI Powered Coding Assisant

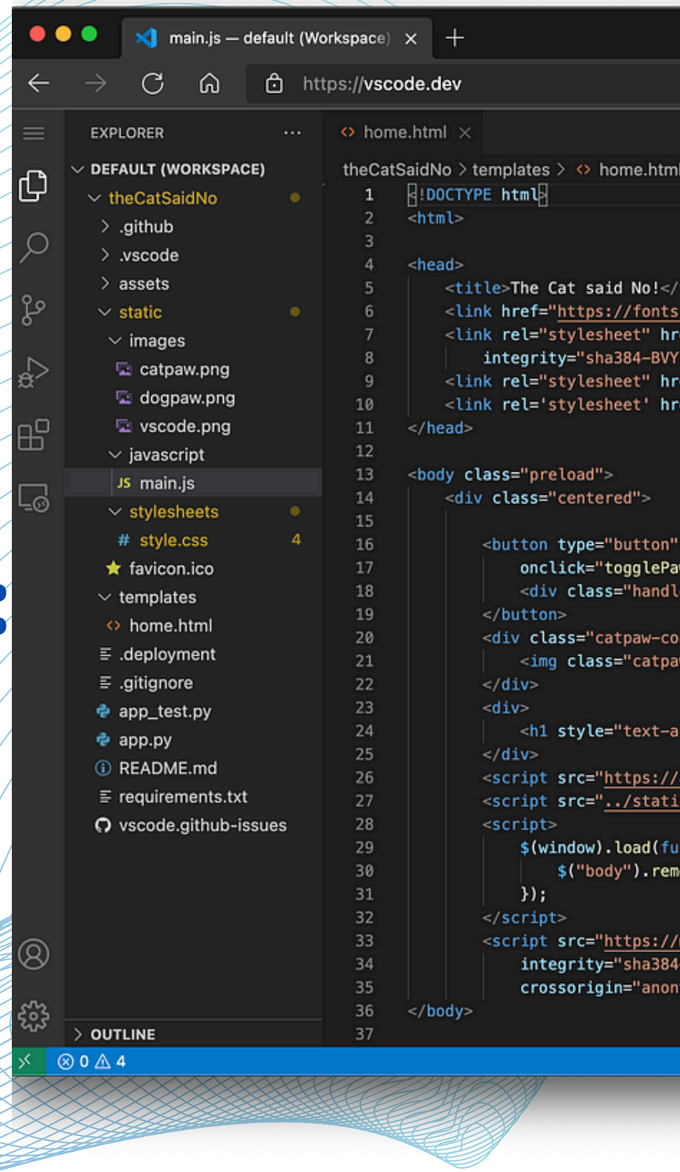
Name : Vandanapu Ashwita
Group Name : G413
Mentor Name : Sripooja

DEEPSEEK-CODER OVERVIEW:

DeepSeek-Coder is an open-source LLM trained on 2 trillion tokens, supporting 87 languages. It uses advanced techniques like next-token prediction and Fill in the middle, with a 16K context window for handling long code files. It outperforms models like GPT-3.5 Turbo and supports both academic and commercial use.

BY SEMESTER'S END, CODEGENI WILL:

- Auto-generate code snippets from prompts.
- Reduce repetitive tasks.
- Suggest clean, error-free code.
- Help developers stay focused by minimizing searches.
- Provide real-time, intelligent code recommendations



BUSINESS PROBLEM :

- Developer inefficiency due to repetitive coding tasks, context switching, and manual writing of boilerplate code.
- Slow development speed, increased coding errors, and reduced productivity, impacting the software development lifecycle for tech companies.

SOLUTIONS :

- Auto-generates code snippets based on prompts, streamlining development.
- It automatically finishes common functions and code for you, saving time and making you more productive.
- Minimizes syntactical errors by offering clean, functional code suggestions.
- Enhances coding accuracy and efficiency with intelligent, real-time code recommendations.

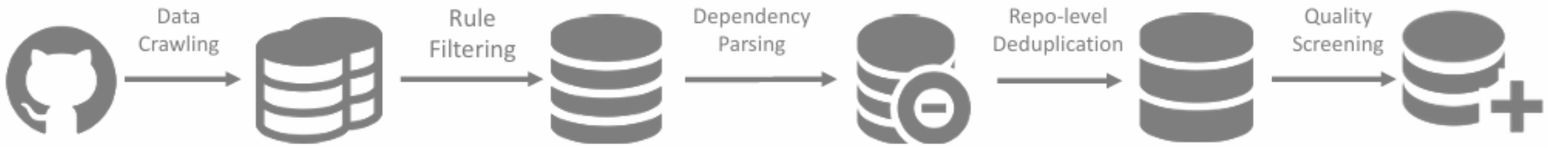
WHAT DATASET ARE USED ?

The project will use a pre-trained DeepSeek Coder model for code generation, which is trained on 2 trillion tokens of data:

- 87% from source code (GitHub, GitLab, Bitbucket)
- 10% from English code(API docs, tutorials, Stack Overflow)
- 3% from Chinese non-code articles

Optional fine-tuning with custom datasets (e.g., code snippets, templates) can be done for further customization.

DATA PREPROCESSED



- Raw code and documents are collected from public repositories and datasets.
- Low-quality, irrelevant, and malicious data is filtered out automatically.
- Dependencies are parsed, and duplicate files or projects are removed.
- Sensitive information like API keys and passwords is anonymized.
- Cleaned data is tokenized and validated before being used for model training

USERS :

- Software developers
- companies and enterprises
- Students
- Open-source contributors

ROLES

Developer

Admin

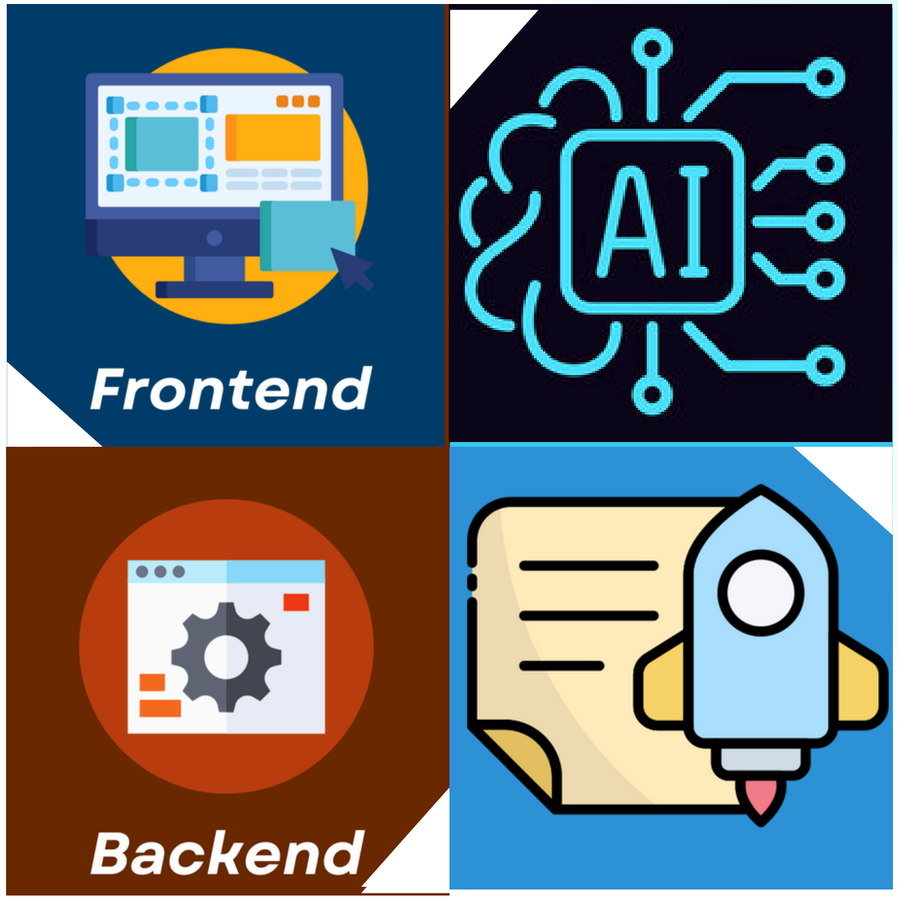
Trainer

USERS INTERACTION :

Input	Preprocessing	Output
User types in Vs code prompt.	Deepseek-coder analyzes context and generate required code.	Generated code appears in the Vs code

TECH STACK :

Language : TypeScript
UI : HTML, CSS, React
Framework :VS Code Extension API



DeepSeek-Coder(LLM)
GPU Acceleration : CUDA(to leverage RTX 4090 GPU
Hugging Face Transformer Library

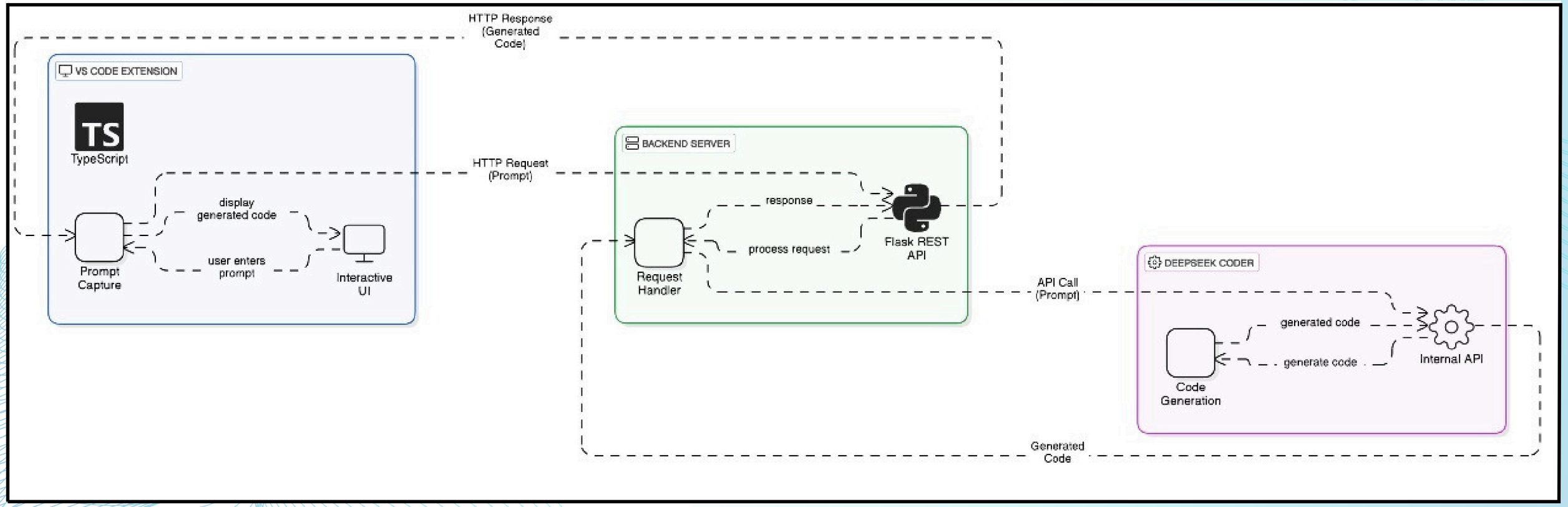
Flask API
Framework :VS Code Extension API
Integration : REST API

Deploy : vscode
Version Control : Git & GitHub

WILL THE APP DEPLOYED ON THE CLOUD ?

Yes, app deploys locally with code suggestions via the DeepSeek API, accessible through AWS, GCP, or Azure, and can be hosted on private servers for privacy.

PROJECT ARCHITECTURE



THANK TOU