International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 07, July 2026)

Unified Multi-Model AI Platform

Er. Syed Kaiser Mehraj[1], Er. Umer Farooq[2], Fahem Noor[3], Tabish Mohammad[4], Mohammad Inaam ul Haq[5]
[1,2]*Assistant Professors, Department of Computer Science, SSM College Of Engineering Pattan, Baramulla, Jammu & Kashmir, India*
[3,4,5]*Students SSM College of Engineering Pattan, Baramulla, Jammu & Kashmir, India*

***Abstract*-- New technologies in Artificial Intelligence (AI) led to the creation of strong language and multimodal models provided by OpenAI, Google, and Anthropic. Nevertheless, the use of various AI services is usually accompanied by difficulties associated with various APIs, authentication model, pricing model, and integration procedures.**

The proposed Unified Multi-Model AI Platform will overcome these issues by offering a centralized web-based platform that will combine various AI models, such as OpenAI GPT, Google Gemini, and Anthropic Claude. The platform supports intelligent model choice, live interaction with AI, image generation, text-to-speech translation, the creation of AI avatars, wallet management, payments, and administrative features.

It is developed with Fast API, Postgre SQL, Redis and Celery on the back-end, and React, TypeScript and Tailwind CSS on the front-end. A smart routing system automatically chooses the best AI model depending on the task of the user, increasing the quality of responses and efficiency of activities. WebSocket streaming is used to support real-time communication.

The system allows integrating many AI services into one platform, making it easier to access more advanced AI technologies and minimize the complexity of integration. The platform offers a scaling platform to future growth into a commercial AI Software-as-a-Service (SaaS) solution.

***Keywords*-- Artificial Intelligence, Multi-Model AI Platform, Large Language Models, Fast API, React, Intelligent Routing, WebSocket Streaming, AI Media Generation, SaaS Platform.**

1. Introduction

Artificial Intelligence (AI) is considered to be one of the most radical technologies of the modern world that has transformed the life of such industries as healthcare, education, finance, software development, and content creation. Recent developments in massive language models (LLMs) and multimodal artificial intelligence systems have allowed machines to do complex computations such as natural language understanding, code generation, reasoning, image generation, speech synthesis, and conversational interaction. Top AI vendors like OpenAI, Google, Anthropic, and others have launched powerful models with specialized features, providing users with access to more advanced AI-based services.

Even with such developments, the AI ecosystem is still very fragmented. All AI providers work on their own platforms, API architecture, authentication, pricing scheme, and service constraints. As a result, users and organizations that want to use several AI models have to operate different accounts, subscriptions, integrations and workflows. Such fragmentation adds complexity to operations, development, and overall costs and restricts the leverage of the capabilities of various AI systems in a cohesive environment.

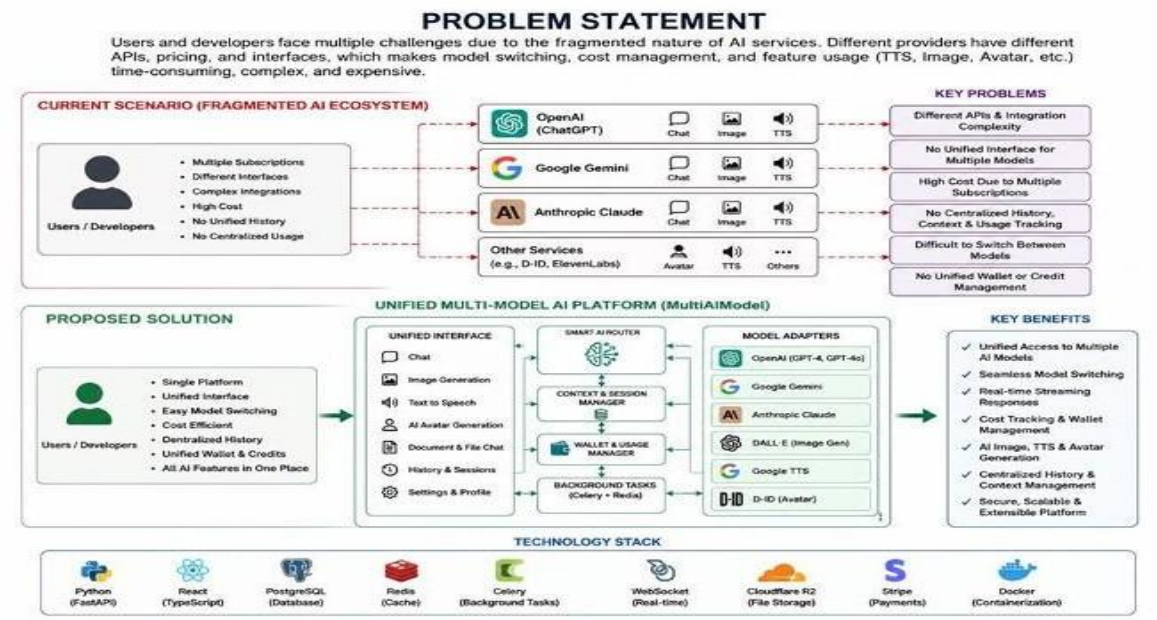
In order to overcome these issues, the idea of a Unified Multi-Model AI Platform has acquired a lot of relevance. This platform allows users to use various AI models at one interface, without the necessity to communicate with various providers separately. With all AI services connected to one unified system, users can easily alternate among models, compare the results, and select the most suitable AI technology to use in a particular task.

The proposed Unified Multi-Model AI Platform is a web-based application that combines various high-performance AI models, such as Open AI GPT, Google Gemini, and Anthropic Claude. The platform will integrate smart model routing, live AI communication, image generation, text-to-speech translation, AI avatars, credit management, payment, and administrative features. The intelligent routing mechanism will automatically choose the most appropriate AI model according to the character of the user request, thus enhancing the quality of responses, efficiency and cost-efficiency.

It is built on the latest technologies of Fast API, Postgre SQL, Redis, Celery, React, TypeScript, and Tailwind CSS. These technologies are scalable, secure, high performance and offer a better user experience. Moreover, the architecture is optimized to enable its further growth, i.e. the possibility to integrate more providers of AI and sophisticated AI processes.

The main goal of the research is to create and design a large-scale platform that would make it easier to access various AI technologies and make the integration process more complex. The proposed system will enhance usability, resource utilization, and establish a base of next-generation AI-powered applications and Software-as-a-Service (SaaS) solutions by offering a centralized environment of AI services.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 07, July 2026)



Need For The Project

The recent growth and high usage of Artificial Intelligence have boosted the need to have a central platform that will effectively coordinate and combine various AI services. Students, researchers, businesses, and content creators among other users often need a variety of AI functions, but it can be complicated, time-intensive, and expensive to manage multiple platforms and integrations.

The proposed Unified Multi- Model AI Platform can overcome these challenges through offering a single space of accessing and managing different AI technologies. This type of platform makes the use of AI easier, quicker and more efficient in operations, as well as a variety of applications.

The main advantages of the platform are:

- Easy integration and less development.
- One-stop access to various AI models and services.
- Improved user experience by having a unified interface.
- Efficient use of AI resources by means of smart model selection.
- Rapid design, testing, and implementation of AI-powered applications.
- Scalability, flexibility, and maintenance of the system are enhanced.
- Multimodal AI features, such as text, image, speech, and video processing.

Scope Of The Project

The suggested Unified Multi-Model AI Platform is a proposed web-based platform that will include numerous AI services under one platform. The platform will enable effective access and management as well as the use of the advanced AI technologies and grant scalability and usability.

The project scope comprises:

- Integration of multiple AI providers and models through a unified interface.
- Intelligent model selection and routing based on user requirements.
- Real-time AI communication with streaming responses.
- AI-powered image generation capabilities.
- Text-to-speech conversion services.
- AI avatar and video generation features.
- Secure user authentication, including Google O Auth integration.
- Wallet-based credit tracking and management.
- Online payment processing through Stripe integration.
- Administrative dash boards and usage analytics.
- Cloud storage for AI-generated media and user assets.
- Web Socket-based real-time communication.
- Redis-supported session and cache management.
- A synchronous task execution for resource-intensive operations.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 07, July 2026)

Objectives Of The Project

The key goals of the project are:

1. To create a web-based application that will be centralized to access various AI models through one dashboard.
2. To combine Open AI GPT models, Google Gemini, and Anthropic Claude with a provider adapter architecture.
3. To adopt a smart routing process that will automatically choose the most appropriate AI model based on user prompting and task difficulty.
4. To create a scaled back Fast API server that will be capable of asynchronous processing and real time WebSocket communication.
5. To develop a responsive frontend with React, TypeScript, and Tailwind CSS to make it more interactive.
6. To achieve real-time AI response streaming with WebSocket technology.
7. To incorporate AI image generating services in creating multimedia content.
8. To incorporate text-to-speech services to produce voice outputs powered by AI.
9. To apply the AI avatar generation to transform AI-generated responses into the animated video content.
10. To apply secure user authentication with the help of JWT authentication and Google OAuth.
11. To create wallet and credit management systems to monitor AI usage.
12. To incorporate Stripe payment services to subscribe and purchase using credit.
13. To use Redis caching to manage sessions and conversations effectively.
14. To execute background tasks asynchronously using Celery workers.
15. Implementation of Cloud flares R2 cloud storage that will be used to scale media storage and delivery.
16. To create an administrative dashboard with analytics, user management, and package management.
17. To offer the scalable architecture that can facilitate the future AI integrations and enterprise deployment.

III. Literature Survey

Recent advancements in Artificial Intelligence include the creation of large language models (LLMs) and multimodal AI systems that can handle tasks like natural language processing, reasoning, code generation, summarization, image generation, and conversational interaction. The major companies, such as OpenAI, Google, and Anthropic, have launched powerful AI models that have diversified the field of AI use in a variety of fields.

The GPT models of OpenAI have proven their abilities in content generation, programming support, educational support, and business automation. In the same way, Gemini models by Google offer advanced multimodal processing, large-context understanding and efficient reasoning which are applicable in both analytical and content-generation tasks. The Claude models of Anthropic are based on safe interactions with AI, understanding contexts, and code support, which provides dependable performance on complex reasoning tasks.

With the growing use of AI services, the demand for scalable and efficient system architectures has arisen. The latest backend technologies like FastAPI are capable of supporting high-performance API communication, asynchronous processing, and WebSocket-powered real-time interactions. Other technologies like Redis and Celery can also improve the scalability of the system with the help of caching, management of sessions, and the execution of background tasks.

Live communication is now an essential need in AI usage. WebSocket provides streaming responses and minimizes latency and enhances user interaction in AI-based dialogues. React and Tailwind CSS are popular frameworks on the frontend that are used to create responsive and interactive user interfaces.

The use of AI in media generation has also become popular. The technologies of image generation, text-to-speech synthesis, and the creation of AI avatars allow creating multimedia content rich based on user inputs. Also, cloud storage and web-based payment systems are significant towards the scalable AI Software-as-a-Service (SaaS) solutions.

The literature review demonstrates the increased significance of bringing together several AI services into a single platform. The suggested Unified Multi-Model AI Platform expands on them and integrates conversational AI, media generation, intelligent routing, cloud storage, and payment management into one scalable and extensible platform.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 07, July 2026)

IV. Requirement Analysis

Introduction

The requirement analysis is one of the most important steps in the software development lifecycle since it defines both functional and non-functional requirements of the planned application. The Unified Multi-Model AI Platform is perceived as a hub where a range of artificial intelligence technologies are integrated into one unified platform. The platform will support real-time AI communications, smart model choice, multimedia content creation, digital wallets, and scaled cloud services.

The step will help identify the hardware and software requirements, user requirements, system constraints, and technical requirements to make the platform design and implementation a success. The solution proposed will provide elevated performance, scalability, security, and effective coordination of different AI services and a user-friendly experience.

Functional Requirements

Functional requirements are used to state the basic functionalities and services that the system should have.

1. User Authentication System

The site must have secure registration and authentication services. It should be authenticated using both email-password and Google OAuth-based authentication. JWT (JSON Web Token) authentication should also be employed and this is to provide security in communication between clients and the application programming interfaces (APIs).

2. Multi-Model AI Chat System

The system must allow users to converse with various AI models, such as OpenAI GPT and Google Gemini, as well as Anthropic Claude, using a single conversational interface.

3. Intelligent AI Routing

The platform must be in a position to analyze user prompts and automatically forward requests to the best suited AI model, depending on the nature of the task like programming, logical reasoning, content summarization or creative content generation.

4. Real-Time AI Response Streaming

To promote user interaction and reduce delays, the system must provide real-time AI response delivery with streaming technology based on WebSockets.

5. AI Image Generation

Users must be able to generate AI generated image by giving descriptive text prompts.

6. Text-to-Speech Generation

The platform is supposed to turn AI-generated textual answers into audio speech via integrated text-to-speech technologies.

7. AI Avatar Generation

This system must enable the development of avatar based videos through synthesized speech coupled with avatar images to create interactive visual mediums.

8. Wallet and Credit Management

The platform must uphold personal user wallets, track attainable credits, and automatically deduct usage credits in accordance with the AI services utilized.

9. Stripe Payment Integration

They should also include an online payment option using Stripe in which users can afford subscriptions, credits, or service packages.

10. Chat History and Session Management

The session management and conversation storage should allow the user to review past conversations and resume past interaction.

11. File Upload and Processing

The platform is to be able to upload and process a variety of file formats, such as PDF documents, text files, and source code files, to engage in AI-assisted analysis and interaction.

12. Administrative Dashboard

It should have an administrative panel to handle user accounts, subscription plans, service packages, and system analytics, and general operation of the platform.

Non-Functional Requirements

Non-functional requirements are used to outline the quality standards and functional attributes that the system should meet.

1. Scalability

The platform must be in a position to support an increasing number of users and artificial intelligence requests without a significant drop in the overall performance.

2. Security

There should be strong security in terms of secure authentication, transmission of data in encrypted forms, verification of tokens and role-based access control mechanisms.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 07, July 2026)

3. Reliability

The system must deliver reliable and continuous AI services, making sure that it is available when required and that it operates steadily.

4. Performance

The platform must provide low-latency and quick response time by using asynchronous processing and optimization of the system architecture.

5. Maintainability

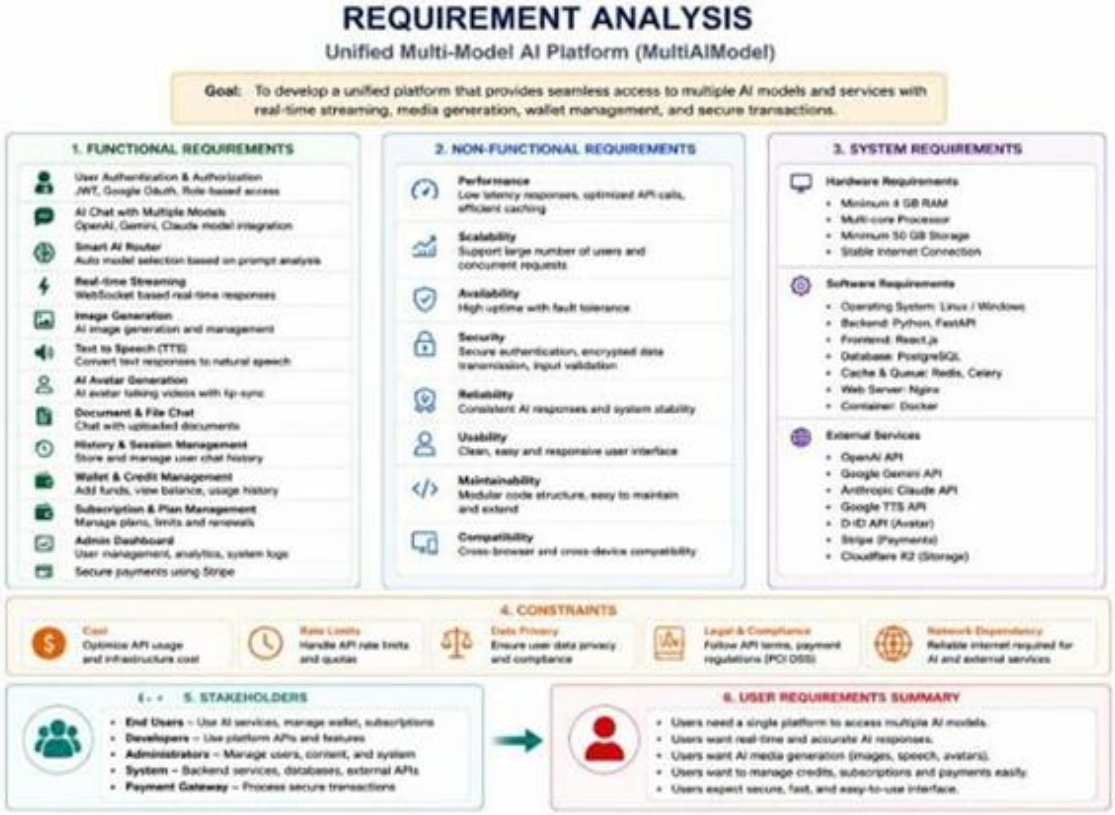
A modular architecture should be used to develop the software, which will make it easier to maintain, update, and upgrade in the future.

6. Availability

This system must be in operation and always available to the users to deliver the services at all times.

7. Cloud Compatibility

It must be able to deploy and execute on cloud computing environments to provide scalability and effective management of resources.



Hardware Requirements

To create and implement the Unified Multi-Model AI Platform, a combination of hardware resources that are able to facilitate efficient application execution and AI service integration is necessary. The hardware requirements will be:

- *Processor:* Intel core i5 or more powerful processor.
- *Memory (RAM):* 8 GB RAM minimum and 16 GB RAM suggested improving performance.
- *Storage:* Solid State Drive (SSD) having a storage capacity of at least 256 GB.
- *Internet Connectivity:* A stable high-speed internet connection to ensure seamless communication with cloud-based AI services.
- *Graphics Processing Unit (GPU):* Optional but beneficial for advanced AI computations, machine learning tasks, and multimedia processing.

Software Requirements

The software environment comprises of backend technologies, frontend technologies, AI service integrations, and infrastructure tools needed to develop and operate the platform.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 07, July 2026)

Backend Technologies

The backend architecture is developed on the basis of modern frameworks and tools that enhance the scalability, security, and efficient data processing.

- *Python 3.11*–Entry programming language in the back end.
- *Fast API*–API framework based on high-performance construction of RESTful APIs.
- *PostgreSQL*–Relational database management system- This is used to store structured application data.
- *Redis* –In-memory data store, used to store data in memory, used in caching and session management.
- *Celery*–A synchronous background processing distributed task queue.
- *SQL Alchemy*–Object Relational Mapping (ORM) of database interactions.
- *Pydantic*–Library of data validation and data settings management to ensure data integrity.

Frontend Technologies

It is built on the latest web technologies and thus offers its user interface is interactive and responsive.

- *React19*–JavaScript library to create dynamic user interfaces.
- *Type Script*–Strongly typed programming language that enhances Java Script development.
- *Tail wind CSS*–Utility-first CSS framework for responsive and customizable interface design.
- *Axios*–HTTP client that is employed to communicate between frontend and backend services.
- *React Router*–Library of navigating and routing through the application.
- *SWR*–Library is a data-fetching library that enhances performance with caching and automatic data synchronization.

AI Service Integrations

The platform combines several AI services to offer high-order conversational, multimedia and content-generation services.

- *Open AIAPIs*–Used for natural language processing, text generation, and conversational AI.
- *Google Gemini APIs* – Provide multimodal AI capabilities and advanced reasoning features.
- *Anthropic Claude APIs* – Support intelligent conversational interactions and content generation.
- *Google Text-to-Speech APIs* –Turn textual results into natural speech.
- *D-ID Avatar APIs*–A video of an avatar created by AI with synthesized speech and images.

DevOps and Infrastructure Tools

The infrastructure technologies used to provide reliable deployment, scaling, and management of the system include:

- *Docker*–Containerization platform for consistent application deployment across environments.
- *NGINX*–Web server and reverse proxy used for load balancing and traffic management.
- *Cloud flareR2*–Cloud-based objects to rage service for storing media files and application assets.
- *Stripe APIs* – Payment processing services used for subscriptions, wallet recharges, and financial transactions

V. Proposed Methodology

Introduction

The suggested Unified Multi-Model AI Platform is based on a modular and scalable design that unites a central system that incorporates various AI technologies. It is based on real-time AI communication, multimedia generation, and efficient AI orchestration. Asynchronous architecture will be used in order to provide high performance, responsiveness and scalability.

System Architecture

The system is divided into frontend, backend, AI orchestration, media generation, database, infrastructure layers. The frontend interacts with the backend via REST APIs and Web Socket links. The backend provides authentication, AI routing, wallet transactions and task processing.

The data is stored with the PostgreSQL system and the Redis is utilized to maintain caching and session management. Celery workers perform resource-consuming activities like generating images and avatars in the background.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 07, July 2026)

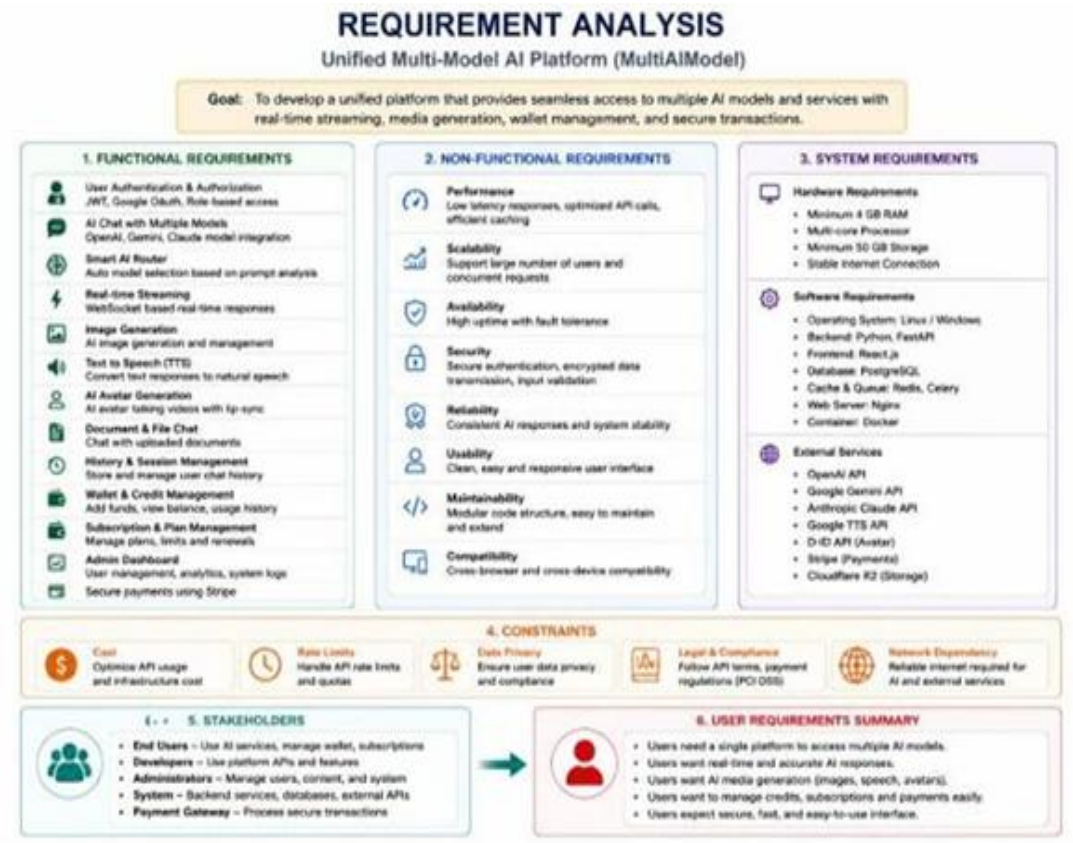


Figure 2: High-Level Architecture of the Unified Multi-Model AI Platform

Frontend Architecture

Reacts 19, Type Script, and Tailwind CSS are used to write the frontend ensuring a responsive, interactive user interface. It consists of AI chat, image creation, avatar creation, wallet management, authentication and administrative modules.

The navigation is performed with React Router, API communication with Axios, and data fetching and caching with SWR which makes it easy to use and efficient in interaction with the user.

VI. Expected Outcomes And Future Work

Expected Outcomes

The Unified Multi-Model AI Platform proposed is likely to provide a unified and effective platform through which various AI services can be accessed via a single interface. The project main deliverables are:

- Single access to various AI models
- Real-time AI response streaming.
- Intelligent AI model routing.
- AI-based image generation
- Text-to-speech conversion
- AI avatar generation
- Secure user authentication
- Wallet and payment management
- Administrative monitoring and management tools
- Scalable cloud-based deployment

The platform will be able to streamline the incorporation of AI, improve operational efficiency, and offer a better user experience over individual AI applications.

VII. Future Work

Other features and sophisticated functions to be added to the platform include:

1. Creation of Android and iOS mobile applications.
2. An addition of other AI providers and open-source language models.
3. Installation of voice-assisted AIs and calling features.
4. Introduction of domain-specific fine-tuned AI models.
5. State-of-the-art analytics and AI performance monitoring dashboards.

6. Multilingual AI interactions.
7. Solutions of enterprise-level deployment and API.
8. Creation of AI agents and workflow automation systems.

VIII. CONCLUSION

The entire goals set in the process of planning and development of the project were met and confirmed by the end delivery of the system. The platform that has been developed successfully combines the essential features, such as AI Chat, Image Studio, Avatar Studio, Billing and credit Management, User Management, Package Management, Dashboard Analytics, Login Authentication, and Account Settings. The screenshot of the system indicate that it is possible that all these modules were successfully operated and integrated, which proves the functionality, reliability, and the overall success of the proposed solution

REFERENCES

[1] Anthropic. (2024). *Claude 3 model card*. Anthropic Research.

[2] Amazon Web Services. (2024). *AWS cloud storage services documentation*. AWS.

[3] Celery Project. (2024). *Celery distributed task queue documentation*. Celery Project.

[4] D-ID. (2024). *D-ID developer documentation*. D-ID.

[5] Google Cloud. (2024). *Cloud Text-to-Speech documentation*. Google Cloud.

[6] Meta Platforms, Inc. (2024). *React documentation*. React.

[7] Stripe. (2024). *Stripe API documentation*. Stripe.

[8] Tailwind Labs. (2024). *Tailwind CSS documentation*. Tailwind CSS.

[9] Team, G., Anil, R., Borgeaud, S., Alayrac, J. B., Yu, J., Dai, A. M., Hauth, A., et al. (2024). *Gemini: A family of highly capable multimodal models*. *arXiv*. https://arxiv.org/abs/2312.11805

[10] OpenAI. (2023). *GPT-4 technical report*. *arXiv*. https://arxiv.org/abs/2303.08774

[11] Ramírez, S. (2023). *FastAPI documentation*. FastAPI.

[12] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical text-conditional image generation with CLIP latents*. *arXiv*. https://arxiv.org/abs/2204.06125

[13] Sanfilippo, S. (2022). *Redis documentation*. Redis Labs.

[14] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). *Language models are few-shot learners*. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

[15] Fette, I., & Melnikov, A. (2011). *The WebSocket protocol (RFC 6455)*. Internet Engineering Task Force (IETF). https://www.rfc-editor.org/rfc/rfc6455