Supplementary File – Team Hackuccino

1. Project Overview

Project Name: Clarifile – Cognitive AI for File Management
Team Name: Hackuccino
Category: AI for Core Applications

Problem Statement:
With the exponential growth of digital content, users today manage thousands of files across multiple devices such as smartphones, laptops, and cloud storage platforms. These files—ranging from documents and presentations to images, videos, and downloads—often accumulate in an unstructured manner.
Traditional manual sorting is inefficient and inconsistent, leading to:

- Redundancy caused by duplicates and scattered versions.
- Reduced productivity due to time wasted in manual organization.
- Fragmented user experience across devices with inconsistent folder structures.

This unstructured accumulation of data contributes to digital clutter, which significantly impacts efficiency, usability, and overall user satisfaction. At the same time, users are increasingly concerned about privacy and data security, making them hesitant to adopt solutions that require uploading sensitive files to external servers. There is a growing need for an AI-driven file management solution that can intelligently analyze file content, context, and usage patterns to automatically create meaningful folder structures and ensure consistency across devices—while preserving user privacy through on-device or privacy-first processing

Solution Summary:
Clarifile is an AI-powered cognitive assistant that helps users automatically categorize, summarize, search, and declutter their files across formats — PDFs, Word docs, images, and audio. It seamlessly integrates with Google Drive through a Chrome Extension, allowing one-click intelligent organization, all while maintaining on-device privacy and security.

Clarifile provides a single, unified dashboard for file insights, semantic search, duplicate cleanup, and category management — empowering users to focus on their work instead of managing files.

Links

- **Demo Video:** Watch Here
- **Project Report:** View Here
- **OAuth and Extension Setup Guide (PPT):** View Here
- **Troubleshooting Guide:** View Here
- **Gemini API Keys:** Access Here
- **Sample File Database:** Access Here
- **Compiled Submission:** Access Here

2. System Architecture

Clarifile follows a **modular microservice architecture**, ensuring scalability and separation of concerns.

Frontend (React + TypeScript)

- Interactive dashboard built with Tailwind CSS.
- Tabs for Dashboard, Files, Duplicates, Categories, AI, and Search.
- Real-time file stats, semantic insights, and batch file operations.

Backend

- **FastAPI (Python):** Handles text extraction, OCR, audio transcription, and AI categorization.
- **Node.js Gateway (Express):** Manages routing, Google Drive API calls, and OAuth2 token handling.
- **FAISS Vector Database:** Enables lightning-fast semantic and similarity searches.

AI/ML Services

- **Gemini API** for contextual understanding and summarization.
- **Sentence Transformers** for semantic embeddings.

- **spaCy** for entity recognition and topic extraction.
- **Tesseract OCR** for text extraction from images.

Integration

- **Chrome Extension** for Drive integration and real-time synchronization.
- **OAuth2 Authentication** with secure token refresh.

---

3. Technical Stack

Layer	Technology	Description
Frontend	React, TypeScript, Tailwind CSS	Responsive, interactive UI
Backend	FastAPI (Python), Node.js (Express)	Handles parsing, routing, API communication
AI/ML	Gemini API, Sentence Transformers, spaCy, Tesseract OCR	NLP, embeddings, document understanding
Database	FAISS	Vector-based semantic indexing
Integration	Chrome Extension, Google Drive API	Secure file access and automation

---

4. Key Features

1. **AI Document Summarization** – Automatically summarizes and extracts insights from long documents.
2. **Hybrid Search Engine** – Combines semantic and keyword search across text, OCR, and audio content.
3. **Duplicate Detection & Resolution** – Identifies exact, near, and semantic duplicates with side-by-side comparison and undo support.

4. **Smart Categorization** – Suggests document categories based on content and user feedback using incremental learning.

5. **Google Drive Integration** – One-click organization directly from Google Drive via Chrome Extension.

6. **Privacy-First Processing** – All AI operations can run locally for sensitive data protection.

---

5. Use Cases & Impact

Clarifile delivers real value across multiple domains:

For Medical Professionals

Doctors and healthcare administrators can:

- Instantly retrieve patient records, prescriptions, and scans.
- Use semantic search to find all files related to a specific diagnosis or patient.
- Categorize documents automatically by patient or department.
- Maintain privacy since all processing can remain local.

Impact: Reduces admin overhead and response times, enabling better patient care and compliance.

For Businesses and Enterprises

Organizations can:

- Auto-sort client contracts, invoices, and IDs.
- Use hybrid search to find specific clients or project documents.
- Resolve duplicates and maintain clean, compliant data.

- Sync organized folders back to Google Drive automatically.

Impact: Saves hours of manual sorting and reduces human error in document management workflows.

For Students and Researchers

Students and research professionals can:

- Summarize academic papers and extract timelines or key terms.
- Auto-categorize files by subject or topic.
- Identify and delete duplicates (e.g., old versions of assignments).
- Create flashcards or key point summaries for revision.

Impact: Enables smart, time-efficient study and research organization.

---

6. Challenges Faced

- **OAuth2 Integration with Chrome Extension:** Managing tokens, refresh logic, and user session persistence.
- **Handling Multi-Modal Data:** Processing PDFs, images, audio, and text uniformly.
- **Real-Time Orchestration:** Coordinating FastAPI and Node.js services efficiently during large file operations.
- **Rollback Safety:** Implementing undo-safe actions during duplicate resolution.
- **Ensuring Privacy:** Designing the architecture to support on-device AI processing for sensitive data.

---

7. Future Enhancements

- **Integration with OneDrive and Dropbox** to expand ecosystem reach.
- **Fine-Tuned Local LLM** for fully offline document Q&A.
- **Collaborative Workspace Dashboard** for team-based file sharing and approvals.
- **Mobile Companion App** for quick access and voice-assisted organization.

---

9. Summary

Clarifile redefines digital file management by fusing AI-driven intelligence, intuitive user experience, and real-world practicality.
It empowers professionals, students, and enterprises to work smarter, declutter faster, and maintain control over their digital ecosystem — all with privacy at its core.

Clarifile — Bringing AI-driven clarity to your digital world.

Team Members

- Nayana Jagadish Raikar
- Prasasthi Sanjana Chekuri
- Prathiksha R
- S Harshitha
- Sampada G Kulkarni