
AnyDocAI

From litun nayak <litunnayak09@hotmail.com>

Date Sun 4/6/2025 4:26 PM

To Litun Nayak <nayaklitun9@gmail.com>; DIBYAKANTA NAYAK <dibyakantanayak2@gmail.com>

What is AnyDocAI?

AnyDocAI is an AI document assistant that lets you chat with all your files — PDFs, Word docs, Excel sheets, PowerPoint presentations, Charts, Graphs and text files — all in one place.








Ask questions, get instant summaries, extract data, and find insights across one or many files at once. Powered by GPT-4 Turbo and Claude 3, AnyDocAI understands your documents like a human expert — but responds in seconds.

Whether you're a student, researcher, analyst, or part of a busy team, AnyDocAI turns your documents into a living, searchable conversation.

Core Capabilities

- Chat with Any File: Just upload and ask – works with .pdf, .docx, .txt, .xlsx, .pptx, and more.
 - Ask Across Multiple Files: Chat with one document — or many at once.
 - Instant Summaries: Condense long reports, papers, or slide decks into simple takeaways.
 - Extract Key Data: Ask for names, dates, decisions, numbers — anything.
 - Smart Search: Search inside hundreds of pages instantly, across all your uploads.
 - Organized Workspaces: Group files by project or topic with folders, tags, and chat history.
 - Collaboration Ready: Share chats and documents with teammates securely.
 - Chat Memory: Save and revisit your conversations anytime — with full context.
-

Why Use AnyDocAI?

Feature	What It Delivers
 Multi-File AI Chat	Ask questions across multiple uploads at once
 GPT-4 + Claude 3	Choose your favorite LLM for better results
 All File Types	Supports PDFs, Word, Excel, PPTs, and plain text
 API Access	Build custom workflows with our upcoming developer API
 Team Features	Invite teammates, set roles, and collaborate
 Secure by Design	Files never leave your workspace; stored on Wasabi (S3)
 Affordable Plans	Free tier available, Pro starts at ₹999/month

Who is it for?

- Lawyers & Analysts → Understand long contracts or reports in seconds
 - Researchers & Students → Review and query papers, notes, and slides
 - Teams & Startups → Organize, search, and extract insights from company docs
 - Anyone Drowning in Docs → Turn your file mess into a smart knowledge base
-

Real-World Use Cases

- “Summarize this 40-page research paper in 3 bullet points.”
 - “What are the key risks mentioned in these financial statements?”
 - “Compare numbers across these Excel sheets.”
 - “Find action items in the last 5 client proposals.”
 - “What was discussed about pricing in these PPT decks?”
-

Under the Hood

- Chunked Retrieval → Only the most relevant content is sent to AI for fast, accurate answers
 - Background Workers → Smart processing, embeddings, and real-time indexing
 - Secure Storage → Stored privately on Wasabi S3-compatible storage
 - Scalable Architecture → Built with Redis, PostgreSQL, and distributed queues
-

AnyDocAI – Master Plan (2025 Edition)

AI-powered, team-ready document assistant for fast search, summarization, and Q&A on any file.

1. Core Tech Stack (Battle-tested & Cost-effective)

Layer	Tech Used	Purpose
Chat AI	OpenAI GPT-4-Turbo (RAG-based)	Language model
Backend	Python + FastAPI, Celery	API + background jobs
Frontend	Next.js 14, TailwindCSS, shadcn/ui	UI framework
File Parsing	unstructured, PyMuPDF, python-docx, pandas, pdf2image	File parsers & pdf2image for PDF visual graphs (in future)
Storage	Wasabi S3	File storage
Databases	PostgreSQL (Supabase), MongoDB	Relational + dynamic storage
Embeddings	text-embedding-3-small / Instructor-XL	Vector embeddings
Vector Store	Weaviate (preferred)	Vector search engine

AI Integration	LangChain	Model routing, RAG chaining
Rate Limiting	Vercel AI SDK / Wavelet	User-level rate limiting + batching
Auth	Supabase Auth / Clerk	Authentication, JWT, OAuth
Billing	Stripe / Razorpay	Per-seat subscriptions
Monitoring	LangSmith, Sentry, Grafana (optional)	Tracing & logs

2. Folder Structure (Cursor-Ready)

Backend

```

/app
  /api
    routes.py          # FastAPI endpoints
    schemas.py         # Pydantic request/response models
  /services
    file_parser.py     # PDF/DOCX/Excel parser logic
    chunker.py         # Splitting files into chunks
    embedder.py        # Embedding logic (LangChain)
    query_engine.py    # Search + prompt builder
  /workers
    tasks.py           # Celery background jobs (upload → chunk → embed)
    utils.py           # Progress updater, logs
  /models
    db_models.py       # SQLAlchemy ORM models (File, Chunk, User, etc.)

/scripts
  init_weaviate.py    # Sets up schema, indexes
  test_queries.py     # Quick GPT test script

/config
  config.py           # All keys and settings
  celery_worker.py    # Worker launcher

main.py               # FastAPI entrypoint

```

Frontend

```

/app
  /dashboard
  /auth
/components
/lib
/styles

```

3. AI Model Setup & Usage

- **Default Model:** GPT-4 Turbo
 - All plans
 - Used for summaries, Q&A, search

- **Optional Upgrade:** Claude 3.5 Sonnet

- Ultimate users only
- Toggle model in settings

```
const llm = user.model === "claude"
  ? new ChatAnthropic({ model: "claude-3-sonnet" })
  : new ChatOpenAI({ model: "gpt-4-turbo" });
```

4. Teams + Pricing Model

Plan	Price (INR/user/mo)	Teams	Claude Access	Summary Limit	Viewer Role
Free	₹0	No	✗	Limited	✗
Pro	₹999	Yes	✗	Fair usage	✓
Ultimate	₹1799	Yes	✓	Higher	✓
Enterprise	₹2999+	Unlimited	✓	Custom	✓

- Per-seat pricing (like Notion)
- Roles: Viewer, Editor
- Fair usage: ~5K tokens/day or 100K/month (adjustable)

5. MVP Feature Rollout Plan

✓ Phase 1: Upload + Ask (Weeks 1–3)

- Upload PDF/DOCX/PPTX/XLSX
- Parse, chunk, embed
- Ask questions with GPT-4 Turbo

✓ Phase 2: Summary + Multiple Files (Week 4–5)

- Generate summaries
- Ask across multiple files
- Save chats, tag/folder organization

✓ Phase 3: SaaS Platform (Weeks 6–7)

- Signup/Login (Google OAuth)
- Billing (Stripe/Razorpay)
- File manager dashboard

✓ Phase 4: Teams + Claude Support (Week 8–9)

- Workspaces, per-seat billing
- Model toggle: GPT vs Claude

- Viewer/Editor roles

6. Performance, Efficiency & Cost Controls

Optimization

Strategy

Free tier throttle	Vercel AI SDK / Wavelet for rate limits
Background jobs	Celery + Redis for parsing + embedding
Embed token caps	Tier-based limits
Smart queries	Use top 5–10 chunks only
Reuse embeddings	Only reprocess on file update
One model per call	Saves cost, avoids double billing

7. Marketing & Positioning

USP

Pitch

All file types	"Chat with any doc, instantly."
Team-ready	Share, and collaborate on documents
Claude + GPT	Choose your assistant (Ultimate+)
Developer API	Soon: build on top of AnyDocAI
Fair pricing	Free to start. Upgrade when needed.

8. Monetization Plan

Plan	Price	Claude?	Uploads	Daily GPT Limit	API?	Notes
Free	₹0	✗	10/day	~2K tokens/day	✗	Rate-limited
Pro	₹999	✗	Unlimited	~5K tokens/day	✗	Priority queue
Ultimate	₹1799	✓	Unlimited	~15K tokens/day	✓	Claude + API
Enterprise	₹2999+	✓	Unlimited	Custom	✓	SLA, white-labeling

9. Hybrid Supabase + MongoDB Setup (via Drizzle ORM)

Layer	Tech Used	Why
Auth	Supabase Auth	Google login, JWT, OAuth
Storage	Wasabi Storage	PDF, DOCX uploads (S3-compatible)
Billing	Supabase Postgres + Drizzle ORM	Team plans, quotas, metadata
Chat Data	MongoDB	Flexible nested structure
Embeddings	Weaviate	Chunk storage + vector retrieval
Metadata	MongoDB	File status, summaries, logs,

Workflow:

1. **Login/Register** via Supabase Auth
2. **Upload** file to Wasabi Storage
3. **Metadata saved** to MongoDB
4. **Celery background job** parses, chunks, embeds

5. **Chunks stored** in Weaviate

6. **User chat** triggers GPT-4 answer using top relevant chunks

7. **Billing** handled via Stripe/Razorpay + Supabase hooks

10. Tools & Integrations

- @supabase/supabase-js – frontend SDK
 - Drizzle ORM – for typed Supabase access
 - LangChain – chaining, retrieval, routing
 - Celery + Redis – background jobs
 - Stripe / Razorpay – subscriptions + webhooks
 - LangSmith – AI trace + debug (optional)
 - Vercel AI SDK / Wavelet – rate limiting
-

Features (Phase-Wise Development Plan)

Phase 1: MVP – Single & Multi-file Q&A (2-3 weeks)

Goal: Upload unlimited files → Chat with AI

User uploads any type of doc (PDF, DOCX, XLSX, PPTX)

Files parsed and chunked in background (Celery)

Chunks embedded to Weaviate

Chatbot retrieves relevant context and sends to GPT-4-Turbo

UI (optional) or use Swagger/Postman initially

Metadata: file title, type, upload date, user ID

Backend Ready for Unlimited Files & Pages

Phase 2: Summarization + Multi-file Intelligence (Week 4-5)

Goal: Smartly summarize & analyze multiple related files

Summarize per file, per folder, or entire upload history

“Find insights across my last 5 files” feature

Group files by tag/project/folder

Save chat history with references to files/chunks used

Phase 3: User Dashboard + SaaS Setup (Week 6-7)

Goal: Build a real SaaS platform

Auth (email or Google)

File manager (delete, rename, tag)

Token usage tracking per user

Stripe or Razorpay integration

Plans + API quota system (via environment config)

Usage dashboard

Phase 4: Team Collaboration + Chat Memory (Week 8-10)

Goal: Make it a long-term doc companion

Shared workspaces or folders

Multi-user chat with shared files

Long-term memory (store AI conversations, re-use context)

Audit logs for compliance (for B2B angle)

Scalability Strategy (Backend-First)

Use chunking + retrieval to avoid re-calling GPT for all data

Async workers = fast ingestion no matter size

GPT used only after most relevant context is known

Auto queue load handling (retry, fail-safe)

You can host 10,000+ files per user without trouble if embedding & retrieval are optimized.

Other Revenue Streams

Usage-Based Pricing: 1M token = ₹30

And we will use cursor's built-in sync