# DAY -1

**🔧 What is uv?**

uv is a next-gen Python package manager from Astral. It’s:

* ⚡️ **Ultra fast** (written in Rust)
* ✅ Compatible with requirements.txt or pyproject.toml
* 🧼 Replaces pip, venv, pip-tools, virtualenv, poetry, etc.
* 💻 Used internally at OpenAI

**📦 What is pyproject.toml?**

pyproject.toml is the **modern standard configuration file** for Python projects.  
Think of it like the package.json of Python.

It defines:

* Your **project metadata** (name, version, description)
* Your **dependencies**
* Your **build system**
* Your **tooling config** (e.g., for uv, black, ruff, pytest, etc.)

Here’s a clean summary of everything we’ve done so far — you can paste this directly into your notes:

**✅ Step 2: Production-Grade Python Environment Setup (with uv and pyproject.toml)**

**1. 📦 Installed uv**

Used uv as our modern, fast dependency and virtualenv manager.

Install command:

curl -Ls https://astral.sh/uv/install.sh | sh

**2. 🧪 Created .venv using uv**

Created a virtual environment **inside the project folder** (production-grade approach):

uv venv

source .venv/bin/activate

**3. 🛠️ Created pyproject.toml**

Defined all runtime and dev dependencies in one clean config file.

**Runtime dependencies:**

* fastapi
* uvicorn[standard]
* sounddevice
* numpy
* pydantic

**Dev tools:**

* black → autoformatter
* ruff → linter/static checker
* pytest → testing framework

[project]

name = "scribe-ai"

version = "0.1.0"

description = "Real-time voice-to-note AI scribe"

readme = "README.md"

requires-python = ">=3.9"

dependencies = [

"fastapi",

"uvicorn[standard]",

"sounddevice",

"numpy",

"pydantic"

]

[project.optional-dependencies]

dev = [

"black",

"ruff",

"pytest"

]

[tool.uv]

virtualenvs.in-project = true

[tool.black]

line-length = 88

[tool.ruff]

line-length = 88

target-version = "py39"

select = ["E", "F", "I"]

[tool.pytest.ini\_options]

addopts = "-ra -q"

testpaths = ["tests"]

**4. ⚙️ tool.\* in pyproject.toml**

These are **config sections** for dev tools:

* tool.black → sets line length for autoformatting
* tool.ruff → sets linter rules for unused imports, formatting, etc.
* tool.pytest → controls pytest behavior
* tool.uv → makes sure .venv/ is local to the project

**5. 📥 Installed all dependencies (main + dev tools)**

uv pip install .[dev]

This installs everything:

* Runtime dependencies from [project.dependencies]
* Dev tools from [project.optional-dependencies].dev

A **WebSocket** is a special communication protocol that allows **real-time, two-way interaction** between a **client** (like a web browser or mobile app) and a **server**.

Unlike normal HTTP requests (which are one-way: client asks, server replies), WebSockets keep the connection **open**, so **both** sides can send data anytime — perfect for **live updates** like speech apps, chat apps, games, or notifications.

🧠 Think of it like a **phone call**, not sending letters:

* HTTP = letter-based communication (slow, one-way)
* WebSocket = live phone call (fast, two-way)

In a **speech app**, you need real-time interaction. Here’s why HTTP isn't enough:

* **HTTP**: Every time you want to send data (like a chunk of audio), you must open a new connection, send it, and close it. That creates **delays** and **overhead**.
* **WebSocket**: You open **one connection**, and then you can stream data **continuously** in both directions — **perfect for sending live audio and getting live transcripts** back without waiting.

💡 For example:

* Your app sends audio chunks as the user speaks (WebSocket → Server).
* Server replies with transcribed text instantly (Server → WebSocket).

This **back-and-forth flow** is what makes speech feel live

Your **FastAPI app will act as a WebSocket server**. That means:

* You define a **special route** like /ws/audio-stream
* But unlike REST routes (@app.get, @app.post), you use @app.websocket
* This route does **not** expect a request → response flow — it stays open

📌 Inside this WebSocket route:

* You’ll **wait for audio chunks** from the client (await websocket.receive\_bytes() or .receive\_text()).
* You **process** each chunk (e.g., speech-to-text using Whisper/OpenAI).
* You **send back transcription** using await websocket.send\_text().

🔹 Step 6: What is WebRTC? (And how is it different from WebSocket?)

WebRTC (Web Real-Time Communication) is a protocol for direct peer-to-peer communication between browsers or devices. It's used for:

🔊 Voice calls (e.g., Google Meet, Zoom)

🎥 Video streaming

📁 File sharing (without server)

🧠 WebRTC is optimized for media (audio/video) and includes:

Echo cancellation

Latency control

NAT traversal (to go through firewalls)

Encryption (by default)

**At the simplest level, a mic is just a hardware device that listens to sound and converts it into a digital signal your computer or microcontroller can understand.**

So we need to understand the hardware setup better and also start coding to use to laptop mike as a starting point

The task is nice since we need a device that will capture audio.

**So we need a speaker phone**

<https://www.amazon.in/Raspberry-Model-LAN-mHDMI-Supply-MicroSD/dp/B0C7N5DJC1?ref_=v_sp_product_dpx&th=1>

so I have ordered this I think this would be cool to play around and learn , so that we do and cook amazing stuff …intersting times ahead

Note – A design choice has been made for now we need to learn rasbeery pie first and then we need to play the device bit of it

But we need to cook the v1 for system a doctor can open the app and listen start processing and send notes to self and patient

# DAY -2

Started with a small recap on design I mean static method and class method were used in one project just saw that , back to work now.

So feature 1 will be something like this

**🩺 Feature 1: Patient Follow-Up on Telegram (Post-Consultation)**

**🎯 Goal:**

Allow patients to **ask doubts or follow-up questions** on Telegram after their consultation, and **get instant, personalized replies** — powered by your system's backend intelligence.

**🧩 System Flow:**

1. **Session Completed by Doctor**
   * The doctor records and completes a session using your app.
   * Your backend transcribes the audio and stores it with patient metadata (e.g., name, phone number, session ID, summary).
2. **Patient Gets Summary**
   * A message is sent to the patient on Telegram via your bot with:
     + The summary of the consultation.
     + Advice, prescriptions, red flags.
     + A note: “You can ask your follow-up questions here anytime.”
3. **Patient Sends a Message**
   * The patient types a question on Telegram, like:
     + “Can I take this medicine after food?”
     + “I still have a sore throat, should I be concerned?”
4. **Your Backend Handles It**
   * Identifies patient using Telegram ID or linked phone number.
   * Retrieves the session summary + prescription + red flags.
   * Feeds all of that + patient’s question into an LLM API.
   * Receives a smart, patient-friendly answer.
5. **Bot Replies Back**
   * Sends the response to the patient on Telegram.
   * Ends with a safety disclaimer (e.g., “This is an AI response. Please call your doctor if symptoms worsen.”)

**🏗️ 1. High-Level System Design Overview**

**🔊 Input**

* Mobile app or web client (using WebRTC or WebSocket) for capturing audio from doctor (or both doctor & patient).
* Audio streamed to backend **in real-time**.

**⚙️ Backend Core (FastAPI)**

* Audio ingestion via **WebSocket**.
* Realtime transcription (e.g. **OpenAI Whisper** or **Deepgram**).
* LLM summary + prescription generation.
* Message delivery via **Telegram API** or WhatsApp Business API.

**☁️ Infrastructure**

* Host on **Azure Cloud** with:
  + Azure App Service or AKS for backend
  + Blob Storage for audio
  + CosmosDB / MongoDB for session data
  + Azure Queue for async processing
  + Redis for temporary memory/cache

**📶 2. Streaming Audio Design**

| **Component** | **Tech** |
| --- | --- |
| Audio Source | Browser/Mobile Mic |
| Stream Protocol | WebSocket (or WebRTC) |
| Format | audio/webm or wav chunks |
| Transcription | OpenAI Whisper API or Deepgram Streaming API |
| Buffering | Circular buffer in Python / Redis |
| Resilience | Auto-reconnect, retries, chunk validation |

**🧠 3. AI/LLM Processing**

1. **Transcription**
   * POST /transcribe auto-triggered after stream ends.
   * Or done in real-time for streaming APIs like Deepgram.
2. **LLM Summary**
   * Context-aware summarization using OpenAI (GPT-4-turbo).
   * Use custom system prompts (e.g., SOAP-style outputs).
3. **Prescription Generator**
   * Separate LLM prompt/chain to generate:
     + Medicines
     + Dosage
     + Follow-ups
     + Investigations

**📦 4. Scalable Architecture (Cloud-native)**

* **App Server**: FastAPI behind **Gunicorn** + **Uvicorn workers**.
* **Message Queue**: Azure Queue or RabbitMQ for:
  + LLM jobs
  + Transcription
  + Messaging jobs
* **Worker Pool**: Celery workers or FastAPI background tasks.
* **Storage**:
  + Audio files → Azure Blob Storage
  + Summaries → MongoDB or CosmosDB
* **Cache**: Redis (for user session / intermediate transcripts)

**📲 5. Messaging + Telegram Integration**

* One-time bot setup per doctor.
* Transcripts sent via Telegram using sendMessage API.
* Patients can reply → webhook hits POST /telegram/receive
* Use RAG or memory DB to answer follow-ups.

**🧪 6. Testing & Monitoring**

* **Logging**: Use structlog or loguru for production logs.
* **Monitoring**: Azure Monitor or Grafana.
* **Rate Limiting & Auth**:
  + Use API Keys / JWT for doctor access.
  + Rate limit audio ingestion endpoints.

**✅ 7. Security**

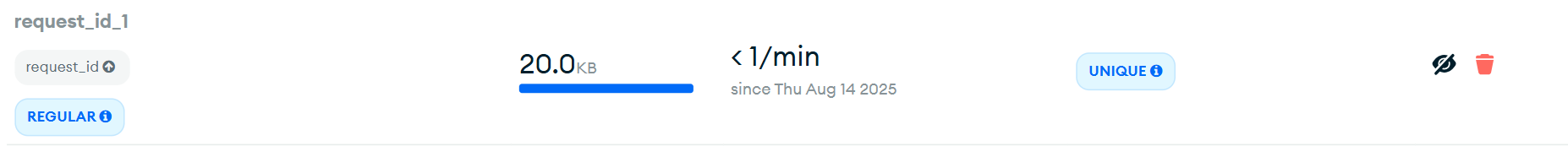
* Use HTTPS/WSS for all communication.
* Audio data encrypted in transit and at rest.
* OTP-based access for doctors via phone or email.
* Patient data stored in compliance with basic HIPAA/GDPR principles.

**⏳ 8. Scalability Notes**

| **Feature** | **Scalable With** |
| --- | --- |
| Audio | Chunked upload, queue buffering |
| LLM Processing | Async via Celery / Queue |
| Messaging | Rate-limited, async |
| Frontend | SSR React/Next.js or Streamlit-lite |
| Logging | Logstash / Azure Monitor |

**Todo: we need this product on whats app I do not know how get a whats app business number register and see how u can cook it**

**So the prime concern is user push message we need to do using what’s app.**



**What we did**

1. Clear schemas (Pydantic v2)

* SessionCreateRequest (input), SessionDB (storage), SessionResponse (output)
* Used model\_dump / model\_validate everywhere (v2 style)

1. Strict fields

* SessionStatus enum: active | closed | archived (no free-form strings)

1. Correct timestamps

* utc\_now() returns tz-aware UTC
* Stored created\_at and updated\_at in BSON datetime

1. Clean inserts & reads

* Insert with exclude\_none=True (no null spam)
* Mongo reads exclude "\_id" so Pydantic doesn’t choke

1. Idempotent create

* Client sends request\_id (UUID) in body
* Code checks for existing request\_id and returns existing session
* request\_id stored in SessionDB

1. DB guarantee

* Unique index on request\_id (sparse, unique) → race-safe idempotency
* Unique index on session\_id for fast lookups

1. HTTP correctness

* POST /v1/sessions/create returns 201 Created
* Sets Location: /v1/sessions/{session\_id}

1. Robust handlers

* Centralized log\_error on failures
* Health check route

Why this is important (quality + longevity)

* Data integrity: Enums, UTC datetimes, and clean dumps stop silent corruption.
* Consistency across services: Typed models + DB constraints mean any producer/consumer sees the same shape.
* Idempotency = real-world reliability: Retries (mobile flaky networks, timeouts) don’t create duplicates.
* Race-condition safety: Unique index turns logic guarantees into DB guarantees.
* **Maintainability: Separate request/DB/response models keep concerns clean and future changes localized.**
* **Observability & auditing: updated\_at, consistent IDs, and structured errors make issues traceable.**
* **Standards compliance: 201 Created + Location plays nicely with clients, SDKs, and tooling.**

**Quick checklist (you’re done ✅)**

* **Pydantic v2 used correctly**
* **Enum status**
* **UTC created\_at/updated\_at**
* **exclude\_none=True on insert**
* **Exclude "\_id" on read**
* **Idempotency via request\_id**
* **Unique sparse index on request\_id**
* **201 Created + Location header**
* **Error logging + health route**

# DAY -3

1.Enforced schema validation at schema layer also

2.so for now skiping transtions check and ttl implemenation at database level,

3.so design will be ->

**Doctor Session: Doctor creates session → Records audio → Real-time transcription → Audio stored in Blob → Transcript chunked → Embeddings created →**

**Chunks stored in MongoDB with vectors → Session completed → Patient notified via WhatsApp Patient Query: Patient asks question via WhatsApp → Query embedding created → Vector search in MongoDB (filter by session\_id) → Relevant transcript chunks retrieved → LLM generates response with context → Reply sent to WhatsApp**

**sessions: {session metadata}**

**transcript\_chunks: {**

**session\_id,**

**patient\_whatsapp\_number, // for filtering**

**chunk\_text,**

**embedding: [vector],**

**timestamp,**

**chunk\_index**

**}**

**conversations: {whatsapp message history}**

**🎯 FEATURE OVERVIEW**

**Core Feature: Real-time audio recording with live transcription during doctor consultations**

**User Journey: Doctor opens existing session → Clicks "Start Recording" → Speaks during consultation → Sees live transcript appear → Clicks "Stop Recording" → Session automatically completed with full transcript stored**

**✨ KEY FEATURES**

**Feature 1: Recording Session Control**

* **Start Recording - Initiate recording for a specific session**
* **Stop Recording - End recording and finalize transcript**
* **Recording Status - Check if session is currently recording**
* **Session Validation - Ensure only one recording per session**
* **Error Recovery - Handle recording failures gracefully**

**Feature 2: Real-Time Audio Streaming**

* **Live Audio Transmission - Stream audio from mobile to server in real-time**
* **Audio Chunking - Process audio in small segments for responsiveness**
* **Connection Management - Maintain stable WebSocket connections**
* **Auto-Reconnection - Recover from network interruptions automatically**
* **Buffer Management - Handle audio data efficiently without memory issues**

**Feature 3: Live Transcription**

* **Real-Time Processing - Convert audio to text as doctor speaks**
* **Partial Updates - Show transcript updates immediately on mobile screen**
* **Final Correction - Replace rough live transcript with accurate final version**
* **Multi-Language Support - Handle different languages if needed**
* **Quality Optimization - Balance speed vs accuracy for medical context**

**Feature 4: Audio Storage & Management**

* **Secure Storage - Save audio files to encrypted cloud storage**
* **File Organization - Organize audio by session with proper naming**
* **Metadata Tracking - Store duration, format, and quality information**
* **Access Control - Ensure only authorized users can access recordings**
* **Cleanup Policies - Automatic deletion of old audio files per compliance**

**Feature 5: Session Integration**

* **Status Updates - Update session status throughout recording lifecycle**
* **Transcript Storage - Save final transcript to session record**
* **Timeline Tracking - Record start/stop times and duration**
* **Data Integrity - Ensure all session data remains consistent**
* **Completion Workflow - Automatically mark session as complete when recording ends**

**🔄 FEATURE WORKFLOW**

**Pre-Recording Phase**

1. **Doctor selects existing session from list**
2. **System validates session is ready for recording**
3. **Mobile app requests microphone permissions**
4. **Server prepares recording infrastructure**

**Recording Initiation**

1. **Doctor taps "Start Recording" button**
2. **System updates session status to "recording"**
3. **WebSocket connection established for audio streaming**
4. **Mobile begins capturing and streaming audio**
5. **Real-time transcription pipeline activated**

**Active Recording Phase**

1. **Audio continuously streams from mobile to server**
2. **Live transcription appears on doctor's screen in real-time**
3. **Audio chunks saved incrementally to cloud storage**
4. **System monitors connection health and quality**
5. **Error recovery mechanisms active for interruptions**

**Recording Completion**

1. **Doctor taps "Stop Recording" button**
2. **Final audio processing and complete file assembly**
3. **High-quality final transcription generated**
4. **Session status updated to "completed"**
5. **All recording data properly stored and indexed**

**Post-Recording**

1. **Complete transcript available in session record**
2. **Audio file securely stored with proper metadata**
3. **Session ready for patient notification workflow**
4. **Recording data prepared for future RAG queries**

# DAY -4

So we have started working on transcription engine, we need to do some planning in order to see what happens and how.

 MongoDB → session + metadata + transcription text.

 Blob Storage → raw audio chunks.

 WebSocket → keeps the pipe open for real-time flow

# DAY -5

Issues Identified in Web socket implementation-

**Biggest gaps / risks (and quick fixes)**

1. **Blocking WS while transcribing.**  
   receive() → transcribe → respond means the client can’t send the next chunk until the previous finishes. Move transcription to a background worker (Celery/Dramatiq) and ACK storage immediately; push transcript\_update later.

**My comments -**

**Yes u r right.**

We will use asyncio.create\_task() so that it runs in background and we can keep receiving chunks of audio

* Concurrent processing = no guaranteed order

So we need to create a buffer and make sure we send eveything in an particular order.

Even this way u won’t be able to move ahead and do stuff since u need to understand 100 users ,what will happen u will have a bottle neck both cpu and rate limits as well.

So the solution is using async with background tasks , ensuring limits are being setup with Semaphore

Enusring we have a response buffer

* non-blocking WebSocket

1. **Sequence handling & resumes.**  
   You track expected\_sequence in memory and reset to 0 on reconnect; DB already knows what you’ve stored. On WS connect, **compute next expected** from audio\_chunks and reject duplicates/out-of-order with an explicit error. Also enforce that the **binary** frame’s sequence matches the most recent metadata.
2. **Race conditions on transcript append.**  
   update\_session\_transcript reads → concatenates → writes; parallel chunks can interleave. Prefer **per-chunk transcripts** (array by sequence) and build the full transcript on read/finalize, or use a single atomic update (e.g., aggregation-pipeline update) with sequence guards.
3. **Status transitions not fully used.**  
   You define ending but never set it; /end jumps straight to completed. Consider ending when the doctor clicks end, finalize transcripts, then completed. Also make /end **idempotent** (200 if already completed).
4. **Auth & abuse controls.**  
   WS accepts any caller with a session id. Add **JWT / API-key** auth, origin checks, size/timeouts, and per-session rate limits. Enforce max\_chunk\_size\_bytes you advertise in ConnectionConfirmed.
5. **Model mismatch in comments.**  
   Comment says “gpt-4o-transcribe” but code calls whisper-1. Align comments, or parameterize the model name.

**Fixed**

1. **Health check heaviness.**  
   models.list() and list\_containers() can be slow/privileged; cache or gate with an env flag, and avoid doing this on hot paths or tight probes.
2. **modified\_count checks can be misleading.**  
   If fields already had those values, modified\_count is 0 even though the write matched. Consider checking matched\_count (and errors) instead.
3. **Content type and container setup.**  
   You upload .webm with audio/webm; if you rely on automatic codec inference (Opus), document it and ensure the mobile side always sends Opus. Also ensure the container audio-chunks exists at boot.