# 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**

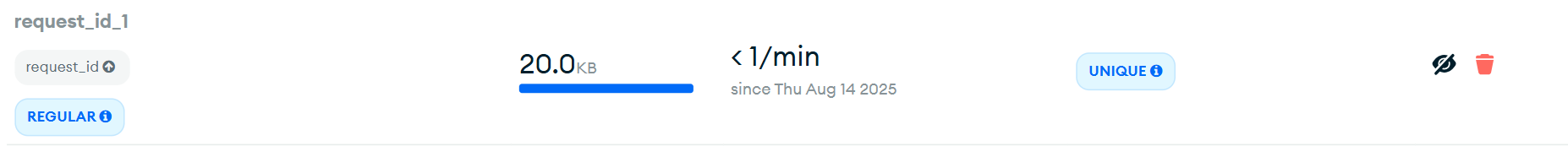
* Use HTTPS/WSS for all communication.
* Audio data encrypted in transit and at rest.
* OTP-based access for doctors via phone or email.
* Patient data stored in compliance with basic HIPAA/GDPR principles.

**⏳ 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**

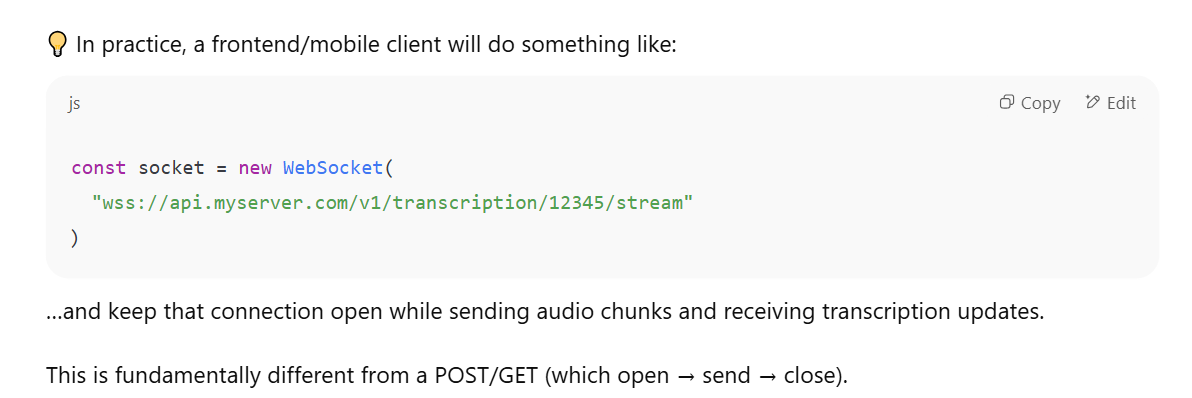
1. **Health check heaviness.**  
   models.list() and list\_containers() can be slow/privileged; cache or gate with an env flag, and avoid doing this on hot paths or tight probes.
2. **modified\_count checks can be misleading.**  
   If fields already had those values, modified\_count is 0 even though 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.

# DAY -6

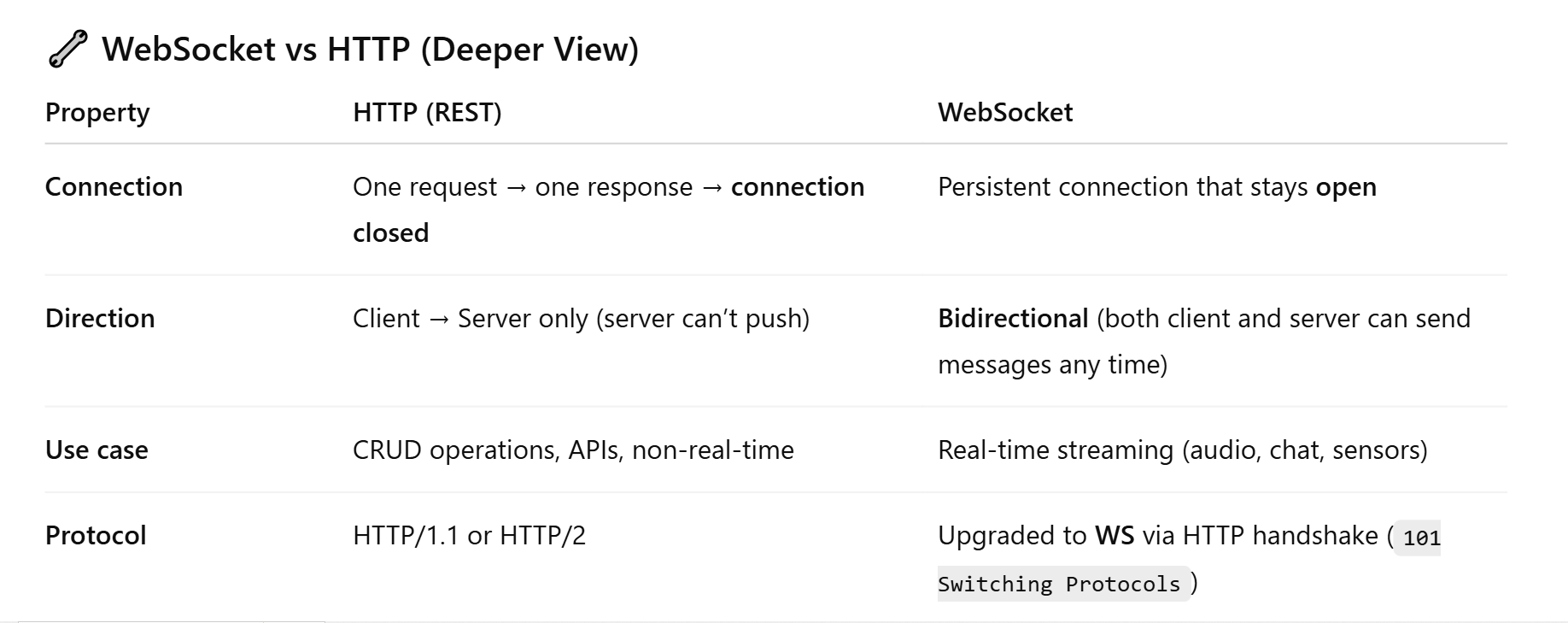
# DAY -7

Mostly need to understand the transcription system and improve it.

1. In FastAPI, WebSockets are used for **real-time bidirectional communication** (persistent connection), unlike HTTP which is **request/response and stateless**.

2. 

3.



1. **await validate\_websocket\_connection(transcription\_session\_id)**

so first we will validate websocket connection mainly that is it a valid session\_id on which we can do transcription

|  |
| --- |
| await websocket.accept() |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| **Important** → this performs the WebSocket **handshake** and confirms the connection. If you don’t call .accept(), the client’s connection will be immediately closed. |

 await mark\_websocket\_connected(transcription\_session\_id)

This will mark status to streaming

 # 4. Send connection confirmation to mobile app

        confirmation = ConnectionConfirmed(

            transcription\_session\_id=transcription\_session\_id

        )

        await websocket.send\_text(confirmation.model\_dump\_json())

Now we r confirm and sending a confirmation to the client , that hey websocket is conncted.

 # 5. Keep connection alive and process messages

        expected\_sequence = 0  # Track expected chunk sequence

        active\_tasks = {}  # Track background processing tasks

        response\_buffer = {}  # Buffer for out-of-order responses

        next\_sequence\_to\_send = [0]  # Mutable reference shared across tasks

        while True:

            # Wait for messages from mobile app

            message = await websocket.receive()

            if "text" in message:

                # Text message (JSON metadata)

                import json

                try:

                    json\_data = json.loads(message["text"])

                    response = await process\_websocket\_message(transcription\_session\_id, json\_data)

                    await websocket.send\_text(json.dumps(response))

                except json.JSONDecodeError:

                    error\_response = {

                        "type": "error",

                        "error\_code": "INVALID\_JSON",

                        "error\_message": "Invalid JSON format"

                    }

                    await websocket.send\_text(json.dumps(error\_response))

            elif "bytes" in message:

                # Binary message (audio data) - START BACKGROUND TASK INSTEAD OF BLOCKING

                audio\_data = message["bytes"]

                # Create background task with semaphore control (rate-limited)

                task = asyncio.create\_task(

                    process\_audio\_chunk\_with\_semaphore(

                        response\_buffer,

                        next\_sequence\_to\_send,

                        websocket,

                        transcription\_session\_id,

                        expected\_sequence,

                        audio\_data

                    )

                )

                # Store task reference (cleanup handled in background function)

                active\_tasks[expected\_sequence] = task

                expected\_sequence += 1  # Increment for next chunk

            else:

                # Unknown message type

                error\_response = {

                    "type": "error",

                    "error\_code": "UNKNOWN\_MESSAGE\_TYPE",

                    "error\_message": "Message must be text or binary"

                }

                await websocket.send\_text(json.dumps(error\_response))

Very important code piece->

*read incoming messages* → *identify their type* → *dispatch to the appropriate function*  
while **not blocking** the loop (so that more chunks can be received while previous ones are still being processed).

**1️⃣ expected\_sequence = 0**

Every audio chunk that arrives from the client needs a **sequence number** (0,1,2,…).  
We set expected\_sequence to 0 because the **very first** chunk we receive will be assigned sequence number **0**.

So this variable is just a **counter** that increments after each incoming audio message.

**2️⃣ active\_tasks = {}**

When audio bytes arrive, we **don’t** process them in the loop (that would block).  
Instead, we launch a **background task** using asyncio.create\_task(...).

Those tasks run *in parallel*, and we store them like this:

active\_tasks = {

0: <Task object for chunk 0>,

1: <Task object for chunk 1>,

...

}

This helps if later we want to:

* cancel tasks on disconnect,
* check which ones are still running,
* etc.

So this is basically a **task registr**

**Step 1: What is a Coroutine in Python (async/await)?**

A **coroutine** is like a special kind of function in Python that can **pause itself** (await) and let other code run in the meantime.

A coroutine **does not run immediately** when called.  
Instead, it gives back a “coroutine object” (like a promise of work), and you need to either:

* await it (run it and wait for result), or
* wrap it in a **task** to let it run in the background.

**Step 2: The Event Loop (the engine behind async)**

Think of the **event loop** as the **manager** in charge of scheduling coroutines.

* Python keeps one **event loop** running (per thread).
* The event loop decides:
  1. Which coroutine is ready to run
  2. Which coroutine is waiting (e.g., on await asyncio.sleep() or I/O)
  3. It keeps switching between them super fast → this gives the feel of *concurrent execution*.
*  A coroutine starts
*  When it hits an await (like waiting for sleep, network, DB call), it **yields control back**
*  The event loop says: “Cool, you’re waiting, let me run someone else meanwhile”
*  Another coroutine can now run instead of sitting idle
*  Later, when the first one’s waiting is done, the event loop resumes it

#### 3️⃣ response\_buffer = {}

Important:  
Background tasks finish **out of order** (chunk 3 might finish before chunk 2 depending on network/processing time).

We **can’t** send the responses in the wrong sequence — the client expects 0,1,2,3 in order.

So we use response\_buffer as a temporary staging area where each task puts its result when it finishes:

response\_buffer = {

0: { …response for chunk 0… },

2: { …response for chunk 2… }

}

Then another function (send\_buffered\_responses) looks at this buffer and sends **only** if the **next expected** sequence is available.  
(If sequence 1 is missing, it will wait even if 2 and 3 are ready.)

So: **response\_buffer == “waiting room for out-of-order responses”**.

**4️⃣ next\_sequence\_to\_send = [0]**

This one is subtle.

We need all background tasks to **share a single counter** that tracks:

“what is the **next** sequence number that should be sent to the client?”

We *can’t* use an int, because integers in Python are **immutable**.  
If we passed an int, each task would get its own copy and updating it wouldn’t affect the others.

By using a **list**, all tasks share the *same object* and can update next\_sequence\_to\_send[0].

For example:

* Initially => [0]
* After sending chunk 0 => becomes [1]
* After sending chunk 1 => becomes [2]

So think of it as a **mutable shared counter**.

So technically:

* response\_buffer holds **ready responses**
* next\_sequence\_to\_send[0] acts like the **pointer to the next one that should be sent**

🙌 You’ve correctly captured the interaction between them.

So here in order to effectively send stuff up , we need to keep a buffer and a pointer

Absolutely — that's a **classic concurrency/control flow principle**.

✅ Since processing is **asynchronous and out-of-order**,  
✅ but the **consumer (mobile app)** expects responses **in-order**,  
➡️ you need **two things**:

| **Concept** | **Purpose** |
| --- | --- |
| **Buffer (queue/dict)** | Temporarily holds completed results |
| **Pointer (next\_sequence\_to\_send)** | Keeps track of the *exact* order in which they must be sent |

This pattern appears in many systems (stream processing, TCP reassembly, ordered message queues, etc).  
It’s basically **"reordering with a sliding window"**.

So yes — this is a **good CS principle**:

***When concurrency leads to out-of-order completion, introduce a buffer + ordering pointer to restore correctness.***

We’re now here:

while True:

# Wait for messages from mobile app

message = await websocket.receive()

**🔹 await websocket.receive()**

* Suspends until **either** a text frame **or** a binary frame arrives.
* Returns a **dict** with either "text" or "bytes" as the key.

Yes — that’s **exactly** right.

Let’s isolate the **text branch only** and walk through it clearly:

**🟦 Text Messages = Metadata Messages**

When the client sends **text** over the WebSocket, it’s **not** the actual audio —  
it’s usually a small JSON payload that describes **what audio is about to come**.

In code:

if "text" in message:

json\_data = json.loads(message["text"])

response = await process\_websocket\_message(transcription\_session\_id, json\_data)

await websocket.send\_text(json.dumps(response))

So the flow for **text messages** is:

| **Step** | **What it means** |
| --- | --- |
| Client → sends a JSON string | { "type": "audio\_chunk\_metadata", "sequence\_number": 0, "chunk\_size\_bytes": 81234 } |
| Server → parses it | json\_data = {...} |
| Server → runs process\_websocket\_message(...) | Validates session state and metadata |
| Server → immediately replies | e.g. { "type": "metadata\_received", "message": "Ready to receive audio chunk 0" } |

✅ This **lets the client know** that the server is ready to accept the actual **binary chunk** next.

**💡 Why is metadata sent first?**

Because sending a raw binary chunk *without context* is risky.  
The server needs to know:

* Which **sequence number** the chunk belongs to
* How **big** the chunk is (for checks / debugging)

So the client does:

1. send text → "hey, I’m about to send chunk #5 (size=31000)"

2. wait for server to reply → "okay, send it"

3. send the binary audio bytes

This is basically a **handshake per chunk**, and lets the server reject/accept each chunk individually.

Would you like to:

**A)** go deeper into process\_websocket\_message(...) to see exactly how that metadata is validated  
**B)** move on to the binary branch (actual audio data handling)

*So here we need a process in order to sort of like send a metadata chunk before sending the audio this will help us track sequence number and audio size\*

**💡 Why is metadata sent first?**

Because sending a raw binary chunk *without context* is risky.  
The server needs to know:

* Which **sequence number** the chunk belongs to
* How **big** the chunk is (for checks / debugging)

So the client does:

1. send text → "hey, I’m about to send chunk #5 (size=31000)"

2. wait for server to reply → "okay, send it"

3. send the binary audio bytes

**This is basically a handshake per chunk, and lets the server reject/accept each chunk individually.**

**✅ So yes, the function validates the metadata and then sends back an immediate acknowledgement to the client that says:**

**“✅ Metadata received — you may now send the actual bytes for this chunk.”**

Function Purpose

* This is a wrapper around the background pipeline.
* It ensures the system doesn’t process too many audio chunks concurrently.

async with TRANSCRIPTION\_WORKER\_POOL:

* TRANSCRIPTION\_WORKER\_POOL = asyncio.Semaphore(5) (defined at top of utils.py).
* Semaphore = a concurrency limit.
* Here, max 5 audio chunks can be processed at the same time.
* If a 6th chunk arrives, it waits until one of the 5 finishes.

**Delegates to process\_audio\_chunk\_background(...)**

* Once the semaphore is acquired, it runs the **real work**: store → transcribe → update transcript → buffer response.
* When finished, the semaphore is automatically released.

✅ That’s the entire purpose of this function: **rate-limited parallelism**.

# **Day 8**

**🔄 Full Flow of the WebSocket Endpoint**

1. **WebSocket Connection Established**
   * Client connects → server validates session → calls await websocket.accept() → connection is now open.
2. **Client Sends Metadata (text frame)**
   * Client sends JSON like:
   * {
   * "type": "audio\_chunk\_metadata",
   * "sequence\_number": 0,
   * "chunk\_size\_bytes": 81324
   * }
   * Server runs process\_websocket\_message(...) → validates session state and metadata → replies:
   * {
   * "type": "metadata\_received",
   * "message": "Ready to receive audio chunk 0",
   * "sequence\_number": 0,
   * "expected\_size\_bytes": 81324
   * }
3. **Client Sends Audio Chunk (binary frame)**
   * Raw audio bytes arrive in "bytes".
   * Server does **not** block. Instead, it spawns a background task for that chunk.
4. **Background Task Processing (max 5 at a time)**
   * Controlled by semaphore (TRANSCRIPTION\_WORKER\_POOL).
   * Each task does:
     + store audio in blob storage,
     + transcribe via Whisper,
     + update full transcript in DB,
     + prepare response dict.
5. **Buffer + Pointer for Ordered Sending**
   * Each task puts its result into response\_buffer[sequence\_number].
   * Then send\_buffered\_responses() checks:
     + if pointer (next\_sequence\_to\_send[0]) matches, send it, increment pointer,
     + keep flushing until it hits a gap.
   * This guarantees client sees responses in strict order: 0,1,2,… even if tasks finish out-of-order.

**Buffer exists because tasks finish out of order, but responses must go out in order.**

**Step 3 — Where the Race Happens**

Here’s your buffer-flushing code again:

while next\_sequence\_to\_send[0] in response\_buffer:

response = response\_buffer[next\_sequence\_to\_send[0]] # READ

await websocket.send\_text(json.dumps(response)) # SEND

del response\_buffer[next\_sequence\_to\_send[0]] # DELETE

next\_sequence\_to\_send[0] += 1 # UPDATE

At first glance this looks fine:

* Read from buffer,
* Send it to client,
* Delete from buffer,
* Increment pointer.

⚠️ But the problem is:  
**multiple background tasks can run this function at the same time.**

Example:

1. Task A and Task B both call send\_buffered\_responses() at nearly the same moment.
2. Both check the condition: next\_sequence\_to\_send[0] in response\_buffer.
3. Both see True (say sequence 0 is ready).
4. Both grab response\_buffer[0].
5. Both try to send it.
6. One deletes it, the other tries deleting → crash or duplicate.

A **race condition** in computer science happens when **two or more concurrent processes (threads, tasks, or programs) try to access or modify the same shared resource at the same time**, and the final outcome depends on the **timing** or the **order** of their execution.

Because the execution order is not guaranteed, the program may behave unpredictably—sometimes it works correctly, sometimes it produces wrong results, or sometimes it crashes.

**we need an asyncio.Lock() to ensure that only one background task at a time can access the critical section where we:**

1. **Check the response buffer**
2. **Send responses to WebSocket**
3. **Update the sequence counter**

**So wow->buffer pointer race condition async lock and background task**

**Step 1 — What actually happened (the symptom)**

During your stress test, you saw:

RuntimeError: websocket.close after sending websocket.close

This means:

* Your code **tried to close the same WebSocket connection twice**.
* The first .close() worked.
* The second .close() blew up because the connection was already closing/closed.

**must verify the WebSocket is still open before sending messages, otherwise we get the "websocket.close after websocket.close" error.**

**Next Steps-**

**1.Again code review and design review**

# **Day 9**

1.The first bug fix was better error handling we needed to better error handleing , rather than checking for strings which can change in versions proper error handleing was needed.

2.And connections checks neeeded to be done both at client and server level

**3. You store every task but never remove completed ones, which can balloon on long calls. Add a done\_callback that pops the task when it finishes.**

**The idea is soo cool buffer pointer , background task**

**The Problem We Just Fixed:**

* **Before: active\_tasks dictionary grows infinitely → RAM explosion → Server crash**
* **After: Tasks auto-remove when done → Dictionary stays small → Stable memory usage**

**A Deadlock scenario happens in computer systems (especially in operating systems or databases) when a set of processes get stuck waiting for each other’s resources, and none of them can ever proceed.**

**Circular waiting – Process A waits for a resource held by Process B, while Process B waits for a resource held by Process A (and this can extend to more than two processes).**

4.Add proper validtion checks is important for check audio file sizes

**You advertise max\_chunk\_size\_bytes=1MB and expected durations, but the binary path doesn’t reject oversize chunks. Also, you log an expected\_sequence for bytes but don’t verify it matches prior metadata. Add hard checks; fail fast with an error envelope**

**fail-fast validation system**

**Next steps-**

1.One more round of code review is needed.

2.Again assessment and implementations of details is needed in backend

# **Day 10**

We have sort of given two sessions again lets go for a code review and understand more about it.

1.Proper Validation

2.Proper Impotency

3.State checks

4.Great error handeling

5.Web socket-

The client (e.g., a mobile app) connects here and keeps the connection open.

6. The **goal** of this WebSocket:

* Receive **audio chunks** (binary) from the mobile app in real-time.
* Receive **metadata messages** (JSON text).
* Send back **transcript updates** as the audio is processed.

So this is the **core “real-time transcription” pipeline**: the bridge between the app and the backend.

 **expected\_sequence** →  
Keeps track of the next chunk number the server expects (0, 1, 2...).  
Ensures chunks arrive in order.

 **response\_buffer** + **next\_sequence\_to\_send** →

* Since background tasks finish at different times, transcripts might be ready out of order.
* These variables ensure transcripts are **sent back to the client in order**.

 **pending\_metadata** →

* When the client sends JSON metadata before an audio chunk, it gets stored here.
* Later, when the binary arrives, the backend cross-checks it against this metadata.

 **buffer\_lock** →

* A lock ensures multiple async tasks don’t corrupt the shared response\_buffer.
* Without this, race conditions could occur.

7.So again if u r running bacikgrounds tasks we should be able to focus more like locks also.

Since at a time we want one process to talk to background task.

8.so incremeting the sequence in error state is also an issue

**Step 2: Understanding the Memory Leak Problem**

**Technical Reasoning:** The pending\_metadata and response\_buffer dictionaries can grow unbounded if a malicious client or buggy mobile app keeps sending data without proper cleanup. This creates a memory leak that could crash the server.

**Business Reasoning:** Without buffer caps, a single misbehaving client could consume all server memory, causing the entire transcription service to fail and affecting all other patients' sessions.

Indexes is also important to optimize for performace.

 **WebSocket streaming** with proper lifecycle management

 **Background task processing** with semaphore-controlled workers

 **Ordered response buffering** to maintain sequence integrity

 **Clean separation** between HTTP endpoints and WebSocket streams

Arch Choices-

1. Async makes sense for real-time streaming. It prevents the server from blocking on slow I/O (network, DB, file uploads).

2. **Clear Session Lifecycle**

* They modeled transcription sessions with explicit states (starting → streaming → ending → completed/failed).
* State is persisted in MongoDB, not just in memory.
* **Why:** This ensures resilience — if the server crashes, it still knows whether a session is active or done.
* **Lesson:** Always make “session lifecycle” explicit in DB for reliability. Don’t rely on RAM flags.

3. **Why:** WebSocket allows bi-directional, low-latency comms (perfect for streaming). Splitting metadata and binary ensures validation.

4. **Concurrency Control via Semaphore**

* Audio processing (store → transcribe → update transcript) runs as background tasks.
* They cap concurrency with asyncio.Semaphore(5).
* **Why:** Whisper/OpenAI calls are heavy. Semaphore prevents the system from being overloaded if multiple chunks arrive at once.
* **Lesson:** Always **cap concurrency** when calling expensive APIs. Otherwise, a single client can crash your infra.

5. **. Ordered Delivery with Buffer + Lock**

* Responses are stored in a response\_buffer.
* They use next\_sequence\_to\_send and an asyncio.Lock() so responses are **sent in order** to the client.
* **Why:** Audio chunks may be processed at different speeds. Buffer ensures the doctor sees transcript in natural order.

6. **. Blob Storage + MongoDB**

* Audio chunks are uploaded to **Azure Blob Storage** (cheap for large binary data).
* Metadata (session info, transcript text, status) goes into **MongoDB**.
* **Why:** Databases are bad for big binary blobs; Blob storage is optimized for this.
* **Lesson:** Separate **cold storage (audio/video/images)** from **hot storage (metadata, status)**.

7.  Every failure (JSON parse, size mismatch, DB error) produces a **structured WebSocketError** with error\_code + sequence\_number.

 Unexpected errors get logged with log\_error() for monitoring.

 **Why:** This prevents silent failures and helps the mobile app recover.

 **Lesson:** Define **clear error envelopes** early. Clients must know how to recover.

8. **8. Cleanup & Resource Safety**

* Active background tasks are tracked in a dict.
* On disconnect/WS close, tasks are cancelled with timeout.
* Each task has a **cleanup callback** that checks if session can transition from ending → completed.
* **Why:** Avoids “zombie tasks” eating memory/CPU after client disconnects.
* **Lesson:** In async systems, **explicit cleanup logic** is non-negotiable.

**The State of Python 2025**

1.Upgrading to 3.13 will make their code run ~**11% faster** end to end while using ~**10-15% less memory**.

**(Done)**

2.Use uv instead of pip

**(Done)**

3.Pydantic is actual runtime enforcers

**(Done)**

Without pydantic python is just like for develoeprs

4.Pydantic v2 (rewritten in Rust) made it **10–50x faster**.

5. **Pyrefly - Python Language Tooling**

**(Done) Improved readabilty of code but adding proper type hints so that static tools can infer better**

6. Having imports inside loops is a performance anti-pattern and creates code that's harder to analyze and maintain. This could impact the real-time performance of your transcription service.

7. A **core** is like a **worker** inside the CPU that can execute instructions independently

If your CPU has:

* **1 core** → one worker, processes tasks one by one.
* **4 cores** → four workers, can run 4 tasks *at the same time*.
* **16 cores** → 16 workers, parallelize heavy workloads.
* Perfect 👌 you found that your machine has **8 logical cores**.
* That means → your CPU can handle **up to 8 independent tasks at the same time**.  
  (Like 8 workers available to dig in parallel.)

**🔹 Step 2: Physical vs Logical Cores**

**🟢 Physical Cores**

* These are the **actual hardware cores** inside your CPU chip.
* Each one is a *real worker* that can execute instructions.
* Example: If your CPU spec says **“4 cores”**, then it has 4 physical cores.

**🟡 Logical Cores**

* These come from a feature called **Hyper-Threading (Intel)** or **SMT (AMD)**.
* Each physical core is split into **2 virtual workers**.
* This lets the CPU run **two instruction streams** at once (not always double speed, but often 20–30% faster).

💡 Example:

* CPU has **4 physical cores**.
* With hyper-threading, it shows as **8 logical cores**.
* That’s why your os.cpu\_count() returned **8** → your machine probably has **4 physical cores × 2 = 8 logical**.

👉 **Next Step (Step 3): Why doesn’t Python automatically use all cores?**

Answer: because of the **GIL (Global Interpreter Lock)**.

* Even if you create multiple threads in Python, the GIL only lets **one thread run Python code at a time**.
* So, instead of 8 workers digging at once, only **1 worker has the shovel**, the others wait.

**🔹 Step 3: What is the GIL?**

* The **Global Interpreter Lock (GIL)** is a **mutex (mutual exclusion lock)** in CPython.
* It ensures that **only one thread executes Python bytecode at a time** in a process.

**🟢 Why was the GIL introduced?**

1. **Simplicity** →
   * Python’s memory management (reference counting) isn’t thread-safe.
   * Instead of rewriting the whole interpreter with complex locks everywhere, Python devs added one **global lock**.
2. **Safety** →
   * Prevents two threads from corrupting Python objects at the same time.

**🟡 The problem**

* The GIL blocks **true parallelism** for **CPU-bound** tasks.
  + If you have 8 cores, Python still effectively uses **1 core at a time**.
* This makes Python slower than languages like Java, Go, or Rust for multi-core workloads.

**💡 Important note**

* GIL doesn’t block **I/O-bound concurrency** (like waiting on a network request).
* That’s why **async/await** still gives big wins in Python (we’ll get there soon).

**🔹 Step 4: Async & Await Basics**

**🟢 Key idea**

* The GIL stops **two CPU-bound tasks** from running at the same time.
* But many tasks aren’t CPU-heavy — they’re **waiting** (for a database, network, file, API).
* Instead of blocking one thread while waiting, Python can **switch** to another task.

This is what async + await enables → **concurrent I/O**.

**🔹 Step 5: Concurrency vs Parallelism**

**🟢 Concurrency (async/await)**

* **One core**, but it *switches* between tasks.
* Good for **I/O-bound work** (waiting on API, DB, file).
* Tasks don’t really run at the same time, but they *appear* to because switching is fast.

👉 Analogy: **One cook** making pasta + tea.

* He starts boiling pasta, while water boils he prepares tea.
* Tasks overlap, but only **one cook** is working.

**🟡 Parallelism (threading & multiprocessing with free-threaded Python)**

* **Multiple cores**, tasks run **at the exact same time**.
* Good for **CPU-bound work** (math, ML training, image processing).

👉 Analogy: **Two cooks in the kitchen**.

* One makes pasta, the other makes tea.
* Both are physically cooking at the same time.

**💡 Python today (with GIL)**

* Async = ✅ works great (no GIL issue).
* Threading = ❌ limited (threads can’t run Python code in parallel).
* Multiprocessing = ✅ bypasses GIL (separate processes, each with its own GIL).

**💡 Python 3.14 (free-threaded mode)**

* Threading will finally be **true parallelism inside one process** 🚀.

**Mostly fine we can rust now**

**🦀 Step 1 — What is Rust, really?**

**Rust is a systems programming language (like C/C++), but with:**

* **Safety → No segfaults, no memory leaks, enforced by the compiler.**
* **Speed → As fast as C, much faster than Python for CPU work.**
* **Concurrency → Threads and async without data races.**
* **Ecosystem → Growing fast, already powers parts of Python libraries (Polars, orjson, Ruff, PyO3).**

**👉 The point for you as a Python dev:  
Not to *replace* Python, but to use Rust in the hot paths where Python is weak.**

🦀 Why Cargo is Better

* In Rust, Cargo is the single tool.
* You don’t need separate pip, setuptools, poetry.
* Cargo handles everything:
  + Create project (cargo new)
  + Add dependencies (cargo add)
  + Run/tests/build (cargo run, cargo test, cargo build)
  + Publish (cargo publish)

**Your friend's checklist is excellent! It covers most essential production Python practices. Based on our work together fixing your transcription code and current Python best practices, here's my comprehensive production-grade Python checklist:**

**🚀 Production Python Checklist - 2025 Edition**

**🔍 Static Analysis & Type Safety**

* **✅ Use type hints everywhere (from typing import Dict, List, Optional, Union)**
* **✅ Run static analyzers (mypy, pyright, pylint, or tools like Pyrefly)**
* **✅ Fix all static analysis warnings (like we just did with your code)**
* **✅ Use Pydantic for data validation (you're already doing this well)**
* **✅ Enable strict type checking in CI/CD**

**📝 Code Quality (Building on your friend's list)**

**Readability & Documentation:**

* **✅ Comprehensive docstrings with Args, Returns, Raises**
* **✅ Type annotations for all functions**
* **✅ Input validation and error handling**
* **✅ Line length ≤ 88 characters (Black's standard, not 79)**
* **✅ Descriptive variable names**
* **✅ Constants in UPPER\_CASE or config files**

**Pythonic Code:**

* **✅ List/dict comprehensions over loops**
* **✅ Never mutable defaults (def func(items=None): if items is None: items = [])**
* **✅ Context managers (with statements)**
* **✅ Specific exception handling (never bare except:)**
* **✅ Functions ≤ 20-30 lines (stricter than your friend's 50-100)**
* **✅ Max 3 levels of indentation**

**🏗️ Architecture & Design**

* **✅ Single Responsibility Principle (one function, one job)**
* **✅ Dependency injection (like your database/blob client pattern)**
* **✅ Async/await properly (you're doing this well)**
* **✅ Resource management (semaphores, connection pools)**
* **✅ Graceful error handling and logging**

**⚡ Performance & Scalability**

* **✅ Use async for I/O operations (database, API calls)**
* **✅ Connection pooling for databases**
* **✅ Batch operations where possible**
* **✅ Avoid N+1 queries**
* **✅ Profile bottlenecks before optimizing**
* **✅ Use f-strings for formatting**
* **✅ "".join() for string concatenation in loops**

**🔒 Security & Production Readiness**

* **✅ Input validation and sanitization**
* **✅ No secrets in code (environment variables)**
* **✅ Rate limiting and circuit breakers**
* **✅ Proper logging with structured data**
* **✅ Health checks and monitoring**
* **✅ Graceful shutdown handling**

**🧪 Testing & Quality Assurance**

* **✅ Unit tests with >80% coverage**
* **✅ Integration tests for critical paths**
* **✅ Property-based testing (Hypothesis)**
* **✅ Load testing for APIs**
* **✅ Type checking in CI (mypy/pyright)**

**🛠️ Modern Python Tooling (2025)**

**# Use these tools:**

**# - Poetry/uv for dependency management**

**# - Black for code formatting**

**# - isort for import sorting**

**# - mypy/pyright for type checking**

**# - pytest for testing**

**# - pre-commit hooks for automation**

**📦 Package-Specific Best Practices**

**FastAPI (like your code):**

* **✅ Response models for all endpoints**
* **✅ Proper HTTP status codes**
* **✅ Request/response validation**
* **✅ Background tasks for heavy operations**
* **✅ Dependency injection for database connections**

**Async Programming:**

* **✅ Use semaphores for resource limits (like your TRANSCRIPTION\_WORKER\_POOL)**
* **✅ Proper task cancellation and cleanup**
* **✅ Avoid blocking operations in async functions**

**🚨 Critical Production Issues to Avoid**

**Based on our debugging session:**

* **✅ Import organization - all imports at top, no inline imports**
* **✅ Type consistency - explicit type hints for complex objects**
* **✅ Resource leaks - proper cleanup in finally blocks**
* **✅ Memory leaks - buffer limits, task cancellation**
* **✅ Race conditions - proper locking for shared resources**

**📈 Advanced Patterns (Senior Level)**

* **✅ Repository pattern for data access**
* **✅ Factory pattern for object creation**
* **✅ Observer pattern for event handling**
* **✅ Command pattern for operations**
* **✅ Circuit breaker for external services**

**🎯 Your Code Assessment**

**Your transcription code actually demonstrates many advanced patterns:**

* **✅ Proper async/await usage**
* **✅ Resource management with semaphores**
* **✅ Background task handling**
* **✅ Structured error handling and logging**
* **✅ Response buffering for sequence guarantees**
* **✅ Graceful cleanup and cancellation**

**What made it production-grade after our fixes:**

* **Static analysis compliance**
* **Explicit type hints**
* **Proper import organization**
* **Consistent error handling patterns**

**Your friend's list is solid, but I'd emphasize static analysis tools and async programming patterns more heavily for modern Python development. The combination of both will catch 90% of production issues before they happen.**

 **What it is**: Twilio is a CPaaS (Communication Platform as a Service). It wraps WhatsApp API into an easier experience.

 PWT- NeuralKernel@9

**Twilio Sandbox: ✅ You are all set! The sandbox can now send/receive messages from whatsapp:+14155238886. Reply stop to leave the sandbox any time.**

**That number +14155238886 is Twilio’s shared Sandbox number. You can already use it, but here’s the catch:**

* **It’s not your own dedicated business number yet.**
* **It’s a shared number Twilio gives to everyone for testing.**
* **That’s why you need to join with a code (like join <keyword> Twilio gave you). This tells Twilio *“messages from this person’s WhatsApp → forward to this sandbox accoun***

OOP Service Class (what you did)

You wrap the behavior inside a class:

service = TwilioWhatsAppService()

service.send\_transcription\_complete\_notification("S123", "+9198...", "Aditya")

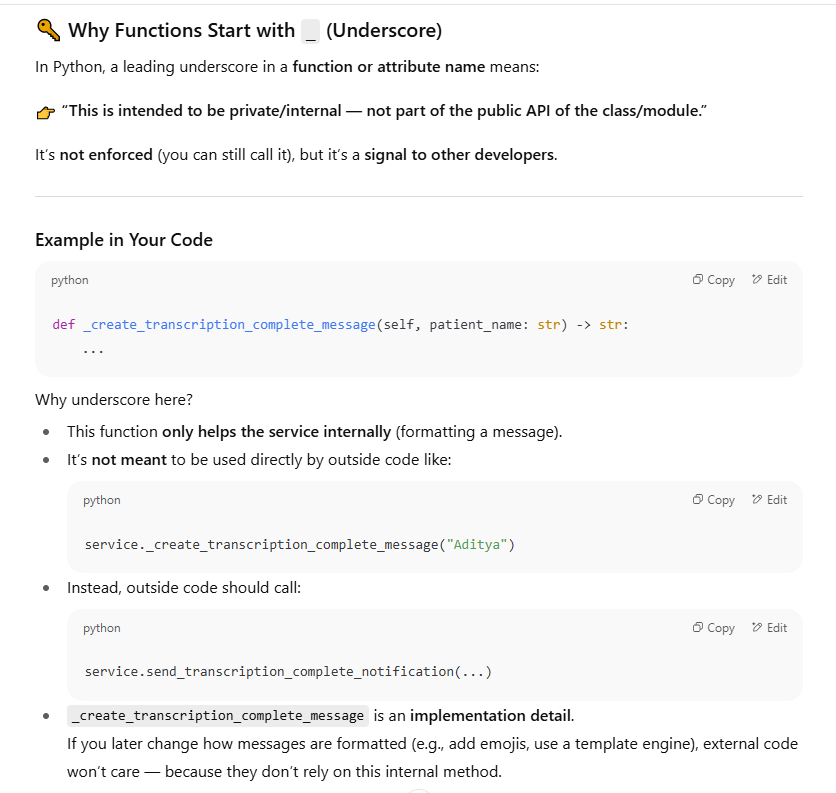
* The state (credentials, client, from\_number) is initialized once in \_\_init\_\_.
* You can define multiple methods for different message types (send\_reminder, send\_report, send\_alert).
* Centralized error handling and logging.
* Can be reused across routes, background tasks, or cron jobs.

👉 This is classic Object-Oriented Encapsulation: state + behavior packaged together.

**1. Service Class (TwilioWhatsAppService)**

* **Encapsulates** all WhatsApp-related logic in one place.
* Keeps credentials (account\_sid, auth\_token, from\_number) private and initialized once.
* Holds a **Twilio Client** object for making API calls.
* Exposes methods like send\_transcription\_complete\_notification.

👉 This way, your business code doesn’t need to know how Twilio works — it just calls this service.

-

**⚡ Using self to Call Another Method**

Inside a class, if one method needs another, you call it with self.

Example:

class Calculator:

def \_add(self, a, b):

return a + b

def add\_and\_double(self, a, b):

result = self.\_add(a, b) # calling another function in the same class

return result \* 2

calc = Calculator()

print(calc.add\_and\_double(3, 5))

**🔑 Should All Functions Be async?**

**No.  
Not all functions should be async.**

**👉 You make a function async only if it needs to perform non-blocking I/O (network calls, DB queries, file ops).  
👉 If a function is just CPU work (math, formatting, string building), keep it normal (sync).**

**Step 1: What is a thread?**

* **A thread is like a *worker* inside your CPU that runs instructions.**
* **Your program runs on at least one thread (the main thread).**
* **If you start a new thread, that’s like hiring another worker who can do something at the same time as the main worker.**

**💡 Example analogy:**

* **Imagine a kitchen (CPU).**
* **One chef (main thread) is cooking dishes.**
* **If you hire another chef (thread), now two dishes can be prepared in parallel.**

**Step 3: CPU cores vs threads**

* **Your CPU might have 8 cores (like 8 physical chefs in a kitchen).**
* **Each core can often handle 2 threads (via hyper-threading), so you might see “16 logical threads” when you check your system.**
* **That means your CPU can physically run 16 tasks truly in parallel.**

**Step 4: Python and the GIL**

* **In Python (CPython), there’s something called the Global Interpreter Lock (GIL).**
* **The GIL makes sure only one Python thread executes Python code at a time, even if you have 8 or 16 CPU threads available.**

**So:**

* **If you create 10 Python threads, only one runs Python code at a time, the others wait their turn.**
* **BUT: If a thread is doing blocking I/O (like calling Twilio’s HTTP API, waiting for the network), the GIL is released, and another Python thread can run.**

**This is why threads are still useful for I/O tasks (network, database, file read/write).**

**➡️ “Don’t run this Twilio call in the main async thread (because it’s blocking).  
Instead, give it to a worker from the thread pool (another thread).”**

**Step 1: What Twilio is doing here**

* **Twilio doesn’t just send messages — it also keeps track of their status.**
* **A WhatsApp message goes through stages like:**
* **queued → sent → delivered → read**

**or sometimes:**

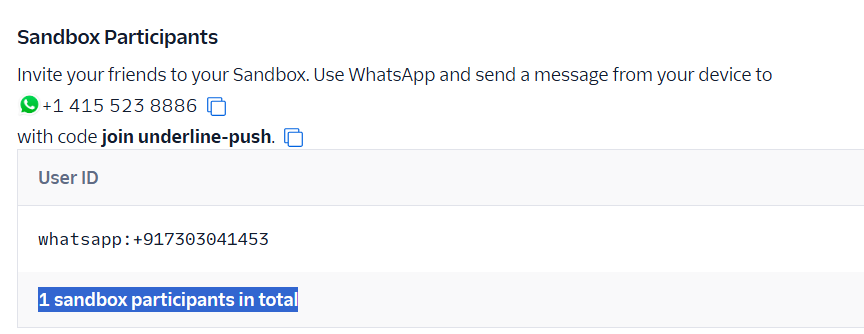
**queued → failed**

* **To let your server know what happened, Twilio makes an HTTP POST to a URL you register.**
* **This is called a webhook (Twilio pushes the event to you, instead of you pulling it).**

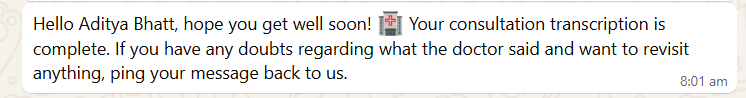
# **Day 11**

So we were stuck in twillo notification error issue just few quick things to know in an sandbox envo for twillo to send notifation what’s app number must be upapproved.

# **Day 12**



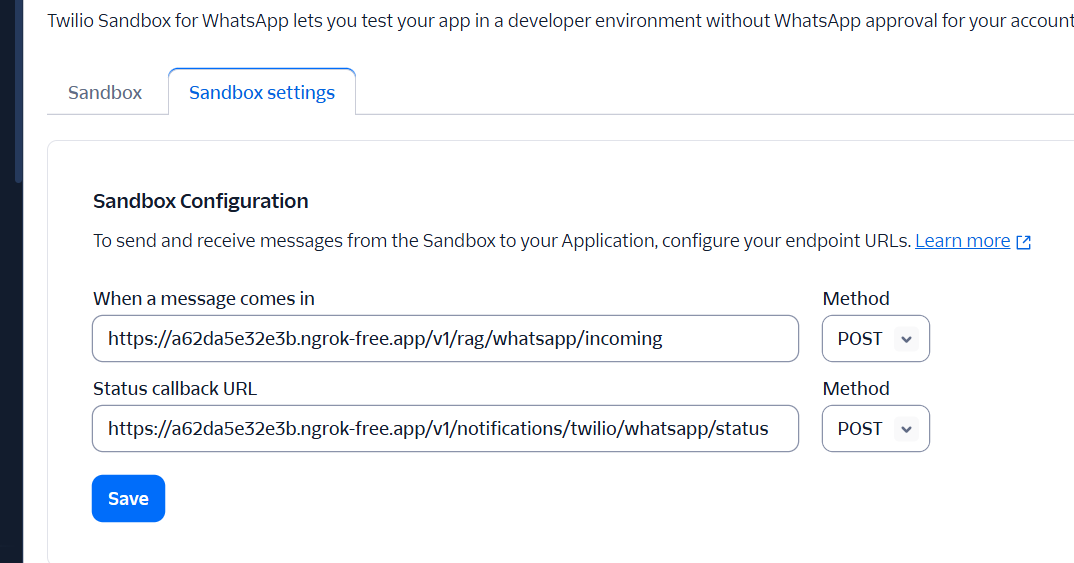
This was resolved , fixing the client for better testing.



So now we are buliding a incoming hook that twillo will call , whenever it recives a what’s message

So we need to figure out this place where we have enterted the when messages comes in

, so twillo will call this



Ngrok is a tool that lets you expose a local server (running on your computer or a private network) to the internet through a secure tunnel.

Think of it this way: normally, if you run an app on your laptop (say http://localhost:8000), only you can access it. Ngrok gives you a **public URL** (like https://abc123.ngrok.io) that forwards requests directly to your local app.

**Why it’s useful**

* **Testing webhooks**: For example, if you’re working with Twilio, Stripe, or WhatsApp APIs, they need to send callbacks to your server. Ngrok lets them reach your local server without deploying to the cloud.
* **Sharing local apps**: You can quickly show your work to teammates or clients by giving them the temporary ngrok link.
* **Bypassing firewalls/NAT**: It creates an outbound tunnel, so you don’t need to mess with router port forwarding.
* **Debugging traffic**: Ngrok provides a web interface where you can inspect all requests/responses passing through the tunnel.

**How it works**

1. You run a local app, e.g., uvicorn main:app --port 8000.
2. You start ngrok with:
3. ngrok http 8000
4. Ngrok gives you a public HTTPS URL.
5. All requests to that URL are tunneled to http://localhost:8000.

👉 In short: **ngrok = instant secure tunnel from internet → your localhost**.

Do you want me to also show you **step-by-step setup** (installation + first tunnel) so you can try it right away?

Top of Form

Bottom of Form

so we need to desgin the rag design before coding also we need to test transcription engine as well the backend needs to be very robust.

We need to plan this better before coding anything

# DAY -13

So here we go, we need to take a design call we can got for both approaches but we don’t know how much will be scale we need to move little faster and and models have a 1 million tokens for now we can move ahead I guess

# DAY -14

So again as a matter of fact when we r cooking background tasks and stuff we should check a lot for race condtions this is getting more important having a proper control over code base is the key.

So we have a working system now the next step is just working on optimizing stuff no new backend work we will work on optimizing

# DAY -15

We will just understand the transcription engine and optimize that nothing else

**Start session (HTTP)** – POST /v1/transcription/start  
Creates a TranscriptionSession document in MongoDB with status "starting"

 **Stream audio (WebSocket)** – /{transcription\_session\_id}/stream  
The mobile app:

* Sends **metadata** (sequence\_number, chunk\_size\_bytes, etc.) for each audio chunk.
* Then sends **binary audio data** (8–10 sec WebM/Opus chunks).
* Server validates, stores audio in blob storage, queues transcription, and sends back transcript updates in order

routes

utils

.

 **End session (HTTP)** – POST /v1/transcription/end  
Marks session as "ending", waits for background tasks to drain, then moves to "completed"

routes

utils

.  
A WhatsApp notification is sent to the patient once done

So in simple terms:  
Doctor starts → Patient speaks (streamed chunks) → System transcribes → Doctor ends → Patient gets a final notification.

**Step 2: WebSocket Streaming Basics**

When the mobile app connects to /{transcription\_session\_id}/stream:

1. **Validation** → ensure session exists + status is "starting" or "streaming"

utils

.

1. **Handshake** → server accepts, marks DB "streaming", sends connection\_confirmed with limits (1 MB chunk, 8s expected)

routes

schemas

.

1. **Message loop** → two kinds of input:
   * JSON text = metadata (sequence\_number, size).
   * Binary = actual audio chunk.  
     Both are paired and validated before processing.

Want me to go **deeper into how metadata + binary pairing works** (the critical guardrail against corruption), or into **how background tasks process audio chunks**?

Step 3: Metadata ↔ Binary Pairing

The client must send two messages per chunk:

Metadata (text) → JSON with sequence\_number, chunk\_size\_bytes, duration\_seconds

schemas

.

Stored in pending\_metadata[sequence\_number].

Capped at 100 entries → prevents runaway memory use

routes

.

Binary (bytes) → the actual audio.

Server checks:

Does metadata exist for expected\_sequence?

If missing → error MISSING\_METADATA, chunk skipped

routes

.

Does binary size match chunk\_size\_bytes?

If mismatch → error SIZE\_MISMATCH

routes

.

Does size exceed 1MB hard limit?

If too big → error CHUNK\_TOO\_LARGE

routes

.

✅ Only after all checks, metadata entry is popped, and the chunk is queued for processing.

This ensures ordering + integrity, but also introduces waiting pressure: the server won’t process until both metadata and binary arrive.

This ensures ordering + integrity, but also introduces **waiting pressure**: the server won’t process until both metadata and binary arrive.

**Abuse risk in merged design**

If you let clients send [header][audio] in one shot:

* An attacker could **lie in the header** (say chunk\_size\_bytes = 100 KB but actually send 50 MB).
* Your server would accept the frame, then blow memory or disk before catching the mismatch.
* This is why you had the JSON-first design: you can reject an oversized chunk *before* the bytes arrive
* So yes — the principle is: **validate early, reject cheap**. First at infra (cheaper), then again in Python (safer).

**Step 6 — What is an API Gateway?**

An **API Gateway** is like the **front door** to your system.

* Every request from a client (mobile, web) goes through it first.
* It decides: *“Should I let you in? How much can you carry? Where should you go?”*

Think of it as a **traffic cop** + **security guard** rolled into one.

**What it does (core jobs)**

1. **Request size limits**
   * “Don’t bring more than 1 MB into the club.”
   * If someone sends 50 MB audio → reject before Python ever sees it.
2. **Authentication & authorization**
   * Check if JWT/API key is valid.
   * Block bad actors right at the edge.
3. **Rate limiting / throttling**
   * “Only 100 requests per minute per user.”
   * Stops spam/flooding.
4. **Routing / load balancing**
   * Send /transcription/\* to your FastAPI pods.
   * Send /analytics/\* to another service.
5. **Monitoring & logging**
   * Gives you metrics: latency, error rates, request counts.

Nginx (say it like “engine-x”) is a **web server** and **reverse proxy**.

* **Web server mode**: it can serve files directly (HTML, images, etc.).
* **Reverse proxy mode**: it sits in front of your app servers and forwards requests.

When used as an **API Gateway**, Nginx is mostly a **reverse proxy**:

* Client talks to Nginx.
* Nginx checks request size, auth, rate, etc.
* Nginx then forwards the request to your Python FastAPI app if it passes.
* proxy\_pass → tells Nginx to forward to your FastAPI app (running at 127.0.0.1:8000).
* The proxy\_set\_header lines forward important request info.

1. Save the file.
2. Reload Nginx to apply changes:

✅ You’ve just turned Nginx into a **reverse proxy**.  
This is the foundation of an API Gateway setup.

**How the big players do it**

* **OpenAI, Meta, Google, etc.** almost never let you hit their app servers (Python, Go, Rust, whatever) directly.
* Instead, traffic goes through a **reverse proxy / gateway layer** first:

Client → (CDN / Load Balancer) → Reverse Proxy (Nginx / Envoy) → App servers (FastAPI, Django, etc.)

* **Reverse proxy (Nginx, Envoy)**:
  + Enforces request size limits (e.g., no >1 MB WebSocket frame).
  + Handles TLS/SSL (HTTPS).
  + Does load balancing → spreads requests across 100s or 1000s of app servers.
  + Can apply rate limiting and caching.
* **App servers (FastAPI, Node, Go)**:
  + Focus only on **business logic** (like transcribing audio, updating Mongo, calling OpenAI APIs).
  + Don’t worry about traffic floods, bad clients, or SSL termination — the proxy already filtered that.

Yes 💡 — Nginx is both a **reverse proxy** *and* a **load balancer**.

Think of it like this:

* **Reverse proxy** → “I stand at the front and forward traffic to *one* backend.”
* **Load balancer** → “I stand at the front and spread traffic across *many* backends.”

**Example: Nginx as load balancer**

Suppose you run 3 FastAPI servers:

* 127.0.0.1:8000
* 127.0.0.1:8001
* 127.0.0.1:8002

Your Nginx config could look like this:

http {

upstream fastapi\_servers {

server 127.0.0.1:8000;

server 127.0.0.1:8001;

server 127.0.0.1:8002;

}

server {

listen 80;

server\_name localhost;

location / {

proxy\_pass http://fastapi\_servers;

proxy\_set\_header Host $host;

proxy\_set\_header X-Real-IP $remote\_addr;

proxy\_set\_header X-Forwarded-For $proxy\_add\_x\_forwarded\_for;

}

}

}

* upstream fastapi\_servers { ... } defines your pool of app servers.
* proxy\_pass http://fastapi\_servers; tells Nginx to distribute requests.
* Default strategy = **round robin** (first request to 8000, second to 8001, third to 8002, then repeat).

This is why Nginx is so popular — it’s not just a guard at the door, it can also act like a **traffic manager** that spreads load evenly.

**The philosophy big companies use**

* **Nginx / API Gateway**: quick & cheap auth (API key, IP allowlist, TLS termination).
* **App server (FastAPI)**: real identity/permission checks (JWTs, OAuth2, RBAC).

So Nginx is the **first filter**, FastAPI does the **deep inspection**.

 **FastAPI + Nginx on a VM**

* What you just did.
* Nginx = reverse proxy, TLS termination, request size limits, load balancing (if you add more FastAPI instances).
* This gives you the *real feel* of how big companies front their apps.

**Later (Kubernetes world)**

Instead of multiple FastAPI ports on one VM, you’ll have multiple **pods** across many VMs.  
Nginx (Ingress) does the same round robin — just across pods instead of ports.

**Why this helps you**

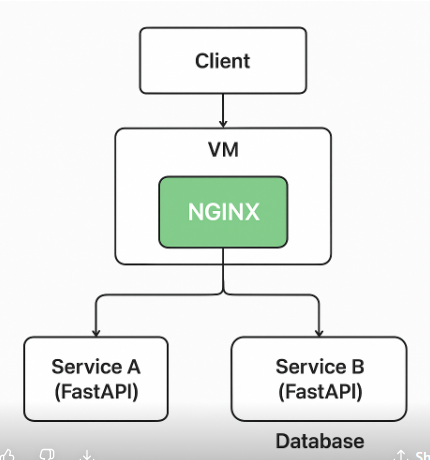
* **Load balancing**: If you run multiple FastAPI workers, Nginx spreads requests.
* **Request shaping**: Block uploads >1MB *before* they reach your WebSocket code (instead of letting Python reject them later).
* **TLS termination**: Clients connect with HTTPS → Nginx handles certs. FastAPI only sees clean HTTP traffic.
* **Rate limiting**: Protects your /v1/transcription/{id}/stream endpoint from spam floods.
* **Observability**: Access/error logs without touching Python code.

**ChatGPT said:**

Exactly 👍 — multiple apps can all **sit behind the same Nginx**, even if they’re on the same VM.

Think of Nginx as a **traffic cop at a big junction**:

* Cars (requests) come in.
* Nginx looks at the sign (URL path, hostname, etc.).
* Nginx decides → “this one goes left (Service A), this one goes right (Service B).”



**Step 15 — What is Azure Front Door?**

Think of it as a **global reverse proxy + load balancer + CDN** that Microsoft runs for you.

* It sits at the **edge of Microsoft’s global network** (not just your VM).
* It routes client traffic to your backend (VMs, App Services, Kubernetes, etc.).
* It adds built-in features like:
  + **Global load balancing** → if you have servers in India and Europe, it routes a user to the nearest healthy server.
  + **SSL/TLS termination** → you don’t have to manage certs, Azure does.
  + **Web Application Firewall (WAF)** → blocks malicious traffic (SQL injection, bot floods).
  + **Caching** → static responses can be cached close to users.
  + **Health probes** → checks if your backend is alive, reroutes if not.

Inside a single VM, you can totally run **all the components together**:

* **FastAPI** → your web app (HTTP + WebSocket APIs).
* **Celery** → background task queue (processes heavy jobs like transcription, RAG calls).
* **Redis / RabbitMQ** → Celery’s broker (where jobs get queued).
* **MongoDB / Postgres** → your database.
* **Nginx** → front door reverse proxy.

All in one machine, wired together by local localhost:PORT connections.

**Why this setup works**

* **Learning-friendly** → You touch all the real production pieces (queue, broker, reverse proxy, DB).
* **Cheap** → One VM bill instead of many.
* **Realistic** → Same architecture as a scaled system, just collapsed into one box.

**Why companies later split it out**

* If Redis crashes, you don’t want it to take down the webserver.
* Scaling → sometimes Celery workers need more CPU, DB needs more RAM.
* Reliability → running DBs in managed services (like Mongo Atlas, Azure Redis) is safer.

But for **a solo developer in India cooking AI backends**, this “all-in-one-VM” setup is **perfect**: it’s cheap, real-world enough, and teaches you how the pieces talk.

So yes for now to be honest I do not know the scale of this thing we can like focus on presgenie.ai and it’s launch and making sure we have a good lanch rather than improving our scribe.ai(to be hnest for a demo it is ready)

**1. R&D (Research + Core Models)**

* People who live in **PyTorch/TensorFlow/JAX**.
* Train/fine-tune models (LLMs, vision, multimodal).
* Read papers, try out new architectures (MoE, RoPE, FlashAttention).
* Outputs: **checkpoints, weights, evaluation metrics**.
* Example role: “Research Scientist” or “Applied Research Engineer.”

**2. Product (Application Layer)**

* People who build **features and apps** on top of models.
* Your FastAPI services (/transcription, /rag, /notifications

routes

routes

routes

).

* Glue AI into workflows (WhatsApp bots, doctor apps, reporting dashboards).
* Outputs: **APIs, apps, user features.**
* Example role: “Product Engineer” or “AI Application Developer.”

**3. AI Infrastructure (MLOps / Backend Infra)**

* People who make sure everything **runs at scale, fast, reliable**.
* Manage **VMs, Nginx, Kubernetes, Celery, Redis, databases**.
* Set up **monitoring, logging, scaling, deployments, cost controls.**
* Outputs: **systems that don’t crash under load.**
* Example role: “MLOps Engineer,” “AI Infra Engineer.