GROUPBRAIN LABS — TALK

AI Agents for Calling Usecases

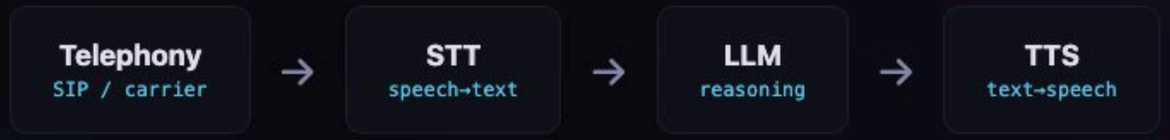
Building a production voice agent, one layer at a time.

August 2026

How a chat LLM works

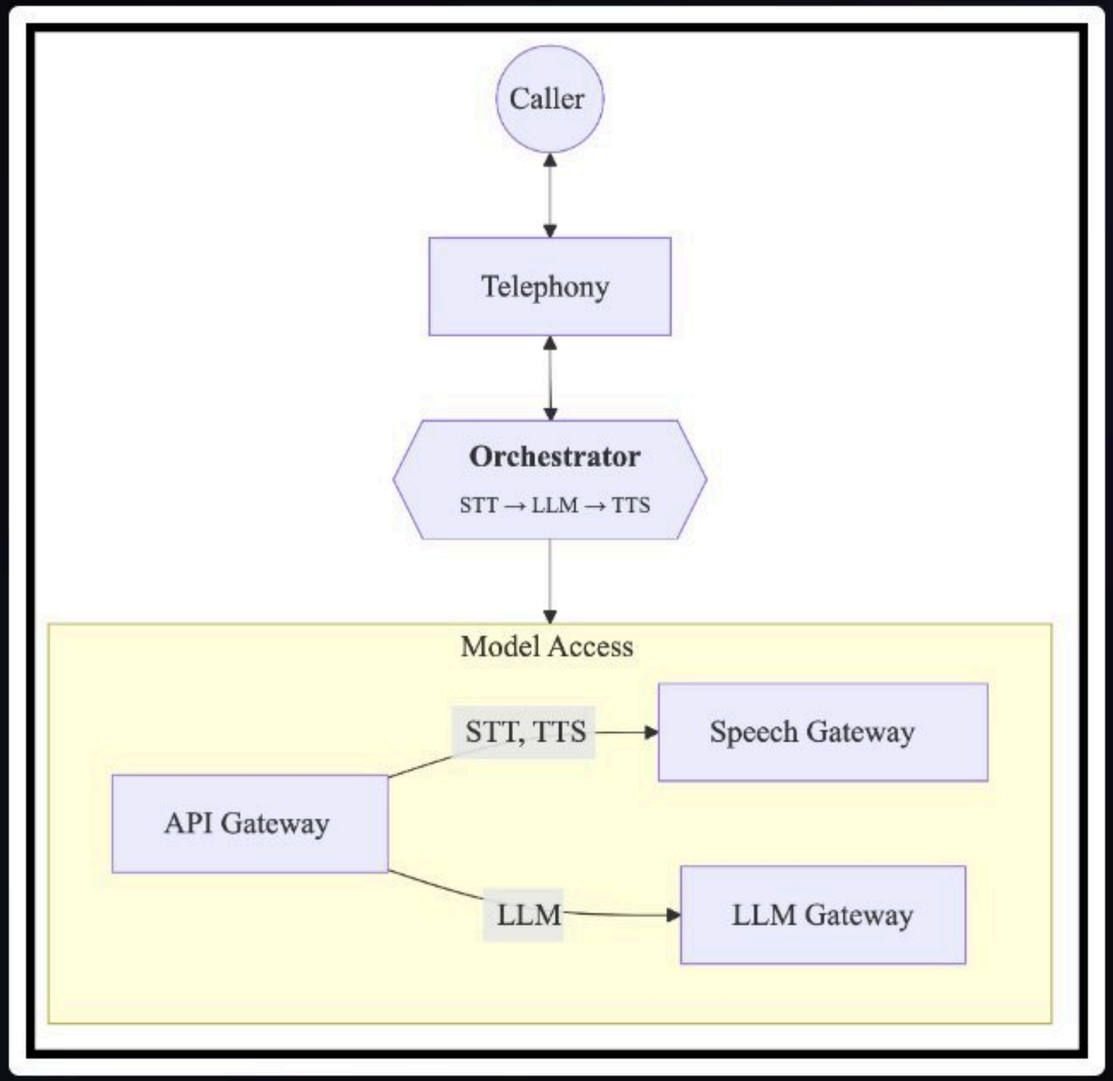
Request → response. No clock, no audio, no urgency — the model can take 3 seconds or 3 minutes.

How a calling agent works

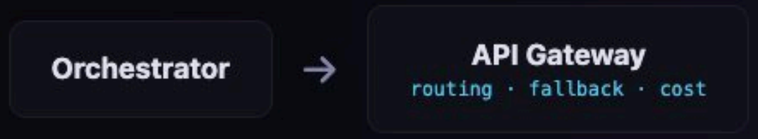


Now there's a human on the line waiting in real time — every millisecond in this chain is felt.

Model access goes through an API Gateway



Model access goes through an API Gateway



SPEECH GATEWAY

STT + TTS

Routes speech-in and speech-out calls to whichever provider is configured — swap STT/TTS vendors without touching the Orchestrator.

e.g. Whisper · Deepgram · ElevenLabs

LLM GATEWAY

LLM

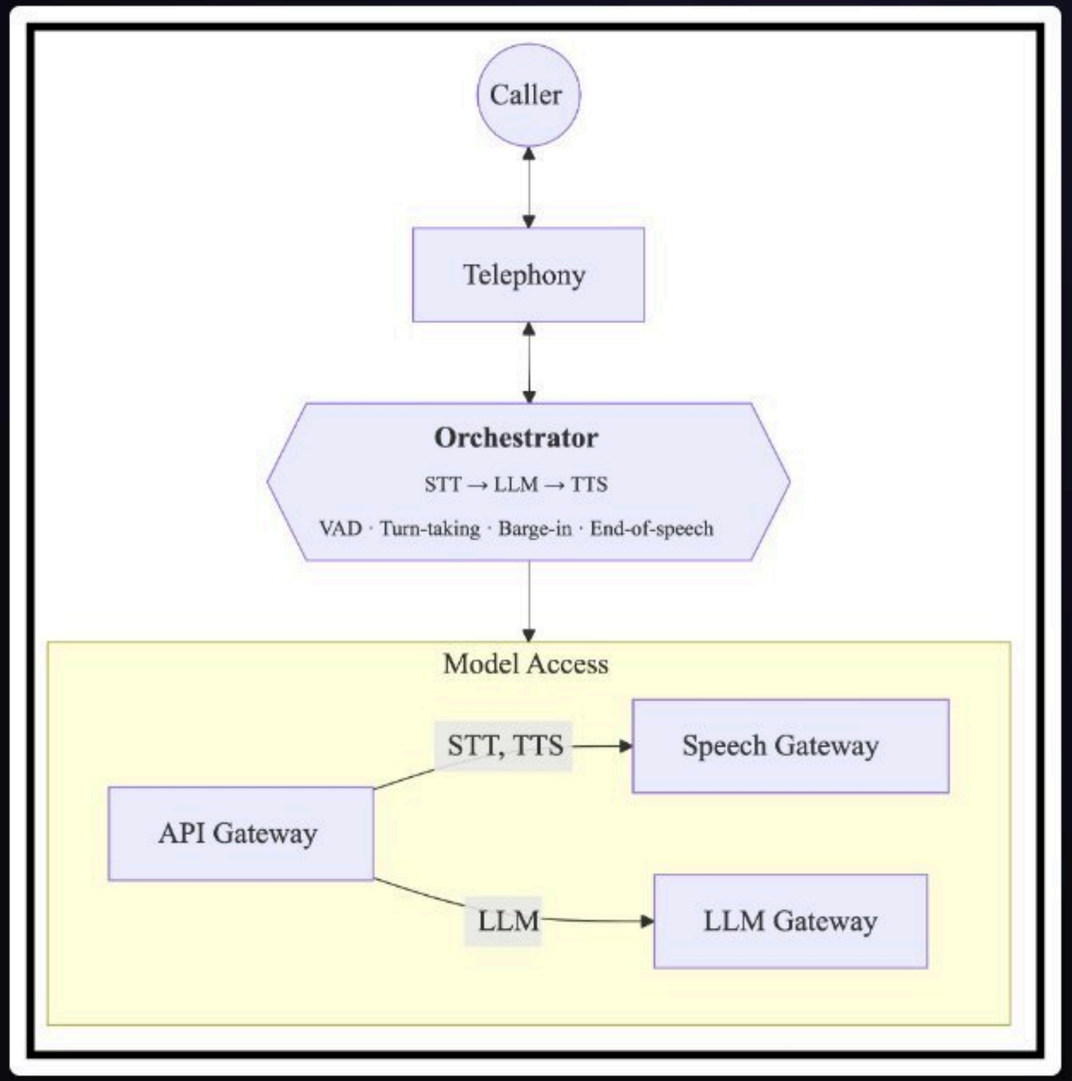
Routes reasoning calls to whichever LLM provider is configured — model fallback, rate limits, and cost tracking live here.

e.g. GPT · Claude · Gemini

Turn-taking

Stages don't wait for each other to finish — they stream and overlap.

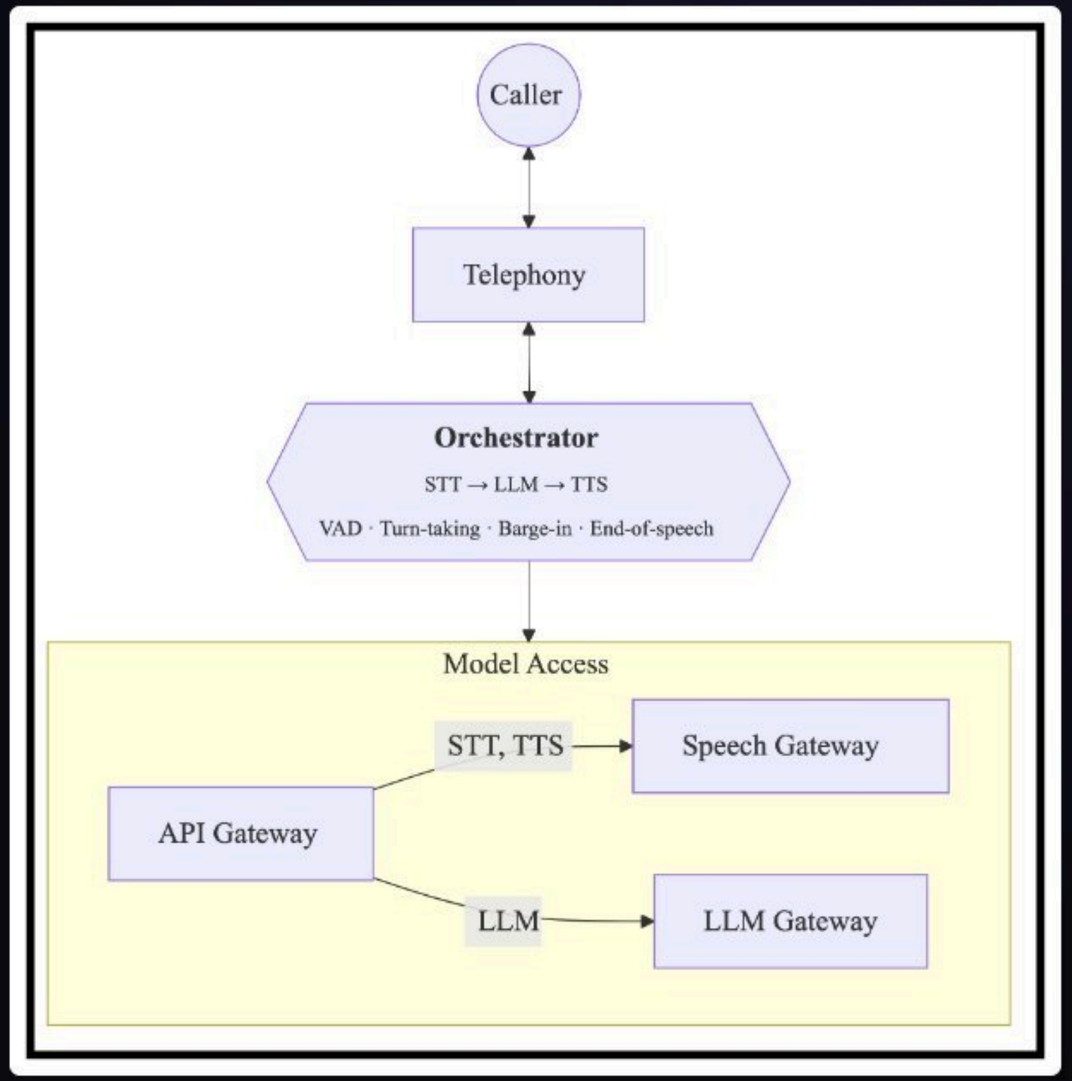
STT emits tokens continuously. The LLM starts drafting on partial transcripts. TTS speaks partial output as it's generated.



Turn-taking

TTS only starts once STT detects end-of-speech.

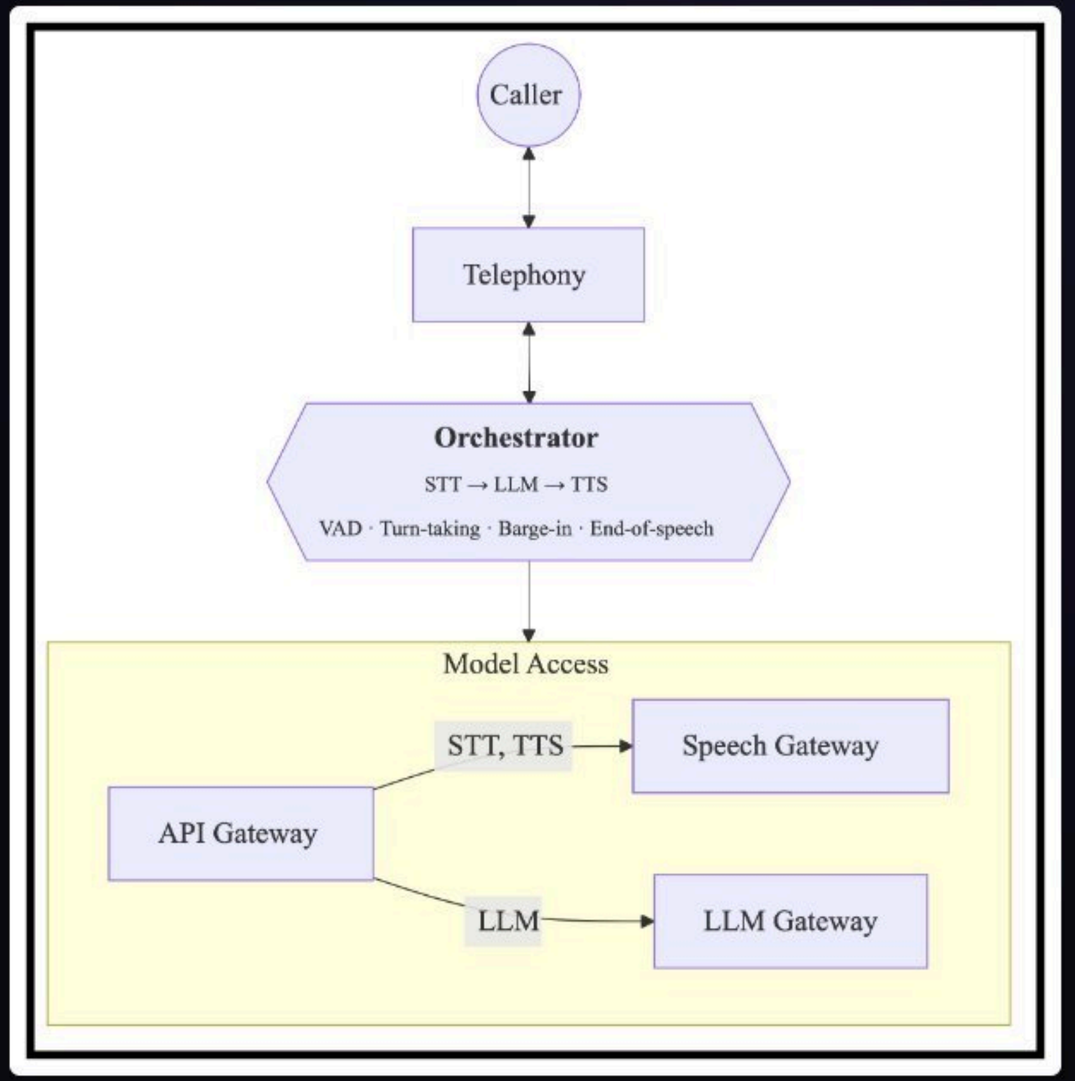
The pause that means the caller is done talking.



Turn-taking

If the caller starts talking while the agent is speaking, the agent stops instantly.

Humans expect a 200–500ms turn-taking window — this is what makes it feel like a real conversation.



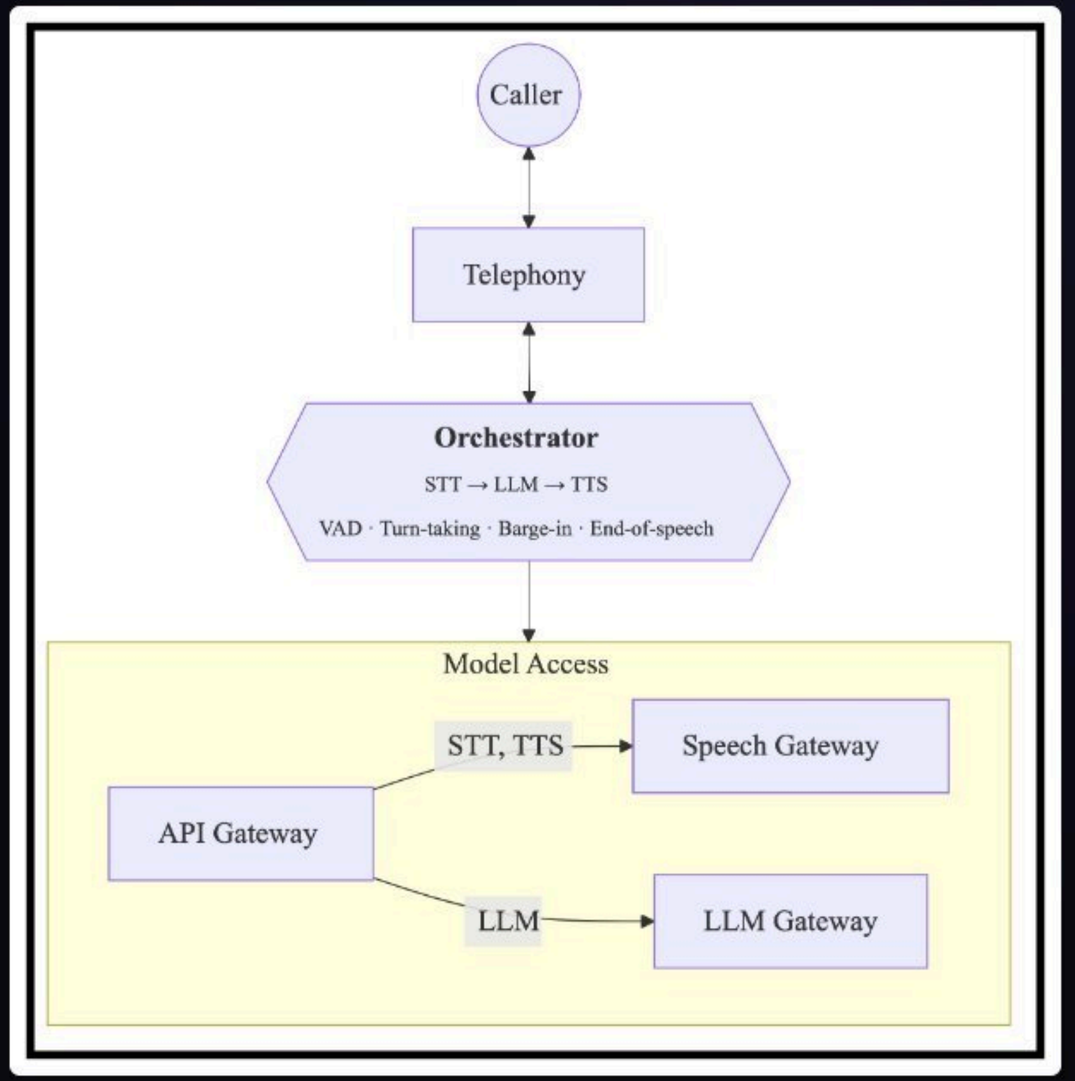
Turn-taking

On interrupt, context must reflect only what was **actually spoken** — not what the LLM generated.

Otherwise the agent "remembers" saying things the caller never heard.

NOTE

We record what the LLM generated, what TTS actually spoke before the interrupt, and what ends up in context — useful signal later for post-deployment evals (more on evals soon).



09 — A REAL SCENARIO

"I want to cancel my policy"

Authentication happens at the *account level*, not the policy level — phone number or email, verified with an OTP, before any policy is even looked up.

OPTION A

RAG

Semantic search over policy documents. Great for open-ended questions — "what does clause 4.2 cover?"

Wrong fit here

OPTION B

Tool-use

Deterministic lookup — send an OTP, verify it. This is structured business logic, not a knowledge question.

Use this

Verifying the caller

1 **Caller provides phone number or email**

A general account identifier — not tied to any specific policy yet.

→ **2** **LLM calls `send_otp()`, caller confirms it**

A real tool call against the auth system — proves the caller owns that phone/email.

Fetching the caller's policies

✓

Caller authenticated

Verified by phone/email + OTP — no policy has been touched yet.

→ 3

LLM calls `get_policies(identity)`

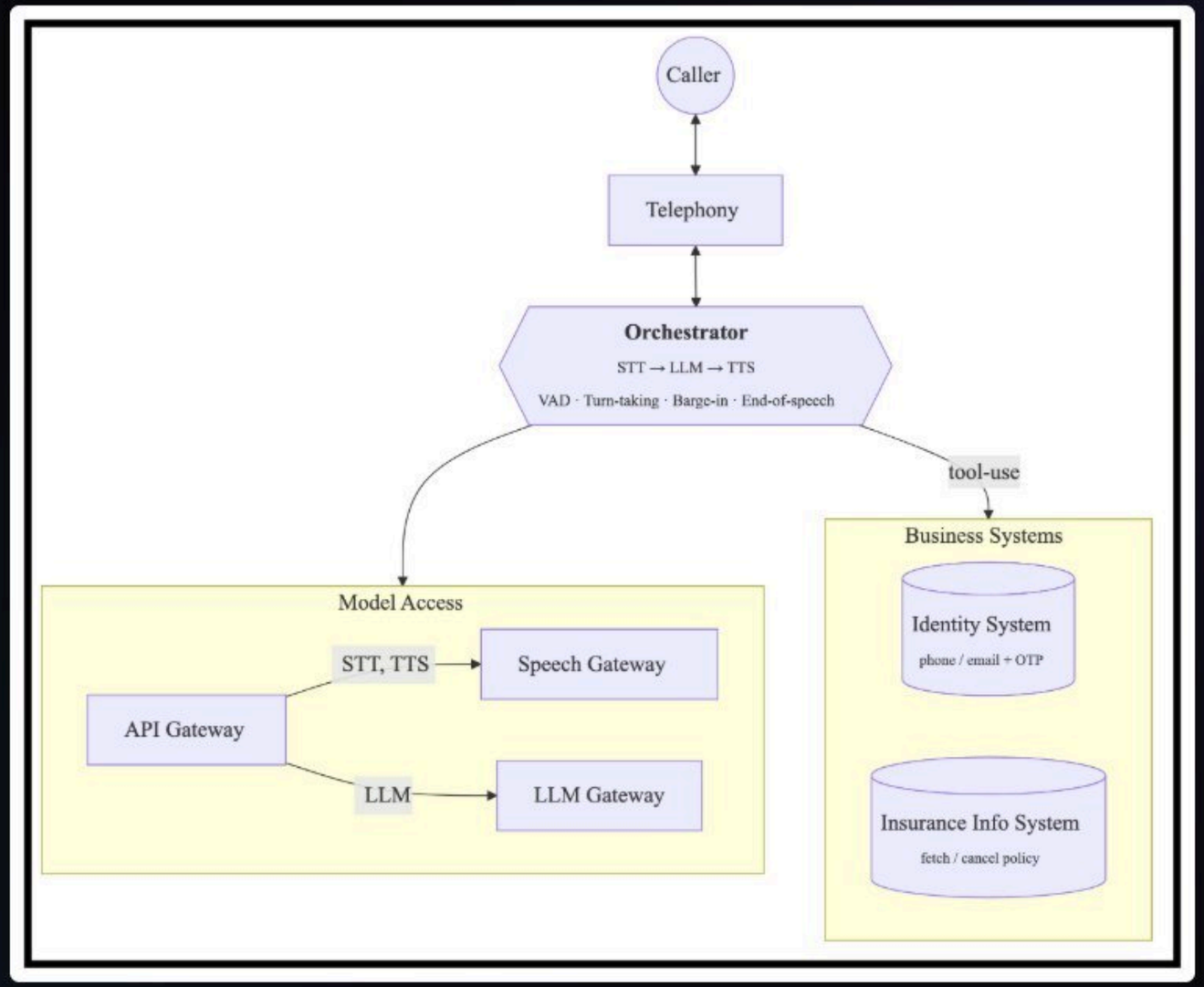
Returns every policy linked to that phone/email, plus any other details the caller gave — nobody else's.

Cancelling the policy

4 **"Which policy would you like to cancel?"**
Agent reads back options, caller confirms one.

→ 5 **LLM calls `cancel_policy(id)`**
A real side-effect — this is where mistakes get expensive.

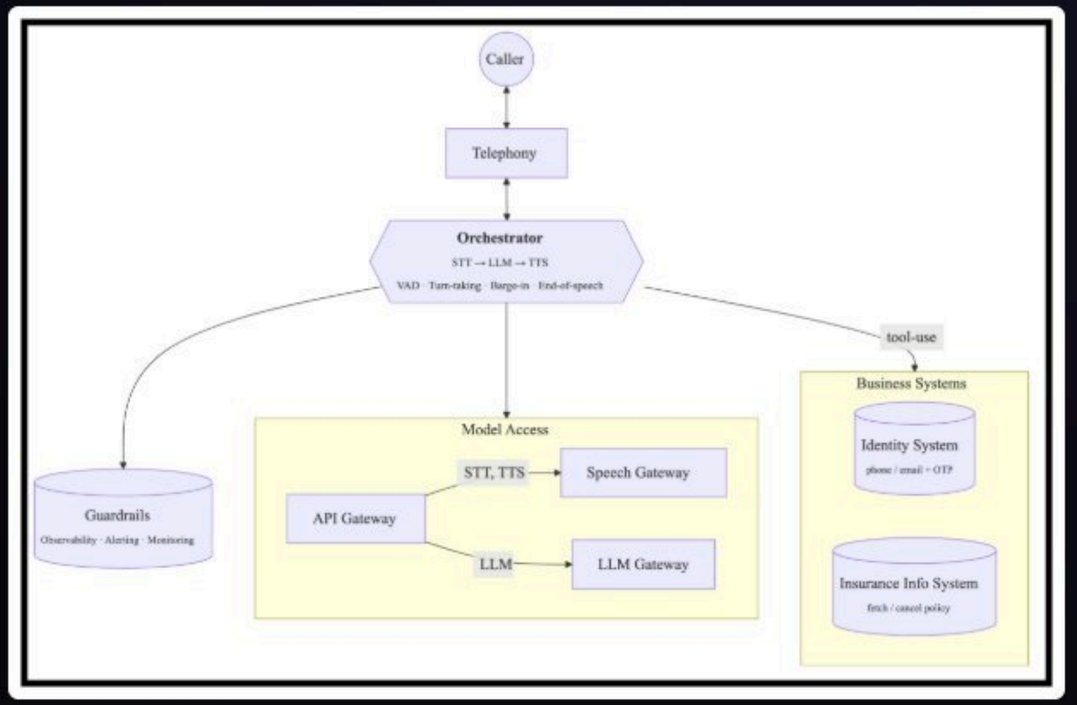
Everything the Orchestrator can reach



Guardrails

Always confirm before destructive actions.

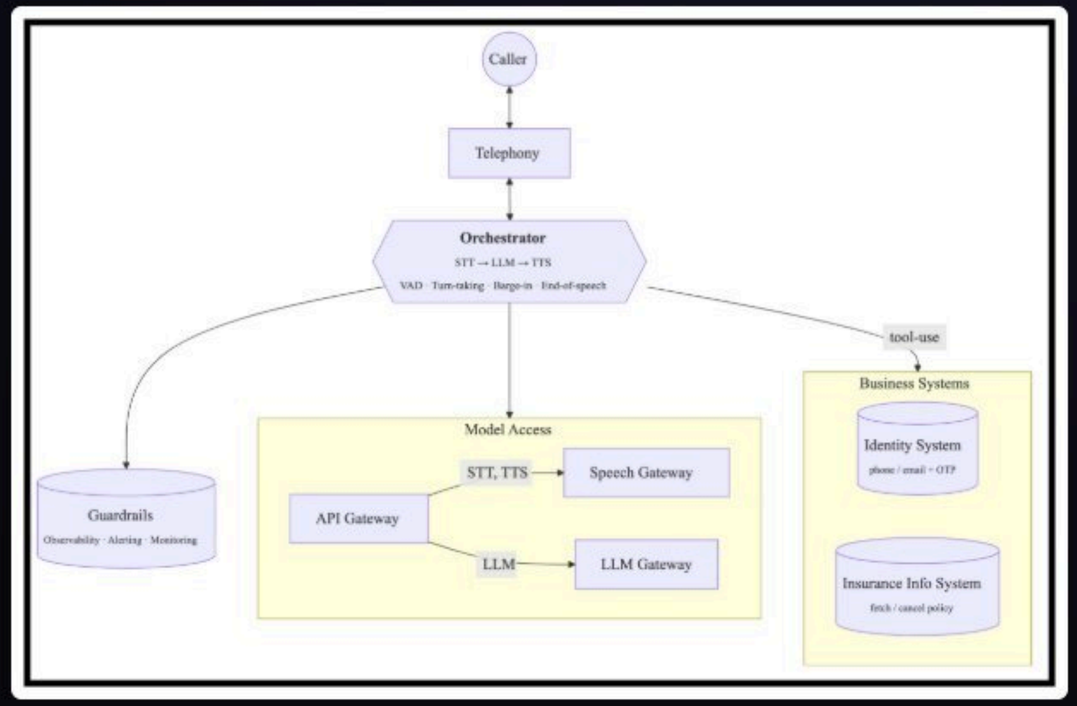
The agent repeats back the exact policy before calling `cancel_policy()`.



Guardrails

Prompt-level rules spell out what the agent must never do.

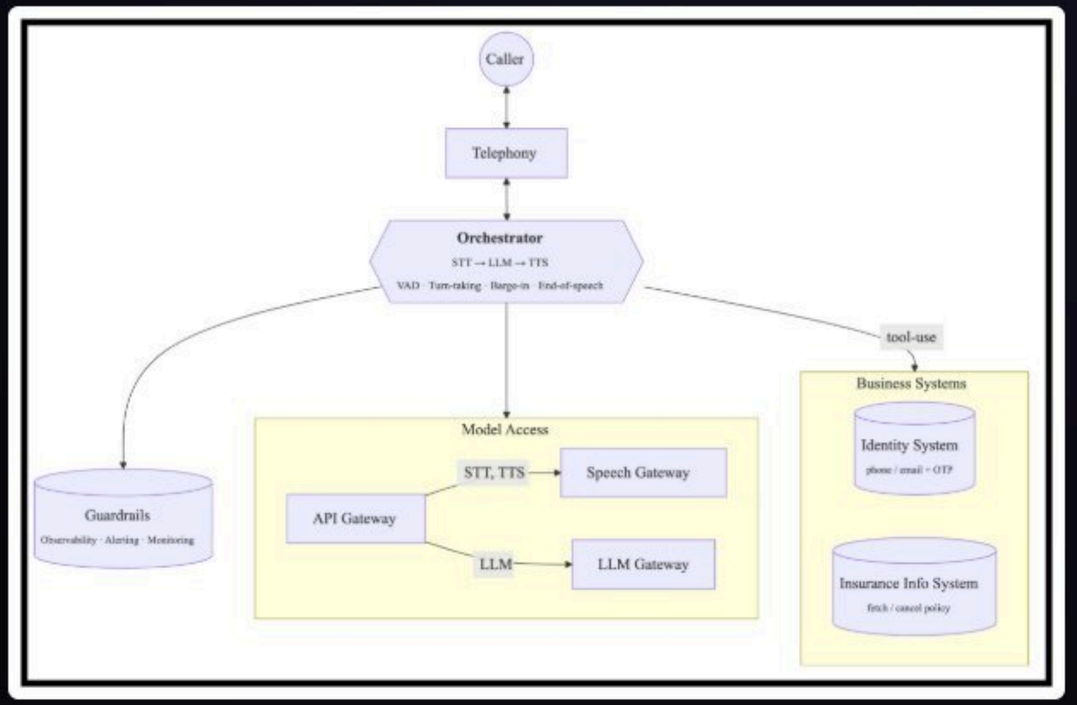
E.g. never cancel without confirmation. Explicit, written instructions the model is told to follow.



Guardrails

Least-privilege tool scoping — the strong one.

The LLM's tool session is scoped only to *this verified caller's* records. Even a hallucinated or injected tool call structurally cannot reach someone else's policy.



Observability

Per-stage latency
Telephony / STT / LLM / TTS — find where time is actually going.

Tool-call outcomes
Success / failure / timeout rates for every function call.

Hallucination flags
Agent claiming something not grounded in a tool response.

Transcription accuracy
Sampled WER — is STT actually hearing people right?

Sentiment / frustration
Signal for escalation before the caller gives up.

Cost per call
LLM tokens + STT/TTS + telephony minutes, tracked per call.

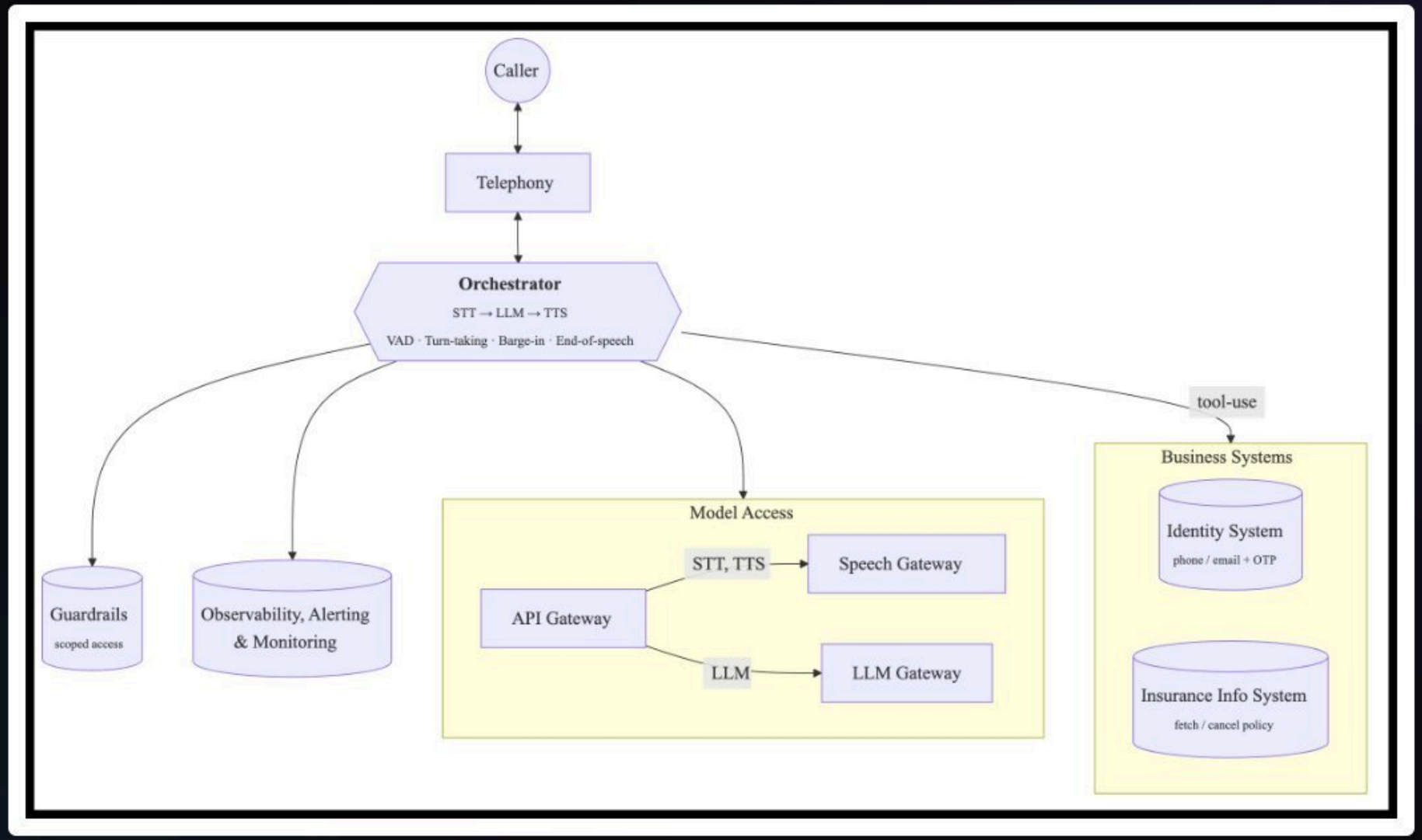
Alerting & Monitoring

- **Tool-call failures** — Identity or Insurance Info System down or erroring.
- **Elevated latency** — STT, LLM, or TTS running slower than threshold.
- **Hallucination-flag spikes** — the agent claiming things not grounded in a tool response, more than usual.
- **Negative sentiment / frustration spikes** — callers sounding upset more often than baseline.

Alerting & Monitoring

- **Real-time push to on-call** — Slack/PagerDuty-style, with severity levels so not everything pages someone at 2am.
- **Dashboards for passive monitoring** — the stuff from "Observability," always there to check, vs. active alerts that interrupt someone.

Observability joins the picture



Human handoff

Low confidence, repeated tool failures, or detected frustration trigger an alert.

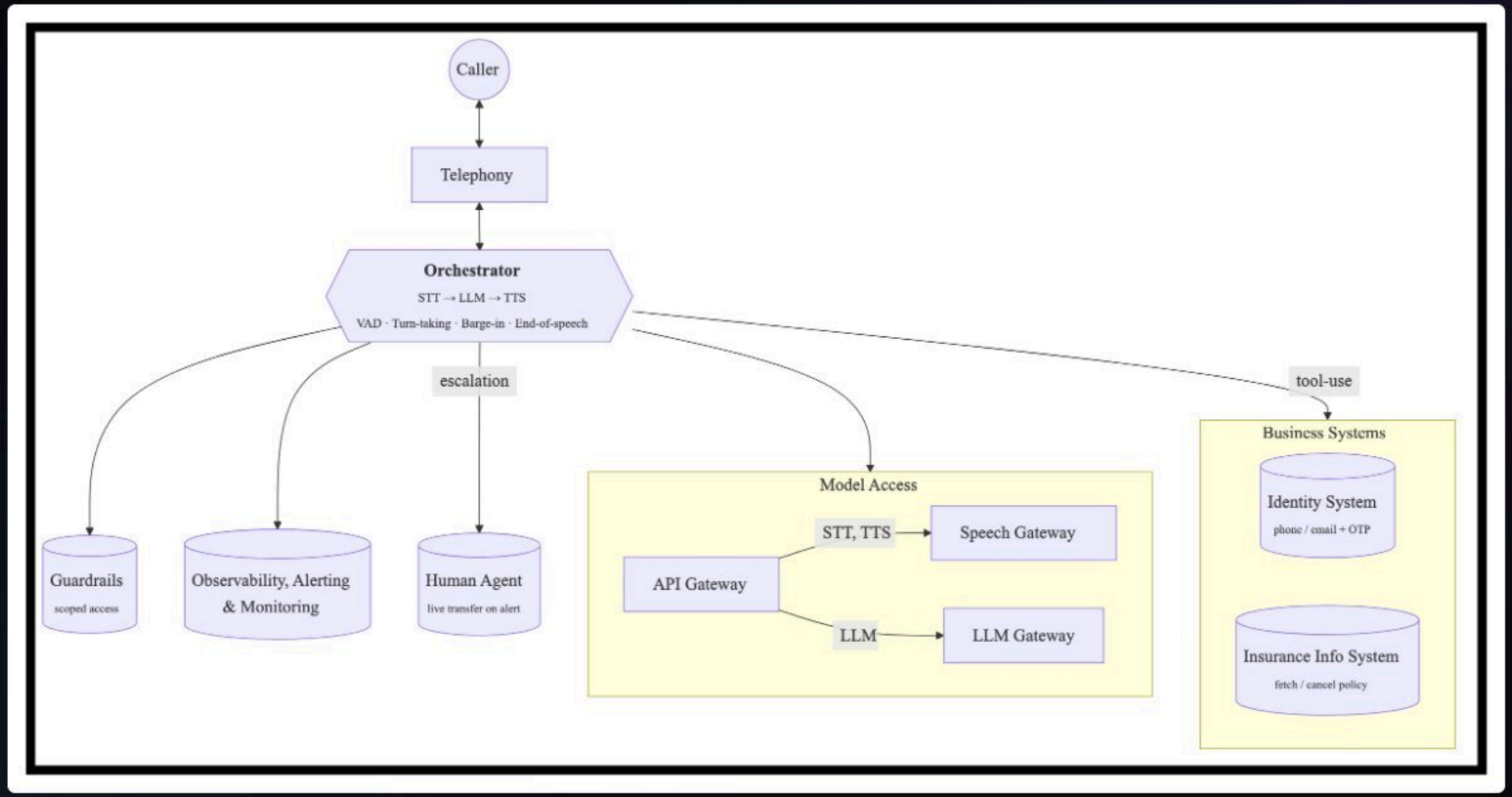
The system flags itself instead of pushing through uncertain.

Human handoff

The call is transferred live to a human agent.

With the conversation context carried over — the caller doesn't repeat themselves from zero.

Human Agent joins the picture



Data collection

- Every conversation is stored — with **PII redacted** (names, policy numbers, phone numbers) before it's used for anything downstream.
- This becomes the raw material for the next slide.

Evaluations

BEFORE LAUNCH

Red-teaming & scenario tests

Adversarial prompts, edge cases, tool-misuse attempts — caught before a real caller hits them.

AFTER LAUNCH

Continuous quality scoring

Drift detection on real conversation data — did the agent get worse this week? Feed it back in.

RECAP — THE FULL PICTURE

Everything we just built