

Voice AI infrastructure you own.

Every capability in this deck exists in three places: the dashboard, the API, and an MCP tool.

22+

Indian languages

multilingual

Hinglish

70+

global languages

< 300 ms

response latency

149+

countries you can call

850+

integrations supported

Advanced call quality controls

01 Predictive Pre-Processing

The reply is drafted while the caller is still talking.

03

02 Audio Caching

Fixed lines are synthesised once and replayed on every call.

04

03 Multilingual Agents

Language detection, up to four languages, and voices to match.

05, 06

04 Hinglish Control

Handled by a platform-level word map applied after generation.

07

05 Fallback LLMs and Transcribers

A second provider per layer, triggered on latency.

08

The reply is predicted before the caller answers

Agent says

 Did you get a chance to visit our clinic?

The agent drafts a reply to each likely answer while the caller talks, so the response is ready the moment they stop.

Likely answer

Yes, last Tuesday.

No, not yet.

Which clinic?

I went elsewhere.

Drafted reply

How did it go?

Shall I hold a slot this week?

The Indiranagar branch.

How was that visit?

Caller says

Not yet, I've been travelling. 

Agent says · matched draft, instant

 Shall I hold a slot this week?

Delivered in 212 ms · about 500 ms sooner than generating fresh

Say it once, reuse it on every call

Turning text into speech is the slowest part of a reply. After about 100 calls, most sentences an agent needs are already covered, so the saved audio for that phrase is played back instead of being generated again.

- **Most of a call is the same every time**

Greetings, confirmations and sign-offs never change. Synthesised once, then stored.

- **The cache warms up as you dial**

Within about 100 calls, over 80% of turns replay from cache.

- **Anything new goes to TTS**

Names, amounts and dates are synthesised live, then stored.

- **A stored line saves a trip**

A stored line saves the trip to ElevenLabs, Cartesia or whichever voice model is in use, getting the reply out roughly 500 ms sooner.

Hi, this is Priya from Nexa Clinic.

FROM CACHE · 272 MS

Yes, speaking.

Is now a good time to talk?

FROM CACHE · 291 MS

Go ahead.

Your appointment is on 18 Sep at 11 AM.

Synthesised live · 708 ms

That works, thank you.

Thanks for your time, have a good day.

FROM CACHE · 264 MS

The agent follows whichever language the caller uses

- **Detection happens on the call**

A multilingual agent identifies the language the caller is speaking and continues in it.

- **Switching mid-call is fine**

If the caller changes language part way through, the agent changes with them.

- **Within the languages you configured**

Detection is scoped to the primary and secondary languages set on the agent.

22 Indian languages supported

Assamese Bengali Bodo Dogri Gujarati Hindi
Kannada Kashmiri Konkani Maithili Malayalam Manipuri
Marathi Nepali Odia Punjabi Sanskrit Santali Sindhi
Tamil Telugu Urdu

English Hindi Marathi Tamil

Hello, this is Priya from Nexa Clinic.

Primary · English

हाँ बोलिए, क्या यह मेरे appointment के बारे में है?

Caller · Hindi

जी हाँ, आपका appointment 18 September को सुबह 11 बजे है।

Secondary · Hindi

நன்றி, அது சரி.

Caller · Tamil

நன்றி, வணக்கம்.

Secondary · Tamil

Name up to four languages, or open the agent to all

- **Name your languages for finer control**

One primary and up to three secondary. The agent speaks only these, and you prompt it separately in each one to shape how it narrates.

- **Or go multilingual for full coverage**

The agent handles any supported language, but you lose the per-language prompting and control.

- **You pick the voice**

The library has thousands of voices to choose from. You can also clone your own, or import any voice from ElevenLabs or Cartesia.

- **Change it any time**

Swap a language or a voice in the editor and the next call picks it up. No rebuild, no redeploy.

×

Select languages

Choose the language your agent speaks. Select secondary languages if you want your agent to be able to speak in multiple languages.

Primary language

 English

▼

Secondary languages (optional)

 Hinglish

▼

×

 Marathi

▼

×

 Tamil

▼

×

21 voices support English, Hinglish, Marathi and Tamil

Cancel

Continue to voices

Hinglish is how India actually talks on the phone

01 Models default to textbook Hindi

Callers say loan and income. A model trained on formal text says ऋण and आय, and the call starts sounding official.

02 Prompting alone cannot fix it

LLMs will always hallucinate. Told to keep it casual the model usually complies, then drifts, and you only find out after the call.

03 So the swap runs after generation

Every mapped word is replaced on the way to speech: a deterministic find-and-replace, not another instruction.

04 3,000+ words shipped, plus your own

The map ships with 3,000+ code-switched words. Add your brand, product and scheme names on top, and your entries take priority.

 **Hinglish Map** ×

3598 words added

ARCHAIC		CASUAL
आय	→	इनकम
ऋण	→	लोन
अतः	→	इसलिए
एवं	→	और
चयन	→	सेलेक्ट
तथा	→	और
यदि	→	अगर
लाभ	→	बेनिफिट

ADD WORD

Archaic... → Modern... + Add

A slow provider should not become a slow call

Every model provider has slow minutes and bad days. If OpenAI is down or lagging, Gemini takes the turn, and the caller hears no difference.

- **Every agent has a backup**

A second LLM and transcriber, from a different provider, on every agent.

- **You decide how long to wait**

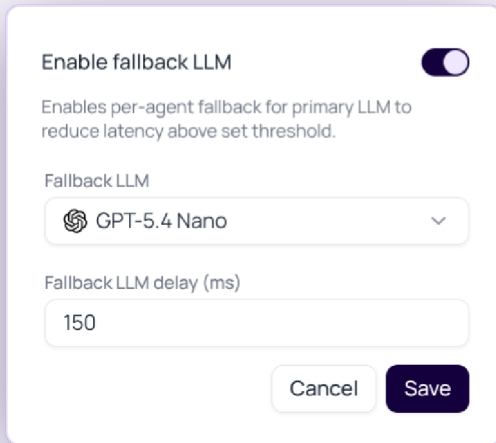
Set in milliseconds: how long the primary gets before the backup is asked.

- **The caller never hears the switch**

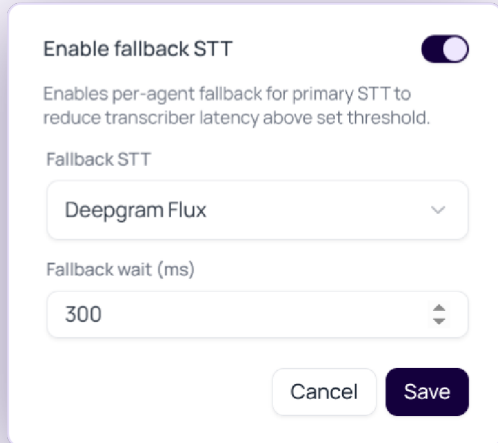
Miss the window and the backup answers, well inside the pause a caller would notice.

- **Set it to zero and run both**

Both models fire in parallel and the fastest reply is used.



Fallback LLM



Fallback transcriber

Four controls, all in the agent editor

Cut the wait

Pre-processing and audio cache

Speak their language

Detection, four languages, native voices

Sound local

Hinglish word map

Stay up

Fallback LLM and transcriber

Previous

Build Your First Voice AI Agent

Building, configuring, publishing and deploying a single prompt agent.

Next

Specialized agents

Flow agents and node settings, warm transfers, speech-to-speech, recorded announcement agents and IVR agents.

Ready to Start?

Pratik Kapasi

Co-Founder, DialNexus

kapasi@dialnexus.com · +91 97414 13670