

Mistral releases Voxtral 3 speech models and a voice agent SDK

Source copy kept by wheresthe.ai for the story "Mistral releases Voxtral 3 speech models and a voice agent SDK" (<https://wheresthe.ai/d/fr-mistral-releases-voxtral-3-speech-models-and-voice-agent-sdk-2026-10/>).

Source: <https://mistral.ai/news/voxtral-3>

Text saved 24 Oct 2026 (the page as published).

SHA-256 of the saved text: 219b9e6154016a57d4e1db446bdbbb4db79f3c66cee5bdd58e1611c314912234

Agent

Sure. What's the reference number?

Tool call

`get_referral(ref="4420 1187")`

Result

Merci de recevoir ce patient pour un avis dermatologique. Traitement en cours : crème à l'hydrocortisone. Dr Martin, Clinique du Parc.

Agent

It says, "Merci de recevoir ce patient pour un avis dermatologique." Your doctor is asking a dermatologist, Dr Martin, to see you. The next slot is Tuesday at 5:45 pm. Shall I book it?

Tool call

`book_appointment(slot="Nov 3 17:45")`

Result

confirmed

Agent

Done. Bring the letter, your carte Vitale, and I'll text you the details in English.

Voxtral 3

Take the call

Béatrice

Hi, my doctor gave me a referral letter in French. Can you book it for me?

Béatrice

4402... sorry, 4420 1187.

Béatrice

Yes, please.

Thinking Summary

Voxtral 3 introduces a cutting-edge voice agent SDK and state-of-the-art speech models designed for real-time, multi-step conversations across 50+ languages. It combines accurate transcription, natural speech synthesis with emotional adaptability, and deep tool integration for seamless workflow automation. Available via Mistral's EU-hosted API or self-deployment, it ensures data sovereignty while delivering low-latency, human-like interactions for contact centers, virtual assistants, and hands-free copilots.

Speech transcription and voice generation have come a long way, but building end-to-end voice agents that can listen, reason, interact, and handle complex workflows remains difficult. A reliable voice agent needs to do four things very well (and simultaneously): Listen accurately across languages, accents, speakers, and interruptions,

Reason deeply about what users mean in specific contexts,

Speak naturally with emotion and expression, and

Take action over multiple steps through connected systems.

Getting all four to work together reliably in real time remains a challenge. To address this, we introduce Voxtral 3: a new family of state-of-the-art speech-to-text and text-to-speech models, alongside a voice agent SDK that brings listening, thinking, tool calling, and speaking into a single real-time streaming pipeline. Voxtral 3 is built to create natural conversational experiences, from contact-center agents and service desks to virtual assistants and hands-free copilots. With support for dozens of languages, expressive speech, complex reasoning, and deep integration with your own tools and data, agents built with Voxtral can interpret nuanced intent, reason through sophisticated requests, and execute multi-step actions across connected systems. As with other Mistral models, the Voxtral 3 family is designed for sovereign deployment. Available today through Mistral's EU-hosted API or for self-deployment on your own infrastructure, it gives organizations control over where audio is processed, how data is handled, and whether conversations are retained.

Highlights Accurate transcription across 50+ languages . The speech-to-text model, Voxtral Small Transcribe 3 , is state-of-the-art in multilingual transcription, identifying who said what (diarization), using keyphrases to improve its accuracy (context-biasing). It detects when someone has finished speaking so agents can respond without delay (end-of-turn detection).

Natural speech that adapts to the conversation . The text-to-speech model, Voxtral Mini TTS 3 , introduces new voices that adjust their emotion, pace, and emphasis based on conversational context, with support for 35+ languages and 50+ dialects.

Human-like intelligence at your (spoken) command . The new voice agent SDK combines listening, reasoning, tool calling, and speaking into a real-time experience. Agents can retrieve information, interact with business systems, and complete tasks during live conversations over phone or web.

Pro Voice for distinctly accurate voice adaptation . For professional use cases where basic voice cloning does not suffice, Voxtral's Pro Voice adaptation is able to reproduce a speaker's distinctive character, emotional range, and speaking style across languages, using just 30 minutes of audio.

Your voice AI, on your infrastructure . Use Mistral's EU-hosted API or deploy Voxtral 3 on your own infrastructure. Keep sensitive conversations within your environment and control whether any audio is retained.

Try Voxtral now | **Learn more** It hears 50+ languages, with state-of-the-art accuracy. Voxtral Transcribe 3 represents a significant milestone across transcription, diarization (who spoke when) and context biasing capabilities. The supported list of languages is expanded to 50, including 24 official EU languages and many low-resource Asian and Indic languages. The Small version achieves state-of-the-art results across diarization and multilingual transcription, including in English, French and Spanish. It outperforms Whisper, Scribe v2 and Gemini 3.5 Transcribe. The Mini variant is 4B parameters, enabling fast and accurate transcriptions at a price of \$0.0035/min. Both models support context biasing, where key terms and free-form text can be given to bias the transcription and improve recall on technical terms. This enables them to be used in production settings with domain-specific vocabulary without the need for fine-tuning. Multilingual Transcription

Diarization

Context Bias

Artificial Analysis

FLEURS 45 Forced vs Inferred Word Error Rate — lower is better

[Download PNG](#) [Download SVG](#)

Forced

Inferred

Macro-average WER over the intersection of 45 languages in the FLEURS benchmark supported by all seven models. WER is used for all languages except Japanese, Korean, Thai, and Chinese, which use CER. Lower is better. Forced = language specified; Inferred = language auto-detected. Languages: AR, BN, BG, CA, ZH, CS,

DA, NL, EN, ET, FI, FR, DE, EL, GU, HI, HU, IS, ID, IT, JA, KN, KO, LV, LT, MK, MS, ML, MR, NO, FA, PL, PT, RO, RU, SK, SL, ES, SV, FIL, TE, TH, TR, UK, VI

Diarization Concatenated Minimum Permutation WER — lower is better

[Download PNG](#) [Download SVG](#)

Macro-average concatenated minimum permutation WER (cpWER) over public diarization benchmarks: AMI-IHM, AMI-SDM, Switchboard, Switchboard CallHome, AISHELL-4. For maximum compatibility of API baselines, we use 30 minute versions of AMI and AISHELL. Lower is better. Diarization benchmark, Concatenated Minimum Permutation Word Error Rate (cpWER).

Keyphrase Recall With and without contextual prompting — higher is better

[Download PNG](#) [Download SVG](#)

Without Context

With Context

Voxtral Small Transcribe 3 is highlighted in brand colours. Keyphrase recall on the Earnings 22 benchmarks. The context bias contains keyterms and a randomly sampled number of distractors, ranging from 0 to 100. Bars are superposed absolute values — the outer bar is recall with context, the inner bar is recall without context; the visible band is the uplift from contextual prompting. Higher is better. Keyphrase recall benchmark. Context = domain-specific vocabulary and speaker names supplied to the model.

Artificial Analysis Benchmark Word Error Rate — lower is better

[Download PNG](#) [Download SVG](#)

Artificial Analysis WER Index, a weighted average over VoxPopuli, Earnings-22 and AgentTalk. Lower is better.

Voxtral Realtime Transcribe 3 is a realtime transcription model built for agentic use-cases. It achieves high-quality transcriptions at sub-500ms latency in over 35 languages, including with diarization and context biasing. FLEURS 35 Realtime transcription Word Error Rate — lower is better

[Download PNG](#) [Download SVG](#)

Macro-average WER over 35 languages from the FLEURS benchmark. WER is used for all languages except Japanese, Korean, and Chinese, which use CER. Lower is better. *Voxtral Mini Transcribe 3 is an offline baseline. The remaining models are realtime. Languages: AR, BN, BG, CA, CS, NL, EN, ET, FI, FR, DE, EL, GU, HI, ID, IT, JA, KN, KO, LV, MK, MS, ML, MR, NO, FA, PL, PT, RO, RU, SK, ES, FIL, TE, TR End-of-turn prediction is natively supported, incorporating both acoustic and semantic information to accurately predict when a user is finished speaking. This allows for low-latency voice agent applications, while minimizing false cut-off rates when the user is prematurely interrupted. It answers with the right emotion, in your or your customer's brand's own voice and local dialect. Voxtral Mini TTS 3.0 , available in batch and realtime, achieves state-of-the-art expressiveness and instruction following for 35+ languages and 50+ dialects. At \$13 per 1M characters, Voxtral TTS costs less to run than comparable models from OpenAI TTS-1, Deepgram, ElevenLabs Flash and Gemini Flash. The speech model implicitly interprets the semantic meaning of context better, producing natural speech across a diverse range of emotions and languages. New and improved features include voice cloning from 30+ minutes of sampled audio, custom prompt instruction following (IF), contextual TTS, and a vocabulary library, all of which serve the goal of speech reliability in deployment. Variables for Explicit IF include emotion , pacing , and emphasis .

English French Italian German

{example.title}

--> Prompt Let's end the debate here: it's GIF. Let's end the debate here. It's GIF.

Let's end the debate here. It's GIF.

Default

Non-standard phoneme hint: dʒ ɪ f

{example.title}

--> Prompt Attends... Ce n'est pas possible. Ce tableau vient de Reims ? Mais alors... C'est celui que ma grand-mère cherchait depuis cinquante ans ! Attends, ce n'est pas possible? Ce tableau vient de Reims? Mais alors, c'est celui que ma grand-mère cherchait depuis 50 ans?

Attends, ce n'est pas possible? Ce tableau vient de Reims? Mais alors? C'est celui que ma grand-mère cherchait depuis 50 ans.

Default

Non-standard phoneme hint: ʁ e i m s

{example.title}

--> Prompt Il nostro nuovo assistente si chiama Quivello. Il nostro nuovo assistente si chiama Quivello.

Il nostro nuovo assistente si chiama Quivello.

Default

Non-standard phoneme hint: tʃ i v ε l l o

{example.title}

--> Prompt Hallo! Ich bin Xyra, deine persönliche Assistentin. Hallo, ich bin Xyra, deine persönliche Assistentin.

Hallo, ich bin Xyra, deine persönliche Assistentin.

Default

Non-standard phoneme hint: z aɪ ɪ ə

All natural-sounding speech relies on emotion, which determines its prosody, rhythm, pacing, emphasis, and transitions. Voxtral TTS natively interprets a prompt's text context to determine the emotion and voice style, making the model inherently better at natural speech without requiring additional instructions or voice prompts.

Voice Code Bench

Pronunciation control

Multilingual

Expressiveness

Voice Code Bench: quality per dollar Voice Code Bench made by BeSimple AI is focused on enterprise customer support failure states related to numbers and unique names.

[Download PNG](#) [Download SVG](#)

Other TTS

Mistral

Up and to the left is better. Voice Code Bench (BeSimple AI): 300 English prompts, 2 voices per system (female and male); judge gemini-3.8-flash; prices from vendor pricing pages; generated 2026-10-20.

Pronunciation control: quality per dollar Share of clips where the target word is pronounced as intended, with its phonemes

[Download PNG](#) [Download SVG](#)

Mistral

Other TTS

Internal eval. Up and to the left is better. Systems without a documented pronunciation control are listed in the data, not plotted. 300 English sentences, each with one rare or made-up word (150 easy, 150 hard), 2 voices per system; judge gemini-3.8-flash hears the clip with the word's intended IPA; prices from vendor pricing pages; generated 2026-10-20.

Multilingual: quality per dollar Mean judge score across 11 languages, 1 native voice per language and system

[Download PNG](#) [Download SVG](#)

Other TTS

Mistral

Internal eval. Up and to the left is better. Systems covering fewer than 11 languages are listed in the data, not plotted. 377 prompts selected for release, 11 languages, unweighted mean of per-language mean judge scores, 1 to 5 scale; judge gemini-3.8-flash; prices from vendor pricing pages; generated 2026-10-20.

Expressive English: quality per dollar Mean judge score on one_eval, 237 prompts selected for release

[Download PNG](#) [Download SVG](#)

Other TTS

Mistral

Internal eval. Up and to the left is better. Expressive English (one_eval): 237 prompts selected for release, 2 voices per system (Voxtral TTS: Sarah and James), 1 to 5 scale; judge gemini-3.8-flash; prices from vendor pricing pages; generated 2026-10-20.

Illustrating Voxtral hill-climbing in expressiveness, numbers and unique names.

Customer Support Personal agent Excited Angry

{example.title}

--> Prompt Thank you for calling Mistral! I understand you're having difficulty logging into AI Studio. Can I confirm your email address? I have here elena.vance82@yahoo.com, is that correct? Thank you for calling Mistral I understand you're having difficulty logging into AI Studio. Can I confirm your email address? I have here elena.vance82 @ yahoo.com. Is that correct?

Thank you for calling Mistral. I understand you're having difficulty logging into AI Studio. Can I confirm your email address? I have here elena.vance82 @ yahoo.com. Is that correct?

Voxtral 1

Voxtral 3

{example.title}

--> Prompt Gosh that was quick, I've found a flight for your upcoming travel. How about the Emirates flight at 21:00pm on 10/10 returning 10/25 at 22:00pm for £1881? Say the word, and I'll book a window seat now. Gosh, that was quick. I've found a flight for your upcoming travel. How about the Emirates flight at 21,000pm on 10 over 10, returning 10 over 25 at 22,000pm for £1,881? Say the word and I'll book a window seat now.

Gosh, that was quick. I found a flight for your upcoming travel. How about the Emirates flight at 21pm on October 10th, returning October 25th at 22pm for £1,881? Say the word and I'll book a window seat now.

Voxtral 1

Voxtral 3

{example.title}

--> Prompt Wow I can't believe you're pregnant! This is so exciting, congratulations Corina, I'm so happy for you! Wow, I can't believe you're pregnant. This is so exciting. Congratulations, Karina. I'm so happy for you.

Wow, I can't believe you're pregnant. This is so exciting. Congratulations, Karina. I'm so happy for you.

Voxtral 1

Voxtral 3

{example.title}

--> Prompt What do you mean my passport won't be finished this week -- I've told you several times how I need to go to Paris for work. This is ridiculous. I can't even begin to describe how infuriating this has been. What do you mean my passport won't be finished this week? I've told you several times how I need to go to Paris for work. This is ridiculous. I can't even begin to describe how infuriating this has been.

What do you mean my passport won't be finished this week? I've told you several times how I need to go to Paris for work. This is ridiculous. I can't even begin to describe how infuriating this has been.

Voxtral 1

Voxtral 3

Speech sounds local in 35 languages and over 50 dialects. Voxtral TTS supports over 35 languages and regional dialects. Our multilingual evals adhere to dialect differences in pronunciation, formal vs. informal register, spelling, verbosity, and locale conventions (time, dates, temperatures, addresses, currencies). Voxtral 3 multilingual library

Europe

Americas

Middle East and Africa

Asia-Pacific

English

French

Spanish

Catalan

Your browser does not support the audio element. Dutch

Your browser does not support the audio element. German

Your browser does not support the audio element. Hungarian

Your browser does not support the audio element. Italian

Your browser does not support the audio element. Norwegian

Your browser does not support the audio element. Polish

Your browser does not support the audio element. Romanian

Your browser does not support the audio element. Russian

Your browser does not support the audio element. Turkish

Your browser does not support the audio element. Ukrainian

Your browser does not support the audio element. English

Portuguese

Arabic

Your browser does not support the audio element. Chinese

Hindi

Your browser does not support the audio element. Japanese

Your browser does not support the audio element. Korean

Your browser does not support the audio element. Vietnamese

Your browser does not support the audio element.

British

Your browser does not support the audio element. France

Your browser does not support the audio element. Spain

Your browser does not support the audio element. American

Your browser does not support the audio element. Brazilian

Your browser does not support the audio element. Beijing

Your browser does not support the audio element. Sichuan

Your browser does not support the audio element.

Voice Presets are available in 18 languages: Arabic (Emirati, Modern Standard), Brazilian (Portuguese), Catalan, Chinese (Mandarin), Dutch, English (American, British), French (France), German, Greek, Hindi, Italian, Korean, Norwegian, Russian, Spanish (Spain), and Vietnamese. The voice agent is in control of the whole conversation. Voxtral 3 introduces a voice agent SDK, built on LiveKit, that runs speech in, reasoning, and speech out as one streaming pipeline. Voxtral Mini Transcribe Realtime 3 listens and detects when the caller has finished speaking, a general model—like Mistral Small—reasons and calls tools, and Voxtral TTS replies in the tone of the conversation so far. No stage waits for the one before it, so each turn completes in under one second, and the agent stops talking when the caller cuts in. During a call, the agent can verify a caller's identity, check a bill, or book an appointment, using function calling, web search, and retrieval over your own systems. It answers and places calls over SIP, runs in browsers and apps over WebRTC, and connects to the telephony, IVR, and contact-center software you already use.

Each stage is a separate model, so you can change the LLM, the voice, or the language without rewriting the agent, and run the whole stack on Mistral's API in Europe or on your own infrastructure. Get started. Voxtral 3 is available today through the Mistral API, hosted through our European, US, and global endpoints. You can test out voice agents, audio transcription, speech generation, and voice cloning in the Studio playgrounds. Try Voxtral now | [Learn more](#)

Model Price

Voxtral Mini Transcribe 3 (batch) \$0.0035/min

Voxtral Mini Transcribe Realtime 3 \$0.008/min

Voxtral Small Transcribe 3 \$0.0040/min

Voxtral Mini TTS 3 \$13/1M chars

Voice agent API with Mistral Small 4 for reasoning 0.04\$/min

Voice agent API with larger model 0.08\$/min

To run the full stack on-premises or in a private cloud, talk to our team .

Voxtral Small Transcribe 3 New

Highest accuracy speech transcription model. State-of-the-art multilingual transcription and diarization, with support for over 50 languages and context biasing. Transcription

Audio

\$0.004

/min

Voxtral Mini Transcribe 3 New

Efficient, low-latency speech transcription model. Supports multilingual transcription, diarization and context biasing in over 50 languages. Transcription

Audio (Batch)

\$0.0035

/min

Voxtral Mini Transcribe Realtime 3

A realtime transcription model built for agentic use-cases, delivering high-quality transcriptions at sub-500ms latency across 40+ languages. Transcription

Audio

\$0.008

/min

Voxtral Mini TTS 3

State-of-the-art expressiveness and instruction following across 35+ languages and 50+ dialects Voice

Input

\$13

/1M chars

0%