MAI-Transcribe-1 Model Card

Date: April 2, 2026

Model summary

MAI-Transcribe-1 is a best-in-class speech-to-text model, designed for real-world audio. It provides consistently strong transcription accuracy across accents, speaking styles, and noisy environments, giving developers a strong foundation for building high-quality voice understanding into their applications.

Key capabilities

About this model

MAI-Transcribe-1 is a speech-to-text model built in-house by the Microsoft AI Superintelligence team, designed to deliver reliable transcription across 25 languages. It powers a wide range of use cases, including video captions, meeting transcription, accessibility tools, call analysis, content creation workflows, and powering voice agents. The model is optimized to be robust across diverse accents, dialects, and real-world acoustic conditions, giving developers a transcription system they can rely on.

MAI-Transcribe-1 is actively under development, with new capabilities coming soon, including real-time transcription, diarization and context biasing.

Key model capabilities

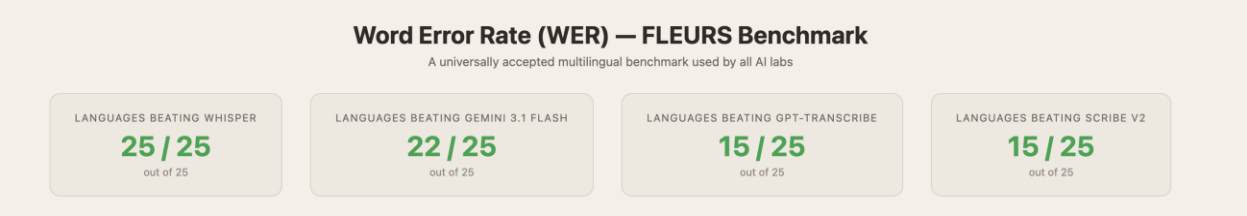
- Best-in-class accuracy across 25 languages: English, French, German, Italian, Spanish, Hindi, Portuguese, Czech, Danish, Finnish, Hungarian, Dutch, Polish, Romanian, Swedish, Japanese, Korean, Chinese, Arabic, Indonesian, Russian, Thai, Turkish, and Vietnamese.
- Robust to real-world noisy situations.
- Automatic Language identification.

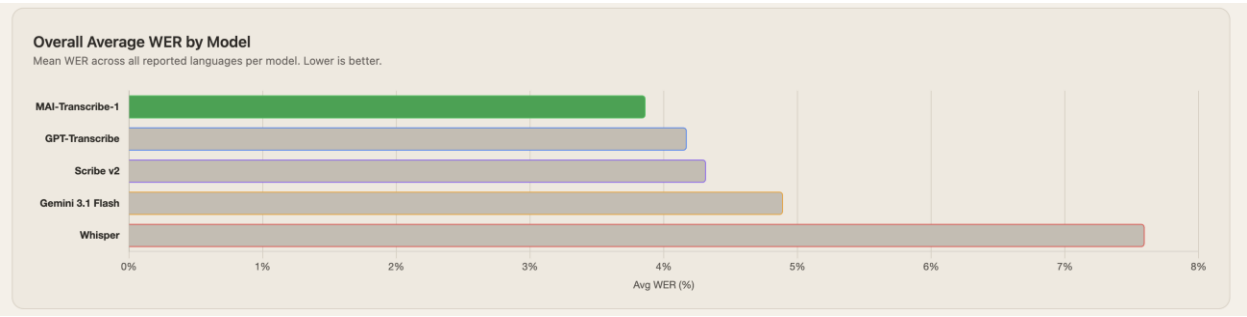
Performance and quality

Best-in-class accuracy on FLEURS

MAI-Transcribe-1 achieves the lowest Word Error Rate against competitive speech-to-text models. On FLEURS (25 languages), it outperforms Scribe v2, Whisper-large-V3, GPT-Transcribe, and Gemini 3.1 Flash.

Word Error Rate (WER) – FLEURS Benchmark





Consistent quality across 25 languages

The model maintains competitively high accuracy across all 25 supported languages, making it adaptable for global products and resilient to a wide range of accents or speaking styles.

Raw WER Data (lower is better)

MODEL	IT	JA	PL	ES	BR	EN	DE	FR	ID	RO	RU	NL	KO	VI	CS	HI	TR	ZH	SV	TH	FI	NB	HU	DA	AR	AVG
MAI-Transcribe-1	1.2%	1.9%	2.1%	2.2%	2.5%	2.7%	2.7%	3.0%	3.0%	3.4%	3.4%	3.5%	3.6%	3.9%	4.0%	4.0%	4.3%	4.5%	4.6%	4.6%	4.7%	5.2%	5.3%	6.2%	10.1%	**3.86%**
Scribe v2	1.3%	2.3%	3.8%	2.7%	3.2%	2.7%	3.5%	3.6%	2.9%	2.9%	4.4%	3.9%	3.3%	3.7%	4.3%	10.2%	4.0%	5.6%	4.9%	6.3%	3.2%	5.4%	5.0%	5.4%	9.4%	**4.32%**
Gemini 3.1 Flash	1.5%	3.9%	4.1%	2.4%	3.0%	3.6%	3.5%	4.4%	3.0%	4.2%	3.6%	4.6%	3.5%	4.0%	6.3%	4.6%	4.6%	6.3%	6.5%	4.9%	5.8%	5.9%	10.0%	8.9%	9.2%	**4.89%**
Whisper	2.2%	5.3%	4.9%	2.8%	4.0%	4.2%	4.1%	5.5%	6.3%	9.6%	5.0%	5.7%	4.6%	9.0%	10.7%	17.5%	6.4%	7.0%	8.0%	8.6%	7.9%	8.8%	13.8%	13.2%	14.8%	**7.60%**
GPT-Transcribe	1.4%	2.8%	2.8%	2.0%	2.6%	2.4%	2.5%	3.1%	3.5%	4.1%	3.2%	3.8%	3.5%	4.0%	4.5%	6.1%	4.1%	5.5%	4.6%	4.3%	3.6%	5.2%	6.9%	6.9%	10.9%	**4.17%**

Use cases

Key use cases

Use case	Scenario	Solution
Live captions	A virtual event platform provides real-time captions for webinars.	Chunk audio and transcribe spoken content into captions displayed live during the event.
Call center transcription	A call center wants accurate, fast transcriptions of customer calls to empower their customer service agents.	Transcribe calls in real time, enabling agents to better understand and respond to customer queries.
Video subtitling	A video-hosting platform needs to generate subtitles for uploaded videos.	Transcribe the full video audio to produce a complete subtitle track.
Accessibility	An organization needs to make audio content accessible to deaf or hard-of-hearing users.	Transcribe audio from meetings, announcements, or media to provide text alternatives that support compliance and inclusive access.
E-learning	An e-learning platform provides transcriptions for video lectures.	Process prerecorded lecture videos, generating text transcripts for students.
Media archiving	A media company needs subtitles for a large archive of videos.	Transcribe video files in bulk, generating accurate subtitles for each video.
Market research	A research firm analyzes customer feedback from audio recordings.	Convert audio feedback into text, enabling easier analysis and insights extraction.

Out of scope use cases

Real-time transcription, diarization, and context biasing aren’t supported yet; these capabilities are planned for an upcoming release.

Pricing

MAI-Transcribe-1 is priced at $0.36 per hour of audio

Input formats

LLM Speech: WAV, MP3, FLAC

Supported languages

English, French, German, Italian, Spanish, Hindi, Portuguese, Czech, Danish, Finnish, Hungarian, Dutch, Polish, Romanian, Swedish, Japanese, Korean, Chinese, Arabic, Indonesian, Russian, Thai, Turkish, and Vietnamese.

Supported Azure regions

Global access enabled, but for now the resources need to point to East US and West US. We'll be scaling out to additional regions soon.

Sample JSON response

{ "durationMilliseconds": 4000, "combinedPhrases": [{ "text": " You transcribe results will appear here" }], "phrases": [] }

Distribution

You can access MAI-Transcribe-1 via Azure Speech.

In some cases, you may not be able to use the Speech SDK. In those cases, you can use REST APIs to access the Speech service. For example, use REST APIs for LLM Speech.

More information

Learn more in the full Azure Speech Service documentation.