

MAI-Transcribe-2

Date: September 3, 2026

Model summary

The second generation of our speech-to-text model family: more accurate, faster, and built to handle a wider range of real-world audio. MAI-Transcribe-2 transcribes reliably across accents, speaking styles, and noisy environments, and now covers 60 languages with automatic language detection for multilingual recordings. This generation adds speaker diarization and word-level timestamps, along with keyword biasing for domain-aware transcription that better recognizes industry and scientific terms, proper names, and other domain-specific terminology.

Overview

About this model

MAI-Transcribe-2 is a speech-to-text model built in-house by the Microsoft AI team, delivering reliable transcription across 60 languages. It powers video captions, meeting transcription, clinical and call center documentation, accessibility tools, content creation workflows, and voice agents. Beyond raw transcription, this generation adds speaker diarization to separate and label multiple speakers in a single recording, and word-level timestamps that pinpoint exactly who spoke and when. Keyword biasing makes transcription domain-aware, improving recognition of specialized vocabulary.

Key capabilities

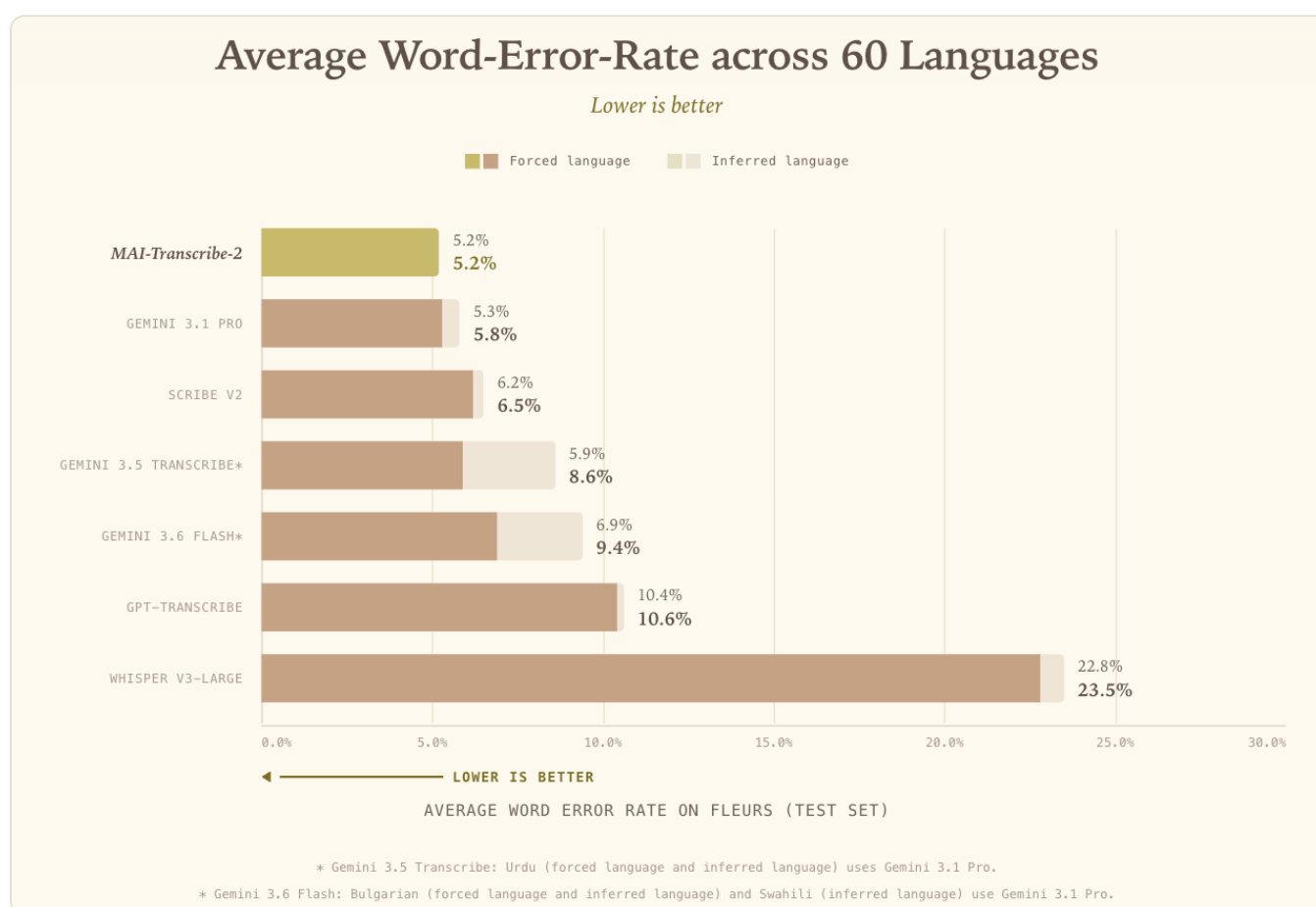
- Best-in-class transcription accuracy across 60 languages.
- Speaker diarization: identifies and separates multiple speakers within the same audio recording.
- Word-level timestamps: precise timing for every word, enabling alignment, search, and editing workflows.
- Robust in noisy, real-world conditions.
- Faster inference with substantially lower latency, especially for long-form audio, with up to 10× faster processing than competitors.
- Automatic language identification.
- Keyword biasing to handle domain specific terminology, abbreviations or disambiguating terms.
- Configurable transcription style: **verbatim** captures speech exactly as spoken, including fillers and false starts, for compliance and analysis workloads; **clean** removes disfluencies for readable captions, notes, and published transcripts.

Performance and quality

Consistently high quality across 60 languages

MAI-Transcribe-2 transcribes with high accuracy across all 60 supported languages, making it particularly interesting for global products. This is reflected by the FLEURS benchmark where MAI-Transcribe-2 achieves the lowest average word error rate (WER) across languages when compared to comparable transcription systems, without needing the language to be specified in advance.

FLEURS Benchmark results (test split, inferred language, across 60 languages):



Improved noise robustness

Even on common languages like English, there were notable improvements driven by targeted quality gains and better noise robustness. This is demonstrated by the English-only speech-to-text benchmark of Artificial Analysis (AA), where MAI-Transcribe-2 ranks #2 on the [Artificial Analysis speech-to-text leaderboard](#).

Boosting transcription quality via keyword biasing

Transcribing and formatting acronyms or names correctly is very challenging. MAI-Transcribe-2 features keyword biasing which allows users to improve transcription accuracy by informing the model about domain-specific terminology or names. By adding important terms and names to the model's keyword context, we

ensure that the model detects them in the audio and transcribes them correctly.

Faster long-form transcription

Thanks to improvements on how the model handles long audios, it transcribes long-form much faster, for example, 1 hour of audio in about 10 secs.

Use cases

Key use cases

Use case	Scenario	Solution
Contact center transcription	A contact center needs accurate transcripts of every customer call for quality assurance, agent coaching, and compliance.	Transcribe calls with diarization to separate agent and customer turns, and apply keyword biasing for product, policy, and account terminology.
Clinical note taking	A healthcare provider wants clinicians to spend less time on documentation after patient visits.	Transcribe consultations with keyword biasing for medications, procedures, and clinical terms, using diarization to distinguish clinician from patient before generating structured notes. (example)
Meeting transcription	An enterprise wants searchable, attributed records of internal meetings.	Transcribe recordings with speaker labels and word-level timestamps, producing transcripts ready for summarization and action-item extraction.
Sales conversation intelligence	A revenue team analyzes sales calls to track objections, competitor mentions, and coaching opportunities.	Transcribe calls with speaker separation and word-level timing so keywords and talk-time ratios can be measured per participant.
Voice agent transcripts and evaluation	A team operating voice agents needs reliable transcripts to evaluate, debug, and improve agent behavior.	Transcribe recorded agent sessions with speaker labels and timestamps to score task completion, interruptions, and error cases. (example)
Video subtitling and localization	A video platform needs subtitles for uploaded content and a base for translation.	Transcribe full video audio with word-level timestamps to produce precisely aligned subtitle tracks ready for localization.
Accessibility and compliance	An organization must make audio content accessible to deaf and hard-of-hearing users.	Transcribe meetings, announcements, and media to provide text alternatives that support accessibility regulations and inclusive access.
Media archiving and search	A media company needs searchable metadata across a large video and audio archive.	Transcribe files in bulk, generating time-aligned, speaker-attributed transcripts that make archives searchable down to the word.

Use case	Scenario	Solution
E-learning and lecture capture	An education platform provides transcripts and captions for course content.	Process prerecorded lectures into text transcripts, using keyword biasing to capture technical and scientific terminology.
Podcast and content production	A studio needs transcripts to power show notes, clips, and text-based audio editing.	Generate speaker-labeled transcripts with word-level timing so editors can cut, quote, and repurpose audio directly from text.
Legal and financial documentation	A legal or financial services provider needs accurate records of hearings, depositions, and advisory calls.	Transcribe with keyword biasing for legal and financial terminology and diarization to attribute each statement to the correct speaker.
Dictation and voice notes	A productivity application lets users capture notes, emails, and reports by voice.	Transcribe short-form audio quickly and accurately, with keyword biasing for user-specific names and jargon.
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.
Market research	A research firm analyzes customer feedback from interviews and focus groups.	Convert recordings into speaker-separated text, enabling easier analysis and insights extraction.
Multilingual content operations	A global team processes recordings in many languages without knowing the language in advance.	Rely on automatic language identification to route and transcribe each file across 60 supported languages.

Pricing

MAI-Transcribe-2 will be priced at \$0.10 per hour as a limited-time offer.

Input formats

WAV, MP3, FLAC

Max file size: 300 MB / 2 hours.

Supported languages

Afrikaans, Arabic, Armenian, Assamese, Azerbaijani, Bengali, Bosnian, Bulgarian, Cantonese, Catalan, Chinese, Czech, Danish, Dutch, English, Estonian, Filipino, Finnish, French, Galician, German, Greek, Gujarati, Hebrew, Hindi, Hungarian, Icelandic, Indonesian, Italian, Japanese, Kannada, Kazakh, Korean,

Latvian, Lithuanian, Macedonian, Malay, Malayalam, Marathi, Nepali, Norwegian Bokmål, Odia, Persian (Farsi), Polish, Portuguese, Punjabi, Romanian, Russian, Slovak, Slovenian, Spanish, Swahili, Swedish, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, Vietnamese.

Per-language word error rate

Lower is better

FLEURS test set, forced language · All 60 languages · The lowest score in each language is shown in bold.

Language	MAI-Transcribe-2	Gemini 3.1 Pro	Gemini 3.5 Transcribe*	Scribe v2	Gemini 3.6 Flash*	GPT-transcribe	Whisper v3-large
Afrikaans AF	13.1%	11.7%	12.7%	10.6%	16.4%	18.8%	33.0%
Arabic AR	8.7%	9.8%	8.9%	9.6%	10.1%	10.3%	14.7%
Armenian HY	12.6%	6.1%	6.9%	6.1%	8.9%	11.2%	48.1%
Assamese AS	8.7%	8.7%	10.3%	11.4%	11.1%	21.5%	108.9%
Azerbaijani AZ	9.4%	6.6%	7.4%	9.1%	8.2%	10.9%	21.3%
Bengali BN	4.8%	5.1%	5.0%	6.9%	6.0%	8.9%	46.4%
Bosnian BS	4.7%	4.8%	14.1%	4.9%	6.9%	53.4%	13.5%
Bulgarian BG	5.1%	4.5%	5.2%	4.5%	4.5%	6.7%	13.1%
Cantonese YUE	6.6%	5.6%	9.9%	21.2%	13.1%	7.8%	21.9%
Catalan CA	2.8%	2.6%	3.4%	2.9%	5.3%	3.8%	5.0%
Chinese ZH	4.5%	9.2%	6.1%	5.6%	8.9%	6.2%	7.0%
Czech CS	3.9%	3.3%	5.3%	4.3%	6.1%	5.3%	10.7%
Danish DA	5.4%	5.3%	7.5%	5.3%	9.3%	8.2%	13.2%
Dutch NL	3.6%	3.4%	3.8%	3.7%	4.8%	4.0%	5.7%
English EN	2.7%	2.6%	2.8%	2.6%	3.0%	3.2%	4.2%
Estonian ET	5.9%	4.6%	6.6%	6.6%	7.1%	8.3%	18.8%
Filipino FIL	5.9%	7.7%	6.3%	7.6%	10.9%	8.9%	11.0%
Finnish FI	3.9%	3.1%	5.4%	3.1%	5.5%	4.9%	7.9%
French FR	2.8%	2.9%	3.7%	3.6%	3.7%	3.7%	5.4%
Galician GL	3.9%	4.1%	4.8%	4.4%	5.9%	6.5%	13.6%
German DE	2.8%	2.8%	2.9%	3.6%	3.3%	3.2%	4.1%
Greek EL	4.2%	5.1%	7.0%	4.8%	9.6%	6.5%	12.2%
Gujarati GU	4.9%	5.3%	5.4%	6.5%	6.1%	12.6%	28.0%
Hebrew HE	13.7%	13.9%	18.0%	16.0%	19.7%	22.2%	26.8%
Hindi HI	3.7%	4.3%	4.0%	9.5%	4.7%	6.2%	17.4%
Hungarian HU	5.1%	5.0%	7.6%	4.9%	8.6%	7.0%	13.8%
Icelandic IS	4.9%	4.3%	7.6%	6.1%	6.4%	15.6%	32.9%
Indonesian ID	2.5%	2.8%	2.5%	2.9%	3.5%	4.5%	6.3%
Italian IT	1.1%	1.4%	1.1%	1.3%	1.6%	1.9%	2.2%
Japanese JA	1.8%	2.5%	3.1%	2.3%	3.7%	3.9%	5.3%
Kannada KN	4.2%	6.5%	5.2%	7.2%	6.3%	9.5%	26.1%
Kazakh KK	5.7%	4.6%	5.6%	5.6%	7.3%	22.2%	33.1%
Korean KO	3.4%	3.3%	3.8%	3.3%	4.1%	4.0%	4.6%
Latvian LV	5.6%	3.5%	5.1%	3.9%	5.6%	7.4%	19.2%
Lithuanian LT	6.4%	5.6%	7.5%	6.5%	10.5%	11.4%	25.3%
Macedonian MK	5.3%	3.7%	3.6%	4.0%	5.3%	6.5%	15.4%
Malay MS	4.0%	3.1%	3.6%	3.8%	4.3%	6.3%	7.9%
Malayalam ML	5.0%	5.6%	5.5%	8.6%	6.7%	8.4%	103.6%
Marathi MR	4.9%	5.6%	5.1%	7.7%	6.7%	9.3%	34.4%
Nepali NE	8.5%	9.9%	8.6%	11.2%	11.2%	16.8%	40.9%
Norwegian Bokmål NB	4.5%	4.7%	5.9%	5.4%	7.8%	5.5%	8.9%
Odia OR	6.0%	6.4%	8.4%	8.4%	7.9%	20.3%	106.1%
Persian FA	10.3%	16.0%	8.8%	10.5%	10.3%	10.0%	30.3%
Polish PL	2.0%	2.8%	3.3%	3.8%	4.7%	3.3%	4.9%
Portuguese PT	2.5%	2.8%	2.7%	3.1%	3.2%	3.4%	4.0%
Punjabi PA	6.4%	6.1%	7.2%	10.6%	8.1%	38.7%	53.2%
Romanian RO	3.1%	3.0%	3.0%	2.7%	4.6%	4.2%	9.5%
Russian RU	3.1%	3.5%	3.4%	4.2%	3.6%	4.3%	5.0%
Slovak SK	3.0%	2.8%	3.8%	3.1%	3.9%	4.7%	9.5%
Slovenian SL	5.0%	6.0%	7.8%	6.9%	9.3%	9.8%	19.6%
Spanish ES	2.1%	2.3%	2.2%	2.6%	2.7%	2.3%	2.8%
Swahili SW	11.3%	10.5%	9.2%	11.2%	11.6%	19.8%	35.4%
Swedish SV	4.0%	4.2%	5.4%	4.8%	6.3%	5.9%	8.0%
Tamil TA	5.2%	8.7%	6.7%	10.9%	8.8%	12.0%	19.1%
Telugu TE	6.7%	7.1%	7.4%	11.8%	8.2%	11.3%	99.2%
Thai TH	3.4%	4.7%	3.8%	6.1%	5.1%	5.4%	8.7%
Turkish TR	3.9%	4.1%	4.3%	4.0%	4.5%	4.6%	6.4%
Ukrainian UK	3.2%	3.2%	3.6%	3.1%	6.1%	4.5%	6.5%
Urdu UR	13.3%	12.1%	12.1%	16.3%	12.7%	131.2%	22.3%
Vietnamese VI	3.5%	2.9%	3.2%	3.6%	3.9%	4.4%	8.9%

* Gemini 3.5 Transcribe: Urdu uses Gemini 3.1 Pro. * Gemini 3.6 Flash: Bulgarian and Swahili use Gemini 3.1 Pro.
Source: MAI-Transcribe FLEURS release comparison, forced language

Sample JSON response

```
{
  "durationMilliseconds": 4000,
  "combinedPhrases": [
    {
      "text": "Gotcha.\nThat makes sense.\nDifferent classification.\nYeah, they they kind of..."
    }
  ],
  "phrases": [
    {
      "speaker": 0,
      "offsetMilliseconds": 80,
      "durationMilliseconds": 339,
      "text": "Gotcha.",
      "locale": "en",
      "confidence": 0
    },
    {
      "speaker": 1,
      "offsetMilliseconds": 800,
      "durationMilliseconds": 460,
      "text": "That makes sense.",
      "locale": "en",
      "confidence": 0
    },
    {
      "speaker": 0,
      "offsetMilliseconds": 1360,
      "durationMilliseconds": 860,
      "text": "Different classification.",
      "locale": "en",
      "confidence": 0
    },
    {
      "speaker": 1,
      "offsetMilliseconds": 2320,
      "durationMilliseconds": 1200,
      "text": "Yeah, they they kind of...",
      "locale": "en",
      "confidence": 0
    }
  ]
}
```

Supported Azure regions

MAI-Transcribe-2 can be accessed globally. The model is currently served from following regions

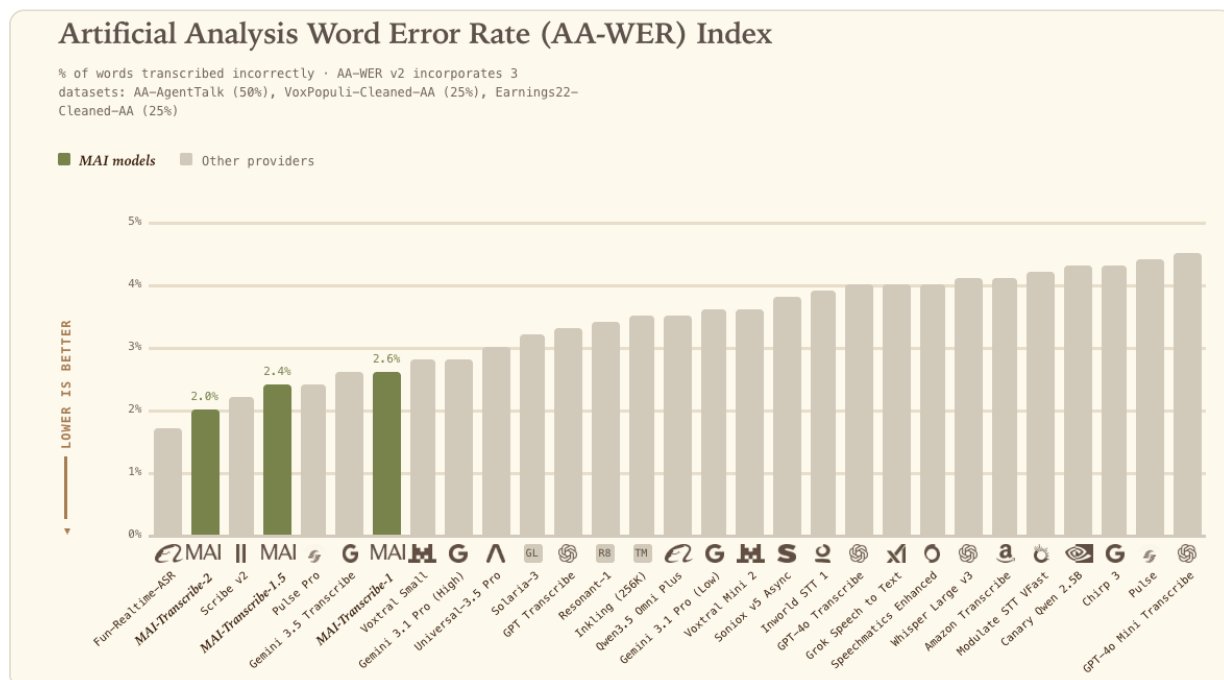
- East US
- West US
- West US2
- North Europe
- Central India
- Southeast Asia

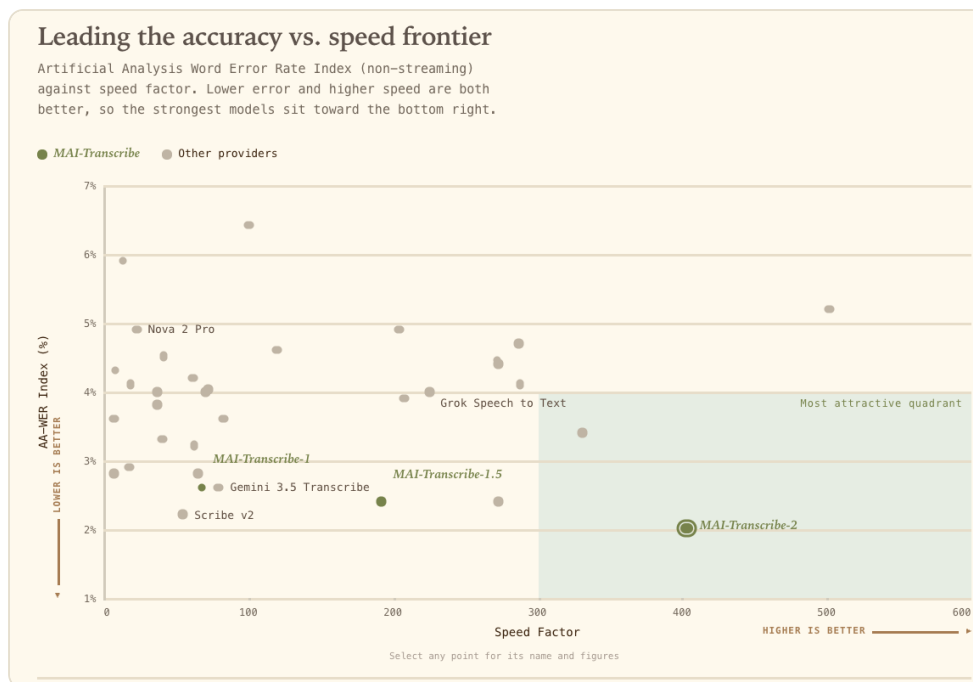
to which the requests are routed.

Distribution

You can access MAI-Transcribe-2 via Azure Speech Fast API or via OpenRouter (coming soon).

Artificial Analysis Evaluations





More information

Learn more in the full [Azure Speech Service documentation](#).