

M3 · RESEARCH WHITE PAPER

Multilingual meeting intelligence.

The M3 system powers MeetriX: speech, participant context and a traceable record of the conversation.



Capture

Speech in context

Connect

People, terms and time

Review

A source-linked record

System design, meeting workflows and measured Arabic and English transcription.

01 / ABSTRACT

Measured in Arabic. Measured in English.

A multilingual system, with transcription evaluated on real business meetings.

1.62%

Arabic CER
After lexical normalization

1.83%

English CER
After English scoring rules

AR + EN

Separate evaluations
Matched inputs within each language

Speech that stays connected to its context

M3 brings multilingual transcription, available participant identity, timing and enterprise terminology into one meeting record. MeetriX can use that record to prepare summaries, decisions and follow-up actions.

M3 recorded the lowest character error rate against every external ASR configuration in each language-specific evaluation. Google Chirp 3 was the closest comparison: **2.26% in Arabic** and **3.18% in English**, versus 1.62% and 1.83% for M3.

28%–83% lower Arabic CER

Relative reduction versus the external configurations on the Arabic evaluation material.

42%–77% lower English CER

Relative reduction versus the external configurations on the English evaluation material.

From system design to inspectable evidence

Product scope and meeting architecture	03–08	Inputs, configurations and scoring	09–12
Results and concrete transcripts	13–22	Operation and evidence boundaries	23–26

Read each result with its method. Arabic and English use separate material and language-specific normalization. This comparison measures transcription; external speaker diarization was not tested. [Read the evaluation protocol ↗](#)

02 / MULTILINGUAL SCOPE

Built for real meetings. Across languages.

Language, speaker identity and enterprise context belong in the same meeting record.

Real meetings move between short replies, overlapping voices, dialects and business vocabulary. A useful transcript keeps those contributions connected to the people, products and decisions being discussed.

40 languages

32 Arabic dialects

MeetriX product coverage. This paper evaluates Arabic and English transcription separately.

Languages and dialects

Automatically select the primary meeting language and handle Arabic–English switches within one M3 system.

Overlapping speakers

Keep the available participant identity connected to each contribution as the conversation moves.

Enterprise vocabulary

Recognise business terminology and resolve references to people, products and platforms.

Conversational detail

Preserve brief replies and words at speech boundaries, with links back to their original audio.

What

Spoken content and enterprise terminology

Who

Available participant identity and context

When

Meeting time and the original audio

A connected meeting record makes company knowledge easier to review, find and use.

02 / SYSTEM DESIGN

Eleven components. One M3 system.

Three preprocessing models and eight connected components carry speech from capture to the meeting report.

01 Capture with speaker identity

Channel separation

Keep participant audio distinct.

01

Speaker diarization

Carry participant-linked identity.

02

02 Prepare the speech

3 preprocessing models

Speech & silence

Find spoken content.

03

Short turns

Keep brief contributions.

04

Speech boundaries

Protect words at the edges.

05

03 Recognise speech in context

Transcription

Turn speech into text.

06

Terminology

Unify business terms.

07

Entity matching

Resolve people and products.

08

04 Produce a traceable record

Source alignment

Connect speaker, time and audio.

09

Post-meeting optimization

Refine the meeting record.

10

Per-speaker reporting

Organise each contribution.

11

Functional architecture. Speech & silence, short turns and speech boundaries describe proposed preprocessing roles, not official model names. Available platform channels and speaker-activity information determine how participant identity is established.

Context stays connected.

Speaker identity, meeting time, enterprise terminology and source audio remain linked through live and post-meeting workflows.

03 / ENTERPRISE CAPABILITIES

The complete meeting record.

Transcription quality creates value when the words remain connected to people and context.

MEASURED OUTPUT

Accurate Arabic and English transcripts

Lower normalized transcription error supports a clearer record for review, search and analysis. Human correction time was not measured.

Evidence. M3 achieved 1.62% Arabic CER and 1.83% English CER, the lowest value in each matched-language comparison. The results chapters provide the complete tables.

MEASURED OUTPUT

Cleaner delivered text

Delivered transcripts avoid many of the filler-like insertions, repeated punctuation patterns, and repetition loops that reduce readability.

Evidence. Observed in the reviewed outputs from the Arabic evaluation material; examples are provided in the accompanying comparison page.

MEASURED OUTPUT

Participant-linked records

Decisions, statements, and follow-ups can remain associated with the available meeting participant rather than becoming anonymous text.

Evidence. Platform-provided speaker attribution remained correctly linked throughout the evaluated Arabic and English meetings.

MEASURED OUTPUT

Searchable terminology

Normalised business and technical terms improve enterprise search, knowledge retrieval, and analytics across many meetings.

Evidence. Terminology handling was reviewed on the Arabic evaluation material and is illustrated in the matched-audio examples.

The practical result. The reviewed outputs connect a lower transcription error rate to details that make a record usable: spoken content, participant context and consistent terminology. Human correction time was not measured.

03 / ENTERPRISE CAPABILITIES

From delivery to controlled deployment.

A configurable meeting workflow, with separate evidence for each operating claim.

DELIVERY AND OPERATION

Post-meeting summary and delivery

Participants receive a shared record of the topic, attendees, summary, key points, recommendations, and next steps without someone manually composing the follow-up email.

Evidence. Product delivery workflow described in the meeting-intelligence chapters; not part of the CER benchmark.

DELIVERY AND OPERATION

Fast batch processing

Large meeting backlogs can be processed quickly, and completed meetings can become available while they are still operationally relevant.

Evidence. Batch transcription processes an hour of recorded audio in 45 seconds on one L4 GPU. This is a recorded-audio throughput measurement.

DELIVERY AND OPERATION

Concurrent meeting capacity

Multiple meetings can share a single accelerator within the operating configuration.

Evidence. A separate operating measurement supports 82 concurrent meetings on one L4 GPU. Capacity and batch speed were measured separately.

DELIVERY AND OPERATION

Customer-controlled processing

Sensitive financial, contractual, employee, and customer conversations can remain within a controlled deployment boundary.

Evidence. The evaluated transcription path called no external transcription service.

Separate dimensions. Transcription error, qualitative output behavior, speaker-context continuity and operating performance describe different aspects of M3. Summary quality and operational value require their own evaluation; CER is not a measure of employee hours saved.

04 / MEETING INTELLIGENCE

From transcript to participant follow-up.

A structured meeting record makes the conversation easier to review and act on.

After transcription, M3 combines the transcript with available participant identities and the meeting timeline. MeetriX uses that structured record to prepare a concise summary and the participant follow-up.



What the structured summary can contain

Meeting context Title, date and language	Participants Attendees and available participant details
Summary A concise account of the discussion	Key points Important topics and supporting context
Decisions Recommendations and agreed direction	Next steps Action items and follow-up work

One record, a configurable delivery workflow

The summary template can match the organisation's workflow. The recipient list is derived from meeting participant information, so the identity context used in the transcript also identifies the intended follow-up recipients.

Delivery policy remains explicit. Email distribution to registered participants follows the organisation's access, consent and retention policies. Summary content and delivery are product workflow capabilities, separate from the transcription benchmark.

05 / ENTERPRISE DATA

A record that stays connected to its context.

Meeting knowledge becomes more useful when people, terminology and source audio stay linked.

A flat text transcript loses part of the meeting. M3 preserves relationships between the meeting, original audio segment, available participant identity, timeline position, transcript and generated summary.

The connected meeting record

Audio + participant + time + transcript + summary

Find

Consistent terminology supports knowledge retrieval.

Trace

Return from a statement to its source audio.

Use

Feed approved knowledge and workflow systems.

PEOPLE AND DECISIONS

Accountability

Keep decisions and commitments connected to the participant and moment from which they came.

TOPICS AND PATTERNS

Operational insight

Analyse topics, recurring issues, customer signals and meeting activity over time.

TERMINOLOGY AND SEARCH

Knowledge retrieval

Canonical business terms make the record easier to find in search and retrieval workflows.

DOWNSTREAM WORK

Approved automation

Structured records can feed a CRM, case record, knowledge base or compliance archive without reconstructing context from flat text.

Customer-controlled deployment. The evaluated transcription path made no external transcription call. Output formats and downstream integrations are configured for each deployment; storage, access, retention, data residency and consent remain deployment responsibilities.

06 / EVALUATION MATERIAL

Language-specific material.

A shared reference within each test.

Each language has its own source recordings. Every compared system is scored against the same references within that language.

Evaluation	Material	What it supports
Arabic	Real, multi-dialect Arabic business meetings with English technical terms.	Normalized Arabic CER and qualitative review of the matched system outputs.
English	Real English-language business meetings with human-verified references.	English CER and a separate qualitative review of the matched outputs.
Product workflow	Participant context, timestamps, structured records and delivery.	Descriptions of meeting capabilities; these are not extra CER results.

Reference preparation

A human reviewer checked each scored reference against the source audio. The reference retains spoken content, names, numbers and business terminology. Only segments with a verified reference enter the CER calculation. Non-speech material can inform diagnostic review without becoming an additional scored speech result.

Evaluation-data separation

The meetings and human-verified references used in both the Arabic and English evaluations were not used to train M3. References were used only for scoring and were not supplied to any system during transcription.

Separate comparisons. Arabic and English use different source material. Neither score establishes performance in every supported language, dialect, acoustic environment or organization.

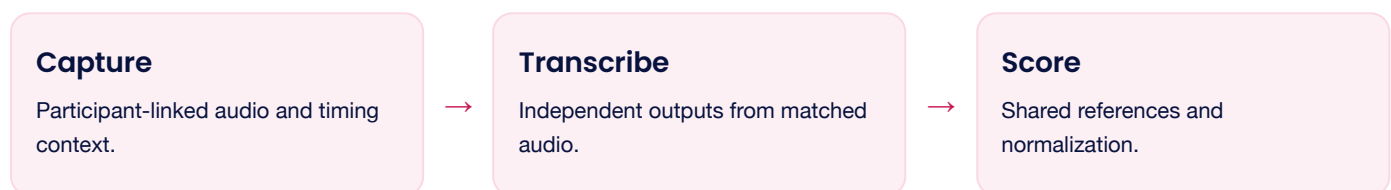
[Read the evaluation protocol ↗](#)

06 / COMPARISON DESIGN

Hold the audio constant. Isolate transcription quality.

M3 is a meeting system compared with external ASR configurations, including distinct Google configurations.

The M3 system and the external ASR configurations received the same source audio and segment boundaries within each language evaluation. Every system was explicitly configured for Arabic or English, as appropriate, and scored against the same human-verified reference. Automatic language selection was evaluated separately from CER. External results came from API runs on the evaluation audio.



Participant identity is not diarization accuracy

The capture workflow already supplied participant-separated audio. External speaker diarization—the step that separates a mixed recording by who spoke—was not used or evaluated. M3 kept every speaker correctly linked to their transcript throughout the evaluated Arabic and English meetings. This checks continuity of available identity context, not independent speaker recognition.

Short utterances stay inside the review

Brief acknowledgements, greetings and short questions can carry a business decision. A speech-bearing reference paired with an empty output contributes deletions; unsupported additional output contributes insertions. Reviewed examples make these behaviors visible without creating a separate aggregate short-utterance score.

Fair input boundary. The comparison holds participant channels, source segments and reference rules constant. It does not measure how the external systems would separate speakers from mixed audio.

06 / CONFIGURATIONS AND ENGLISH SCORING

Explicit language settings.

Clear scoring rules.

The published configurations define the benchmark; they do not establish a universal ranking of providers.

Evaluated entry	Arabic setting	English setting
M3	Arabic explicitly selected; participant-linked capture	M3 + soxr + 1 s silence; <code>--language en</code>
Google Chirp 3	<code>chirp_3 · ar-XA</code>	<code>chirp_3 · en-US</code>
Google Latest Long	<code>latest_long · ar-SA</code>	<code>latest_long · en-US</code>
ElevenLabs Scribe v1	<code>scribe_v1 · ara</code>	<code>scribe_v1 · eng</code>
Cohere Transcribe Arabic	<code>cohere-transcribe-arabic-07-2026 · language=ar</code>	<code>cohere-transcribe-arabic-07-2026 · language=en</code>
Deepgram Nova-3	Nova-3; Arabic explicitly selected	<code>nova-3 · en</code>
Sonix	Arabic explicitly selected	<code>en</code>

English normalization

English CER counts character substitutions, deletions and insertions against the reference. Punctuation, letter case and whitespace are ignored, except that a missing space which fuses separate words counts as one error. Remaining spelling differences count. The Arabic lexical-equivalence policy is not applied to the English comparison.

Why Scribe v1 remains in the comparison

The same Arabic evaluation audio was rerun with Scribe v2 under the same input and scoring conditions. It did not improve accuracy on this material; v1 performed slightly better. The table therefore retains v1. This observation does not rank the model versions for other workloads.

06 / ARABIC SCORING

Arabic: define equivalence before counting errors

Arabic lexical normalization is part of the Arabic result. English uses the separate rules stated in its configuration chapter.

CER after defined lexical normalisation. Character error rate is calculated as substitutions plus deletions plus insertions, divided by the number of characters in the human reference. Equivalent ways of expressing the same lexical content are treated consistently before character-level alignment:

CER after the defined normalisation steps

$$\text{CER} = \frac{\text{substitutions} + \text{deletions} + \text{insertions}}{\text{characters in the human reference}} \times 100$$

- 01 An Arabic word is counted as correct when it is transcribed in either a dialectal or Modern Standard Arabic form that carries the same meaning.
- 02 An English word is counted as correct whether it is written in Latin characters or represented phonetically in Arabic script.
- 03 An English word translated into an equivalent Arabic word is also counted as correct.
- 04 Equivalent numeric expressions are counted as the same value whether they are written with Western digits, Arabic-Indic digits, or words; for example, 15 , ١٥ , and خمسة عشر are treated as equivalent.
- 05 Punctuation marks are excluded from error scoring, whether they are present, absent, or written differently.

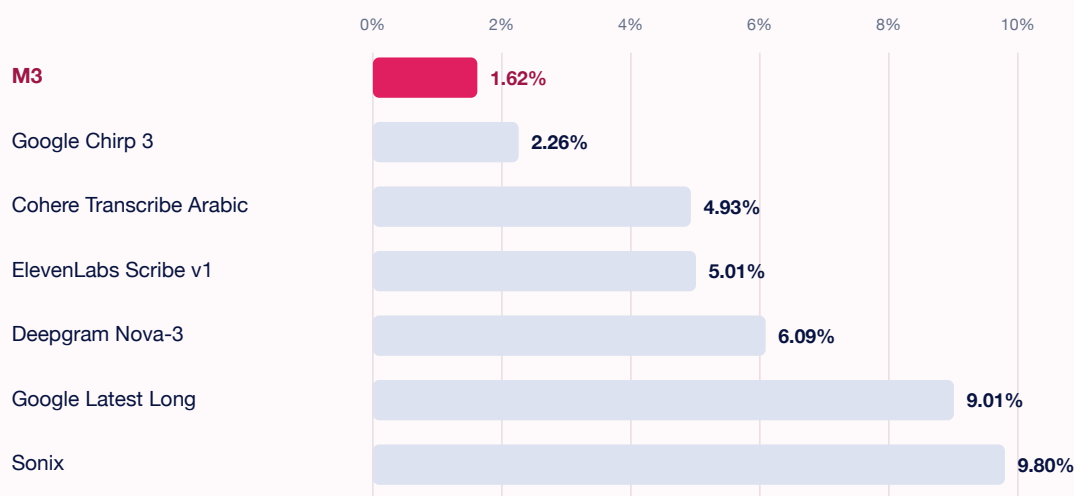
Arabic letter variants such as the different forms of alef are also normalised consistently so that equivalent forms are treated as the same character. Latin text is case-insensitive, so uppercase and lowercase differences are not counted as errors. Whitespace is otherwise excluded from error scoring; however, when a missing word-boundary space joins two separate words, that missing boundary is counted as an error. After all accepted equivalence rules and normalisation steps are applied, every remaining missing, additional, or substituted character is counted as an error. Numbers and English business terms remain part of the evaluated content.

07 / ARABIC RESULTS

Arabic transcription measured against the same inputs

Character error rate on matched Arabic recordings and human-verified references. Lower is better.

ARABIC BENCHMARK · CHARACTER ERROR RATE



Evaluated entry	CER	Character accuracy	Gap vs M3 (points)	M3 relative reduction
M3	1.62%	98.38%	—	—
Google Chirp 3	2.26%	97.74%	0.64	28%
Cohere Transcribe Arabic	4.93%	95.07%	3.31	67%
ElevenLabs Scribe v1	5.01%	94.99%	3.39	68%
Deepgram Nova-3	6.09%	93.91%	4.47	73%
Google Latest Long	9.01%	90.99%	7.39	82%
Sonix	9.80%	90.20%	8.18	83%

Character accuracy = 100 – CER. Gap and relative reduction compare each entry with M3; relative reductions are rounded to whole percentages. Arabic lexical-equivalence rules apply.

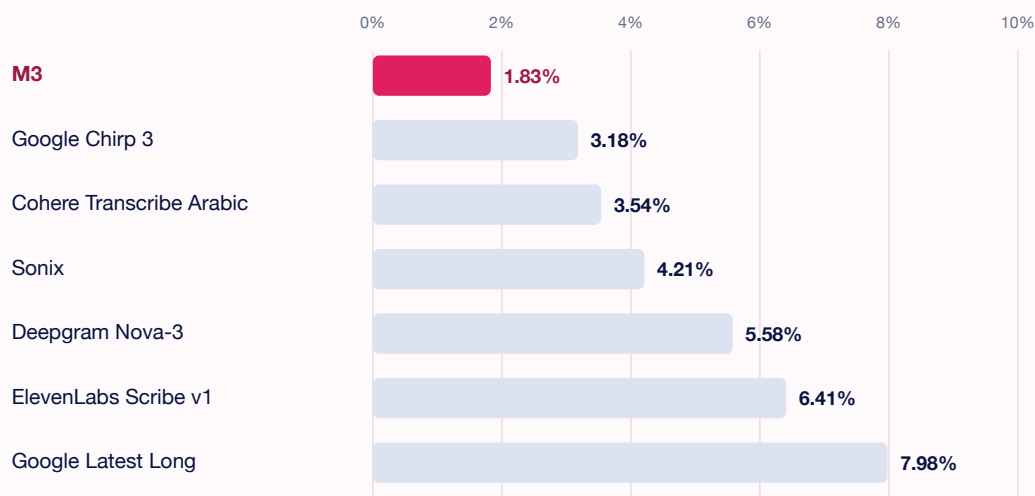
Within-language comparison. Arabic and English use separate source material. These results compare systems within each language; they do not rank one language against the other.

07 / ENGLISH RESULTS

English transcription measured against the same inputs

Character error rate on matched English recordings and human-verified references. Lower is better.

ENGLISH BENCHMARK · CHARACTER ERROR RATE



Evaluated entry	CER	Character accuracy	Gap vs M3 (points)	M3 relative reduction
M3	1.83%	98.17%	—	—
Google Chirp 3	3.18%	96.82%	1.35	42%
Cohere Transcribe Arabic	3.54%	96.46%	1.71	48%
Sonix	4.21%	95.79%	2.38	57%
Deepgram Nova-3	5.58%	94.42%	3.75	67%
ElevenLabs Scribe v1	6.41%	93.59%	4.58	71%
Google Latest Long	7.98%	92.02%	6.15	77%

Character accuracy = 100 – CER. Gap and relative reduction compare each entry with M3; relative reductions are rounded to whole percentages. English case, punctuation and whitespace rules apply; remaining spelling differences count.

Within-language comparison. Arabic and English use separate source material. These results compare systems within each language; they do not rank one language against the other.

08 / ENGLISH QUALITATIVE REVIEW

How the English outputs behaved

These review levels describe observed output behavior. They are not additional error-rate measurements.

Criterion	M3	Cohere Transcribe Arabic	ElevenLabs Scribe v1	Deepgram Nova-3	Google Latest Long	Google Chirp 3	Sonix
Overall English transcription	Strong	Strong	Weak	Mixed	Weak	Strong	Mixed
Short utterances	Strong	Strong	Mixed	Mixed	Weak	Mixed	Mixed
Text omission	Strong	Mixed	Strong	Mixed	Mixed	Strong	Weak
Punctuation	Mixed	Mixed	Weak	Mixed	Mixed	Mixed	Mixed
Names, numbers and terminology	Strong	Strong	Mixed	Weak	Mixed	Mixed	Weak
Word repetition	Strong	Strong	Weak	Weak	Mixed	Mixed	Mixed
Speaker identification	Strong	Not rated	Not rated	Not rated	Not rated	Not rated	Not rated
Unsupported fillers and insertions	Strong	Strong	Weak	Mixed	Mixed	Mixed	Mixed

How to read the levels. Strong indicates a clear positive observation; Mixed a qualified result; Weak a recurring or concentrated defect; Not rated means the criterion was not evaluated. Only the overall-transcription row uses CER bands: Strong at $\leq 4\%$, Mixed above 4% and below 6%, Weak at $\geq 6\%$.

The other rows summarize the reference-error review, including approximate token alignment. They do not establish statistical significance or independent aggregate rates. M3's Strong speaker rating reflects correct participant-to-transcript linkage throughout the evaluated English meetings. External speaker identification was not evaluated; the comparison uses participant-separated clips.

08 / ENGLISH TRANSCRIPT EXAMPLE

Names that remain recognisable in the record.

English example · Core42 and Compass in a brief contribution.

HUMAN-VERIFIED ENGLISH REFERENCE

access to core 42 compass models, right?

Evaluated entry	Exact transcript
M3	access to core 42 compass models, right?
Google Chirp 3	Access to code 42 combos, model 3.
Cohere Transcribe Arabic	access to core 42 compass models, right?
ElevenLabs Scribe v1	Access to Core 42 combo small drive
Deepgram Nova-3	Access to code 42 Compass model drive.
Google Latest Long	Access to Code 42. Composed model Drive.
Sonix	Access to code 42 campus models, right?

Why this matters. M3 and Cohere match the reference exactly. Other outputs change one or more product terms or the closing phrase. The difference affects what a reader can retrieve from the meeting record.

Short turns and names

M3 and Cohere matched the reviewed short references. M3 showed consistent preservation of names and terminology; Deepgram and Sonix had recurrent losses.

Readability

Punctuation is excluded from CER. Repeated punctuation and unsupported repetition remain part of the qualitative review.

Omissions and insertions

M3, Chirp 3 and ElevenLabs showed the lowest concentration of omissions. ElevenLabs also showed recurrent filler-like insertions.

Further inspectable examples

The viewer also contains “clear” versus “good” or “great,” and the role “note taker” with the participant name “Adam.”

Listen and compare the full English case ↗

This selected case illustrates the evaluated outputs; it does not replace aggregate CER or establish a general vendor ranking.

09 / ARABIC TRANSCRIPT EXAMPLE

Technical terms. Inside a Saudi-dialect update.

Arabic example · A routine status update with Kubernetes, scaling and paper.

Human-verified reference

صباح الخير كنت أمس أشتغل على موضوع في الـ Kubernetes من ناحية scaling كان في مشكلة بسيطة قاعد أدور لها حل وأيضاً قاعد أشوف الموضوع البيرر ويعطيكم العافية

M3

صباح الخير كنت أمس أشتغل على موضوع في الـ Kubernetes من ناحية scaling كان في مشكلة بسيطة قاعد أدور لها حل وأيضاً قاعد أشوف الموضوع البيرر ويعطيكم العافية

Google Chirp 3

صباح الخير، كنت أمس اشتغل على موضوع في الكبرنتيز من ناحيه السكيلينج كان في مشكله بسيطه قاعد ادور لها حل، وايضا قاعد اشوف الموضوع البيرر ويعطيكم العافيه.

ElevenLabs Scribe v1

صباح الخير، آآ، كنت أمس أشتغل على موضوع في الـ Kubernetes من ناحية scaling، كان فيه مشكلة بسيطة قاعد أدور لها حل، وأيضاً قاعد أشوف الـ موضوع الـ paper، ويعطيكم العافية.

Why this matters. M3 matches the verified reference and keeps Kubernetes and scaling in Latin script. ElevenLabs also preserves the English terms, with added filler-like content. Chirp 3 uses phonetic Arabic forms, which the Arabic scoring rules accept when equivalent.

Readable terminology and CER are related but different: a canonical spelling can help search even when its phonetic equivalent is accepted during scoring. The full case retains every compared output, including omissions and substitutions in other configurations.

[Listen and inspect every Arabic system output ↗](#)

09 / ARABIC QUALITATIVE REVIEW

What the outputs showed in practice

Observed behaviours explain the measured comparison without creating separate category scores.

The evaluated Arabic meeting outputs were reviewed for content omission, handling of names and English terminology, filler-like insertions, unexpected language changes, repeated punctuation, and repetition loops. These observations explain visible output behaviours and are supported by matched-audio examples. They are not presented as separately scored aggregate metrics unless a numerical result is explicitly stated.

Table 3. CER and qualitative output observations for all systems evaluated on the Arabic meeting comparison material.

System	Character error rate (CER)	Observed output characteristics
M3	1.62%	Lowest measured CER, with reliable preservation of business-critical content, numbers, and technical terminology.
Google Cloud Speech-to-Text V2 (chirp_3)	2.26%	Strong handling of English technical terminology within Arabic speech, with low content loss in the reviewed output.
Cohere	4.93%	Highly accurate on Arabic speech. The review identified omissions, inconsistent terminology, repeated punctuation and repetition loops.
ElevenLabs Scribe v1	5.01%	Preserved English technical terminology particularly well. Frequent spurious filler-like vocalisations such as “iij” were observed, and one segment switched to an unexpected language or script.
Deepgram Nova-3	6.09%	Frequent errors were observed in participant names, together with substantial content omissions, particularly at the beginning of source segments.
Google Cloud Speech-to-Text V1 (latest_long)	9.01%	Correctly returned no transcript for many non-speech segments, but missed a small number of speech-bearing segments. English technical terms were frequently omitted and numbers were occasionally omitted.
Sonix.ai	9.8%	Frequent errors were observed in names and English technical terms, together with segment-start omissions and excessive sentence-final punctuation.

Evidence scope. CER was calculated for every system using the same meeting recordings, human-verified reference, and scoring rules. The vendor-specific output characteristics are qualitative observations from that same evaluation set; separate aggregate rates were not calculated for those categories.

09 / ARABIC DETAILED REVIEW

M3 and Google Chirp 3

M3 and Google Cloud Speech-to-Text V2 (chirp_3).

Detailed qualitative criteria. Category observations describe the reviewed outputs; aggregate rates were not calculated unless explicitly stated.

Review criterion	M3	Google Cloud Speech-to-Text V2 (chirp_3)
Arabic transcription	Highly accurate.	Accurate.
English inside Arabic	Preserved English technical and business terminology more consistently within Arabic speech.	English technical terminology was preserved strongly within Arabic speech.
Text omission	Preserves business-critical content more reliably and reduces the loss of important numbers, terminology, and organisation or product names.	Low content loss was observed, with substantially fewer empty outputs than the V1 configuration.
Punctuation	Produced clearer sentence boundaries without the same concentration of repeated-punctuation artifacts.	Not separately scored; no recurring punctuation defect was highlighted in the qualitative review.
Spelling and terminology	Produces more consistent business terminology for reading and search.	English technical terms were rendered reliably within Arabic speech.
Word repetition	Showed substantially fewer repetition problems in the reviewed output.	No recurring repetition issue was isolated in the qualitative review.
Filler words and event tags	Did not show the same concentration of filler-like, event-tag, or repeated-punctuation artifacts in the evaluated meeting review.	No recurring filler-word or event-tag pattern was highlighted in the qualitative review.

All evaluated entries produced Arabic transcripts across the dialects represented in this evaluation set. External speaker-identification capabilities were not evaluated. Normalised CER results appear on the Arabic results chapter.

09 / ARABIC DETAILED REVIEW

Cohere and ElevenLabs Scribe v1

Cohere and ElevenLabs Scribe v1, on the same meeting evaluation set.

Detailed qualitative criteria. Category observations describe the reviewed outputs; aggregate rates were not calculated unless explicitly stated.

Review criterion	Cohere	ElevenLabs Scribe v1
Arabic transcription	Highly accurate.	Accurate.
English inside Arabic	English technical and business terms were sometimes omitted, mistranscribed, or rendered inconsistently within Arabic speech.	English technical terminology was preserved particularly well in the reviewed output.
Text omission	May omit or distort business-critical content, including numbers, technical terms, and organisation or product names.	No separate omission rate was calculated; the principal qualitative issue was inserted filler-like content rather than a recurring segment-start omission pattern.
Punctuation	Repeated punctuation patterns that reduce readability were observed in the reviewed output.	Not separately scored; readability was affected more by filler-like insertions than by a recurring punctuation defect.
Spelling and terminology	Technical and business terms may appear inconsistently or in phonetic form.	English technical terms were handled well, although filler-like insertions reduced overall transcript cleanliness.
Word repetition	Token and phrase repetition loops were observed in the reviewed output.	No separate repetition rate was calculated; frequent filler-like vocalisations were the more prominent insertion issue.
Filler words and event tags	Filler-like insertions, event-style tags, and repeated-punctuation artifacts were observed in the reviewed output.	Frequent spurious filler-like vocalisations such as “” were observed, and one segment switched to an unexpected language or script.

All evaluated entries produced Arabic transcripts across the dialects represented in this evaluation set. External speaker-identification capabilities were not evaluated. Normalised CER results appear on the Arabic results chapter.

09 / ARABIC DETAILED REVIEW

Deepgram and Google Latest Long

Deepgram Nova-3 and Google Cloud Speech-to-Text V1 (latest_long).

Detailed qualitative criteria. Category observations describe the reviewed outputs; aggregate rates were not calculated unless explicitly stated.

Review criterion	Deepgram Nova-3	Google Cloud Speech-to-Text V1 (latest_long)
Arabic transcription	Accurate.	Many errors.
English inside Arabic	English-term performance was not scored separately; reviewed examples showed inconsistent rendering alongside segment-start omissions.	English technical terms were frequently omitted.
Text omission	Substantial content loss was observed, particularly at the beginning of source segments.	Correctly rejected many non-speech segments, but missed a small number of speech-bearing segments and sometimes omitted English terms or numbers.
Punctuation	Not separately scored; no recurring punctuation defect was isolated in the qualitative review.	Not separately scored; no recurring punctuation defect was isolated in the qualitative review.
Spelling and terminology	Participant names were frequently transcribed incorrectly.	English technical terms were frequently missed, and some numbers were omitted.
Word repetition	No recurring repetition issue was isolated in the qualitative review.	No recurring repetition issue was isolated in the qualitative review.
Filler words and event tags	No recurring filler-word or event-tag pattern was isolated in the qualitative review.	Handled many non-speech segments correctly, although a small number of speech-bearing segments were also rejected.

All evaluated entries produced Arabic transcripts across the dialects represented in this evaluation set. External speaker-identification capabilities were not evaluated. Normalised CER results appear on the Arabic results chapter.

09 / ARABIC DETAILED REVIEW

Sonix and shared conditions

The final system profile and the common dialect and speaker-context conditions.

DETAILED QUALITATIVE CRITERIA / CONTINUED

Review criterion	Sonix.ai
Arabic transcription	Accurate.
English inside Arabic	Names and English technical terminology were frequently transcribed incorrectly or omitted.
Text omission	Content loss was observed particularly at the beginning of source segments.
Punctuation	Punctuation quality was weak, with excessive sentence-final marks observed in the reviewed output.
Spelling and terminology	Performance was weak on participant names and English technical terminology.
Word repetition	No recurring word-repetition issue was isolated; punctuation was the more prominent readability problem.
Filler words and event tags	No recurring event-tag pattern was isolated; excessive punctuation was the more prominent artifact.

Dialect coverage

All evaluated entries produced Arabic transcripts across the dialects represented in this evaluation set.

Speaker identification

Not evaluated for the external systems. Each system received the same participant-separated audio segments prepared by the M3 capture workflow and was scored only on transcription.

M3: Uses participant-linked audio channels and platform participant information. Platform-provided speaker attribution remained correctly linked throughout the evaluated Arabic and English meetings.

10 / OPERATION

Fast processing. Connected meeting context.

Batch speed and concurrent capacity are separate measurements on one L4 GPU.

BATCH TRANSCRIPTION

45_s

For an hour of recorded audio

Processing time on one L4 GPU.

CONCURRENT CAPACITY

82

Concurrent meetings

In a separate operating measurement on one L4 GPU.

These figures describe different workloads. They are not a combined speed-and-capacity benchmark or an end-to-end live-latency measurement.

Participant identity travels with the audio

M3 uses participant-linked audio channels and available platform information to keep identity attached to transcript segments. Each speaker remained correctly linked to their transcript throughout the evaluated Arabic and English meetings. External diarization and a separate speaker-attributed error rate were not evaluated.

Speaker

Available participant identity

Time

Position in the meeting

Source

The original audio segment

From processing to a usable record

Fast batch processing makes recorded conversations available for review sooner. Participant context and source alignment support navigation and follow-through. The measured CER results establish transcription accuracy under the stated scoring rules; summary quality and delivery time are separate evaluation questions.

Operating scope. Performance depends on the workload and deployment configuration. The batch and capacity figures are not provider-by-provider latency comparisons or guarantees for every deployment.

11 / AUTOMATIC LANGUAGE SELECTION

One system.

The right language, automatically.

Automatic primary-language selection, with Arabic–English switching inside the same meeting.

Before transcription, M3 identifies the meeting’s primary language from its audio and applies the appropriate language handling. Participants do not need to choose a language manually.

Listen

Use the meeting audio to identify its primary language.

Adapt

Apply the appropriate language handling within M3.

Continue

Handle Arabic–English switches as the discussion moves.

Arabic and English. One M3 system.

Arabic and English are handled by the same M3 system, rather than separate models. Selecting the primary language does not prevent M3 from handling the other language when it appears in the conversation.

Automatic selection check

M3 identified the correct primary language for every evaluated Arabic and English meeting. This was a separate workflow check on the same meeting material.

Controlled transcription test

For CER, M3 and every comparison system were explicitly configured for the language being evaluated. Automatic selection did not contribute to the CER score.

What the result establishes. Correct primary-language selection across the evaluated meetings. It does not establish a universal language-detection rate across all languages or meeting conditions.

Evaluate on the meetings that matter

Customer evaluations can measure language selection, transcription, participant attribution and meeting intelligence separately, using matched inputs and human-verified references.

12 / EVIDENCE BOUNDARIES

Results with clear boundaries

Keep benchmark findings, product coverage and operational capabilities distinct.

Finding	Evidence boundary
Arabic transcription: M3 1.62% CER	Matched Arabic audio and human-verified references, after the stated lexical equivalences.
English transcription: M3 1.83% CER	Matched English audio and human-verified references, after English normalization.
Qualitative output behavior	Criterion-level review of actual outputs. No separate aggregate rates or significance tests.
Automatic primary-language selection	Correct for every evaluated Arabic and English meeting in a separate workflow check; outside CER.
Participant linkage	Correct participant-to-transcript linkage throughout evaluated Arabic and English meetings. External diarization was not tested.
Batch processing: 45 seconds per recorded hour on one L4 GPU	Recorded-audio transcription throughput. Separate from concurrent capacity, live latency and downstream delivery.
Concurrent capacity: 82 meetings on one L4 GPU	A separate operating measurement; not a combined batch-speed or live-latency benchmark.
Product coverage: 40 languages and 32 Arabic dialects	MeetriX product scope. The CER results in this paper apply to the separate Arabic and English evaluations.
Customer-controlled deployment	The evaluated transcription path called no external transcription service. Deployment policy and controls must be assessed in their own scope.

Not established by CER. Summary factuality, action-item accuracy, independent speaker recognition, employee time saved and performance on every deployment are separate evaluation questions.

The paper brings together methods, measured results and inspectable transcript evidence.

12 / INSPECT THE EVIDENCE

Hear the input. Read the exact output.

Inspect Arabic and English examples alongside their human-verified references.

01 Listen to the audio

Source clips make the reviewed examples inspectable. The same clip is the input for every system shown. Selected examples do not become a new aggregate benchmark.

02 Compare exact transcripts

Each case keeps the human reference and named system outputs. Corrected Arabic display examples preserve the existing benchmark results; the corrections do not add new scored audio or alter the aggregate CER.

03 Apply the right scoring rule

Arabic uses the stated lexical equivalences. English ignores punctuation, case and whitespace while retaining the merged-word exception. Readability and searchability differences can remain even when a variant is equivalent under CER scoring.

What full reproduction still requires

A complete rerun needs pinned system builds and provider requests, a canonical audio/reference manifest, executable normalization rules and per-segment edit records. Those inputs are not all present in the supplied research bundle.

The published evaluations establish the matched-language results shown here. The source material does not support confidence intervals, general multilingual error rates or independent certification of the benchmark.

[Open audio and transcript comparisons ↗](#) [Read the evaluation protocol ↗](#)