Exbabel translation research

> **Exbabel model configurations.** Exbabel Flagship Model denotes the OpenAI-backed speech engine. Exbabel Translation Model denotes the Gemini-backed speech engine. Exbabel uses custom prompting in its translation system. The 180-clip comparison measured the underlying engines directly. The isolated worker probes tested Exbabel routing separately; one French clip compared the worker with and without a terminology instruction. The study did not estimate a general improvement from custom prompting.

How the recordings compared with UN interpreters

Both providers produced scorable audio and passed the reference alignment rule on 66 of 180 assigned excerpts. The two alignment checks agreed sufficiently on 66 interpreter passages. A separate automated reference audit supported both boundaries on 55 of those 66 passages.

Direction	Aligned/30	Pairedn	OpenAICOMET	GeminiCOMET	chrF++O/G	BLEUO/G
EN→FR	7	7	0512 .	0508 .	363/364 . .	54/59 . .
FR→EN	18	18	0658 .	0661 .	485/469 . .	110/93 . .
ES→EN	10	10	0595 .	0639 .	380/413 . .	68/81 . .
AR→EN	7	7	0693 .	0569 .	444/413 . .	108/75 . .
RU→EN	11	11	0656 .	0715 .	539/610 . .	168/260 . .
ZH→EN	13	13	0640 .	0640 .	441/421 . .	105/76 . .

Each direction has its own denominator. Higher scores indicate closer agreement with the ASR transcript of the professional interpretation. COMET uses the floor transcript as source context. chrF++ and BLEU measure textual overlap. Differences can arise from paraphrasing, ASR errors, or incorrect boundaries. The scope is the automatically alignable subset.

Both providers received the same source clips. Exbabel Flagship Model returned 180 recordings and Exbabel Translation Model returned 176 after retries. Four Exbabel Translation Model absences remain visible. No Exbabel Translation Model primary session supplied the expected final-turn event, even where the recording included the final idea.

The workbook includes separate native-caption scores, paired COMET differences with delegation-cluster intervals, and a sensitivity check restricted to references supported by a further automated audit. The per-direction samples contain only one to three delegations, limiting uncertainty estimates. In Mandarin, the 13-clip COMET scores were nearly identical. On the 10 clips supported by the additional audit, Exbabel Flagship Model scored 0.692 and Exbabel Translation Model 0.633. Human judgments remain pending.

Passage alignment and last-phrase checks

For each source clip, the interpreter search began eight seconds before the floor window and extended 22 seconds beyond it. Whisper-1 produced source-language and interpreter-language transcripts. Two text models independently proposed target-word boundaries without access to the provider hypotheses. The rule required both to mark the passage usable and at least 0.72 intersection over union across their word spans.

The retained reference is the timestamped interpreter text from the longer recording. A second ASR pass on the cropped audio checks sensitivity to the cut. The separate reference audit can retain fewer samples. No human verified a reference for this release. Source STT word or character error rates require verbatim source transcripts and remain unavailable for these 180 clips.

System	Endingpresent	Missingorchanged	Uncertain	Noaudio
Exbabel Flagship Model	142	34	4	0
Exbabel Translation Model	118	51	7	4

The ending check used an anonymized transcript comparison for every delivered WAV. It flags whether the last recoverable source idea appears in the generated-audio transcript. The classifier was Gemini Flash-Lite. The source and output ASR used OpenAI Whisper. Shared vendor dependencies, ASR mistakes and the fragmentary clips limit independence. Review flagged passages by ear.

Spain, GA80-ES-8090-es-en: the source ends with “no menores”. The interpreter says “just as important”, Exbabel Flagship Model says “no less important”, and Exbabel Translation Model says “minor ones” in both its caption and output ASR. The proposed interpreter span also omits the opening sentence. This case illustrates both a polarity flag and the limits of automatic boundary agreement.

The listening page includes all 180 source clips, interpreter context, and every available model recording. Reviewers have separate blind IDs, proposed reference cuts, and forms for correcting boundaries or recording an unalignable passage.

EXBABEL EVENTS / INTERNAL STUDY UN speech translation pilot

20 November 2026 | gpt-realtime-translate and gemini-3.5-live-translate-preview

180 original-floor excerpts from 14 delegations were tested across six language directions. Exbabel Flagship Model returned audio for 180 of 180 primary clips. Exbabel Translation Model returned audio for 176 of 180 after recovery. These counts measure delivery. Bilingual assessment of diplomatic meaning is pending.

Studycomponent	Observedevidence
Authenticspeech	30clipsperdirection Twostandaloneproviderssawidenticalcliphashes . .
Arabicspokentarget	Exbabel Translation Model returned 30/30 English -to-Arabic outputs The Arabic-source direction was .
	evaluatedseparately .
Productionworker	7/7 local worker probes emitted audio A recovered Exbabel Translation Model -routed worker output was .
	identifiedasArabicbyASR Listenerdeliveryremainsunmeasured . .
Writtenmaterial	35UNParallelCorpusitemsastextandrecordedreadings Twotextsystemsalso .
	translated12passagesmatchedtotheactualspeeches .
Reviewstatus	Reference-based metrics available for written material No adjudicated diplomatic .
	accuracyorpass/faildecision .

Interpretation for the presentation

A working speech path can still change a named country or lose a final phrase. The report separates first-attempt reliability, captured target audio, reference-based text quality and pending human judgments.

EXBABEL EVENTS / INTERNAL STUDY Authentic sources and sampling

United Nations Transcription Corpus, General Debate session 80, September 2025. Each selected clip links to the pinned source MP3, delegation, UTC-independent broadcast timecode and hashes. The project owner confirmed written consent for internal provider testing and a presentation to the UN. Media stayed in the local evidence folder except for authorized processing by the benchmark providers.

Direction	OpenAIaudio/planned	Geminiaudio/planned
ENtoFR	30/30	30/30
FRtoEN	30/30	30/30
EStoEN	30/30	29/30
ARtoEN	30/30	30/30
RUtoEN	30/30	28/30
ZHtoEN	30/30	29/30

The floor channel was identified by the archive. Selection used seeded, disjoint windows with low-energy boundary adjustment, automatic source-language identification and a waveform match against the same-language channel. A bilingual listener has yet to check speaker turns and semantic boundaries. The China country row says English. In the accepted floor clips, independent identification and the Chinese-channel waveform both supported Chinese.

A Mandarin excerpt from the China delegation illustrates the review task. The UN interpretation refers to remaining silent for fear of power. The Exbabel Flagship Model English caption instead speaks in the first person about holding power. Independent ASR of its WAV reproduced that opening. A bilingual reviewer should inspect the full surrounding sentence before assigning severity. Case GA80-CN-4728-zh-en.

These are General Debate excerpts from a small set of speakers, with one Chinese delegation and two Russian-language delegations. Performance across meeting types, spontaneous exchanges, accents, or overlapping speakers cannot be inferred from this cohort. The matched UN interpretation channel is included on the broadcast clock. Its natural lag requires alignment before use as a scoring reference.

EXBABEL / INTERNAL EVALUATION 2

EXBABEL EVENTS / INTERNAL STUDY Streaming reliability and latency

Exbabel Translation Model returned audio on 16 primary first attempts; recovery raised coverage to 176. The attempt ledger preserves quota rejections and every timeout. A saved output without the expected final turn signal remains incomplete under the frozen rule. Exbabel Flagship Model closed 180 primary sessions successfully. Audio and caption language checks have separate denominators in the machine-readable tables.

Sourcetotarget	OpenAImedian(s)	n	Geminimedian(s)	n
ENtoFR	085 .	30	318 .	30
FRtoEN	065 .	30	291 .	30
EStoEN	059 .	30	296 .	29
ARtoEN	059 .	30	284 .	30
RUtoEN	055 .	30	288 .	28
ZHtoEN	053 .	30	292 .	29

Time is measured from first input send to first translated audio packet. Silence can precede speech. Each result uses only attempts with a packet, so timing comparisons can favor the provider with more missing responses. See latency_results.csv for per-clip values and language_pair_metrics.csv for p95. Interactive audience playback, listener comprehension and meaningful phrase-level latency require separate measurements.

Supplementalcondition	Provider	Input(s)	Audio	Terminaloutcome
extended	Exbabel Translation Model	80	1	failed
extended	Exbabel Flagship Model	80	1	ok
noise	Exbabel Translation Model	20	1	failed
noise	Exbabel Flagship Model	20	1	ok
smoke	Exbabel Translation Model	20	1	failed
smoke	Exbabel Flagship Model	20	1	ok

Bahamas floor speech supplied the clean and 10 dB additive-noise pair and an 80-second continuous excerpt. The longer excerpt overlaps the short probe. Additional probes used a two-second mute, four seconds of competing speech and two paced-send stalls. All supplemental observations are exploratory.

EXBABEL / INTERNAL EVALUATION 3

EXBABEL EVENTS / INTERNAL STUDY Written material and recorded readings

Thirty-five directional items came from seed-selected, bilingual UNPC development material. Each language direction had five items. Written-only requests used a text translation model per vendor. Real-time speech models received eSpeak NG readings of the same source words. The original bilingual target sentence served as the reference. This controlled experiment uses synthetic voices. It is separate from authentic UN speech.

FrenchtoEnglish	System	n	BLEU	chrF++	COMET
Spoken/ASR	GeminiLive	5	0565 .	18761 .	04287 .
Spoken/ASR	OpenAIRealtime	5	20955 .	52934 .	06874 .
Spoken/caption	GeminiLive	5	1541 .	22781 .	04945 .
Spoken/caption	OpenAIRealtime	5	13868 .	50895 .	06679 .
Writtentext	Exbabel Translation Model Flash -Lite	5	54864 .	77108 .	09108 .
Writtentext	GPT -4o-mini	5	49117 .	77331 .	09148 .

These corpus scores measure agreement with a reference. Each group has at most five sentences. Missing, empty, or absent hypotheses reduce n. Captured captions and ASR from drain-timeout attempts are included as exploratory partial outputs. Text-model and speech-model differences cannot isolate the effect of the voice alone. Full tables include all directions, metric signatures and denominators.

One French reference names Chile. On its synthetic reading, the Exbabel Flagship Model caption named Brazil, and independent ASR also transcribed Brazil from the output WAV. This is a concrete country-name discrepancy. A reviewer must determine its source and diplomatic severity. See UNPC-fr-en-2007 in the listening comparison.

Separately, both text models translated 12 official written passages matched to the recorded speeches, two per source language. All 24 calls returned text. The listening index places each written result beside the real speech. Alignment is provisional. The text ledger records a 404 for gemini-2.5-flash and the switch to gemini-3.1-flash-lite.

EXBABEL / INTERNAL EVALUATION 4

EXBABEL EVENTS / INTERNAL STUDY Review, uncertainty and next decisions

Two independently randomized reviewer packets contain the source, generated audio, native caption, original interpretation where available and the written reference where applicable. The two reviewers score meaning adequacy, entities, terminology, legal force, omissions, intelligibility, naturalness and comprehension. An adjudicator preserves both original judgments. No acceptance thresholds have been approved.

Priority	Evidence	Follow-up
1 Namesandlegal .	Chile/Brazildiscrepancyinacontrolled	Adjudicatesourcespeech nativecaptionandtarget ,
meaning	reading .	WAV Auditdates amounts negationandobligations . , , .
2 Geminicompletionand .	Terminal -event timeouts and quota	Confirmfinalphrasedeliveryunder
quota	rejectionswerelogged .	continuous-stream semantics Stress-test account .
		limitsatintendeddeploymentload .
3 Timingandalignment .	First -event timestamps and nominal	Marksourcephraseends correctinterpretationlag , ,
	interpretercutsavailable .	andmeasuretimetoequivalentmeaning .
4 Samplingandreview .	Smallnumberofdelegationsper	Expandspeakerdiversity preapprovethresholdsand ,
	language;zeroadjudicatedratings .	commissionblindedreview .

Wilson intervals quantify binary audio-return proportions. Paired provider differences and delegation-cluster bootstrap intervals are in paired_results.csv. With few delegations those intervals are unstable, and one delegation yields no cluster interval. No comparison establishes professional-interpreter equivalence.

Evidence index: manifest.csv, attempt_history.csv, results_raw.jsonl, written_metrics.csv, listening.html, human_review_packets and provenance/. Source material: United Nations Transcription Corpus and the UN Parallel Corpus, Ziemski, Junczys-Dowmunt and Pouliquen (2016).

EXBABEL EVENTS / INTERNAL STUDY Audio, terminology and adverse conditions

System	WAVs	MedianWAV(s)	Medianactive(s)	Medianquiettail(s)
exbabel	6	246 .	89 .	23 .
Exbabel Translation Model	176	439 .	138 .	247 .
Exbabel Flagship Model	180	488 .	80 .	270 .

Active time counts 20 ms windows at or above -40 dBFS. Quiet tails include the client drain interval. Total WAV length can therefore substantially exceed the duration of audible speech. Perceptual intelligibility and naturalness await listening ratings.

System	ASRtarget-languagematches	WAVswithevaluableASR
Exbabel Flagship Model	180	180
Exbabel Translation Model	176	176
exbabel	6	6

11 sample-term matches were flagged using UNTERM forms. Context, acceptable variants and critical-error severity need bilingual adjudication. On one French floor clip, the isolated Exbabel Translation Model worker emitted audio with and without a terminology instruction about Ukraine. One paired case cannot establish a glossary effect.

In the three additional conditions (mute, overlap, and send stalls), Exbabel Flagship Model returned audio in 3/3 and Exbabel Translation Model in 0/3. These conditions are individually traceable in the listening index. All classified service failures remain in the denominator.

The workbook includes the complete capability matrix and per-direction results. REMEDIATION.md prioritizes named entities, final-phrase delivery, quota capacity, the Arabic route, and listener timing.