EXBABEL EVENTS · INTERNAL RESEARCH

Exbabel real-time translation of UN diplomatic speech

> **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.

20 November 2026 · Six UN languages · Exbabel internal study for a UN presentation.

What the recordings show

On 55 UN speech excerpts, blinded AI reviewers gave Exbabel Flagship Model 91.3, Exbabel Translation Model 87.1 and the recorded UN interpreter 87.3 out of 100 for conveying the delegate’s meaning. The same professional interpretation also served as a reference for a separate 66-clip comparison. The two experiments answer different questions. Neither produced a percentage of diplomatically correct utterances.

Evidence	Observed result	Interpretation
Primary audio delivery	Exbabel Flagship Model 180/180; Exbabel Translation Model 16/180 on first pass, 176/180 after recovery	Delivery is measured across every assigned clip. Four missing Exbabel Translation Model outputs stay in the denominator.
Paired reference scoring	66/180 clips passed automatic alignment	COMET, chrF++ and BLEU measure agreement with one rendering.
Source fidelity	55/180 clips passed a second automated audit	All three translations were scored against the original speaker. A faithful paraphrase can get full credit.
Consequential differences	Both AI raters flagged at least one material issue on 12 Exbabel Flagship Model, 22 Exbabel Translation Model and 15 interpreter clips out of 55	Flags can arise from errors in the translation, audio cut or transcription. Human adjudication remains pending.
Arabic voice output	Exbabel Translation Model returned 30/30 English→Arabic recordings after recovery	Independent ASR identified Arabic on all 30. Local worker output was also confirmed as Arabic in one routed probe.

The strongest defensible conclusion is specific: both real-time engines frequently conveyed the speaker’s intended meaning on the selected clips. These results test standalone engines under Exbabel’s benchmark harness. The local Exbabel worker had a smaller engineering cohort. Deployed audience audio and independent bilingual quality ratings were outside the completed study.

How 180 clips became two different scored sets

Stage	Count	Decision rule
Selection	180	30 disjoint floor excerpts in each of six directions. Fixed seed 20261120, 15–23 seconds in the selected clips, 3,489.8 seconds total, 14 delegations.
Delivery	180 Exbabel Flagship Model / 176 Exbabel Translation Model	Retain failures, quota rejections, timeouts, and output without confirmed completion.
Two-aligner comparison	66 retained; 108 boundary disagreements; 6 partial or uncertain	Search the interpreter channel from 8 s before to 22 s after the floor interval. Two models independently select inclusive word spans, blind to system translations.
Alignment rule	66 paired	Both must mark coverage usable, word-span intersection over union ≥0.72, with at least four words in the selected span.
Further audit	55 supported; 11 partial	Re-transcribe the isolated interpreter cut and audit both boundaries against the floor transcript.
Blind source fidelity	55 clips × 3 candidates × 2 ratings	Use the isolated interpreter transcript as a third candidate. Every rating has a stored reason and artifact/material flags.

Floor identity was checked against the archive designation, automatic language identification and the same-language channel waveform. For China, the catalog row said English, but the selected floor speech and channel match supported Mandarin. One delegation supplied Mandarin and two supplied Russian speech. The cohort comes from one General Debate session; speaker selection and acoustic conditions limit generalization.

A clock-time cut can omit part of a sentence because an interpreter listens before speaking. Even after automatic alignment, a cut can begin after the delegate’s opening phrase. This is visible in the Spanish “no menores” example. A missing phrase in the excerpt does not, by itself, establish an interpreter omission.

Agreement with UN interpretation

The reference was the machine transcript of the matched UN interpreter passage. For generated speech, the hypothesis was Whisper-1's independent transcript of the saved WAV. Native model captions were scored in a separate condition. Original floor transcripts supplied source context to COMET.

Direction	Paired n	Exbabel Flagship Model COMET	Exbabel Translation Model COMET	Exbabel Flagship Model chrF++	Exbabel Translation Model chrF++
en → fr	7	0.512	0.508	36.3	36.4
fr → en	18	0.658	0.661	48.5	46.9
es → en	10	0.595	0.639	38.0	41.3
ar → en	7	0.693	0.569	44.4	41.3
ru → en	11	0.656	0.715	53.9	61.0
zh → en	13	0.640	0.640	44.1	42.1

COMET uses a learned model to estimate quality from the source, hypothesis and reference. Its score cannot be converted into "percent matched." BLEU measures corpus-level word n-gram precision with a brevity penalty. chrF++ uses character n-gram agreement plus word n-grams. Single-reference overlap can penalize good paraphrases. See the workbook for BLEU, uncertainty intervals and native-caption results.

SacreBLEU 2.6.0 used 13a tokenization for these English/French targets and word order 2 for chrF++. The written suite also configured zh tokenization for Chinese targets where applicable. COMET used Unbabel/wmt22-comet-da via unbabel-comet 2.2.7 on CPU. Checkpoint SHA-256: e213091cde220f97b89f8bdfa750c458cfea741ad62affb455b59900210ff2af.

Per-direction samples are small. For Mandarin, both engines scored 0.640 on 13 aligned clips. On the further-audited 10-clip subset, Exbabel Flagship Model scored 0.692 and Exbabel Translation Model 0.633. The reference eligibility rule can change an apparent ranking.

Fidelity to the source speaker

Three fresh AI subagents worked independently from masked packets. Two agents rated each candidate. Labels and ordering varied by rater. The 0–100 Direct Assessment rubric gave full credit to meaningful paraphrases and harmless compression. Raters separately flagged consequential changes to actors, quantities, polarity, legal force and key propositions. The professional interpreter received a score under the same rubric.

Direction	Clips	Exbabel Flagship Model DA	Exbabel Translation Model DA	Interpreter DA
All directions	55	91.3	87.1	87.3
ar → en	7	91.1	83.4	82.9
en → fr	6	86.1	82.7	81.2
es → en	7	99.1	86.6	92.1
fr → en	14	90.8	92.0	85.6
ru → en	11	91.7	87.2	92.7
zh → en	10	89.2	86.0	87.0

Exbabel Flagship Model minus Exbabel Translation Model averaged +4.1 DA points (exploratory paired 95% interval +1.8 to +6.5). Exbabel Flagship Model minus interpreter averaged +4.0 (+0.4 to +6.1). Exbabel Translation Model minus interpreter averaged −0.2 (−4.1 to +2.1). Intervals resample delegations within directions, with 10,000 paired draws. Few delegations, automatic selection and related raters limit those intervals.

All six rater/control combinations ranked both faithful paraphrases above the deliberate quantity or polarity error. Ratings were within ten points on 162/165 genuine candidates. The three subagents share an OpenAI model family. High agreement among them cannot substitute for independent bilingual listening.

A stringent sensitivity check discarded an entire three-way case if both raters flagged a possible recognition or boundary artifact on any candidate. It retained zero cases. This does not establish that every translation was wrong. It does mean that the study cannot report a separate artifact-free DA estimate.

Written translation and controlled recorded readings

35 directional items (five per direction) were drawn from 28 distinct UNPC bilingual rows. The professionally translated target sentence served as reference. Each source was tested directly as text and separately as an eSpeak NG reading delivered to the speech system. The recording in this experiment is synthetic. Differences include model, voice and transcription effects.

French → English	System	n	COMET	chrF++	BLEU
Output WAV → ASR	Exbabel Translation Model	5	0.4287	18.761	0.565
Output WAV → ASR	Exbabel Flagship Model	5	0.6874	52.934	20.955
Native caption	Exbabel Translation Model	5	0.4945	22.781	1.541
Native caption	Exbabel Flagship Model	5	0.6679	50.895	13.868
Written input	gemini-3.1-flash-lite	5	0.9108	77.108	54.864
Written input	gpt-4o-mini	5	0.9148	77.331	49.117

The written branch returned 65/65 translations. The recorded-reading branch returned 29/30 Exbabel Flagship Model WAVs and 30/35 Exbabel Translation Model WAVs. The initial text-only Exbabel Translation Model alias returned HTTP 404; the separately logged gemini-3.1-flash-lite run supplied successful text results. Those findings use development material, not the inaccessible official fixed test archive.

Twelve passages from actual prepared statements (two per source language) were separately matched to delivered speeches using floor ASR without looking at either system's output. Both text models translated each passage, producing 24 outputs. Prepared and spoken wording can differ. The casebook shows all 12 beside the original floor transcript.

Source word error rate uses edit counts divided by reference word count; Chinese uses character units. A known written script supplied the reference for controlled readings. For five Spanish readings the recognizer made 1 edit in 112 reference words. A verified verbatim source transcript is still required to calculate this measure for authentic speech.

Streaming, reliability and the Exbabel worker

Exbabel Flagship Model's gpt-realtime-translate received paced 24 kHz audio. Exbabel Translation Model's gemini-3.5-live-translate-preview received paced 16 kHz audio from the same 48 kHz source masters. Logs capture local sends, translated-audio packets, caption events, errors and close signals. Quota retries used new attempt directories. The first captured result, or original failure, supplied the canonical sample.

Direction	Exbabel Flagship Model first audio p50 / p95 (s)	Exbabel Translation Model first audio p50 / p95 (s)	Exbabel Translation Model WAVs
ar → en	0.59 / 1.00	2.84 / 3.88	30/30
en → fr	0.85 / 1.10	3.18 / 4.75	30/30
es → en	0.59 / 0.99	2.96 / 4.02	29/30
fr → en	0.65 / 0.90	2.91 / 3.69	30/30
ru → en	0.55 / 0.94	2.88 / 3.26	28/30
zh → en	0.53 / 5.08	2.92 / 3.36	29/30

Exbabel Flagship Model confirmed completion on 180 primary sessions. No Exbabel Translation Model primary session returned the expected final-turn signal before the deadline. The ending classifier nevertheless marked the final recoverable idea present on 142 Exbabel Flagship Model and 118 Exbabel Translation Model assignments; missing or changed on 34 and 51; uncertain on 4 and 7. Four Exbabel Translation Model assignments had no audio. Those are transcript-based flags, pending listening review.

Time to first packet includes silence; it does not measure when the listener understood a translated phrase. The archive retains p50/p95 caption and audio onset, jitter proxies, ASR/caption agreement, active signal levels and exploratory term-anchor timing. A single Bahamas speech supplied clean, 10 dB noise and 80-second continuous probes. Additional probes introduced a two-second mute, four seconds of competing speech, and send delays. These do not determine deployed failover or naturalness.

Seven isolated worker route probes returned speech after recovery, including the explicit Exbabel Translation Model Arabic route. One French clip tested Exbabel Translation Model worker behavior with and without a Ukraine terminology instruction. Both returned audio. This sample cannot estimate glossary benefit. A later expansion of worker testing was stopped at the requester's clarification. Its partial logs are retained for chronology and excluded from these results.

Sources, permissions and reproducibility

The project owner confirmed written permission for internal provider testing and a presentation to the UN. This package is prepared for that scope. Public distribution of source audio, private reviewer keys or unredacted study materials is a separate decision. No UN endorsement or adjudication is implied.

UN Transcription Corpus · Authentic General Debate floor recordings and human interpretation channels. Pinned revision e55f7d9d2e2e7cac4e2f5c6b6c89cd03c87a5b62.

UN Parallel Corpus v1.0 · Ziemski, Junczys-Dowmunt and Pouliquen (2016). Official parliamentary texts, professionally translated. The fixed test archive returned HTTP 403 in this run.

Helsinki-NLP UNPC mirror · The actual bilingual sentence pool. The selected development items came from the mirror’s train split. Rows, revision and URLs are in written.csv.

GA80 official statements · Prepared speeches and session records. The collection includes extracted text for 14 UN-hosted statement PDFs and a Chinese-language original from the Chinese mission.

Chinese-language statement · Supplementary Chinese original used in matching published words with delivered speech.

UNTERM · Official terminology records; includes preferred/variant forms and individual record URLs. Contextual verification remains open.

OpenSTBench · Framework reviewed for speech translation evaluation and metric selection. The implementation used the pinned metric packages individually.

SacreBLEU · Metric implementation used for BLEU and chrF++; per-run signatures are retained in the results.

WMT22 COMET-DA · Learned reference-based translation-quality metric. The pinned model, package version and checkpoint hash are included in the methodology record.

IWSLT 2024 evaluation campaign · Appendix A.2 describes source-based Direct Assessment on a 0–100 scale. This study adapted it for blinded AI reviewers.

Sperber et al. (2024) · Study of segmentation, automatic metrics and human evaluation in speech translation.

The source directory includes the selected rows, original MP3 URL/hash registry, delegation pages, statement URLs, extracted statements, terminology records and the original user-supplied starter files. Source and model revisions are recorded in methodology_and_provenance.json. The casebook reproduces the specific ASR text used for scoring.

What is in the evidence package

Where to look	Contents
Translation-Choices.pdf / .html	All 66 aligned comparisons, 35 written/read-aloud tests and 12 matched statement pairs; 16 focused three-way explanations.
Reports-and-workbooks	Versioned historical feasibility and authentic-pilot reports, presentations, full metrics, source DA workbook and interpreter-comparison workbook.
Methods-and-sources	Study freezes, rubric, source registers, extracted documents, terminology and written source tables, scoring protocols and validation records.
Original-brief	Original project request, protocol and starter benchmark folder including supplied PowerPoint and tracker.
Research-evidence.tar.gz	The preserved benchmark tree: original MP3s, WAVs, machine transcripts, provider events, failures, retries, randomized packets, scripts, raw judgments and historical artifacts.
UN-benchmark-20261120/internal-pilot/listening.html	All sample cohorts in the original pilot. interpreter-comparison/listening.html opens the 180 floor clips and context audio.
Source and metric integrity	Historical validation checked 49 original recordings, 180 disjoint floor clips, 474 sample-provider speech conditions and 461 retained pilot outputs. Ten offline tests passed. Source DA comprises 330 genuine candidate ratings plus 18 control ratings.

To reproduce the analysis, unpack the evidence archive beside the dossier directory, then consult EXECUTION.md for the historical setup, internal-pilot/README.md for the authentic run, interpreter-comparison/README.md for alignment and source-da/protocol.md plus analysis_plan.md for Direct Assessment. Analyze stored recordings offline; live reruns require provider accounts and introduce model, load and latency variation.

To strengthen a public performance claim, agree acceptance criteria before a fresh holdout run, recruit bilingual listeners independent of the vendors, align full meaning units in context, and test the delivered Exbabel audience stream. Prioritize country and actor substitutions, negation, time periods, final-phrase loss, and real-world latency.