Exbabel diplomatic speech translation

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 faithfully did AI convey the delegate’s meaning?

Blinded AI raters compared each recorded translation with its original speaker on 55 automatically aligned UN excerpts.

Recorded output	Mean Direct Assessment / 100	Material issue flagged by both raters
Exbabel Flagship Model	91.3	12 / 55
Exbabel Translation Model	87.1	22 / 55
UN interpretation	87.3	15 / 55

Scores apply to the selected standalone-engine recordings. Bilingual adjudication and deployed Exbabel listener testing remain pending.

What was tested

Identical speech inputs and a preserved record of failures support paired comparisons.

Study component	Scope and outcome
UN General Debate	180 floor clips, 30 per direction, 14 delegations.
Spoken Arabic target	30 English→Arabic Exbabel Translation Model tests; all returned ASR-identified Arabic.
Controlled written test	35 directional items as text and recorded readings; 65 text outputs, 59 WAVs.
Prepared speeches	12 aligned written passages, two text systems, 24 translations.
Isolated Exbabel worker	Seven routed probes emitted speech after recovery.

How the comparison sets were built

Interpreter lag shifts the words heard at a given broadcast timestamp.

Stage	Clips	Evidence
Floor speech	180	Seeded, disjoint excerpts; floor-language and waveform checks.
Interpreter alignment	66	Both models agreed on a usable span. 108 boundary disagreements and six partial or uncertain cases.
Additional audit	55	An automated check supported both boundaries.
Source-based scoring	55	Three masked transcripts per clip; two AI judgments per candidate.

Two metrics answer different questions

The same system can preserve meaning with different words and receive a lower reference-agreement score.

Method	Question	How to read it
COMET	How well does this output agree with the source and one interpreter rendering?	Learned reference-based score on 66 matched clips.
chrF++ and BLEU	How much wording overlaps with that reference?	Character and word n-gram overlap; paraphrase affects results.
Source-based DA	How much of the delegate’s meaning does this rendering preserve?	Raw 0–100 adequacy points. UN interpretation is a third candidate.

Source-fidelity scores by direction

Each row compares all three candidates on the same subset. The clips available in each language differed.

Direction	Paired n	Exbabel Flagship Model	Exbabel Translation Model	UN interpreter
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

Equal weighting across directions gives 91.3, 86.3 and 86.9 respectively. One to three delegations per direction limit generalization.

Interpreter-reference COMET by direction

Independent transcripts of generated speech were compared with the automatically aligned UN reference.

Direction	Paired n	Exbabel Flagship Model COMET	Exbabel Translation Model COMET
en → fr	7	0.512	0.508
fr → en	18	0.658	0.661
es → en	10	0.595	0.639
ar → en	7	0.693	0.569
ru → en	11	0.656	0.715
zh → en	13	0.640	0.640

COMET does not represent the percentage of words matched. Boundary changes can change the ordering on small samples.

Are the other challenges less important?

Recorded-transcript excerpts · GA80-ES-8090-es-en

Rendering	Transcript excerpt	What it conveys
UN interpreter	There are other just as important challenges.	The remaining challenges carry equal weight.
Exbabel Flagship Model	There are other challenges too, no less important.	The different wording preserves the comparison.
Exbabel Translation Model	There are other challenges, minor ones.	Calling them minor reverses the delegate’s point.

Casebook includes the original-language source, full candidate transcripts, per-clip scores and both reviewer notes.

Three ways to express the same political argument

Recorded-transcript excerpts · GA80-BY-9283-ru-en

Rendering	Transcript excerpt	What it conveys
UN interpreter	harsh self-isolation	The speaker portrays EU policies as isolating the EU itself.
Exbabel Flagship Model	harsh self-isolation measures	The same actor and consequence survive the rephrasing.
Exbabel Translation Model	measures, well, of strict self-isolation	The emphasis and delivery differ, but the same causal argument is recoverable.

Casebook includes the original-language source, full candidate transcripts, per-clip scores and both reviewer notes.

A quarter of a century or a single quarter?

Recorded-transcript excerpts · GA80-BY-10982-ru-en

Rendering	Transcript excerpt	What it conveys
UN interpreter	The past quarter century	A 25-year period frames the political claim.
Exbabel Flagship Model	Last quarter	The transcript suggests a much shorter period.
Exbabel Translation Model	The past quarter century	The intended timescale is explicit.

Casebook includes the original-language source, full candidate transcripts, per-clip scores and both reviewer notes.

Did the speaker call for financing or consensus?

Recorded-transcript excerpts · GA80-EG-10503-ar-en

Rendering	Transcript excerpt	What it conveys
UN interpreter	provide concessional financing	Names a specific kind of financing on favorable terms.
Exbabel Flagship Model	provide affordable financing	Conveys the favorable terms, with less technical precision.
Exbabel Translation Model	providing consensus	The output ASR changes the object. The Exbabel Translation Model native caption says “providing concessional financing.” Listen before assigning the discrepancy to the translator.

Casebook includes the original-language source, full candidate transcripts, per-clip scores and both reviewer notes.

Delivery and final-phrase evidence

First-packet latency, returned speech and the last recoverable idea are distinct outcomes.

Measure	Exbabel Flagship Model	Exbabel Translation Model
Primary audio on first attempt	180 / 180	16 / 180
Primary recordings after recovery	180 / 180	176 / 180
Expected completion signal	180 / 180	0 / 180
Final idea present · automated	142 / 180	118 / 180
Final idea missing or changed · automated	34 / 180	51 / 180

Remaining end checks were uncertain (4 Exbabel Flagship Model, 7 Exbabel Translation Model) or had no audio (4 Exbabel Translation Model). Listener-side semantic latency was not measured.

What the written controls revealed

Five French→English items had both written-input and recorded-reading results against professional references.

Input and scored output	Exbabel Flagship Model COMET	Exbabel Translation Model COMET
Written input → written translation	0.915	0.911
Synthetic reading → saved audio ASR	0.687	0.429
Synthetic reading → native caption	0.668	0.495

Different text and speech models were used. In one reading, Exbabel Flagship Model’s caption and output ASR named Brazil where the French source named Chile.

How far the findings travel

The blinded check rewards legitimate paraphrase and preserves adverse results.

Strength of the pilot	What remains open
All 180 assignments and failures retained.	Audience playback, realistic failover and time to equivalent meaning.
Full translation transcripts, provenance and two judgments per scored candidate.	Independent bilingual adjudication of boundaries and high-impact discrepancies.
55 matched three-way cases with control questions.	Strict artifact exclusion retained zero complete three-way cases.
Authentic speech, controlled text and isolated worker probes.	Larger held-out speaker sample, agreed thresholds and product-level scoring.