Exbabel Translate

RESEARCH REPORT

Evaluation of Real-Time Speech Translation

United Nations General Debate

A comparative study of two Exbabel translation models and live UN interpretation



EXBABEL RESEARCH

NOVEMBER 2026

KEY FINDINGS

Exbabel Flagship scored 4 points higher than the UN interpreter in blind review.

MEANING PRESERVED / AVERAGE AI REVIEWER SCORE OUT OF 100

91.3	87.1	87.3
Exbabel **Flagship Model**	Exbabel **Translation Model**	UN **interpreter**

Who scored higher on the same clip?

The Flagship Model scored higher than the UN interpreter on 34 clips. The interpreter scored higher on 21. Each block below represents one clip.



34 Flagship higher **21 Interpreter higher**

The average Flagship advantage was 4.0 points. The estimated 95% interval was +0.4 to +6.1. For the Translation Model, the gap was −0.2 points, with an interval of −4.1 to +2.1. That second result is consistent with a lead for either version.

Intervals were estimated by resampling delegations within each language direction. They reflect variation in the selected speeches. Bias shared by the AI reviewers is a separate uncertainty.

Which speeches did we test?

The UN Transcription Corpus includes the delegate's microphone and a separate channel for the live interpreter. We used both from the September 2025 General Debate of the UN General Assembly.

HOW 180 RECORDINGS BECAME 55 SCORED CLIPS

180

Original clips

15–23 seconds each, from 14 delegations. Thirty clips in each of six language directions.

66

Matched passages

Two text models independently located the interpreter's matching words. Both had to find a usable passage and agree closely on its boundaries.

55

Scored clips

A further automated check supported both boundaries for 55 clips. These are the clips used for every three-way quality comparison.

Why matching the timing matters

Interpreters often finish a thought after the delegate does. A cut at the same second can capture different sentences. Of 180 excerpts, 108 had disputed boundaries and six had uncertain or partial matches. Eleven more had only partial support in the final check.

Recording source

How was meaning judged?

Reviewers saw the speaker's words and an anonymous translation. They could give full credit to any wording that conveyed the same message.

55	3	2	330
source clips	translations each	ratings each	ratings

What the reviewers saw

The UN interpretation and both AI translations were transcribed from audio. Labels and order were shuffled. Two of three AI reviewers scored each candidate against the source transcript, with the producer's identity hidden. Each pair of scores was averaged, then each clip counted equally.

What a score means

0	25	50	75	100
Unrelated or reversed	Topic only	About half survives	Main point, local error	Full meaning survives

Reviewers checked who acted, what happened, quantities, timing, negation and the strength of commitments. They also flagged consequential additions and omissions. The scores measure judged meaning on a 100-point scale.

Reviewers came within 10 points on 162 of 165 translations. All three came from the same model family, so agreement can include shared blind spots.

Method adapted from IWSLT 2024, Appendix A.2

How did each language score?

The biggest spread was in Spanish to English. The Flagship Model averaged 99.1 on seven clips. The other Exbabel model averaged 86.6 on those same clips.

AVERAGE MEANING SCORE / POINTS OUT OF 100

Speaker to listener	Clips	UN interpreter	Exbabel Flagship Model	Exbabel Translation Model
English to French	6	81.2	86.1	82.7
French to English	14	85.6	90.8	92.0
Spanish to English	7	92.1	99.1	86.6
Arabic to English	7	82.9	91.1	83.4
Russian to English	11	92.7	91.7	87.2
Mandarin to English	10	87.0	89.2	86.0
All clips	**55**	**87.3**	**91.3**	**87.1**

How much weight should each row carry?

Each row represents 6 to 14 excerpts. All Mandarin clips came from one delegation. The differences describe these recordings. A broader claim about a language would need more speakers, topics and independent listeners.

The dossier also reports COMET, chrF++ and BLEU comparisons with UN wording. The table above uses blind judgments against the original speaker.

When does a word change the point?

SPANISH to ENGLISH / SPAIN / ES-8090

“No menores”

Spain’s delegate described other challenges as just as important. Listen for the change in their importance across the three versions.

UN interpreter

“There are other just as important challenges.”

Equal importance.

Exbabel Flagship Model

“There are other challenges too, no less important.”

Different wording keeps the comparison.

Exbabel Translation Model

“There are other challenges, minor ones.”

“Minor” reverses the speaker’s point.

MANDARIN to ENGLISH / CHINA / CN-6647

A shorter phrase can work just as well.

The interpreter said “advocating the people-centered approach.” The Flagship Model said “put people first.” The Translation Model said “advocate for a people-centered approach.” All three kept the same priority.

Which details deserve a second listen?

RUSSIAN to ENGLISH / BY-10982

A time span

The interpreter and Translation Model said “the past quarter century.” The Flagship transcript said “Last quarter,” shortening 25 years to three months.

FRENCH to ENGLISH / BJ-5933

A condition

Benin’s delegate said national efforts alone were insufficient without cooperation. The interpreter preserved that condition. Both AI versions made it less clear.

ENGLISH to FRENCH / BB-15548

A country name

The interpreter and Flagship transcripts said “Haïti a besoin” (Haiti needs). The Translation Model transcript read “Ici, il a besoin” (Here, he needs). The audio needs checking.

ARABIC to ENGLISH / EG-10503

A financing term

The interpreter said “concessional financing.” Flagship said “affordable financing.” The Translation Model caption kept the term, but the audio transcript read “consensus.”

12 / 55 **Flagship clips drew a material-error flag from both judges.**

Both judges flagged 22 of 55 Translation Model clips and 15 of 55 UN interpreter clips. A flag calls for listening to the recording. It can reflect the translation, a speech-recognition mistake or where the clip was cut.

Delivery also mattered. Flagship returned 180 of 180 recordings on the first try. The Translation Model returned 16 on the first try and 176 after retries. In a separate English-to-Arabic run it returned 30 of 30 recordings. That run measured delivery only.

What can we conclude?

Both AI systems often carried the speaker's point across languages. The Flagship Model came closest to preserving the full message on average in this pilot.

Different words can serve the listener equally well.

In the Mandarin excerpt, "put people first" expressed the same priority as "advocating the people-centered approach." Scoring against the original speaker gave both phrasings credit. That is why an overlap score alone understates what a good paraphrase can achieve.

Some details carry the weight of the speech.

The Spanish comparison turned on the word "minor." The Russian example turned on the length of a "quarter." A listener could follow the topic in either case and still leave with the wrong understanding. The strongest average left room for consequential errors.

WHAT THE EVIDENCE SUPPORTS

These findings support a broader listening trial. Only 55 of 180 clips had sufficiently clear boundaries for scoring, and AI reviewers read machine transcripts. The main trial tested the underlying engines directly. Exbabel tested its custom prompts separately. The next step is independent bilingual review of full-service audio across longer speeches.