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.

EXBABEL / SOURCE-BASED EVALUATION How faithfully did the translations convey the speech?

Three blinded AI reviewers rated the meaning conveyed by the saved Exbabel Flagship Model, Exbabel Translation Model and UN interpreter recordings. Each candidate was compared with the delegate’s source transcript and received two independent judgments. 55 of 55 selected clips had a complete three-way comparison. Each clip carried equal weight.

Recordedoutput	MeanDA/100	Bothflaggedmaterial	Oneflaggedmaterial
Exbabel Flagship Model	913 .	12/55	11/55
Exbabel Translation Model	871 .	22/55	7/55
UNinterpreter	873 .	15/55	9/55

DA means Direct Assessment. A score is an adequacy rating in points. It measures how well the assessors thought the translation preserved the meaning. A percentage of correct utterances would require a separately defined pass rule. Material flags record a substantive error identified by the indicated number of AI reviewers; bilingual adjudication is pending.

Paired differences

Firstminussecond	Meandifference	Exploratory95%interval
OpenAIminusGemini	+41 .	+18to+65 . .
OpenAIminusUNinterpreter	+40 .	+04to+61 . .
GeminiminusUNinterpreter	02 - .	-4 .1 to +2 .1

The intervals resample delegations within language directions. With few delegations they can understate uncertainty. They omit systematic evaluator and transcription bias. The exercise estimates performance on the selected passages and supports no equivalence claim.

What different word choices looked like

On a Russian-language excerpt about EU policy, each rendering conveyed the speaker’s claim that self-isolation deepened an internal crisis. The ratings were 99 for Exbabel Flagship Model, 98 for Exbabel Translation Model and 100 for the interpreter (GA80-BY-9283-ru-en). In a French excerpt about strategic cooperation, both judges flagged a weakened or reversed condition in the AI transcripts. The interpreter scored 97.5, Exbabel Translation Model 75.5 and Exbabel Flagship Model 71.5 (GA80-BJ-5933-fr-en). These illustrations were chosen after scoring.

TRANSCRIPT-BASED PILOT / BLINDED AI RATINGS 1

EXBABEL / SOURCE-BASED EVALUATION Results by language direction

Sourcetotarget	Clips	Exbabel Flagship Model	Exbabel Translation Model	UNinterpreter
ARTOEN	7	911 .	834 .	829 .
ENTOFR	6	861 .	827 .	812 .
ESTOEN	7	991 .	866 .	921 .
FRTOEN	14	908 .	920 .	856 .
RUTOEN	11	917 .	872 .	927 .
ZHTOEN	10	892 .	860 .	870 .

Differences between languages also reflect speakers, topic, and clip selection. The six directions each had 30 attempted floor clips, but the available three-way comparison covers different numbers in each direction. The English-to-Arabic capability run did not have a matched three-way comparator.

How sensitive are the results?

Aggregation	Exbabel Flagship Model	Exbabel Translation Model	UNinterpreter
Eachdirectionequallyweighted	913 .	863 .	869 .
Artifactsensitivity(n=0)	n/a	n/a	n/a
Rater-standardized points (z)	+023 .	012 - .	011 - .

The artifact check removes a whole paired clip if both raters flag a possible recognition or boundary artifact on any candidate. The z scores adjust for individual raters using different parts of the scale. Both analyses use the same blinded judgments.

Quality and availability answer different questions. Exbabel Flagship Model produced audio on 180/180 original assignments. Exbabel Translation Model produced audio on 16/180 first attempts and 176/180 after recovery, with no confirmed final-turn signal. A high meaning score on returned audio does not resolve final-phrase reliability or listener latency.

TRANSCRIPT-BASED PILOT / BLINDED AI RATINGS 2

EXBABEL / SOURCE-BASED EVALUATION Test design and publication guidance

Why the comparison is fairer

The UN interpreter is a third candidate. A faithful paraphrase can receive full credit. Reviewers rate preservation of intent, actor, attribution, quantities, polarity, obligations and key propositions. Compression and stylistic variation are acceptable when the speaker’s material meaning survives.

All 6 control judgments compared two valid paraphrases with a deliberate error; 6 passed the predeclared ordering check. On genuine candidates the median absolute difference between the two ratings was 2.0 points. Both ratings were within ten points on 162 of 165 candidates.

What the sample establishes

The 55 selected clips came from 180 original UN floor excerpts. Two automatic aligners accepted 66 interpreter cuts. A further automated audit supported 55 of those cuts. This selection may favor easier interpretations. The full pilot archive retains exclusions and surrounding audio. Whisper produced the transcripts, including the isolated UN interpreter cuts. Reviewing an imperfect cut may penalize the interpreter for speech outside the matching window. Punctuation, recognition mistakes, and clipped sentences can still reveal recording characteristics or change perceived meaning.

Each case was scored by two of three fresh subagents with independently shuffled identities. They share a model family and were developed by OpenAI, one of the vendors under test. Their agreement is evidence of repeatability within this exercise. Independent bilingual listeners must validate high-impact claims, ratings of naturalness, and comparisons of professional performance.

For publication

Lead with the observed mean scores, sample count, direction coverage, AI judging method and delivery rates. Describe these as tests of the named standalone engines on authentic UN speech. A claim about the Exbabel product requires scoring output captured from its actual routing and processing. For a confirmatory study, freeze new speeches, sampling and acceptance criteria, include independent bilingual judges, and measure comprehension and delay on the delivered audio.

Method sources

IWSLT 2024 evaluation campaign, Appendix A.2

Speech translation segmentation and direct assessment

COMET score interpretation