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.