

Auravio: A Blinded Physician Evaluation of Multilingual Clinical Translation

Michael Cacioli

Auravio

<https://auravio.org>

Abstract

We evaluated Auravio, a full multilingual clinical interpretation pipeline, against a fixed OpenAI *whisper-1* speech-to-English translation endpoint. We used fifty fictional clinical scenarios that were reviewed by two pilot physicians and then, by native translators, translated into five different languages and recorded. Both systems processed the same recordings, and ten additional physicians independently rated the anonymized final outputs from both systems. This process produced 500 paired physician assessments as well as 1,000 separate severity ratings. Our prespecified primary outcome measure was the proportion of physician ratings identifying major or critical clinical errors in the final English output generated by either the baseline translation system or the complete Auravio pipeline. In terms of serious errors, there were 26/500 (5.2%) ratings identifying these errors for Auravio versus 158/500 (31.6%) for the baseline. The baseline system had a higher error rate than Auravio, with an absolute difference of -26.4 percentage points for Auravio minus baseline (95% CI: -33.4 to -19.4) and a relative reduction of 83.5% (95% CI: 73.3–91.2). There was no language in which the serious-error rate did not favor Auravio. Substantially non-English baseline outputs accounted for 54 major-error ratings. Overall, these results indicate an advantage when using the entire Auravio pipeline for this clinical interpretation benchmark compared to using the configured baseline translation endpoint.

1 Introduction

Clinical translation must preserve details that can change the meaning of an encounter, including medication names, doses, allergies, negation, and timing. A fluent English output can still misrepresent these details. Evaluating a clinical translation system therefore requires attention to clinical meaning as well as linguistic fluency.

In this respect, Auravio recognizes speech in a source language and translates it into English. In addition to its translation function, Auravio’s trust evaluator also provides recommendations, risk bands, and structured flags. This report focuses on the clinical meaning of final English output, assessed through blinded physician ratings.

The comparison system uses a frozen configuration of OpenAI’s *whisper-1* direct speech-to-English translation endpoint. Whisper is a multilingual speech-processing model family described by Radford et al. (2023). This study evaluates two complete systems using different architectures and processing stages. Although this study does not evaluate each component of Auravio separately, it does provide an overall assessment of each complete system.

Therefore, the primary question is whether physicians identify fewer major or critical clinical errors in the output generated by Auravio when both systems use identical recordings. Secondary questions include physician safety preference, agreement between physicians rating the outputs, and results across languages.

2 Methods

2.1 Scenarios and Recordings

The case bank consisted of fifty fictional English-language clinical scenarios that spanned medication and allergy, dosing and numeric information, negation and uncertainty, symptoms and anatomy, timing and history, as well as low-risk controls. The dosing cases included frequency of administration, and the symptom and anatomy cases included laterality. In addition to providing a reference clinical meaning for each case, all cases documented examples of errors at differing levels of severity.

Prior to system output generation, two pilot physicians reviewed the entire case bank. These physicians did not participate in the final blinded

evaluation. One paid native translator-speaker per language (Latin American Spanish, Italian, Brazilian Portuguese, France French, and Standard German) translated and recorded each of the 50 cases. This created a total of 250 source audio files. Each translation was self-reviewed by its creator, rather than reviewed independently by a second translator. All of these audio files were scripted and contained no actual patient encounters.

2.2 Evaluated Systems

Auravio. The configuration that was tested utilized OpenAI Whisper for source-language automatic speech recognition, with the resulting transcript then translated into English via Claude Sonnet 4.6. Both the trust judge and back-translation check were activated. The system produced a single final English output, a risk band, a recommendation, and structured flags.

Baseline Direct Translation. The baseline directly translated each of the provided recordings into English, without a language hint or prompt, utilizing OpenAI’s direct audio translation API endpoint and the model `whisper-1`.¹ No warning or escalation layer was present in the baseline.

Both systems processed the same set of recordings using their respective frozen configurations. Each generated a single successful final output for each recording. Therefore, there were 250 outputs created by each system, resulting in 500 distinct outputs overall. Output length and wording were not restricted to be identical across the two systems. Summary generation, FHIR export, and languages beyond the five that were tested were not separately assessed.

2.3 Blinded Physician Review

The 10 physicians who did not audit the case bank each reviewed 50 anonymized output pairs. Each output pair was evaluated independently by two physicians to yield 500 paired evaluations and 1,000 individual severity ratings.

Each physician viewed 10 recordings per language and exactly one language realization of each master scenario. The identities of the systems were concealed behind A/B labels, with system positions balanced within each language. The presentation order was counterbalanced between the two reviewers evaluating each recording.

¹API endpoint: `/v1/audio/translations`.

Reviewers selected one of four possible severity levels (no clinically meaningful error, minor, major, or critical) for each output. In addition, they chose which output was preferable for clinical safety, whether both were equivalent, or whether neither was safe. Optional written feedback was accepted. The rubric is described in Appendix A.

English was the required final output language. Reviewers were instructed not to translate substantially non-English outputs or use translation tools to interpret them, and to classify these as major errors.

In total, twelve physicians participated: two in the pilot case-bank audit and ten in the final blinded review.

2.4 Endpoints and Analysis

Our prespecified primary outcome measure was the percentage of physician ratings that identified a major or critical clinical error in the final English output, which we referred to as a “serious clinical error.” We kept both ratings for each output without averaging them or adjudicating disagreements. Therefore, we had 500 severity ratings per system, based on 250 distinct outputs per system.

For system s , the serious-error rate was

$$\hat{p}_s = \frac{1}{500} \sum_{i=1}^{500} \mathbb{I}[y_{i,s} \in \{\text{major, critical}\}]. \quad (1)$$

The absolute difference was defined as $\hat{p}_A - \hat{p}_B$, where A denotes Auravio and B the baseline. Relative reduction was $1 - \hat{p}_A/\hat{p}_B$.

Our analysis plan was internally frozen before system identities were revealed. This report does not claim public preregistration. All ratings were locked before unblinding.

We computed confidence intervals through a percentile cluster bootstrap across the 50 master scenarios, using 10,000 replicates and a random seed of 20260812. In each resampled scenario, we kept all five language versions, the two physician ratings for each recording, and both systems. Therefore, our use of resampling to estimate variability preserved the paired nature of the data in addition to the dependence among observations within a given scenario.

The confidence intervals represent variability due to sampling across scenarios using a fixed speaker and physician panel. However, they do not separately estimate uncertainty associated with recruiting new physicians or speakers. Only the

Measure	Auravio <i>n</i> (%)	Whisper-1 <i>n</i> (%)	Difference (pp) 95% CI
Serious: major or critical	26 (5.2)	158 (31.6)	−26.4 (−33.4, −19.4)
No clinically meaningful error	450 (90.0)	263 (52.6)	—
Minor	24 (4.8)	79 (15.8)	—
Major	18 (3.6)	130 (26.0)	−22.4 (−28.8, −16.0)
Critical	8 (1.6)	28 (5.6)	−4.0 (−7.0, −1.4)

Table 1: Clinical error severity: 500 ratings per system from 250 outputs. Serious errors combine major and critical errors. Differences are Auravio minus baseline in percentage points (pp), with 95% scenario-cluster bootstrap CIs. Dashes denote unreported contrasts.

primary outcome measure is confirmatory, while all other analyses are secondary or exploratory, without multiplicity-adjusted significance claims. Finally, we provide raw agreement and Krippendorff’s alpha for the varying reviewer pairs (Artstein and Poesio, 2008).

3 Results

3.1 Primary Endpoint

Serious errors were assigned in 26/500 Auravio ratings (5.2%, 95% CI: 2.8–8.0) and 158/500 baseline ratings (31.6%, 95% CI: 25.2–38.4). The absolute difference was −26.4 percentage points (95% CI: −33.4 to −19.4). The relative risk was 0.165 (95% CI: 0.088–0.267), corresponding to an 83.5% relative reduction (95% CI: 73.3–91.2). Figure 1 presents the primary endpoint. Table 1 gives the severity distribution. The 83.5% reduction applies to the major-or-critical composite, not critical errors alone.

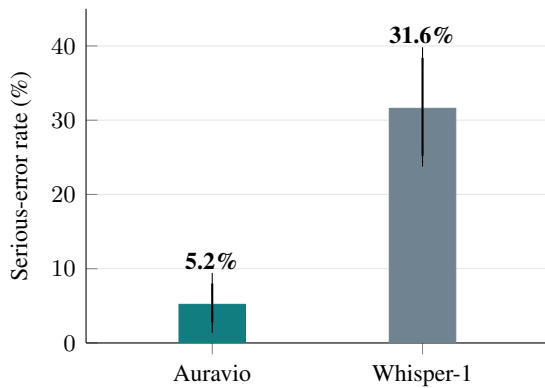


Figure 1: Serious-error rates with 95% scenario-cluster bootstrap CIs. Each system has 500 physician ratings from 250 outputs.

3.2 Language Subgroups

Auravio showed a lower serious-error rate in each of the five languages examined (Table 2), with absolute differences ranging from −16.0 percentage points in French to −33.0 percentage points in Italian. Each language was represented by 50 recordings and 100 ratings per system across all 50 master scenarios. These subgroup analyses are secondary, and no adjustments were made for the possible inflation of Type I error due to multiple comparisons.

3.3 Safety Preference and Agreement

Across 500 total clinical-safety comparisons, physicians preferred Auravio in 218 (43.6%) and the baseline system in 15 (3.0%). Physicians rated the outputs as equivalent in 251 (50.2%) of their assessments; they also indicated that neither output was clinically safe in 16 (3.2%). The difference between the rates at which physicians preferred Auravio and the baseline was +40.6 percentage points (95% CI: 33.6–47.8). Half of all physician comparisons resulted in a rating of equivalent.

In terms of raw percent agreement on serious-error status, Auravio had an agreement level of 95.2% and the baseline had an agreement level of 90.4%. The corresponding nominal Krippendorff’s alpha values for the two systems were 0.514 and 0.778, respectively. Using exact four-level severity agreement as the metric for comparison, the respective levels of agreement were 86.4% and 76.0%. Similarly, using ordinal alpha to assess agreement between physicians on severity ratings for each system, the respective values were 0.356 and 0.818.

Language	Auravio <i>n</i> (%)	Whisper-1 <i>n</i> (%)	Difference (pp) 95% CI
Latin American Spanish	4 (4.0)	36 (36.0)	−32.0 (−45.0, −19.0)
Italian	8 (8.0)	41 (41.0)	−33.0 (−47.0, −19.0)
Brazilian Portuguese	5 (5.0)	25 (25.0)	−20.0 (−32.0, −9.0)
France French	7 (7.0)	23 (23.0)	−16.0 (−28.0, −5.0)
Standard German	2 (2.0)	33 (33.0)	−31.0 (−43.0, −20.0)

Table 2: Serious-error ratings by language: 100 ratings per system from 50 recordings. Differences are Auravio minus baseline in percentage points (pp), with 95% scenario-cluster bootstrap CIs. Comparisons are not adjusted for multiplicity.

3.4 Operational Observations

The median latency was 5.17 seconds per recording processed by Auravio and 1.73 seconds per recording processed by the single-endpoint baseline. These latencies reflect different amounts of processing in Auravio’s multi-step pipeline and the baseline’s single-endpoint process. Per-call API costs were not tracked and are not estimated here.

4 Discussion

Auravio received substantially fewer serious-error ratings compared with the direct translation baseline system. The results were in the same direction across all five languages tested. In addition, safety-related preferences favored Auravio more often than the baseline. The observed advantage reflects overall task performance rather than semantic accuracy alone.

As a result of the fact that the two systems differ in both their models and processing stages, this study cannot determine which components resulted in the differences. Additional experiments will be required to isolate the impact of different aspects of the architectures on overall quality.

In future evaluations, there would be merit in testing Auravio using a broader range of speakers. Translations could be reviewed by independent reviewers, recordings could be made under various environmental conditions, and alternative or differently configured comparison systems could be tested.

5 Conclusion

Auravio received serious clinical error ratings in 5.2% of judgments on this five-language fictional benchmark, versus 31.6% for a frozen Whisper-1 direct translation endpoint, representing an 83.5% relative reduction. This supports additional evaluation of the full pipeline with a greater variety of speech and clinical environments to determine its potential for safe clinical deployment and patient benefit.

Limitations

Materials and Participants. The audio recordings were scripted and fictional, so they do not replicate how people naturally interact with each other (e.g., interruptions and overlapping speech), nor do they represent the various conditions in which recordings may be made. The fact that there is only one speaker per language makes it difficult to distinguish between speaker-based and language-based effects. Translations were not independently reviewed. The final panel consisted of ten physicians, and the benchmark had fifty master scenarios.

Comparative Analysis. This study analyzed a single fixed baseline configuration. It utilized Whisper-1 without a prompt or language hint. Different configurations of these systems could perform differently, so this study does not establish that the baseline was the most effective option available. The systems being compared differ in their architectures and output formats, so we cannot attribute the observed differences to specific components. In addition, results from this study should not be generalized to all languages supported by Auravio.

Assessment and Generalizability. Physician ratings are judgments of translated output, rather than

measures of patient outcomes. Agreement among raters was not perfect. Bootstrap estimates account for clustering among scenarios, but they do not separately estimate uncertainty associated with recruiting different physicians or speakers. Patient benefit, clinical workflow improvement, and suitability for replacing professional interpreters were not evaluated.

Developer Involvement. The author designed and implemented Auravio and coordinated its assessment. Blinded methods and frozen protocols reduce potential developer bias, but they do not completely eliminate it.

Ethics Statement

All of the scenarios in this study were fictional. Real patient records or physician-patient interactions were not involved. No formal ethics or institutional review board (IRB) review was obtained, and no exemption determination is claimed. The physician reviewers are described collectively. None of the reviewers' identities or individual review worksheets are published here. No institutional endorsement from any hospital or university is claimed.

Funding and Competing Interests

This study was self-funded. The author is the founder and lead developer of Auravio and organized the data collection and analysis. Compensation to physicians who completed the ratings was fixed regardless of their ratings or system preference. Translator-speakers were also compensated.

References

- Ron Artstein and Massimo Poesio. 2008. [Survey article: Inter-coder agreement for computational linguistics](#). *Computational Linguistics*, 34(4):555–596.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. [Robust speech recognition via large-scale weak supervision](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

A Rating Instrument

No clinically meaningful error. Clinical meaning is preserved, allowing differences in wording.

Minor error. An error or awkwardness would not change a clinical decision.

Major error. The output could lead to a wrong clinical decision or require stopping and re-asking. Substantially non-English output was classified as major because English was required.

Critical error. Acting on the output could plausibly cause serious harm.

Reviewers could mark multiple error categories (medication, allergy, dosage, frequency, route, number or measurement, negation, uncertainty, symptom, anatomy, laterality, time or duration, omission, unsupported addition, and other). For each pair, they selected “Output A,” “Output B,” “equivalent,” or “neither safe.” Comments were optional. The reviewers' instructions did not permit them to translate substantially non-English output or use translation tools for their reviews.