Speech-to-Text for Cypriot Greek

A WER Benchmark on Real Phone Calls

aseto.ai — Technical Report
Giorgos Kosta, ML Engineer
Evaluation date: 18 May 2026

1. Summary

aseto.ai builds AI voice agents that operate over the telephone in Cypriot Greek (CG). Reliable voice agents depend on reliable transcription: errors in the speech-to-text (STT) layer propagate into intent detection, response generation, and ultimately the caller's experience. This report measures that layer directly.

We compare aseto.ai's in-house STT model, served on our own infrastructure, against five widely used systems: three commercial cloud services — ElevenLabs Scribe v2, Microsoft Azure Speech, and Google Chirp 3 — and two open-weight models run locally — faster-whisper large-v3 and NVIDIA Canary-1b-v2. The evaluation uses Word Error Rate (WER) and Character Error Rate (CER) on a held-out set of 1,027 real, in-the-wild phone-call samples in Cypriot Greek.

Headline result: aseto.ai achieves a WER of 23.9% on Cypriot Greek phone calls, against 29.1% for the strongest competing system (Google Chirp 3) — an 18% relative reduction in word errors. The remaining systems range from 36.5% to 50.4% WER, with aseto.ai roughly halving the error rate of the weakest. The same ordering holds on CER.

2. Why Cypriot Greek is hard for STT

Cypriot Greek is a low-resource variety, and the gap between it and the commercial systems benchmarked here is not an accident of tuning — it reflects how those systems were trained.

It is not Standard Modern Greek. CG diverges from Standard Modern Greek (SMG) in phonology (e.g. palatalisation and consonant gemination), in morphology, and in a substantial body of everyday vocabulary. A model trained predominantly on SMG treats many CG word forms as out-of-distribution, even when the surface script is identical.

It is under-represented in training data. Mainstream STT systems are trained on large multilingual corpora in which CG is, at best, a thin slice folded into a generic "Greek" label. There is little incentive for a general-purpose provider to optimise for it, so CG-specific phenomena are effectively averaged out.

Phone-call audio compounds the problem. Telephony narrows the audio band, adds codec artefacts, and brings spontaneous-speech effects — disfluencies, overlapping speech, background noise. These conditions are already harder than clean read speech, and they interact badly with a dialect the model has barely seen.

Local named entities are absent from training data. Cypriot village names, districts, and local institutions appear rarely or not at all in the corpora behind mainstream STT systems. Faced with one, a general-purpose model has no entry to recognise and falls back on whatever common word sounds closest. For a voice agent, these named entities are often the most important words in the call — a location to route to, an organisation to identify — so an error here is disproportionately costly.

Together these factors mean a general-purpose STT system is solving a noticeably different problem from the one a CG voice agent actually faces. The benchmark below quantifies that difference.

---

3. Methodology

3.1 Evaluation data

The test set consists of 1,027 samples from real phone calls handled in Cypriot Greek, totalling 1 hour 8 minutes of audio (4,112.7 seconds) across more than 100 distinct speakers. The calls come from customer-support and IVR-agent interactions, so the set reflects the conversational, task-oriented speech that aseto.ai's voice agents handle in deployment. Calls were drawn from production-like telephony conditions rather than studio recordings, so the set reflects the acoustic and linguistic variation aseto.ai's agents encounter in practice.

Reference transcripts were produced manually by a native Cypriot Greek speaker and reviewed in a second pass for consistency. The evaluation set is held out and was not used in developing the aseto.ai model.

3.2 Metric

We report two metrics. **Word Error Rate (WER)** is the sum of word substitutions, insertions, and deletions divided by the number of words in the reference. **Character Error Rate (CER)** applies the same calculation at the character level. Lower is better for both. CER is included because it is more robust to the morphological variation of a low-resource dialect: minor inflection or spelling differences inflate WER more than they do CER, so reporting both guards against over- or under-stating the gap.

To compare systems fairly, the same text-normalisation pipeline is applied to every hypothesis and reference before scoring. The pipeline strips any bracketed or angle-bracketed markup (such as tags or annotations), converts text to lowercase, replaces punctuation with spaces, and collapses runs of whitespace to a single space. The identical pipeline is applied to every system's output and to the reference transcripts, so no system is advantaged or penalised by formatting conventions. The pipeline does not transliterate or otherwise alter Greek characters. Holding normalisation constant matters here: CG has spelling and accentuation variation, and inconsistent normalisation can move WER by several points independent of true recognition quality.

3.3 Systems compared

The baselines fall into two groups: three commercial cloud STT services, and two open-weight models run locally. aseto.ai's model is also self-hosted.

System	Category	Notes
aseto.ai STT	**In-house, self-hosted**	**Specialised for Cypriot Greek; served on aseto.ai infrastructure**
ElevenLabs Scribe v2	Commercial cloud service	General-purpose commercial STT
Azure Speech	Commercial cloud service	General-purpose commercial STT
Google Chirp 3	Commercial cloud service	General-purpose commercial STT
faster-whisper large-v3	Local open-weight model	CTranslate2 implementation of Whisper large-v3
NVIDIA Canary-1b-v2	Local open-weight model	1B-parameter general-purpose ASR model

All systems were evaluated on the identical 1,027-sample audio set with the Greek language setting (el). The evaluation was run on 18 May 2026; each provider was accessed through its current SDK with the latest model available on that date. No fixed version strings are reported, so results should be read as representative of each system as of 18 May 2026 and may shift as providers update their models.

---

4. Results

Across all 1,027 samples, aseto.ai's model records the lowest error rate on both metrics. The full results:

System	WER	CER	WER vs. aseto.ai	Inference time (s)
aseto.ai STT	**23.9%**	**13.6%**	**—**	**378**
Google Chirp 3	29.1%	16.1%	+21.7% rel.	1,334
ElevenLabs Scribe v2	36.5%	21.7%	+52.4% rel.	787
faster-whisper large-v3	39.3%	24.3%	+64.4% rel.	401
Azure Speech	47.8%	35.4%	+99.9% rel.	1,132
NVIDIA Canary-1b-v2	50.4%	28.6%	+110.7% rel.	34

Inference time is the total wall-clock time to transcribe all 1,027 samples. It is reported for completeness only and is not a like-for-like speed comparison — see the note in Section 5.

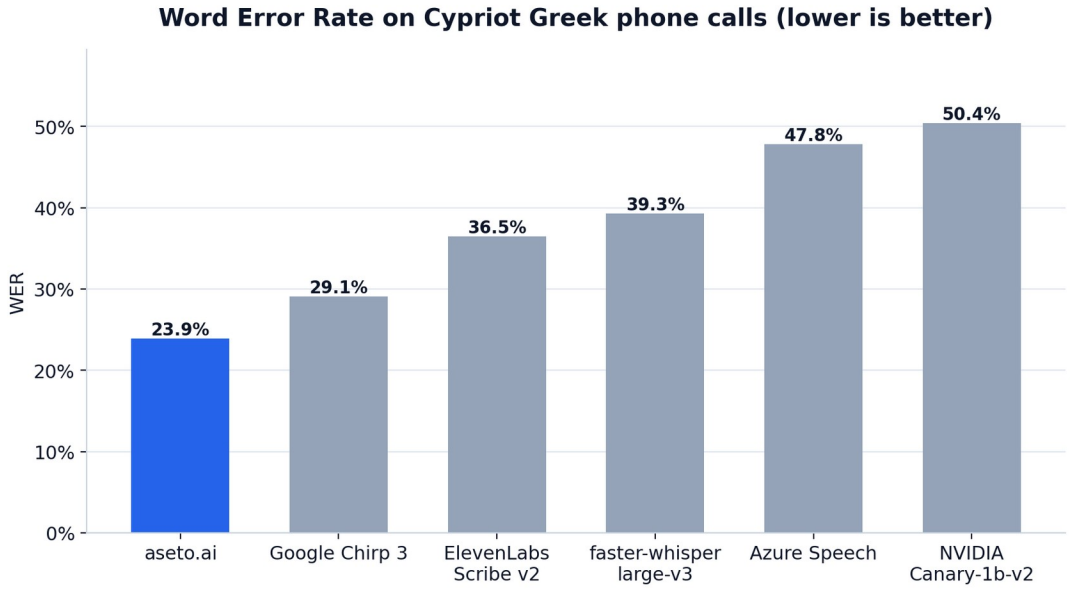

Figure 1. Word Error Rate by system (lower is better).

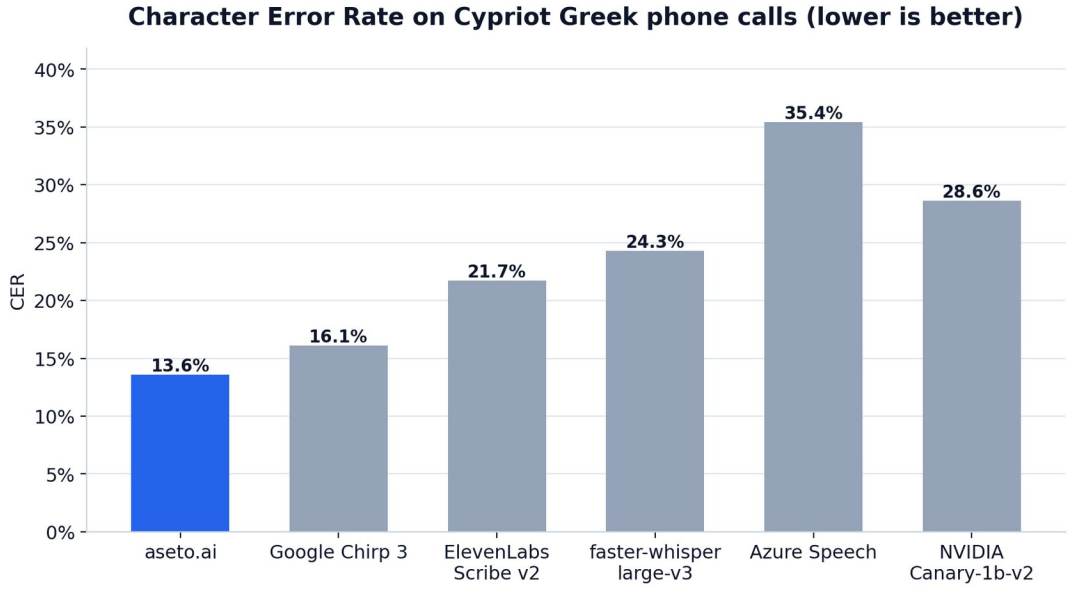

Figure 2. Character Error Rate by system (lower is better).

aseto.ai's 23.9% WER is the best result by a clear margin. The strongest competitor, Google Chirp 3, trails at 29.1% — meaning aseto.ai removes roughly one in five of the word errors Chirp 3 makes (an 18% relative reduction). Every other system sits above 36% WER, and the two weakest — Azure Speech and NVIDIA Canary-1b-v2 — have word error rates of roughly double aseto.ai's. CER tells a consistent story: aseto.ai is again lowest by a clear margin (13.6%), and the top of the ranking is unchanged. The two metrics diverge only at the weak end — Canary's CER (28.6%) is lower than Azure's (35.4%), the reverse of their WER order — which reflects that Canary's errors are more often whole-word breakdowns than character-level

drift. The agreement between the two metrics at the top indicates aseto.ai's lead is not an artefact of morphological scoring sensitivity.

The spread among the general-purpose systems themselves — a WER range from 29.1% to 50.4% — is also informative. It shows that performance on Cypriot Greek is not a solved, commoditised capability: results vary by more than 20 points across major commercial and open systems, and none approaches the accuracy of a model built for the dialect.

4.1 Where the gap comes from

The aggregate numbers are easier to read through concrete examples. The two cases below are single utterances from the test set — one from an IVR interaction, one from an appointment-booking call — comparing aseto.ai against all five baselines. Both illustrate a recurring failure mode: faced with a Cypriot Greek word the system has not been trained on — whether a local place name or a dialect verb form — the general-purpose systems substitute a phonetically similar Standard Greek word or a non-word, changing the meaning of the sentence.

Example 1 — IVR call, caller asking to be transferred to a human agent.

Reference: *με με λειτουργό για το αλεθρικό*

System	Hypothesis	WER	CER
aseto.ai	**με με λειτουργών για το αλεθρικό**	**0.17**	**0.06**
ElevenLabs Scribe v2	με λειτουργ με λειτουργών για το ηλεκτρικό	0.50	0.45
Azure Speech	με λειτουργούν για το αλιευτικό	0.50	0.29
Google Chirp 3	με λειτουργό με λειτουργών για το ηλεκτρικό	0.50	0.48
faster-whisper large-v3	με λειτουργόνια το αλευρικό	0.67	0.19
NVIDIA Canary-1b-v2	με λειτουργών για το αλεύριχο	0.50	0.26

aseto.ai's only error is a single inflectional ending (*λειτουργό → λειτουργών*); the content words, including the place name, are intact. Every other system mis-recognises *Αλεθρικό* — a village in Cyprus. The commercial systems substitute a common Standard Greek word (*ηλεκτρικό*, "electrical", or *αλιευτικό*, "fishing"), while the open-weight models produce non-words (*αλευρικό*, *αλεύριχο*). Several systems also mishandle the repeated word at the start of the utterance. This is a named-entity failure: the village is precisely the information an IVR agent needs to route or answer the call, and only aseto.ai recovers it. A general-purpose system has no exposure to small Cypriot place names, so it defaults to whatever common word sounds closest — and the routing signal is lost.

Example 2 — appointment-booking call, caller describing a symptom.

Reference: *να σου πω πονώ το πόδι μου* (“let me tell you, my foot hurts”)

System	Hypothesis	WER	CER
aseto.ai	**σου πω πονώ το πόδι μου**	**0.14**	**0.12**
Azure Speech	να σου πω πόνο το πόδι	0.29	0.23
ElevenLabs Scribe v2	να σου πω πώς το αποδίδω	0.43	0.38
Google Chirp 3	να σου πω που να το ποδήλατο	0.57	0.35
faster-whisper large-v3	να σε πω πω να το πω δίπλα	0.71	0.46
NVIDIA Canary-1b-v2	λάσομπο πόνο το φοβίζουμε	0.86	0.62

Aseto.ai drops one short function word but preserves the entire clinical content — *πονώ το πόδι μου* (“my foot hurts”). Every other system degrades the part that matters most: ElevenLabs produces *πώς το αποδίδω* (“how I attribute it”), Google produces *ποδήλατο* (“bicycle”), and the two open-weight models break down further still, with Canary producing a largely unintelligible string. In each case the symptom — the actionable information in the call — is lost. The caller here spoke with a heavy Cypriot accent over a phone line, which is what the competing systems failed on; for a Cypriot Greek voice agent, that is not an edge case but the normal operating condition of every call. For a voice agent routing a medical appointment, this is the difference between a usable transcript and a misrouted call.

Across both examples the pattern is consistent with the analysis in Section 2: the systems are not failing on audio quality but on Cypriot-specific language — local place names and dialect word forms — defaulting to Standard Greek vocabulary that the caller did not use. In each case the substituted word carries the information the voice agent depends on, so the error is not cosmetic but operational.

---

5. Discussion

The benchmark is designed to isolate one question: how well does each system transcribe the language aseto.ai’s agents actually operate in? The result — a 23.9% WER against 29.1–50.4% for five major commercial and open systems — indicates that general-purpose STT quality on mainstream languages does not transfer to Cypriot Greek, and that closing the gap requires data and modelling targeted at the dialect itself rather than at Greek in general.

For a voice agent, this gap is not cosmetic. WER at the STT layer sets a ceiling on every downstream component; the 5-point absolute reduction over the closest competitor — and the near-halving of errors against the weakest — translates into fewer misunderstood intents, fewer failed turns, and fewer escalations to a human.

Running the model on aseto.ai's own infrastructure additionally gives control over latency, cost, and data handling — properties that matter for production telephony. Inference time was recorded during evaluation but is not a controlled speed benchmark: the runs were not configured for one. The open-weight models are not comparable on timing — NVIDIA Canary-1b-v2 was run with batching enabled, which alone accounts for its very short wall-clock time. The three commercial cloud services are also subject to network round-trip overhead unrelated to model speed. The report therefore makes no head-to-head speed claim.

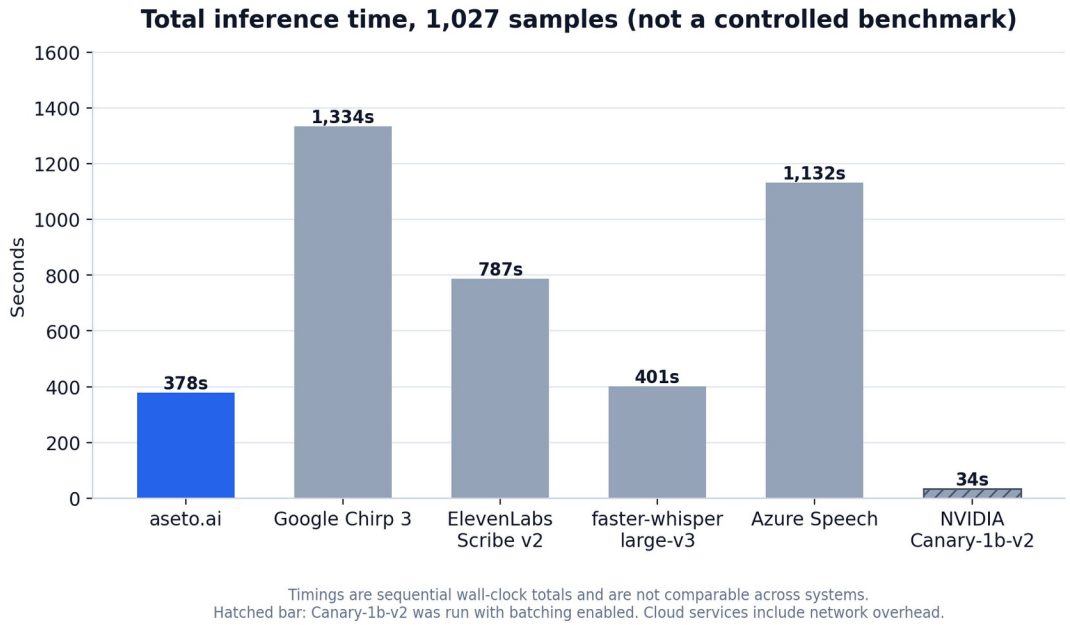

Figure 3. Total inference time over 1,027 samples. Not a controlled benchmark — see text.

Latency is still worth noting, because of where STT sits in a voice agent. Each caller turn passes through a chain — speech-to-text, then a language model, then text-to-speech — and the caller hears silence until the whole chain completes. Transcription latency is the first link in that budget, and time spent there is time the downstream LLM and TTS steps cannot get back. Minimising it is therefore valuable independent of accuracy, and serving the model on aseto.ai's own infrastructure removes the per-request network hop that a cloud STT API adds to every turn.

Limitations. Several caveats should be read alongside these results. The test set is modest in size — 1,027 samples, about 1 hour 8 minutes of audio — and is drawn from customer-support and IVR domains; results may differ on other domains or in different acoustic conditions. aseto.ai's model is purpose-built for exactly this setting, while the five baselines are general-purpose systems not optimised for Cypriot Greek; the comparison measures fitness for this specific task rather than general STT quality. All systems were run with a single Greek language setting and each provider's default configuration, so it is possible a competitor could perform somewhat better under provider-specific tuning. Finally, the commercial systems are

continuously updated: the figures reflect each provider as accessed on 18 May 2026 and may change over time.

---

6. Conclusion

On 1,027 real Cypriot Greek phone-call samples, aseto.ai’s specialised STT model achieves a 23.9% WER — the lowest of all systems tested, an 18% relative improvement over the strongest competitor and roughly half the error rate of the weakest. The comparison against three major cloud services (ElevenLabs Scribe v2, Azure Speech, Google Chirp 3) and two open-weight models run locally (faster-whisper large-v3, NVIDIA Canary-1b-v2) indicates that the advantage comes from treating Cypriot Greek as a first-class target rather than a sub-case of Greek — which is the basis on which aseto.ai builds voice agents for this market.