Roger? Probabilistic Certificates for Safety-Critical ATC Speech Recognition

Ben Wooding *ben.wooding@vanderbilt.edu*
Institute for Convergent Software Integrated Systems, Vanderbilt University, USA
Taylor T. Johnson *taylor.johnson@vanderbilt.edu*
Institute for Convergent Software Integrated Systems, Vanderbilt University, USA

Reviewed on OpenReview: *https://openreview.net/forum?id=9Pj3FcSfgz*

Abstract

When a speech-to-text (STT) model transcribes an air-traffic-control (ATC) clearance, the safety question attached to an individual transmission is: "for *this* message, over the noise the very-high-frequency (VHF) radio channel will add, how likely is the decoded content to change?" We demonstrate an approach to give every message its own distribution-free certificate. For each message, model, specification, and channel quality, we count k violations among N=1000 *i.i.d.* noise draws yielding an exact-binomial upper limit on the probability that channel noise changes the decoded message. We certify decode *robustness* relative to the model's own (possibly incorrect) clean decode, under a specified channel model, rather than recognition accuracy, and we provide 452,544 per-message certificates covering four publicly available STT systems — QuartzNet15x5, DeepSpeech2, Whisper (74M parameters), and Whisper-ATC (1.54B parameters, fine-tuned on ATC speech) — on two public ATC datasets, *none* of which can be handled by sound verification techniques that provide worst-case guarantees. Three specification families are considered to describe the equivalence between sent and received messages (exact equivalence, semantic equivalence, and call-sign equivalence), where relaxations on the exact-equivalence specifications determine a "changed-but-forgiven" probability mass. Semantic specifications are non-formal, delegating *equivalent meaning* to a learned embedding which has not been verified; the relaxation is desirable but, as we show in this work, *currently too dangerous to trust in deployed settings*, as critical call-sign flips can occur undetected by the semantic floor. We show the probability that a draw flips the call-sign *and* passes the cosine floor is exactly measurable from the stored transcripts. Pooled over draws, that audit records that in-domain fine-tuned model produces by far the fewest call-sign flips, yet the flips it does produce are the most forgiven, *i.e.* the least visible to a semantic specification. Because the certificate is reference-free, needing no ground-truth transcript, we can also certify audio that Whisper-ATC was not fine-tuned on: across seven datasets disjoint from the fine-tuning sets, including *live ATC messages*, certified fractions fall from 0.93 to 0.81 at 60 dB on the most closely matched pair. Our certificate framework enables *for the first time* a method to separate the four models by their robustness to a specified channel-noise perturbation.

1 Introduction

Air-traffic control still runs by voice. A controller keys a very-high-frequency (VHF) radio, speaks a clearance, and every party that acts on that transmission must recover the content from noisy narrow-band audio; flight crews and increasingly automatic transcription systems. The transcription systems sit outside the controller-pilot loop: they feed assistance and safety-net tooling (call-sign highlighting, read-back checks, archiving), so *a transcription error propagates silently* to whatever uses the transcript. A safety hazard occurs if channel noise changes the decoded *call-sign*, a clearance would be attributed to the wrong aircraft (Kocour et al., 2021). How hazardous a corrupted call-sign is depends on the traffic. The call-sign matching a different

aircraft risks the wrong aircraft following critical instructions, while a garbled call-sign leads the intended aircraft to ignore a critical instruction believing it meant for a different flight[1].

This is a recognized, recurring hazard rather than a hypothetical one: FAA regulation forbids acting on a clearance issued to another aircraft (14 CFR 91.123(e)) and mandates call-sign read-back precisely to catch the error (AIM 4-4-7) (Federal Aviation Administration, 2024; 2026), and both ICAO and EUROCONTROL track "call-sign confusion" as a named hazard (SKYbrary, 2024). It recurs at scale: EUROCONTROL's voluntary incident reporting logged more than 4,000 call-sign-confusion reports over 2018–2022 (9.2% of all reports) (EUROCONTROL, 2024). Historically, the UK CAA counted 482 in 1997 alone (UK Civil Aviation Authority, 2000), and the classic NASA ASRS study attributed 462 incidents (altitude deviations, runway incursions, and traffic conflicts) to call-sign errors (Monan, 1983). It is a procedural, human-factors hazard, outside the scope of airworthiness directives (which govern equipment), so a *per-message probabilistic certificate* for the robustness of the decoded call-sign is a valuable tool. Crucially, that hazard is a property of an individual transmission.

An operator sending a message does not want to know the receiver model's average behavior over a dataset; they want a statement of the form *"for this specific message, on this channel, the probability that noise changes what the model heard (and above all who it heard) is provably below some number."* We want to know which messages in front of the operator are robust and how robust. We provide in this paper an approach, usable per-message, to quantify the probability that channel noise changes the decoded content of a message (its decode *robustness*). Throughout, the *clean decode* is the model's transcript of the recorded utterance with no added noise; robustness is measured by change relative to that string, while *accuracy* (word error rate against human references) is reported separately.

Three features separate this from the neural-network verification literature it builds on. First, ATC speech recognition has scarcely been studied. Second, unlike the ℓ_p-norm perturbation balls certified in the literature (whose operational meaning around, say, an image is unclear) our perturbations are physically motivated: band-limited Gaussian noise at a target signal-to-noise ratio (SNR) models additive noise in the VHF voice passband. A third difference follows from the domain's data situation. ATC datasets are largely private, and the few public ones are typically used to fine-tune models, so any evaluation that requires ground-truth transcripts is usually confined to data the model has already seen; making them challenging to analyze. Our certificates are *reference-free* — they measure decode robustness against the model's own clean decode, without consulting any transcript, so they can be computed on any audio at all.

Concretely, we certify each of 1,500 ATCOSIM and 857 ATCO2 utterances (Hofbauer et al., 2008; Zuluaga Gomez et al., 2024) under band-limited Gaussian channel noise at eight SNRs, for four publicly available speech-to-text (STT) systems, under a layered family of three robustness specifications — exact, semantic (meaning preserving), and call-sign — from N=1000 i.i.d. noise draws per (message, SNR) pair: 452,544 exact-binomial certificates in total, each valid at confidence $1 - 10^{-6}$. The certificate is a Clopper–Pearson (Clopper & Pearson, 1934; Lehmann & Romano, 2005) upper limit on a per-message violation probability, where the difficulty lies in what the specification should say (an exact-transcript invariant overcounts harmless rewordings; a semantic floor delegates safety to a learned embedding), and in whether the specification itself can be trusted. Our main claims rest on the two specifications that involve no learned oracle: *exact transcript* and *call-sign*.

The semantic relaxation is *desirable*, since it forgives harmless rewordings, but *dangerous*, since its learned oracle admits safety-relevant counterexamples, so we audit it rather than trust it, pairing it with a mandatory call-sign check, and regard it as interesting future work. No sound verification technique can currently handle the ATC STT models we consider (including a 1.54B parameter model), so all guarantees are probabilistic. A little unintuitively, because channel quality is measured as an SNR in decibels, each higher-dB setting injects exponentially less noise energy relative to the speech, leaving the corrupted waveform close to the clean utterance; certificates are therefore easier to obtain in the high-SNR regimes and degrade essentially monotonically for lower-dB settings.

[1]Recently, there was even an incident of two planes with identical call-signs — Funk, Josh. "2 Planes with the Same Flight Number Avoid Possible Disaster Thanks to Quick-Thinking Controller." AP News, Associated Press, 14 Aug. 2026.

Contributions.

- **Scalable per-message exact-binomial certificates.** For each message x, model, specification, and SNR, k violations in N=1000 draws gives an upper bound on the risk of incorrectly determining specification satisfaction (Theorem 1).
- **Comparable specification families.** Exact-transcript, semantic (embedding cosine ≥ 0.90 / 0.95), and call-sign specifications instantiate the same oracle template, with the former relaxed to acquire the latter two. Theorem 3 shows the count difference $d = k_{\text{msg}} - k_{\text{sem}}$ is itself binomially distributed, so we can certify the "changed-but-forgiven" mass; the probability that noise alters the transcript in a forgivable way.
- **The danger with semantic equivalence oracles.** We formalize correctness levels for certificates that depend on a learned oracle, proving an audited upper bound on the *true* semantic violation probability under an audit-representativeness assumption, and then remove that assumption entirely for the ATC call-sign safety check. The per-message probability that a noise draw flips the call-sign *and* passes the cosine floor is exactly measurable from the stored transcripts of the same decode pass, so we certify it. At the safety-grade level (digit/NATO content of the call-sign changed) the audit quantifies precisely the likelihood of call-sign misattribution messages being forgiven.
- **An in-domain vs. out-of-domain comparison beyond word error rate (WER).** Whisper-ATC (a fine-tuned `whisper-large-v3`) is separated from QuartzNet, DeepSpeech2, and Whisper more sharply by certified-message fractions and per-message SNR budgets than by WER, with the oracle audit showing its rare residual call-sign errors are the most likely to hide beneath a semantic floor, *motivating the call-sign specification as a mandatory companion in this domain.*
- **Assessment of Whisper-ATC on messages outside of training set.** Due to dataset scarcity, Whisper-ATC was trained on both ATCOSIM and ATCO2, so we download a *smaller dataset of live ATC messages* from 6 different countries and assess, using our framework, how robust Whisper-ATC is to messages outside of its training data.

2 Related work

Sound neural-network verification. Complete solvers (Katz et al., 2017), abstract interpretation and zonotopes (images of a box under an affine map, propagated layer-by-layer as sound over-approximations of the reachable activations) (Gehr et al., 2018; Singh et al., 2018), interval bound propagation (Gowal et al., 2019), linear relaxation (Zhang et al., 2018), and branch-and-bound frameworks (Xu et al., 2021) answer a per-input, worst-case question: is the decision *invariant* over an entire perturbation set? Star-set reachability extends the reach to some recurrent architectures (Tran et al., 2019), and has recently been carried over from input perturbations to perturbations of the weights themselves (Zubair et al., 2026), relational relaxations to LSTMs (Ko et al., 2019; Ryou et al., 2021), and dedicated relaxations to small transformers (Shi et al., 2020); Li et al. (2023) survey the field and VNN-COMP (Ladner et al., 2026) documents the state-of-the-art toolboxes, each remains orders of magnitude below deployable STT systems. The scaling wall is now being met from the other side, by deliberately weakening soundness: Marzari et al. (2025) tighten linear-relaxation bounds using sampled intermediate reachable sets and retain only a probabilistic guarantee. That trade is the same one we make, but at a different level — they relax the bound propagation and still require it, whereas we discard bound propagation entirely and treat the decoder as a black box, which is what admits a 1.54B-parameter autoregressive model. Appendix B highlights the bottlenecks with attempting sound approaches on our four models.

Probabilistic and statistical certification. The scenario approach (Calafiore & Campi, 2006; Campi & Garatti, 2008; 2018; Garatti & Campi, 2022; Wooding et al., 2026) attaches distribution-free risk bounds to sampled decisions; Anderson & Sojoudi (2023) certify the safety probability of a fixed input under random noise via chance-constrained sampling — the closest prior instrument to ours. Statistical neural-network verification likewise bounds a per-input violation or robustness probability from i.i.d. samples with PAC or

exact-binomial guarantees (Webb et al., 2019; Baluta et al., 2019; Weng et al., 2019; Cardelli et al., 2019); we *adopt* that exact-binomial instrument, with our contribution lying in the specification families we consider and the oracle audit for the mostly unverified domain of ATC STT. Our certificates differ in being exact-binomial bounds on a black-box *decode* predicate, deployed at dataset scale with layered specifications and an audited oracle. The present paper shows the exact-binomial (Clopper–Pearson) limit (Clopper & Pearson, 1934) is a tighter instrument for per-message verification. Reporting verification outcomes per input rather than as a single aggregate is itself an established position: Bosman et al. (2025) study the *distribution* of per-input robustness across a benchmark and argue that the distribution variance can be concealed by summarizing metrics. We share that position; they measure exact robustness radii with complete solvers on small networks, whereas we bound a per-input violation *probability* for decoders beyond sound solvers, providing a confidence statement to an operator to act on. Conformal methods (Vovk et al., 2005; Angelopoulos et al., 2024; Ghosh et al., 2023; Lin & Bansal, 2024; Hashemi et al., 2025) certify coverage of prediction sets rather than per-input violation probabilities of a fixed black-box predicate. Randomized smoothing (Cohen et al., 2019) and flow-matching (Sasaki et al., 2026) certify a smoothed surrogate model, including an automatic speech recognition (ASR) instantiation (Olivier & Raj, 2021); our certificates are statements about the deployed decoder itself, at the cost of being probabilistic over the channel noise rather than adversarial.

ASR robustness and ATC speech. Deployed STT systems are known to be attackable in the worst case (Carlini & Wagner, 2018), and worst-case training defenses (Madry et al., 2018) do not transfer directly to transduction. ATC speech is a (narrow-band, jargon-heavy, safety-critical) distinct regime. The dataset ATCOSIM provides clean non-prompted controller speech (Hofbauer et al., 2008), and ATCO2 provides real noisy channel recordings with transcripts (Zuluaga Gomez et al., 2024). Call-sign recognition has been studied as its own task (Kocour et al., 2021); and Whisper-ATC (van Doorn et al., 2024) fine-tunes `whisper-large-v3` on ATC datasets. We certify QuartzNet15x5 (Kriman et al., 2020), DeepSpeech2 (Amodei et al., 2016), Whisper (Radford et al., 2023), and Whisper-ATC; the semantic specification uses sentence embeddings (Reimers & Gurevych, 2019) over a distilled MiniLM encoder (Wang et al., 2020).

3 Probabilistic certificates under channel noise

Sound verification cannot supply the guarantees we need for the models we study, with the reasons differing per architecture. For the two connectionist temporal classification (CTC) models, the shared digital signal processor (DSP) front end degrades from a usable 4.6× enclosure at 50 dB to a vacuous one at 40 dB and below. The forward set propagation reaches logit radii of 10^{96}–10^{110}, far beyond the margins that decide the greedy decode. For the Whisper family, the issue is architectural, no scalable sound relaxation covers softmax self-attention inside an autoregressive decoder, so there is no bound to measure at all. Even a tight logit enclosure would not finish the job, because CTC collapse and the sampling loop map logit sequences to strings through non-monotone argmax structures, with the call-sign readout adding a further string-level extractor on top. Appendix B gives further details for each model, and the remainder of this paper therefore works with probabilistic certificates that require only sampled forward passes through the model.

A frozen STT model f maps audio to a token sequence; a readout Φ extracts the content a specification cares about (the full normalized transcript, the call-sign span, or an embedding of the transcript). A *message* x is one clean recorded utterance. Deployment corrupts x with additive channel noise: at SNR level s the noise law $\mathcal{P}_{x,s}$ produces δ, band-limited Gaussian noise scaled to the target SNR relative to x. The single primitive we then use is a violation indicator

$$g_\sigma(x,\delta) \;=\; \mathbf{1}\big[\Phi_\sigma\big(f(x+\delta)\big) \neq \Phi_\sigma\big(f(x)\big)\big]\,, \qquad \delta \sim \mathcal{P}_{x,s}, \tag{1}$$

for specification σ, where $x+\delta$ is representative of a perturbed input (not strictly additive). The object we certify is the per-message violation probability

$$V_x^\sigma(s) \;=\; \mathbb{P}_{\delta\sim\mathcal{P}_{x,s}}[g_\sigma(x,\delta)=1]\,. \tag{2}$$

Equation (1) is computed against the model's own clean decode $\Phi_\sigma(f(x))$: the certificate measures decode *robustness* under the channel, deliberately decoupled from accuracy, which we report separately as word

error rate (WER) against reference transcripts. A model can be robust but inaccurate: a stably mis-decoded message still certifies as robust, so a robustness certificate is a safety statement only when read together with an accuracy check on the clean decode (discussed more in Section 7). Two observations support this pairing: (i) deployed STT models report accuracy but rarely robustness (the sensitivity of the model to output changes), so we provide that half, (ii) the accuracy check need not be statistical: in a sender-side deployment, the operator (who knows what they said) can check the locally decoded clean transcript for its accuracy, with our certificates bounding the probability that the receiver decodes something different.

3.1 The exact-binomial certificate

Fix a message x, specification σ, and SNR s; draw $\delta_1, \ldots, \delta_N \overset{\text{i.i.d.}}{\sim} \mathcal{P}_{x,s}$ and count violations $k = \sum_{i=1}^{N} g_\sigma(x, \delta_i)$. Because the draws are *i.i.d.* and the decode is treated as a deterministic function of the draw, k is exactly $\mathrm{Bin}(N, V_x)$-distributed with $V_x = V_x^\sigma(s)$, and the classical one-sided exact-binomial upper confidence limit applies.

Definition 1 (Exact-binomial upper limit)**.** For $N \in \mathbb{N}$, $k \in \{0, \ldots, N\}$, and $\beta \in (0, 1)$,

$$\mathrm{CP}^+(k, N, \beta) \;=\; \inf\Big\{ p \in [0,1] \;:\; \mathbb{P}_{K \sim \mathrm{Bin}(N,p)}\big[K \le k\big] \;\le\; \beta \Big\}, \tag{3}$$

with the convention $\inf \emptyset = 1$ (so $\mathrm{CP}^+(N, N, \beta) = 1$). For $k < N$, $\mathrm{CP}^+(k, N, \beta)$ equals the $1 - \beta$ quantile of the $\mathrm{Beta}(k + 1,\ N - k)$ distribution.

Theorem 1 (Per-message validity)**.** *Fix any message x, model f, readout Φ_σ, SNR s, and any noise law $\mathcal{P}_{x,s}$. Let k be the violation count from N i.i.d. draws. Then*

$$\mathbb{P}^N\big[\, V_x \;>\; \mathrm{CP}^+(k, N, \beta) \,\big] \;\le\; \beta,$$

where $\mathbb{P}^N$ is over the N draws. The bound depends on the data only through (k, N, β); no property of f, Φ_σ, or $\mathcal{P}_{x,s}$ enters.

Proof. All proofs can be found in the appendix. □

Remark 1. Exact binomiality requires $g_\sigma(x, \cdot)$ to be a fixed measurable function of the draw δ_i alone.Our draws are decoded in batches, and reduced-precision GPU kernels can make a decode depend on the batch size as well as on the draw. A zero-noise control (Appendix D) tests this by decoding the unchanged audio in the same batches. Under PyTorch's default TF32 convolutions, the CTC decodes differ from the batch-1 clean decode for 154 of 4,714 (model, message) pairs, always for whole batches, while repeated batch-1 decodes never differ. We therefore disable TF32 for QuartzNet and for the single affected DeepSpeech2 message; 1 near-tie message still depends on fp32 rounding and is flagged. The fp16 models show the same effect for 72 (Whisper) and 33 (Whisper-ATC) of 2,357 messages. All affected messages are near-ties that the added noise also flips: disabling TF32 moves QuartzNet's certified fractions by at most 0.5 percentage points, and certifying the Whisper models against a clean decode taken in the draws' own batch configuration changes no certification decision (Appendix D). All counts compared in this paper are taken within a single pass, and we never compare counts across distinct forward passes.

Throughout we fix the number of samples $N = 1000$ and confidence parameter $\beta = 10^{-6}$, and we introduce threshold $\tau = 0.05$: a message x is *certified robust* at (s, σ) if $\mathrm{CP}^+(k, N, \beta) \le \tau$, i.e. the violation probability for a new unseen draw is below 5% at confidence $1 - 10^{-6}$. The certificate itself is the number $\mathrm{CP}^+(k, N, \beta)$, which is the deliverable an ATC operator can use. The threshold $\tau = 0.05$ is an operational reporting choice, not a certified safety target, and because the k=0 floor is 0.013721 no N=1000 certificate can assert a per-message violation probability below $\approx 1.4\%$; safety-grade guarantees would require a larger N. At $k = 0$ the bound has the closed form $\mathrm{CP}^+(0, N, \beta) = 1 - \beta^{1/N}$, a *floor* below which no N-draw certificate can go (numerically 0.013721 at our N, β); and CP^+ is monotone in k, so a certified message is one whose count k stays at or below a deterministic threshold $k^*(N, \beta, \tau)$, equal to $k = 20$ violations at our chosen (N, β, τ). Each certificate is marginal at β; when consulting a family of K certificates simultaneously, one must pay a union bound $K\beta$ (Dunn, 1961). Concretely, all (message, SNR) certificates of one model and specification

($K = 18{,}856$) retain joint confidence $1 - K\beta > 0.98$; a larger family needs a smaller β and hence a smaller budget k^*. Reported certified *fractions* summarize per-message decisions and are not joint statements.

The wait-and-judge scenario bound $\bar{\varepsilon}(k, N, \beta)$ of Campi & Garatti (2018); Garatti & Campi (2022) certifies the same quantity from the same statistic and is more general in its applications. Per input, that generality costs: k is exactly binomial here, CP^+ is the tightest monotone bound obtainable from (N, k), and the scenario bound is never the tighter of the two (only equal to it when both use $k = N$). Averaged over this paper's certificates with $\mathrm{CP}^+ < 0.999$, the scenario bound is relatively looser by 11.6%. We therefore report CP^+ as the primary bound in this paper.

4 A layered specification family

One decode pass yields the transcript when a message is passed through one of the models. Every specification uses the same transcript, so all specifications are measured *on the same draws*. We consider three specifications.

Exact transcript (*msg*). Φ_{msg} is the full normalized greedy transcript (lower-cased, numerals spelled digit by digit as spoken in ATC, e.g. "81" → "eight one", other characters outside `[a-z' ]` removed); any change in the transcript is considered a violation. This is the strictest available invariant and is assumption-free, but at the cost of over-counting harmless rewordings as violations. Many over-counted changes are alternative textualizations of the same speech rather than rewordings: spelling variants (notably proper nouns), re-tokenizations, and dropped or repeated greetings. Standardized phraseology makes genuine rewording of command content rare; Section 5.1 measures the re-tokenization share of call-sign flips.

Call-sign (*cs*). Φ_{cs} extracts the leading call-sign span from the transcript (Kocour et al., 2021); where flipping a call sign is a safety-critical hazard. We fix Φ_{cs}, a transcript is first normalized as for *msg*. Φ_{cs} returns the leading run of tokens of the normalized transcript, stopping before the first token that belongs to a fixed list of 73 ATC command words e.g. `descend`, `climb`, `maintain` ... (full list released with the code). The first token is always retained, even when it is itself a command word, so the readout is never empty on a non-empty transcript; and the span is capped at 8 tokens, so a transcript containing no command word yields its first 8 tokens. For example, *"klm eight four seven descend flight level one zero zero"* yields *"klm eight four seven"*. The map involves no model, no embedding and no learned parameters, avoiding any oracle assumptions. It is heuristic, a transcript whose call-sign is not utterance-initial will yield spans that are not exactly the call-sign. Again, our specifications can show *robustness* is certified, but do not make any claims about ground truth. A *cs* certificate therefore concerns the stability of this extractor's output, which is the aircraft's identity only when the extractor finds the call-sign. On the 249 call-signs among our ATCO2 messages (Appendix E), the leading span reproduces the call-sign's digits and phonetic letters in 62.6% of cases: 82.1% when the call-sign opens the utterance and 35.6% otherwise, typically in read-backs.

Call-sign identity (*id*). Φ_{id} searches the whole transcript for the two ICAO call-sign forms, a telephony designator with a flight number ("lufthansa four two two three") or a registration in phonetic letters and digits ("oscar kilo papa mike bravo"), and returns the call-sign's identity (designator and ordered digit and phonetic-letter signature), or an empty value when it finds none (Appendix E). It reproduces 85.5% of the annotated signatures (75.0% of non-initial ones); its rules were developed on this small annotation, 79% of which comes from one airport, so these figures are optimistic. As a deterministic function of the transcript, Φ_{id} satisfies Lemma 2 and Theorem 1 unchanged; Table 3 reports certificates for both readouts.

Semantic floor (*sem*$_\chi$). With the sentence-embedding map $e(\cdot)$ from all-MiniLM-L6-v2 (Reimers & Gurevych, 2019), and T_0, T_δ the clean and perturbed transcripts, $g_{\mathrm{sem}_\chi}(x, \delta) = \mathbf{1}[\cos(e(T_\delta), e(T_0)) < \chi]$; we store semantic equivalences $\chi \in \{0.90, 0.95\}$. The semantic equivalence forgives rewordings that preserve embedding-cosine similarity akin to preserving meaning. ATC phraseology is narrow and formulaic, so transcript pairs concentrate near the top of the cosine range; Section 5 audits exactly this compressed regime and shows safety-relevant call-sign changes clearing even the 0.95 floor.

Lemma 2 (Per-draw monotonicity)**.** *Fix a draw δ and compute all readouts from the same decoded transcript T_δ. Then $g_{\mathrm{cs}}(x,\delta) \le g_{\mathrm{msg}}(x,\delta)$ and $g_{\mathrm{sem}_\chi}(x,\delta) \le g_{\mathrm{msg}}(x,\delta)$ for every semantic equivalence $\chi \le 1$. Consequently $V_x^{\mathrm{cs}} \le V_x^{\mathrm{msg}}$, $V_x^{\mathrm{sem}_\chi} \le V_x^{\mathrm{msg}}$, and the corresponding counts satisfy the same inequalities, so a message certified under* msg *is certified under both weaker specifications at the same (N,β,τ).*

The exact transcript robustness is the strictest specification, but the two relaxations are incomparable with each other: a call-sign flip need not breach the cosine floor of the semantic equivalence, so *cs* is a companion specification, not a member of a chain.

Theorem 3 (Nested-specification decomposition)**.** *Fix x, s, a floor χ, and N i.i.d. draws, with all readouts computed from the same decoded transcripts. Let k_{msg} and k_{sem} be the same-pass exact and semantic counts and $d = k_{\mathrm{msg}} - k_{\mathrm{sem}} \ge 0$. Then d is exactly* $\mathrm{Bin}(N,\, D_x)$*-distributed, where*

$$D_x \;=\; V_x^{\mathrm{msg}} - V_x^{\mathrm{sem}_\chi} \;=\; \mathbb{P}_\delta\big[\, T_\delta \neq T_0 \;\wedge\; \cos(e(T_\delta), e(T_0)) \ge \chi \,\big]$$

is the changed-but-forgiven *mass. Hence $\mathbb{P}^N[D_x > \mathrm{CP}^+(d,N,\beta)] \le \beta$, and with probability at least $1-2\beta$ both $V_x^{\mathrm{sem}_\chi} \le \mathrm{CP}^+(k_{\mathrm{sem}},N,\beta)$ and $D_x \le \mathrm{CP}^+(d,N,\beta)$ hold.*

5 Auditing the semantic equivalence oracle

The semantic specification quietly changes what is being certified. Exact match is a fact about strings; the cosine floor is a *learned verdict*. Before reporting semantic certificates at all, we make explicit what they claim, in three correctness levels, and then audit the verdict.

L0 (oracle-free). The exact and call-sign specifications use no learned component beyond the model under test. A certificate at L0 needs no assumption about any oracle.

L1 (oracle-as-specification). A semantic certificate at L1 is a statement *about the oracle's verdict*: with confidence $1-\beta$, the probability that a noise draw makes the cosine fall below χ is at most $\mathrm{CP}^+(k_{\mathrm{sem}},N,\beta)$. This is sound by definition (the specification *is* the oracle), we report all our sem_χ certificates at this level. Its safety relevance, however, is in the trustworthiness of the oracle: if meaning can change while cosine stays high, an L1 certificate can be tight but with high deployment risk.

L2 (audited oracle). Let $g^*(x,\delta)$ be the ground-truth semantic violation indicator — the transcript's operational meaning changed — and $V_x^* = \mathbb{P}[g^*=1]$ the true semantic violation probability. Meaning cannot change without the string changing, so $g^* \le g_{\mathrm{msg}}$ pointwise. The draws the oracle can mislabel one-sidedly are the *forgiven* ones (g_{msg}=1, g_{sem}=0): a false alarm (flagged but harmless) only inflates k_{sem} and keeps L1 conservative, whereas a false pass (forgiven but harmful) is invisible to it.

Assumption 1 (Audit representativeness)**.** An audit procedure supplies $\rho_{\mathrm{FP}} \in [0,1]$ such that $\mathbb{P}\big[g^*{=}1 \,\big|\, g_{\mathrm{msg}}{=}1,\, g_{\mathrm{sem}}{=}0\big] \le \rho_{\mathrm{FP}}$ for the messages to which the bound is applied.

Theorem 4 (Audited semantic bound)**.** *Under Theorem 3 and Assumption 1, for each message x,*

$$V_x^* \;\le\; V_x^{\mathrm{sem}_\chi} + \rho_{\mathrm{FP}}\, D_x, \qquad \textit{and hence} \qquad \mathbb{P}^N\Big[\, V_x^* \;>\; \mathrm{CP}^+(k_{\mathrm{sem}},N,\beta) + \rho_{\mathrm{FP}}\,\mathrm{CP}^+(d,N,\beta)\Big] \;\le\; 2\beta.$$

No semantic specification can currently satisfy Assumption 1, because the forgiven set really does contain harmful draws at arbitrarily high cosine: across our experiments the most forgiving call-sign-flipping draw attains cosine 1.000 (at reported precision), and the counterexamples below are digit-changing call-sign flips whose quoted cosines all clear the tightest 0.95 floor. Any L2 claim therefore inherits whatever epistemics produced ρ_{FP}, and we do not report L2 numbers in this paper; instantiating a defensible ρ_{FP} is an important topic for future study. Instead, for the ATC call-sign flipping hazards we remove this assumption.

5.1 The call-sign oracle audit

The reason Assumption 1 is needed is that g^* is not computable. But the ATC safety core does not need all of g^*: the wrong-aircraft hazard is a *string property* of the stored transcripts. For any flip predicate

$F(T_0, T_\delta) \in \{0, 1\}$ computable from the transcript pair, define the *forgiven-flip mass*

$$\eta_F(x; \chi, s) \;=\; \mathbb{P}_{\delta \sim \mathcal{P}_{x,s}}\big[\, F(T_0, T_\delta) = 1 \;\wedge\; \cos(e(T_\delta), e(T_0)) \geq \chi \,\big]. \tag{4}$$

Proposition 5 (Assumption-free call-sign audit)**.** *Fix x, s, χ, and N i.i.d. draws whose decoded transcripts are stored. The per-draw indicator in Equation* (4) *is a deterministic function of the stored pair (T_0, T_{δ_i}), so the count $m = \sum_{i=1}^{N} \mathbf{1}[F{=}1 \wedge \cos \geq \chi]$ is exactly $\mathrm{Bin}(N, \eta_F(x; \chi, s))$-distributed and*

$$\mathbb{P}^N\big[\, \eta_F(x; \chi, s) \;>\; \mathrm{CP}^+(m, N, \beta) \,\big] \;\leq\; \beta.$$

η_F is the safety-relevant mass that the floor χ *forgives*: the probability that a single noise draw re-addresses the message while the semantic certificate declines to count it. Reporting CP^+ bounds on η_F next to the sem_χ certificate turns "is 0.95 an appropriate floor?" into a measurable per-message quantity.

We run our audit on three detectable message changes:

1. *raw* extractor-string inequality — the experiment's *cs* specification, which counts re-tokenization artifacts (spacing, hyphenation) as flips;
2. *orthographic normalization* — non-alphanumerics dropped before comparison;
3. and *digit/NATO-signature flips* — the digit or phonetic-alphabet content of the call-sign changed (the ordered digit words and phonetic letters in the span), *i.e.* the message now addresses a *different aircraft* whenever the span contains the call-sign (as in Section 4).

The last is the most important for the safety-critical setting. The raw and normalized levels differ on only a small share of flipping draws (1.6% of raw flips pooled over the experiments, at most 3.0% in any model–dataset cell). The digit/NATO predicate is conservative. A single-digit flip such as TAROM 7533 → TAROM 753 yields a plausible call-sign, and whether it misattributes a clearance or causes a missed call depends on which aircraft are on the frequency when the message is sent. Reported confusion incidents involve similar call-signs on the same frequency (SKYbrary, 2024), the class single-digit drops and doublings produce, so we count every digit/NATO change. ATCO2 lets us examine this on real traffic, as each recording lists the aircraft near the airport at recording time (Appendix E). For the 134 Whisper-ATC messages whose own call-sign is on the list, 0 of the 54,002 identity-changing draws named another listed aircraft: 7.2% lost the call-sign and the rest named no listed aircraft. Flips in this traffic are thus mostly missed or garbled calls, which remain hazardous for time-critical instructions.

Table 1 reports, per model and dataset, the number of call-sign-flipping draws (normalized predicate) and the share of them forgiven at each floor, plus the digit-level restriction. Pooled over all models, datasets, and SNRs, 24.4% of call-sign-flipping draws pass floor 0.90 and 11.5% pass floor 0.95; restricted to digit-changing flips the pooled pass rate at 0.95 is 7.5%. In the pooled counts, tightening the floor from 0.90 to 0.95 roughly halves the forgiven share. All four models lead in at least one column of a dataset block in Table 1.

Whisper-ATC produces at least 25× fewer call-sign-flipping draws than the CTC models (225,404 vs. 5,666,852–6,245,336 on ATCOSIM), but *conditional on flipping*, its errors are the most cosine-forgiven: 51.8% pass floor 0.90 and 36.2% pass floor 0.95 on ATCOSIM, versus 8.9–36.7% and 3.8–17.5% across the other model–dataset cells. The digit level is sharper still: 44.0% of Whisper-ATC's digit-changing call-sign flips pass floor 0.95 on ATCOSIM (40,471 draws), versus 1.2–10.7% for the other models. The safety-relevant point is that on these models the semantic floor is at its least informative when the call-sign errors are hardest to detect. We observed the fine-tuned model fails imperceptibly, dropping or doubling one digit inside an otherwise perfect readback, and a sentence embedding understandably regards the result as nearly the same sentence, even though it may refer to a different aircraft. For this reason, we consider the call-sign specification as a necessary companion to any semantic certificates in this domain. We repeat the audit with two further encoders, all-mpnet-base-v2 (Reimers & Gurevych, 2019) and bge-base-en-v1.5 (Xiao et al., 2024), each at the floor that forgives the same share of all transcript changes as MiniLM at 0.95; they forgive 6.5% and 7.0% of signature-changing call-sign flips, against 7.5% for MiniLM (Appendix C).

Table 1: The call-sign audit, by model and dataset, pooled over the eight SNRs: number of call-sign-flipping draws under the orthographically-normalized predicate; percentage of those draws with cosine ≥ floor (i.e. forgiven by the semantic specification); digit/NATO-signature flips and their pass rate at floor 0.95. Bold marks the best (lowest) value in each column of each dataset block, as discussed in the text.

		normalized call-sign flips			digit/NATO flips	
dataset	model	draws	≥ 0.90 (%)	≥ 0.95 (%)	draws	≥ 0.95 (%)
ATCOSIM	QuartzNet	5,666,852	36.7	17.5	1,174,248	10.7
	DeepSpeech2	6,245,336	31.4	14.7	1,063,770	**5.7**
	Whisper	2,633,637	**22.5**	**11.5**	1,254,327	9.8
	Whisper-ATC	**225,404**	51.8	36.2	**91,922**	44.0
ATCO2	QuartzNet	5,301,347	12.7	4.9	943,980	**1.2**
	DeepSpeech2	5,032,629	19.8	9.0	**191,773**	1.8
	Whisper	2,217,397	**8.9**	**3.8**	929,432	2.9
	Whisper-ATC	**716,585**	33.0	16.2	343,154	16.7
pooled (all cells)		28,039,187	24.4	11.5	5,992,606	7.5

Counterexamples. The audit stores the top floor-passing digit-flip examples per model and dataset, two verbatim examples are included below:

- **Whisper-ATC, ATCOSIM, 25 dB** (cosine 0.9976; 621 of 1000 draws): *"tarom **seven five three three** contact reims one three four decimal four bye"* → *"tarom **seven five three** contact reims one three four decimal four bye"*. One dropped digit turns TAROM 7533 into TAROM 753; the sentence is otherwise untouched.
- **Whisper-ATC, ATCOSIM, 20 dB** (cosine 0.9982; 3 of 1000 draws): *"lufthansa **four two two three** good afternoon radar contact trasading"* → *"lufthansa **four two three** good afternoon radar contact trasading"*. Lufthansa 4223 becomes Lufthansa 423 at cosine 0.9982.

Beyond the counterexamples, the audit collects 264,606 floor-0.95-passing call-sign-flip edit records across the experiments; the digit-flip counterexamples above are the adversarial cases the cosine floor does not detect.

6 Experimental setup

Datasets. ATCOSIM (Hofbauer et al., 2008) provides clean, non-prompted controller speech (a 10,078-utterance dataset; we use a fixed 1,500-utterance subset): in-vocabulary phraseology with no channel noise on the recordings. ATCO2 (Zuluaga Gomez et al., 2024) provides real radio-channel transmissions with gold transcripts; our approach uses 857 messages from the roughly one-hour free release. Every message in both datasets is certified; there is no train/test split at the certificate level because nothing is fitted — each certificate is a self-contained statement about one message.

Noise model. $\mathcal{P}_{x,s}$ adds Gaussian noise band-limited to 300–2500 Hz — the aeronautical VHF voice passband — scaled to a target SNR $s \in \{60, 50, 45, 40, 35, 30, 25, 20\}$ dB relative to the clean utterance. Each (message, SNR) pair receives N=1000 i.i.d. draws; every draw's violation string and decoded transcript (unique strings) are stored. Each draw generates i.i.d. Gaussian samples, zeroes their FFT bins outside 300–2500 Hz (at the next power-of-two length) and crops them to the utterance length, giving d; the perturbed input is $x + d\sqrt{P_x/P_d}\,10^{-s/20}$, where P_x and P_d are the powers of the whole utterance and of d. It is neither clipped nor renormalized; each model's front end applies its own per-utterance normalization. Because P_x includes pauses and any noise already on the recording, s is an *added-noise* SNR: ATCO2 recordings already have a median active-speech SNR of about 18.8 dB (Appendix F), so ATCO2 certificates measure the stability

Table 2: Certified-message fractions: fraction of dataset messages with $\mathrm{CP}^{+}(k, N, \beta) \leq \tau = 0.05$, per specification and SNR (dB), with clean WER for context. These are empirical fractions of per-message certificates with each underlying certificate being marginal at confidence $1 - 10^{-6}$. Three representative SNRs are tabulated; the full eight-SNR sweep is in Figure 1.

		exact (*msg*)			call-sign (*cs*)			semantic ($sem_{0.95}$)			
dataset	model	60	40	20	60	40	20	60	40	20	WER
ATCOSIM	QuartzNet	0.5433	0.0567	0.0073	0.6913	0.1667	0.0273	0.7593	0.1313	0.0087	0.573
	DeepSpeech2	0.5413	0.0227	0.0	0.734	0.1153	0.0033	0.798	0.0913	0.0	0.867
	Whisper	0.846	0.3807	0.05	0.9153	0.5787	0.1407	0.892	0.4813	0.0773	0.452
	Whisper-ATC	**0.9687**	**0.9107**	**0.7333**	**0.9867**	**0.962**	**0.8493**	**0.986**	**0.962**	**0.8393**	**0.200**
ATCO2	QuartzNet	0.1645	0.021	0.007	0.3034	0.0525	0.0105	0.294	0.028	0.007	0.842
	DeepSpeech2	0.2112	0.0117	0.0	0.3512	0.0198	0.0	0.4679	0.0198	0.0	1.036
	Whisper	0.734	0.196	0.0047	0.8681	0.4376	0.0537	0.7725	0.2532	0.0082	0.989
	Whisper-ATC	**0.9253**	**0.7071**	**0.2404**	**0.9627**	**0.8215**	**0.4236**	**0.9545**	**0.79**	**0.3687**	**0.295**

of an already noisy recording under a further perturbation. The theory is not tied to this choice: Theorems 1 and 3 and Proposition 5 hold for any sampleable noise law.

Models. The models we consider are: QuartzNet15x5 (Kriman et al., 2020) (19M parameters; NVIDIA NeMo checkpoint; greedy CTC), DeepSpeech2 (Amodei et al., 2016) (87M; `deepspeech.pytorch` V3; greedy CTC), Whisper (Radford et al., 2023) (74M; fp16 greedy decode), and Whisper-ATC (van Doorn et al., 2024) (`whisper-large-v3`, 1.54B, fine-tuned on ATC speech). All four run the same per-message verification framework as presented in this paper. The CTC models run in fp32, with TF32 disabled where it affects the decode (Section 3), and the Whisper models in fp16. The clean decode uses batch size 1; draws are decoded in batches of up to 256 (CTC), 128 (Whisper) and 32 (Whisper-ATC).

Remark 2. Whisper-ATC's fine-tuning data includes the ATCOSIM and ATCO2 datasets we certify on, an unavoidable overlap given how few in-domain ATC STT models and public ATC datasets exist. The certificates remain valid — they are distribution-free statements about decode robustness — but because the model has seen this material its high certified fractions cannot be cleanly separated from memorization. We evaluate this further in the results.

For QuartzNet and Whisper, all specifications come from one decode pass that stores every transcript. For DeepSpeech2 and Whisper-ATC, the *msg* and *cs* counts come from one decode pass; the semantic counts ($\mathrm{sem}_{0.90}$, $\mathrm{sem}_{0.95}$) come from a second pass together with their own same-pass *msg* count. GPU nondeterminism in fp16 models minimally shifts borderline counts between passes, so every semantic-vs-exact comparison in this paper uses the same-pass and no claims mix passes.

Scale. 4 models × 2,357 messages × 8 SNRs × 6 stored specification columns (*msg*, *cs* and *id*, and $\mathrm{sem}_{0.90}$, $\mathrm{sem}_{0.95}$ and the other *msg*) = 452,544 certificates, each a $\mathrm{CP}^{+}(k, 1000, 10^{-6})$ value. WER against reference transcripts is computed once per model and dataset on clean audio, see Table 2.

7 Results

The certificates themselves are strongly bimodal over messages: a message either never violates in 1000 draws, so its certificate sits on the $\mathrm{CP}^{+}(0, N, \beta)$ floor, or it violates in most of them. A dataset-mean quantity such as WER integrates over that bimodality and hides it. Appendix C plots the variance in robustness scores at different perturbations, and Table 5 tabulates the medians for every model–dataset cell.

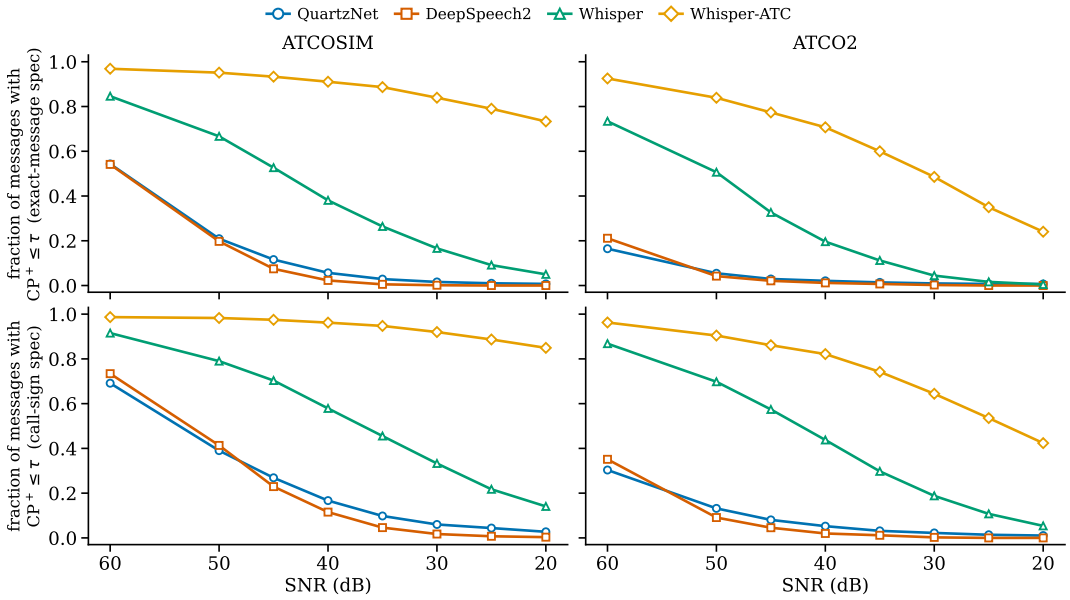


Figure 1: Fraction of messages certified ($CP^+ \leq \tau$) versus SNR, for the exact-message specification (top) and the call-sign specification (bottom), per dataset. Empirical fractions of per-message certificates, not population bounds.

7.1 Certified-message fractions

Table 2 and Figure 1 give the model comparison. At the cleanest channel, Whisper-ATC certifies a fraction 0.9687 of ATCOSIM messages under the strictest specification and still certifies 0.7333 at 20 dB; the call-sign and semantic fractions at 20 dB are 0.8493 and 0.8393. QuartzNet falls from 0.5433 to 0.0073; DeepSpeech2 starts at a similar level and ends at 0.0 at 20 dB. Whisper lies between the two regimes: its certified fraction declines progressively across the scan, without the single sharp drop both CTC models show in Figure 1.

Accuracy does not predict this ordering. QuartzNet beats Whisper on clean ATCO2 WER (0.842 vs. 0.989) yet certifies fewer messages than Whisper in every ATCO2 cell of Table 2, except the exact specification at 20 dB, where both fractions are near zero. DeepSpeech2's ATCO2 WER exceeds one (1.036) as it introduces more errors than tokens in the original string, a setting where WER stops providing any value at all, yet, its certificates remain well-defined because they measure decode robustness not accuracy. Whisper-ATC has both the best WER (0.200/0.295) and the largest certified robust fraction in every SNR. Because our certificates measure robustness of the model's own decode, this is consistent with in-domain fine-tuning improving noise-robustness as well as accuracy, though the fine-tuning/test set overlap means we cannot separate it from memorization of the certified utterances.

The CTC rows on ATCO2 need care. Their low certified fractions at 60 dB are not inference variability (Appendix D): the added noise has relative amplitude 10^{-3}, and decoders whose greedy choices sit near ties change under perturbations of that size. These decoders also mostly decode ATCO2 incorrectly (WER 0.842 and 1.036), so their certificates show sensitivity to small perturbations rather than operational usefulness.

Call-sign identity. Table 3 compares the two call-sign readouts of Section 4. *id* certifies more messages than *cs* for every model, because the span also changes when a greeting or station name does while the identity stays intact. The gap is largest for the CTC models on ATCO2 for a less useful reason: the parser finds a call-sign in few of their decodes (Appendix E), so their *id* certificates mostly certify that no call-sign appears.

Table 3: Certified-message fractions under the two call-sign readouts: *cs* (stability of the extracted span, as in Table 2) and *id* (stability of the parsed call-sign identity, Section 4). For DeepSpeech2 and Whisper-ATC the *id* counts come from the semantic pass.

		span (*cs*)			identity (*id*)		
dataset	model	60	40	20	60	40	20
ATCOSIM	QuartzNet	0.6913	0.1667	0.0273	0.912	0.626	0.2913
	DeepSpeech2	0.734	0.1153	0.0033	0.9747	0.8193	0.5213
	Whisper	0.9153	0.5787	0.1407	0.9453	0.7253	0.3167
	Whisper-ATC	0.9867	0.962	0.8493	0.9887	0.972	0.8867
ATCO2	QuartzNet	0.3034	0.0525	0.0105	0.9277	0.8215	0.6908
	DeepSpeech2	0.3512	0.0198	0.0	0.9988	0.9942	0.9918
	Whisper	0.8681	0.4376	0.0537	0.9475	0.7188	0.3769
	Whisper-ATC	0.9627	0.8215	0.4236	0.9802	0.881	0.5776

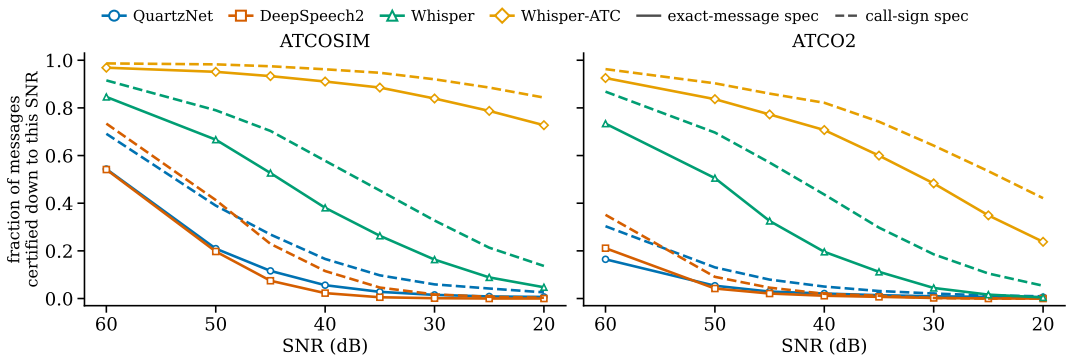


Figure 2: Per-message SNR budgets: fraction of messages certified at *every* scanned SNR down to s (solid: exact-message; dashed: call-sign). A message on the 20 dB point carries a certificate at all eight channel qualities.

7.2 SNR budgets

For deployment the most useful per-message summary is the *SNR budget*: the lowest scanned s such that the message is certified at every scanned SNR down to s, "this message is robust even for s dB." Figure 2 shows the survival curves. For Whisper-ATC on ATCOSIM the curve declines gradually from the fraction 0.9687 of messages certified at 60 dB (*msg* spec), and certification is sufficiently close to monotone in SNR that the certified-down-to-20 dB fraction almost coincides with the at-20 dB fraction 0.7333 (0.8493 for the call-sign). For QuartzNet the certified-down-to fraction falls below 0.5 by the first step of the scan, and only a small fraction of messages are robust to the full noise sweep. Conversely, the set of Whisper-ATC ATCOSIM messages that never certifies at any scanned SNR is small: 3.1% under the exact specification. An ATC operator can use our approach to determine at which channel quality does which fraction of traffic carry an individual certificate. The operator can use this knowledge to determine the risk that channel noise, as modeled, alters the decoded content of a message. Using a budget requires an estimate of the channel SNR, which is measurable from the received signal (the noise floor between transmissions). Precision is not required: certification is close to monotone in SNR, so a conservative estimate selects a stricter row. Additional results appear in Appendix C (including semantic oracle results we showed admit safety-relevant counterexamples) and Appendix G documents how to reproduce this paper's results.

					60 dB		40 dB		20 dB	
dataset	condition	site	region	n	msg	cs	msg	cs	msg	cs
ATCOSIM	clean	simulation	Europe	1500	0.97	0.99	0.91	0.96	0.73	0.85
ATCO2	noisy	multiple feeds	Europe	857	0.93	0.96	0.71	0.82	0.24	0.42
UWB-ATCC	noisy	Czech ANS	Europe	250	0.81	0.86	0.41	0.58	0.06	0.18
Poland	noisy	Kraków (EPKK)	Europe	250	0.85	0.89	0.35	0.48	0.01	0.05
Netherlands	noisy	Amsterdam (EHAM)	Europe	250	0.61	0.68	0.40	0.50	0.04	0.11
Norway	noisy	Trondheim (ENVA)	Europe	250	0.84	0.90	0.36	0.54	0.00	0.09
Sweden	noisy	Göteborg (ESGG)	Europe	250	0.65	0.76	0.34	0.50	0.02	0.09
Canada	noisy	Calgary (CYYC)	N. America	250	0.82	0.90	0.25	0.51	0.05	0.13
United States	noisy	Boston (KBOS)	N. America	250	0.52	0.67	0.28	0.44	0.03	0.12

Table 4: Certified fractions on the two fine-tuning datasets (top) and on seven datasets disjoint from Whisper-ATC's fine-tuning (bottom), under the identical threat model ($N = 1000$, $\beta = 10^{-6}$, certified $\iff k \leq 20$). Certification is possible on all of these datasets only because the certificate is *reference-free*: none of the held-out audio carries a ground-truth transcript.

7.3 Generalization beyond the fine-tuning data

Whisper-ATC is fine-tuned on ATCOSIM and ATCO2, so its certificates on those datasets cannot by themselves distinguish noise robustness from familiarity with the fine-tuning distribution. Because our certificate is reference-free, it can be computed on any audio, including audio for which no ground-truth transcript exists; we therefore reran the identical procedure on seven datasets that are disjoint from the fine-tuning sets. These are UWB-ATCC (Šmídl, 2011), a curated dataset of Czech controller–pilot radiotelephony, and six sets of 250 publicly streamed transmissions each, drawn from Kraków (EPKK), Amsterdam (EHAM), Trondheim (ENVA), Göteborg (ESGG), Calgary (CYYC) and Boston (KBOS), segmented by energy-based voice activity detection and filtered to clips whose clean decode is a plausible transmission. Table 4 reports the results.

The sharpest comparison is ATCO2 against UWB-ATCC, because the two are similar in most metrics but UWB-ATCC is not included in the fine-tuning set: both are curated datasets of real, over-the-air European controller–pilot radiotelephony, and both are spectrally band-limited in the same way (median energy above 4 kHz is 0.33% and 0.04% respectively, with over 91% of energy inside the 300–2500 Hz perturbation band), so the nominal difference in container sample rate is immaterial. Across that single change the exact-message certified fraction falls from 0.93 to 0.81 at 60 dB, from 0.71 to 0.41 at 40 dB and from 0.24 to 0.06 at 20 dB. The pattern holds across the whole held-out family: both fine-tuning datasets reach budgets of 0.73 and 0.24, whereas no disjoint dataset exceeds 0.06. We read this as evidence that a substantial part of Whisper-ATC's apparent robustness is specific to its fine-tuning distribution. The spread among the held-out sets is *not* explained by geography. Canada, which is outside the European region the fine-tuning datasets come from, certifies 0.82 at 60 dB, above the Netherlands and Sweden at 0.61 and 0.65, so these numbers do not obviously reflect an in-region versus out-of-region effect. What varies instead is per-feed audio quality and the quality of automatic segmentation, which is why we treat the curated UWB-ATCC comparison as primary. These results also highlight that there is not a current open-source ATC STT model that should be trusted in deployment at this stage.

More importantly, our contribution is the framework that enables this discussion: the per-message, matched-condition comparison in Table 4 is a comparison our certificates provide, and they can be rerun for any model, dataset and noise law without ground-truth transcripts. The specific conclusion about Whisper-ATC is simply a demonstration of that capability.

8 Limitations

Certificate sweeps. Every certificate is marginal at $\beta = 10^{-6}$; simultaneous use of K certificates costs $K\beta$. Therefore, using large sweeps has one of two implications; (i) lowered confidence, or (ii) lower budgets of k failures to still reach threshold τ.

Oracle assumptions. Semantic certificates are L1 statements about the cosine verdict; the L2 bound requires Assumption 1, which the counterexamples show cannot be discharged from cosine alone. The assumption-free call-sign audit covers the ATC safety-critical setting, but "meaning" beyond the call-sign (altitudes, headings, frequencies) is audited only insofar as the flip predicate captures it; extending the digit/NATO predicate to full clearance semantics, and a more detailed assessment of audio verification applications is interesting future work.

Noise realism. $\mathcal{P}_{x,s}$ is additive stationary Gaussian noise, band-limited to the VHF voice passband, at fixed SNR, a deliberately simple lower bound on real channel stress. Crucially, no theorem depends on this choice: Theorem 1, Theorem 3, and Proposition 5 require only i.i.d. draws from a sampleable noise law, so the framework applies unchanged to far more complex and realistic distributions — recorded VHF/AM channel noise, co-channel interference, fading, squelch artifacts, and speaker effects. $\mathcal{P}_{x,s}$ also has no *convolutional* component: a real channel filters the transmitted speech and AM demodulation adds nonlinear distortion, whereas our draws band-limit only the added noise. Each certificate is therefore a statement under the stated noise law, not an end-to-end channel guarantee, and its SNR is the SNR of the *added* noise relative to the recording, not the total channel SNR (Section 6). The ATCO2 and live-feed recordings did pass through a real VHF/AM channel before perturbation, so those certificates measure robustness to noise added on top of genuine channel distortion — though not to varying the distortion itself. Swapping in such a generator changes the reported numbers but none of the proofs; we also leave this extension for more thorough domain applications to future work.

Held-out evaluation data. Assessing whether a fine-tuned ATC model generalizes beyond its training data is limited by a gap in the publicly available datasets: we could not identify any obtainable dataset of *clean* ATC speech that is disjoint from the ATCOSIM and ATCO2 sets used to fine-tune Whisper-ATC. Every obtainable disjoint dataset of controller–pilot radiotelephony — UWB-ATCC (Šmídl, 2011; Šmídl et al., 2019), LDC-ATCC (Godfrey, 1994), and live VHF feeds — is recorded over the air, and therefore noisy by construction; and the simulation-recorded AcListant data (Ohneiser et al., 2014; Helmke et al., 2015), which would satisfy all three criteria, has not been released as a downloadable or licensable dataset. The one obtainable dataset that is both clean and disjoint, HIWIRE (Segura et al., 2007), is license-gated and covers cockpit command-and-control rather than controller–pilot radiotelephony. This scarcity is a property of the domain rather than an issue in our approach: Zuluaga Gomez et al. (2024) observe that "nearly all ongoing and former projects in the area of ATC prohibit the release of databases, code, and AI models due to privacy issues". Our held-out evaluations are therefore necessarily asymmetric (more noisy data than clean data), and we regard this as an evaluation-infrastructure gap for the field.

9 Conclusion

For the STT systems that actually deploy in safety-critical settings, sound verification is vacuous in practice. This paper derives 452,544 exact-binomial certificates, each a distribution-free bound on the probability that channel noise changes what one model decoded from one message, under a layered family of three specifications measured from sampled forward passes through each model. Our certificates separate deployment candidates across the channel-quality range in a way a single corpus-level accuracy number does not, and could be used by ATC operators to determine the risk that channel noise, as modeled, changes the decoded message; the certificates bound decode changes under the stated noise law (the theory admits any sampleable noise distribution), and are not statements of transcription correctness or end-to-end safety. Accuracy and stability are complementary: models are trained for accuracy, the certificates test stability, consistent decisions under noise. The certificates expose what accuracy alone cannot; call-sign changes that a semantic floor forgives and lower certified fractions outside a model's fine-tuning data. Our experiments considered the in-domain 1.54B parameter Whisper-ATC model, one of the largest models to be certified in any of the

verification literature. While it attains the largest certified fractions at every scanned SNR for messages in the training set, we note model performance degrades for messages captured outside of the training data, an analysis that is only possible with our framework and that can be used more broadly to assess future state-of-the-art audio verification models.

Broader impact statement

This work makes the failure probabilities of deployed speech recognition measurable per message in a safety-critical domain. The main risk is reading the certificates as end-to-end safety guarantees: a certificate limits how likely the specified channel noise is to change a model's decode, but a stably mis-decoded message still certifies, so certificates must be read alongside accuracy; the semantic specification can forgive call-sign changes; and the threat model does not span all operational hazards. No model studied here, including the fine-tuned one, is deployment-ready on this evidence; we read this as a call for more robust ATC-specific STT models and the public data their development requires, with per-message certification as a measurement instrument. We see no dual-use concern beyond what the public ASR robustness literature already enables.

Acknowledgments

The material presented in this paper is based upon work supported by the National Science Foundation (NSF) through grant number 2220401 and the Defense Advanced Research Projects Agency (DARPA) under contract number FA8750-23-C-0518. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of DARPA or NSF.

References

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, and Zhenyao Zhu. Deep speech 2: End-to-end speech recognition in English and Mandarin. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pp. 173–182. PMLR, 2016.

Brendon G. Anderson and Somayeh Sojoudi. Data-driven certification of neural networks with random input noise. *IEEE Transactions on Control of Network Systems*, 10(1):249–260, 2023. arXiv:2010.01171.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024.

Teodora Baluta, Shiqi Shen, Shweta Shinde, Kuldeep S. Meel, and Prateek Saxena. Quantitative verification of neural networks and its security applications. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 1249–1264, 2019.

Annelot W. Bosman, Aaron Berger, Holger H. Hoos, and Jan N. van Rijn. Robustness distributions in neural network verification. *Journal of Artificial Intelligence Research*, 83(20), July 2025. doi: 10.1613/jair.1.18403. Article 20, 41 pages.

Christopher Brix, Stanley Bak, Taylor T. Johnson, and Haoze Wu. The fifth international verification of neural networks competition (VNN-COMP 2024): Summary and results. *arXiv preprint arXiv:2412.19985*, 2024.

Giuseppe C. Calafiore and Marco C. Campi. The scenario approach to robust control design. *IEEE Transactions on Automatic Control*, 51(5):742–753, 2006.

Marco C. Campi and Simone Garatti. The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization*, 19(3):1211–1230, 2008.

Marco C. Campi and Simone Garatti. Wait-and-judge scenario optimization. *Mathematical Programming*, 167(1):155–189, 2018.

Luca Cardelli, Marta Kwiatkowska, Luca Laurenti, Nicola Paoletti, Andrea Patane, and Matthew Wicker. Statistical guarantees for the robustness of Bayesian neural networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5693–5700, 2019.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 1–7. IEEE, 2018. doi: 10.1109/SPW.2018.00009.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2019.

Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, volume 12699 of *LNCS*, pp. 625–635. Springer, 2021.

Olive Jean Dunn. Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56 (293):52–64, 1961. doi: 10.1080/01621459.1961.10482090.

EUROCONTROL. EUROCONTROL voluntary ATM incident reporting (EVAIR) safety bulletin no. 25 (years 2018–2022). Technical report, EUROCONTROL, May 2024. URL `https://www.eurocontrol.int/publication/eurocontrol-voluntary-atm-incident-reporting-evair-safety-bulletin-25`.

Federal Aviation Administration. 14 CFR 91.123: Compliance with ATC clearances and instructions. U.S. Code of Federal Regulations, Title 14, Part 91, Subpart B, 2024. URL `https://www.ecfr.gov/current/title-14/part-91/section-91.123`.

Federal Aviation Administration. Aeronautical information manual (AIM), paragraph 4-4-7: Pilot responsibility upon clearance issuance. Official Guide to Basic Flight Information and ATC Procedures; Basic dated 20 February 2025, incorporating Change 3, 2026. URL `https://www.faa.gov/air_traffic/publications/atpubs/aim_html/chap4_section_4.html`.

Simone Garatti and Marco C. Campi. Risk and complexity in scenario optimization. *Mathematical Programming*, 191(1):243–279, 2022.

Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. AI2: Safety and robustness certification of neural networks with abstract interpretation. In *IEEE Symposium on Security and Privacy (S&P)*, 2018.

Subhankar Ghosh, Yuanjie Shi, Taha Belkhouja, Yan Yan, Janardhan Rao Doppa, and Brian Jones. Probabilistically robust conformal prediction. In Robin J. Evans and Ilya Shpitser (eds.), *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 216 of *Proceedings of Machine Learning Research*, pp. 681–690. PMLR, 2023.

John J. Godfrey. Air traffic control complete. Web Download, 1994. URL `https://catalog.ldc.upenn.edu/LDC94S14A`. LDC Catalog No. LDC94S14A (ATC0).

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. Scalable verified training for provably robust image classification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4841–4850. IEEE, 2019. doi: 10.1109/ICCV.2019.00494. Preprint circulated as arXiv:1810.12715, "On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models".

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 369–376. ACM, 2006. doi: 10.1145/1143844.1143891.

Navid Hashemi, Samuel Sasaki, Ipek Oguz, Meiyi Ma, and Taylor T. Johnson. Scaling data-driven probabilistic robustness analysis for semantic segmentation neural networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1700.

Hartmut Helmke, Jürgen Rataj, Thorsten Mühlhausen, Oliver Ohneiser, Heiko Ehr, Matthias Kleinert, Youssef Oualil, and Marc Schulder. Assistant-based speech recognition for ATM applications. In *Proceedings of the Eleventh USA/Europe Air Traffic Management Research and Development Seminar (ATM2015)*, Lisbon, Portugal, June 2015. Paper 363.

Konrad Hofbauer, Stefan Petrik, and Horst Hering. The ATCOSIM corpus of non-prompted clean air traffic control speech. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). URL `https://aclanthology.org/L08-1507/`.

Zohar Jackson, César Souza, Jason Flaks, Yuxin Pan, Hereman Nicolas, and Adhish Thite. Free spoken digit dataset (FSDD). `https://github.com/Jakobovski/free-spoken-digit-dataset`, 2018. Version v1.0.8, Zenodo, `https://doi.org/10.5281/zenodo.1342401`.

Guy Katz, Clark W. Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In *Computer Aided Verification (CAV)*, volume 10426 of *Lecture Notes in Computer Science*, pp. 97–117. Springer, 2017. doi: 10.1007/978-3-319-63387-9_5.

Ching-Yun Ko, Zhaoyang Lyu, Tsui-Wei Weng, Luca Daniel, Ngai Wong, and Dahua Lin. POPQORN: Quantifying robustness of recurrent neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3468–3477. PMLR, 2019.

Martin Kocour, Karel Veselý, Alexander Blatt, Juan Zuluaga Gomez, Igor Szöke, Jan Černocký, Dietrich Klakow, and Petr Motlicek. Boosting of contextual information in ASR for air-traffic call-sign recognition. In *Interspeech*, pp. 3301–3305, 2021. doi: 10.21437/Interspeech.2021-1619.

Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. QuartzNet: Deep automatic speech recognition with 1D time-channel separable convolutions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6124–6128. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9053889.

Tobias Ladner, Konstantin Kaulen, Stanley Bak, Julien Girard-Satabin, Taylor T. Johnson, Edoardo Manino, ThanhVu H Nguyen, Ben Wooding, and Haoze Wu. The 7th International Verification of Neural Networks Competition (VNN-COMP 2026): Summary and Results, 2026. URL `https://arxiv.org/abs/2611.24114`.

Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, NY, 3 edition, 2005. ISBN 978-0-387-98864-1. doi: 10.1007/0-387-27605-X.

Linyi Li, Tao Xie, and Bo Li. SoK: Certified robustness for deep neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, 2023.

Albert Lin and Somil Bansal. Verification of neural reachable tubes via scenario optimization and conformal prediction. In *Proceedings of the 6th Annual Learning for Dynamics & Control Conference (L4DC)*, volume 242 of *Proceedings of Machine Learning Research*, pp. 719–731. PMLR, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Luca Marzari, Ferdinando Cicalese, and Alessandro Farinelli. Probabilistically tightened linear relaxation-based perturbation analysis for neural network verification. *Journal of Artificial Intelligence Research*, 84 (30), December 2025. doi: 10.1613/jair.1.20808. Article 30, 34 pages.

William P. Monan. Addressee errors in ATC communications: The call sign problem. Technical Report NASA-CR-166462, Battelle Columbus Laboratories, January 1983. URL `https://ntrs.nasa.gov/citations/19840012433`. Prepared for NASA Ames Research Center under Contract NAS2-10060.

Oliver Ohneiser, Hartmut Helmke, Heiko Ehr, Hejar Gürlük, Michael Hössl, Thorsten Mühlhausen, Youssef Oualil, Marc Schulder, Anna Schmidt, Arif Khan, and Dietrich Klakow. Air traffic controller support by speech recognition. In Neville Stanton, Steven Landry, Giuseppe Di Bucchianico, and Andrea Vallicelli (eds.), *Advances in Human Aspects of Transportation: Part II*, Advances in Human Factors and Ergonomics 2014, pp. 492–503, Kraków, Poland, July 2014. AHFE Conference. ISBN 978-1-4951-2098-5. doi: 10.54941/ahfe100712. Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics (AHFE 2014), 19–23 July 2014.

Raphael Olivier and Bhiksha Raj. Sequential randomized smoothing for adversarially robust speech recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6372–6386, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.514.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410.

Wonryong Ryou, Jiayu Chen, Mislav Balunović, Gagandeep Singh, Andrei Dan, and Martin Vechev. Scalable polyhedral verification of recurrent neural networks. In *Computer Aided Verification (CAV)*, volume 12759 of *Lecture Notes in Computer Science*, pp. 225–248. Springer, 2021. doi: 10.1007/978-3-030-81685-8_10.

Samuel Sasaki, Ben Wooding, and Taylor T. Johnson. Learning reach-set geometry for tighter probabilistic neural network verification. *Advances in Neural Information Processing Systems (NeurIPS)*, 39, 2026. (to appear).

José C. Segura, Thibaut Ehrette, Alexandros Potamianos, Dominique Fohr, Irina Illina, P.-A. Breton, V. Clot, Roberto Gemello, Marco Matassoni, and Petros Maragos. The HIWIRE database, a noisy and non-native English speech corpus for cockpit communication. ELRA Catalogue, ELRA-S0293, 2007. URL `https://catalogue.elra.info/en-us/repository/browse/ELRA-S0293/`. Distributed by ELRA/ELDA as ELRA-S0293; ISLRN 934-733-835-065-0. Collected under the IST-EU STREP project HIWIRE. Former project site `http://www.hiwire.org` no longer resolves.

Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. Robustness verification for transformers. In *International Conference on Learning Representations (ICLR)*, 2020.

Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin Vechev. Fast and effective robustness certification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

SKYbrary. Call sign confusion. SKYbrary Aviation Safety, EUROCONTROL / Flight Safety Foundation, 2024. URL `https://skybrary.aero/articles/call-sign-confusion`.

Luboš Šmídl. Air traffic control communication. LINDAT/CLARIAH-CZ digital library, 2011. URL `http://hdl.handle.net/11858/00-097C-0000-0001-CCA1-0`. Licensed CC BY-NC-SA 4.0; 8 kHz, 16-bit PCM, mono.

Luboš Šmídl, Jan Švec, Daniel Tihelka, Jindřich Matoušek, Jan Romportl, and Pavel Ircing. Air traffic control communication (ATCC) speech corpora and their use for ASR and TTS development. *Language Resources and Evaluation*, 53(3):449–464, 2019. doi: 10.1007/s10579-019-09449-5.

Hoang-Dung Tran, Diego Manzanas Lopez, Patrick Musau, Xiaodong Yang, Luan Viet Nguyen, Weiming Xiang, and Taylor T. Johnson. Star-based reachability analysis of deep neural networks. In *Formal Methods – The Next 30 Years: Third World Congress, FM 2019*, volume 11800 of *Lecture Notes in Computer Science*, pp. 670–686. Springer, 2019. doi: 10.1007/978-3-030-30942-8_39.

UK Civil Aviation Authority. ACCESS — aircraft call sign confusion evaluation safety study. CAP 704, UK Civil Aviation Authority, April 2000. URL `https://www.caa.co.uk/our-work/publications/documents/content/cap-704/`.

Jan van Doorn, Junzi Sun, Jacco M. Hoekstra, Patrick Jonk, and Vincent de Vries. Whisper-ATC: Open models for air traffic control automatic speech recognition with accuracy. In *Proceedings of the International Conference on Research in Air Transportation (ICRAT)*, 2024. Checkpoint: HuggingFace `jlvdoorn/whisper-large-v3-atco2-asr-atcosim` (Whisper large-v3 fine-tuned on ATCO2-ASR and ATCOSIM).

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, NY, 2005. ISBN 978-0-387-00152-4. doi: 10.1007/b106715.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, volume 33, 2020.

Stefan Webb, Tom Rainforth, Yee Whye Teh, and M. Pawan Kumar. A statistical approach to assessing neural network robustness. In *International Conference on Learning Representations (ICLR)*, 2019.

Tsui-Wei Weng, Pin-Yu Chen, Lam M. Nguyen, Mark S. Squillante, Akhilan Boopathy, Ivan Oseledets, and Luca Daniel. PROVEN: Verifying robustness of neural networks with a probabilistic approach. In *International Conference on Machine Learning (ICML)*, pp. 6727–6736, 2019.

Ben Wooding, Simone Garatti, Marco C Campi, and Abolfazl Lavaei. Scen-opt: A scenario optimization toolbox for data-driven convex programming. *arXiv preprint arXiv:2610.07846*, 2026.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649, 2024. doi: 10.1145/3626772.3657878.

Kaidi Xu, Huan Zhang, Shiqi Wang, Yihan Wang, Suman Jana, Xue Lin, and Cho-Jui Hsieh. Fast and complete: Enabling complete neural network verification with rapid and massively parallel incomplete verifiers. In *International Conference on Learning Representations (ICLR)*, 2021.

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Muhammad Usama Zubair, Taylor T. Johnson, Kanad Basu, and Waseem Abbas. ModelStar: Reachability analysis-based safety verification of neural networks against model perturbations. *Journal of Artificial Intelligence Research*, 85(37):37:1–37:29, 2026. doi: 10.1613/jair.1.18922.

Juan Pablo Zuluaga Gomez, Karel Veselý, Igor Szöke, Alexander Blatt, Petr Motlicek, Martin Kocour, Khalid Choukri, Iuliia Nigmatulina, Claudia Cevenini, Allan Tart, Jan Cernocký, and Dietrich Klakow. ATCO2 corpus: A large-scale dataset for research on automatic speech recognition and natural language understanding of air traffic control communications. *Journal of Data-centric Machine Learning Research*, 2(5):1–45, 2024. URL `https://data.mlr.press/assets/pdf/v02-5.pdf`.

A Proofs

A.1 Proof of Theorem 1

Write $F_{N,p}(k) = \mathbb{P}_{K\sim\mathrm{Bin}(N,p)}[K \le k]$ and $u(k) = \mathrm{CP}^+(k, N, \beta)$ as in Definition 1. For fixed $k < N$ the map $p \mapsto F_{N,p}(k)$ is continuous and strictly decreasing on $(0, 1)$, with $F_{N,0}(k) = 1$ and $F_{N,1}(k) = 0$; hence $u(k)$ is the unique p with $F_{N,p}(k) = \beta$ when $k < N$, and $F_{N,q}(k) \le \beta$ for every $q \ge u(k)$. Also u is nondecreasing in k because $F_{N,p}(k)$ is nondecreasing in k.

Fix the true violation probability $q = V_x$ and let $S = \{k \in \{0, \dots, N\} : u(k) < q\}$. If $S = \emptyset$ the event $\{V_x > u(k)\}$ is impossible and the claim is trivial. Otherwise, by monotonicity of u, $S = \{0, \dots, k_0\}$ for some $k_0 < N$ (note $u(N) = 1 \ge q$ excludes $k = N$). Since $u(k_0) < q$, there exists p with $u(k_0) \le p < q$ and $F_{N,p}(k_0) \le \beta$; monotonicity of F in p gives $F_{N,q}(k_0) \le F_{N,p}(k_0) \le \beta$. Because $k \sim \mathrm{Bin}(N, q)$ exactly (the draws are i.i.d., and g_σ is treated as a fixed measurable function of the draw),

$$\mathbb{P}^N\big[V_x > u(k)\big] = \mathbb{P}^N\big[k \in S\big] = F_{N,q}(k_0) \le \beta.$$

For the Beta form: the binomial–Beta tail identity $F_{N,p}(k) = \mathbb{P}[B > p]$ with $B \sim \mathrm{Beta}(k + 1, N - k)$ shows $F_{N,p}(k) = \beta$ iff $\mathbb{P}[B \le p] = 1 - \beta$, i.e. $u(k)$ is the $1 - \beta$ quantile of $\mathrm{Beta}(k + 1, N - k)$. □

A.2 Proof of Lemma 2

If $g_{\mathrm{msg}}(x, \delta) = 0$ then $T_\delta = T_0$ as normalized strings. The call-sign extractor is a deterministic function of the transcript, so $\Phi_{\mathrm{cs}}(T_\delta) = \Phi_{\mathrm{cs}}(T_0)$ and $g_{\mathrm{cs}} = 0$. The embedding map is deterministic, so $e(T_\delta) = e(T_0)$ and $\cos(e(T_\delta), e(T_0)) = 1 \ge \chi$, giving $g_{\mathrm{sem}_\chi} = 0$. Contrapositives give the pointwise inequalities; taking expectations gives the V inequalities, and summing over draws gives the count inequalities. □

A.3 Proof of Theorem 3

Define per draw i the indicator $h_i = \mathbf{1}[g_{\mathrm{msg}}(x, \delta_i) = 1 \wedge g_{\mathrm{sem}_\chi}(x, \delta_i) = 0]$. By Lemma 2, $g_{\mathrm{sem}_\chi} \le g_{\mathrm{msg}}$ pointwise, hence $h_i = g_{\mathrm{msg}}(x, \delta_i) - g_{\mathrm{sem}_\chi}(x, \delta_i)$ and $\sum_i h_i = k_{\mathrm{msg}} - k_{\mathrm{sem}} = d$. The h_i are i.i.d. Bernoulli with parameter

$$\mathbb{P}[g_{\mathrm{msg}} = 1,\, g_{\mathrm{sem}_\chi} = 0] = \mathbb{P}[g_{\mathrm{msg}} = 1] - \mathbb{P}[g_{\mathrm{sem}_\chi} = 1] = V_x^{\mathrm{msg}} - V_x^{\mathrm{sem}_\chi} = D_x,$$

using the event inclusion again. Thus $d \sim \mathrm{Bin}(N, D_x)$ exactly, and Theorem 1 applied to the sequence (h_i) gives $\mathbb{P}^N[D_x > \mathrm{CP}^+(d, N, \beta)] \le \beta$. The event inclusion also identifies D_x with the mass of $\{T_\delta \ne T_0 \wedge \cos \ge \chi\}$. The joint statement follows from a union bound over the two β-events (which are functions of the same N draws; independence is not needed). □

A.4 Proof of Theorem 4

Decompose the event $\{g^* = 1\}$ by the oracle verdict and the exact flag:

$$\mathbb{P}[g^*{=}1] = \mathbb{P}[g^*{=}1,\, g_{\mathrm{sem}}{=}1] + \mathbb{P}[g^*{=}1,\, g_{\mathrm{sem}}{=}0,\, g_{\mathrm{msg}}{=}1] + \mathbb{P}[g^*{=}1,\, g_{\mathrm{sem}}{=}0,\, g_{\mathrm{msg}}{=}0].$$

The third term vanishes because $g^* \le g_{\mathrm{msg}}$ (identical strings have identical meaning). The first term is at most $V_x^{\mathrm{sem}_\chi}$. If $D_x > 0$, the second term equals $\mathbb{P}[g^*{=}1 \mid g_{\mathrm{msg}}{=}1,\, g_{\mathrm{sem}}{=}0] \cdot D_x \le \rho_{\mathrm{FP}} D_x$ by Assumption 1; if $D_x = 0$ it vanishes. Hence $V_x^* \le V_x^{\mathrm{sem}_\chi} + \rho_{\mathrm{FP}} D_x$ deterministically. On the intersection of the two CP^+ events of Theorem 3 (which has probability at least $1 - 2\beta$), monotone substitution gives $V_x^* \le \mathrm{CP}^+(k_{\mathrm{sem}}, N, \beta) + \rho_{\mathrm{FP}} \mathrm{CP}^+(d, N, \beta)$. □

A.5 Proof of Proposition 5

The predicate F and the cosine are deterministic measurable functions of the stored transcript pair (T_0, T_{δ_i}), which is itself a deterministic function of the draw δ_i within the pass whose transcripts are stored. Hence the per-draw indicators $\mathbf{1}[F{=}1 \wedge \cos \ge \chi]$ are i.i.d. Bernoulli with parameter $\eta_F(x; \chi, s)$, their sum m is $\mathrm{Bin}(N, \eta_F)$, and Theorem 1 applies verbatim with g_σ replaced by the audited indicator. No labeling beyond the stored transcripts enters, so no representativeness assumption is needed. □

B Why we cannot use sound verification

These ATC speech recognition certificates are per-input, aligning with current sound neural-network verification methods. However, at the noise magnitudes we consider, the sound approaches are vacuous in practice. The QuartzNet and DeepSpeech2 bottlenecks can be *measured*, while the Whisper challenges are *architectural*.

QuartzNet15x5 (19M parameters; convolutional CTC) QuartzNet is the simplest of the four models: its encoder is 15 blocks of time-channel-separable 1-D convolutions with BatchNorm and ReLU (Kriman et al., 2020), squarely inside the affine+ReLU fragment that sound tools support. There are two failures. First, the DSP front end shared by all our connectionist-temporal-classification (CTC) models (power spectrum (x^2), log-mel ($\log x$), and per-utterance instance normalization); as the noise level shrinks the certified variance lower bound, the measured front-end enclosure degrades from a usable 4.6× over-approximation at 50 dB SNR to a vacuous 1000000× at 40 dB and below. Second, ignoring the front end issues, forward set propagation diverges with depth: the measured interval radii grow from 1.9×10^{-3} after the first block to 1.5×10^{96} at the logits, and a rank-r affine-set (zonotope-style) propagation (Singh et al., 2018) still reaches a final logit radius of 5×10^{110}; well beyond the logit margins that decide the greedy CTC decode (Graves et al., 2006), so no decode invariant is decidable. Backward CROWN (Zhang et al., 2018) with sound forward intermediate bounds hits the same 10^{110} threshold, demonstrating an open problem of finding sound but *tight* intermediate bounds at depth.

DeepSpeech2 (87M parameters; bidirectional-LSTM CTC). DeepSpeech2 (Amodei et al., 2016) fails by a different mechanism. Each of its five bidirectional LSTM(1024) layers applies three Hadamard gate products per step, these are bilinear, non-convex, with no useful global Lipschitz constant, and the bidirectionality is the blocking point. Within the standard back-substituted relational relaxation class (Zhang et al., 2018; Ko et al., 2019; Ryou et al., 2021), a bidirectional stack forces enclosure widths $w_L \geq \gamma^L w_0$ with $\gamma > 1$, whereas an equal-width unidirectional control layer would propagate in a stable way. The enclosure's signed correlation structure collapses by fifteen orders of magnitude across the five layers, so the bound is equally useless against the logit margins, and, notably, it is independent of the perturbation size: considering smaller perturbations provides no benefit. A zoo of small CTC models on spoken digits (Jackson et al., 2018) highlights this clearly: star-set reachability (Tran et al., 2019) certifies feed-forward and unidirectional members at high SNR but is vacuous on the bidirectional layer at the same SNR.

Whisper and Whisper-ATC (74M / 1.54B parameters; encoder–decoder transformers). For the Whisper family (Radford et al., 2023; van Doorn et al., 2024) the obstruction is architectural rather than measured, because no sound tool produces any bound to measure: softmax self-attention in the encoder, cross-attention inside an *autoregressive* decoder, LayerNorm's input-dependent reciprocal, and GELU all lie outside scalable sound relaxations. Transformer-specific certification (Shi et al., 2020) reaches neither this depth nor the autoregressive decode loop, and no VNN-COMP entrant handles the class at this scale (Brix et al., 2024). These issues are different to the measured QuartzNet and DeepSpeech2 failures above. There is no sound baseline to compare against, and at 1.54B parameters, Whisper-ATC may be one of the largest models to be considered for verification in the literature.

The decode gap. Even a miraculously tight logit enclosure would not finish the job: CTC collapse and Whisper's sampling loop map logit sequences to strings through non-monotone argmax structures for which no formal verification exists, and the call-sign readout adds a further string-level extractor on top. Sound verification of these systems is not merely expensive, but is vacuous in practice. These obstructions can be ignored by probabilistic certificates, that provide guarantees using only sampled forward passes through the model: the type of certificate we leverage through the remainder of the paper.

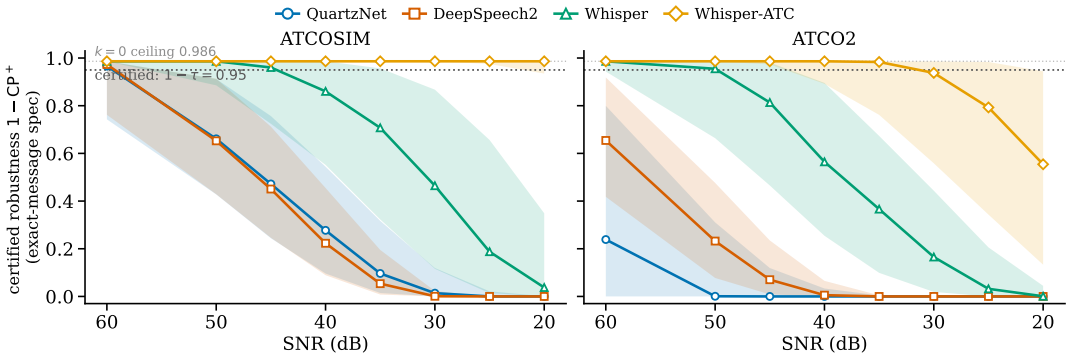


Figure 3: Per-message robustness across the SNR scan, plotted 60 dB to 20 dB: median and interquartile range of the certified robustness $1 - \mathrm{CP}^{+}(k, N, \beta)$ over messages, for the exact-message specification, per dataset. Robustness increases upward; the dotted line marks the certified threshold $1 - \tau = 0.95$ and the faint upper line the k=0 ceiling $1 - \mathrm{CP}^{+}(0, N, \beta)$, the best any N-draw certificate can state. Whisper-ATC's median ATCOSIM message sits on that ceiling at every scanned SNR.

C Additional results

C.1 Per-message bounds across the scan

Figure 3 shows the distribution of the certificates themselves; two structural features dominate. First, the per-message counts are strongly bimodal: a message either never violates in 1000 draws (its certificate sits on the $\mathrm{CP}^{+}(0, N, \beta)$ floor) or violates in most of them (its certificate is near 1), which is why the median robustness falls from the ceiling to the floor of Figure 3 as SNR degrades and why the interquartile bands are wide in the transition region. A dataset-mean quantity such as WER integrates over this bimodality, hiding the wide variance between messages for the same SNR. Second, the models occupy different regimes. Whisper-ATC's median ATCOSIM message keeps its risk value at the floor at every scanned SNR (0.013721 both at 60 dB and 20 dB for *msg*), while QuartzNet's median message is already at 0.03401 at 60 dB and saturates at 1.0 by 20 dB. On the noisy-channel ATCO2, Whisper-ATC's median keeps its certificate at the floor until 40 dB and reaches 0.445685 at 20 dB. Table 5 in Appendix C tabulates the medians for all model–dataset cells.

Table 5: Median per-message certificate $\mathrm{CP}^{+}(k, N, \beta)$ over dataset messages, exact-message and call-sign specifications, at three SNRs. The value 0.013721 is the k=0 floor $1 - \beta^{1/N}$: a median at the floor means the majority of messages never violated in 1000 draws.

		exact (msg)			call-sign		
dataset	model	60	40	20	60	40	20
ATCOSIM	QuartzNet	0.03401	0.722751	1.0	**0.013721**	0.487408	0.998051
	DeepSpeech2	0.027028	0.776927	1.0	**0.013721**	0.493469	1.0
	Whisper	**0.013721**	0.139857	0.962608	**0.013721**	0.021156	0.707722
	Whisper-ATC	**0.013721**	**0.013721**	**0.013721**	**0.013721**	**0.013721**	**0.013721**
ATCO2	QuartzNet	0.761315	1.0	1.0	0.408595	0.998718	1.0
	DeepSpeech2	0.345739	0.994599	1.0	0.154102	0.943258	1.0
	Whisper	**0.013721**	0.435427	0.999272	**0.013721**	0.109403	0.92091
	Whisper-ATC	**0.013721**	**0.013721**	0.445685	**0.013721**	**0.013721**	0.142246

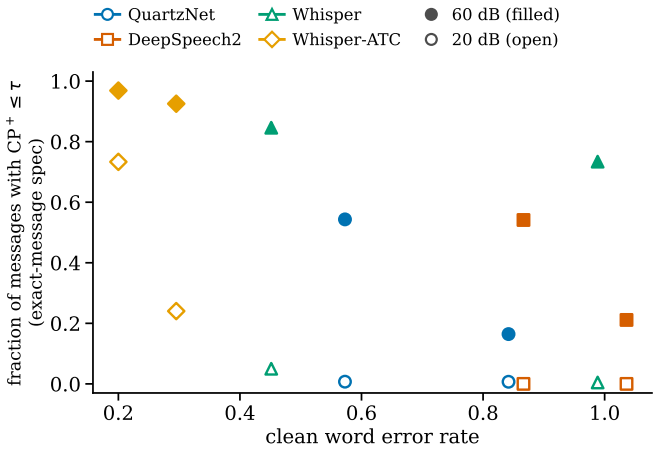


Figure 4: Clean word error rate vs. the fraction of messages whose exact-message certificate clears $\tau = 0.05$, at 60 dB (filled) and 20 dB (open); one point for each of ATCOSIM and ATCO2. Fractions are empirical counts of per-message certificates. On ATCO2, WER ranks Whisper below QuartzNet, yet Whisper certifies more messages at 60 dB.

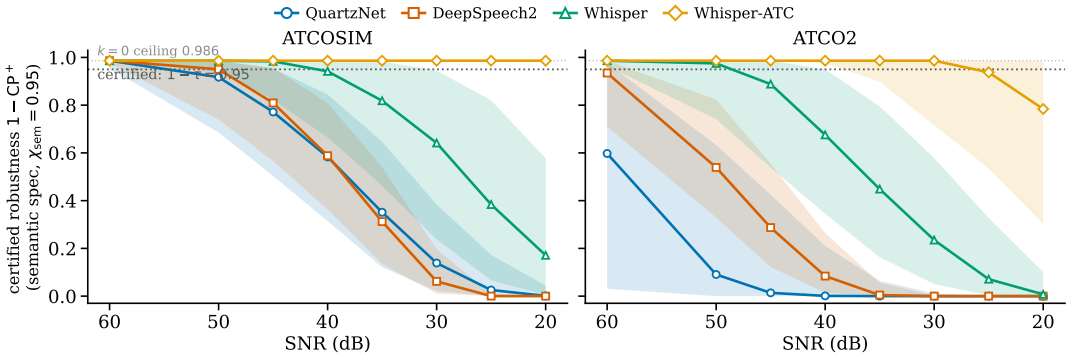


Figure 5: Per-message robustness under the semantic specification at floor $\chi_{\text{sem}} = 0.95$, the counterpart to Figure 3: median and interquartile range of $1 - \text{CP}^+(k, N, \beta)$ over messages, per dataset, 60 dB (left) to 20 dB (right).

Semantic floors at 0.90. For completeness alongside Table 2: at 20 dB on ATCOSIM the $\text{sem}_{0.90}$ certified fractions are 0.0133 (QuartzNet), 0.0 (DeepSpeech2), 0.1207 (Whisper), and 0.896 (Whisper-ATC); the same-pass exact fractions are 0.0073, 0.0, 0.05, and 0.7393.

Accuracy does not predict certifiability. Figure 4 plots clean WER against the empirical certified-message fraction at 60 and 20 dB (exact-message specification; the y-axis is the per-message fraction of Table 2). On ATCO2, Whisper certifies more messages at a *worse* clean WER than QuartzNet at 60 dB; at 20 dB both certify almost none.

C.2 Semantic floors and the forgiven mass

Section 5 explains why the semantic oracle is not trusted. Figure 6 overlays each semantic certified fraction with its same-pass exact companion. The gap is largest for the CTC models in the mid-SNR transition region, and it does not vanish for Whisper-ATC at low SNR: at 20 dB on ATCOSIM the $\text{sem}_{0.90}$ fraction is 0.896 against a same-pass exact fraction of 0.7393. Per message, Theorem 3 certifies this gap: d forgiven draws yield $\text{CP}^+(d, N, \beta)$ on the changed-but-forgiven probability, so an operator reading a semantic certificate can always see how much transcript-changing mass it forgave. Whether that forgiven mass is harmless is *not*

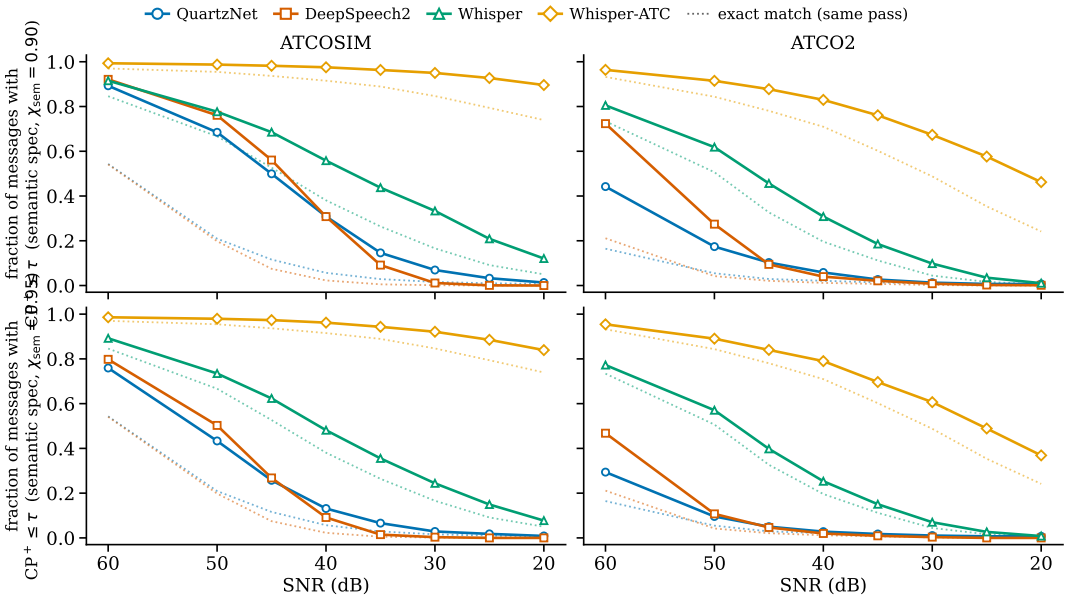


Figure 6: Certified-message fractions under the semantic specification at floors 0.90 (top) and 0.95 (bottom), with the *same-pass* exact-match fraction underneath (dotted). The vertical gap between each solid curve and its dotted companion is the mass of messages certified only because the floor forgives their transcript changes.

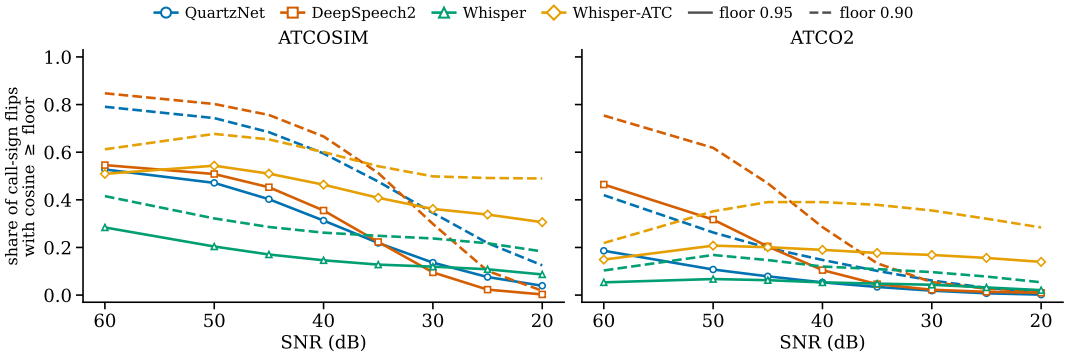


Figure 7: The call-sign audit by SNR: share of call-sign-flipping draws with cosine ≥ floor (solid: 0.95; dashed: 0.90), per dataset. Every point is an exact function of stored transcripts (Proposition 5). Whisper-ATC's forgiven share stays high across the scan; the CTC models' flips become cosine-visible as noise grows.

decidable by the embedding (as discussed in Section 5). Figure 7 shows the call-sign audit resolved by SNR, for the CTC models the forgiven share of call-sign flips falls quickly as noise grows, while for Whisper-ATC it stays high across the scan, i.e. its call-sign errors remain undetectable even in heavy noise.

Other sentence encoders. Cosine scales differ across encoders: bge-base-en-v1.5, for instance, places unrelated sentences far higher than MiniLM does. Fixed floors are therefore not comparable, and for each encoder we use the floor at which it forgives the same share of all changed transcripts as MiniLM at 0.95. The matched floors are 0.949 for all-mpnet-base-v2 and 0.969 for bge-base-en-v1.5. At a fixed 0.95 floor, bge-base-en-v1.5 forgives 13.0% of signature-changing call-sign flips against 7.5% for MiniLM, which is why fixed floors are not comparable.

D Zero-noise control

The control runs the certification pipeline with the noise generator switched off, so any change in the decode is inference variability. For every message it (i) repeats the batch-1 clean decode twice, (ii) decodes N_0 copies of the unchanged audio in the certification batch schedule and precision, final partial batch included, and (iii) decodes one batch whose first and last rows are the clean audio and whose other rows are 60 dB draws, to test batch composition. $N_0 = 1000$, except for Whisper-ATC, where $N_0 = 264$ (eight batches of 32 plus the remainder batch of 8) keeps its batch sizes at a quarter of the cost.

Table 6: Zero-noise control. *Changed*: messages with a zero-noise decode that differs from the batch-1 clean decode. $>k^*$: messages whose zero-noise count alone exceeds the certification threshold (scaled to $N = 1000$), so they could not certify at any SNR. fp32†: PyTorch's default TF32 convolutions.

dataset	model	precision	N_0	changed	$>k^*$
ATCOSIM	QuartzNet	fp32†	1,000	76	76
		fp32	1,000	1	1
	DeepSpeech2	fp32†	1,000	1	1
		fp32	1,000	0	0
	Whisper	fp16	1,000	37	31
	Whisper-ATC	fp16	264	16	16
ATCO2	QuartzNet	fp32†	1,000	77	76
		fp32	1,000	0	0
	DeepSpeech2	fp32†	1,000	0	0
		fp32	1,000	0	0
	Whisper	fp16	1,000	35	31
	Whisper-ATC	fp16	264	17	16

Repeated batch-1 decodes never differ, and a row's decode depends on the batch size, never on the other rows' content. Under TF32, the CTC batched decodes differ from the batch-1 decode for whole batches, so an affected message exceeds k^* at every SNR. The affected messages are near-ties that the added noise also flips: disabling TF32 moves QuartzNet's certified fractions by at most 0.5 percentage points and leaves 1 near-tie message affected, which we flag. Whisper and Whisper-ATC show the same pattern for 72 and 33 of 2,357 messages; certifying them against the clean decode taken in the draws' own batch configuration changes no certification decision at any SNR, so we keep the batch-1 reference throughout.

E Call-sign readouts: definition and validation

Annotation. The ATCO2 release tags call-signs in 279 of its 877 segments. We match our ATCO2 messages to these segments by transcript similarity (all but 4 match) and keep the 273 tagged ones, from 173 recordings at 5 airports; 249 contain a call-sign, taken as the first run of tagged tokens. Both readouts read the same reference text, which isolates the extractor from recognition errors.

Grammar readout. Φ_{id} scans the canonicalized tokens left to right ("alpha"→"alfa", "juliet"→"juliett", "niner"→"nine", "x ray"→"xray") and returns the first match of: (a) a telephony designator of up to three tokens, compared with spaces removed so that "speed bird" matches SPEEDBIRD, followed by an identifier run of digit words, phonetic letters and "double"/"triple", optionally followed by "heavy"; (b) a single unlisted word, not preceded by a value keyword, followed by an identifier run of at least two tokens that contains a digit ("csa one delta zulu"); (c) a registration: an identifier run of at least two tokens that opens with a phonetic letter and is not preceded by a value keyword, where a run opening with three or more letters stops at the first digit, which begins a value; (d) an utterance-initial run of at least three digits (an abbreviated call-sign). Value keywords are words that introduce numbers or letters that are not call-signs (`level`, `heading`, `runway`, `squawk`, `qnh`, `information`, `via`, `final`, ...). The designators are the telephony

column of the OpenFlights airline table together with the spoken designators of the ATCO2 traffic lists. Its value is the pair (designator with spaces removed, ordered signature); word lists ship with the code.

Table 7: Call-sign readouts against the ATCO2 gold annotation (249 reference transcripts with a call-sign, 145 utterance-initial and 104 not; 24 without). *Signature*: the readout's ordered digit and phonetic-letter content equals the gold call-sign's. *Exact*: the readout equals the gold span. *Empty*: share of call-sign-free transcripts on which the readout is empty.

readout	signature	initial	non-initial	exact	empty
Φ_{cs} (leading span)	62.6%	82.1%	35.6%	28.5%	0.0%
Φ_{id} (grammar)	85.5%	93.1%	75.0%	76.3%	79.2%

The leading span is exact on only 49.0% of utterance-initial call-signs although it reproduces the signature on 82.1%: it usually contains the call-sign together with a greeting or station name. Its failures are concentrated on non-initial call-signs, which the positional rule cannot reach.

Traffic context. Each ATCO2 recording lists the aircraft near the airport at recording time, with ICAO code and spoken call-sign (median 29 per recording; 26 of 560 lists are empty). A decoded identity is *consistent* with a listed aircraft when designator and signature agree or, for registrations, when it equals the listed signature, a suffix of it of length at least two, or its first letter followed by such a suffix (the ICAO abbreviation rule). Of Whisper-ATC's matched messages, 624 have a parsed clean call-sign and 134 of these are on the list. Over all messages with a parsed call-sign, 5.6% of identity changes are consistent with some other listed aircraft, an overestimate since short abbreviations match many registrations. The CTC decoders rarely produce a parseable call-sign on ATCO2 (QuartzNet 163, DeepSpeech2 0). Whisper's decodes contain a parseable call-sign more often (296 messages, 18 of them listed), and on those listed messages 2.0% of call-sign flips name another listed aircraft. Half of Whisper-ATC's *cs* flips on ATCO2 (49.6%) leave the parsed identity unchanged, changing only a greeting or station name inside the leading span.

F Recording SNR

Added-noise SNR and total SNR differ because the recordings carry noise of their own. We estimate each recording's own SNR blindly: 20 ms frames with a 10 ms hop; noise power P_n is the mean power of the quietest 10% of frames; frames more than 6 dB above P_n are active; speech power P_s is the mean power of active frames minus P_n; the inherent SNR is $10\log_{10}(P_s/P_n)$, and the total SNR at added-noise level s is $10\log_{10}\left(P_s/(P_n + 10^{-s/10}P_x)\right)$. Short clips with few pauses bias P_n upward, so these are rough lower estimates of the recordings' SNR. Table 8 reports the medians.

Table 8: Estimated inherent active-speech SNR of the recordings (median, with 10th–90th percentiles) and median total SNR at added-noise SNR s. *Clipped*: recordings with peak magnitude at least 0.99.

dataset	inherent (dB)	$s = 60$	$s = 40$	$s = 30$	$s = 20$	clipped
ATCOSIM	27.0 (22.7–30.9)	27.0	26.8	25.4	19.8	35
ATCO2	18.8 (11.7–28.5)	18.8	18.8	18.6	16.8	169

G Reproducibility

Artifacts. For every (model, dataset, message, SNR) cell the experiments store the per-draw violation bits for all five specifications, the set of unique decoded transcripts with occurrence counts, and the CRC32-derived seeds; every certificate in this paper is recomputable from those artifacts with one Beta-quantile evaluation. For QuartzNet and Whisper the decode pass stores ev-

ery draw's raw transcript, so any specification or normalization can be recomputed without inference. The analyses of Appendices D to F are produced by `run_zeronoise.py`, `validate_callsign.py`, `callsign_grammar.py`, `analyze_traffic_context.py` and `estimate_inherent_snr.py`. The number macros in this document (`numbers_pi.tex`, `numbers_semaudit.tex`, `numbers_stt.tex`, `numbers_atc.tex`) and all figures are emitted by the analysis scripts (`analyze_atc_perinput.py`, `emit_semaudit_macros.py`, `make_stt_paper_figs.py`); no certificate-campaign number in this document is typed by hand (the sound-verification measurements of Appendix B predate the campaign and are quoted from that study's logs). The counterexample transcripts of Section 5 are quoted verbatim from the stored counterexample file (`witnesses_semaudit.json`); the audit collects 264,606 floor-0.95-passing call-sign-flip edit records, and the 200 highest-cosine unique edits are released as `semaudit_examples.json`.

Call-sign extraction and audit procedure. For independent reimplementation: a transcript is normalized by lower-casing, spelling numerals digit by digit ("1,000" → "one zero zero zero", a decimal point between digits → "decimal"), deleting characters outside `[a-z' ]`, and collapsing whitespace. Φ_{cs} returns the leading token run, stopping before the first of the 73 released ATC command words, always retaining the first token, capped at 8 tokens (Section 4); Φ_{id} is specified in Appendix E. The three flip predicates compare extractor outputs on the stored pair (T_0, T_δ): *raw* — span inequality; *normalized* — inequality after dropping non-alphanumerics; *digit/NATO* — the ordered sequence of digit words and phonetic letters in the span changed (canonical spellings identified; token lists released with the code). Cosines use all-MiniLM-L6-v2 on the full transcripts of the same pass. All audit quantities are recomputed from the stored per-draw transcripts by `analyze_atc_semaudit.py` and `scan_semaudit_witnesses.py` and formatted by `emit_semaudit_macros.py`; no additional inference is involved.

Determinism. Noise draws are seeded per (model, dataset, message, SNR) via CRC32 of the identifying tuple, so the noise realizations behind any single certificate can be regenerated in isolation (given the same batch schedule, since the GPU random stream is drawn batch by batch). Decoding is greedy throughout; fp16 GPU inference is bitwise nondeterministic across passes, which is why counts are paired within passes. The noise is reproducible but the decode is not, and we quantify the residual rather than assert that it is negligible. The zero-noise control (Appendix D) measures it within a pass. For DeepSpeech2 and Whisper-ATC, whose semantic counts come from a second pass, that pass re-decodes only first-pass violations (an unchanged draw cannot violate a floor), so the comparison is one-sided: it detects violations that disappear, not new ones. Re-decoding those draws changes k in 9.9% of the 37,712 (message, SNR) cells, with a 99th percentile of 11 draws and a maximum of 81 out of 1000; the certification decision itself flips in 76 cells, or 0.20%. At the aggregate level the largest shift in any reported certified fraction is 0.0074. This is small relative to the differences between models that we interpret, but it is not zero, and it exceeds some of the differences *within* a model across adjacent SNRs, so single-cell counts should be read with that tolerance in mind.

Execution environment. All certification runs were executed on a single workstation: one NVIDIA GeForce RTX 5090 (34 GB VRAM, compute capability 12.0), 63 GB system RAM, Windows 11 (build 10.0.26200), Python 3.10.11, PyTorch 2.11.0 with CUDA 12.8, `transformers` 5.12.1, `sentence-transformers` 5.6.0, NumPy 1.26.4, SciPy 1.15.3 and `soundfile` 0.14.0 (libsndfile 1.2.2). Perturbation and decoding are batched on the GPU under a fixed memory budget; the held-out experiments reported in Section 7.3 took approximately 38 GPU-hours in total. The main certification campaign took approximately 217 GPU-hours: 3.1 h for QuartzNet, 9.6 h for DeepSpeech2, 21.9 h for Whisper and about 7.6 days of compute for Whisper-ATC, whose 1.54B-parameter autoregressive decode dominates the total. The semantic-audit pass added a further 47 h. The zero-noise control took about 29 GPU-hours.

Formalization. The statements of Appendix A are formalized and machine-checked in Lean 4 / Mathlib (de Moura & Ullrich, 2021); the development is `lean/Roger.lean` in the paper repository, built against Mathlib `v4.29.0`. Table 9 maps each result to its Lean declaration. Every proof is `sorry`-free (audited with `#print axioms`); the two standard analytic facts about the binomial tail — that $F_{N,p}(k) = \mathbb{P}[\mathrm{Bin}(N,p) \le k]$ is antitone in p, and the discrete probability-integral-transform bound $\mathbb{P}[F_{N,q}(K) \le \beta] \le \beta$ — are isolated as

the clearly-labeled axioms `binCDF_antitone` and `cdf_transform`. Every other result (Lemma 2, Theorem 4, and the structural parts of Theorem 3) is proved from Lean's foundations alone.

Table 9: Correspondence between the paper's formal results and the Lean 4 formalization in `lean/Roger.lean`. All declarations are `sorry`-free. "+ ax." marks results that, beyond Lean's foundational axioms (`propext`, `Classical.choice`, `Quot.sound`), additionally invoke the two labeled classical binomial axioms.

Paper result	Lean declaration(s)	Extra axioms
Definition 1	`CPp`	—
Theorem 1	`CPp_coverage`	+ ax.
Lemma 2	`gcs_le_gmsg`, `count_mono`, `V_mono`	none
Theorem 3	`hgap_indicator`, `decomp_coverage`, `union_two`	+ ax. (coverage)
Theorem 4	`audited_bound`, `audited_bound_prob`	none
Proposition 5	`csaudit`	+ ax.
Labeled classical axioms: `binCDF_antitone` (binomial tail antitone in p), `cdf_transform` (discrete probability-integral-transform bound).		

Compute. The experiments consists of forward passes only: N=1000 draws per (message, SNR) for 2,357 messages and eight SNRs per model. Certificate computation from stored counts takes minutes on a CPU; the dominant cost is model inference, which is a quantity procurement already tracks.