The Anatomy of an ASR Hallucination

Anonymous ACL submission

Abstract

Modern speech recognition systems hallucinate under distribution shift, emitting fluent text disconnected from the underlying audio. At what depth of the network does grounded text become irrecoverable? We investigate this through a study of two Conformer-based recognizers — one with a CTC head, one with an RNN-T head — that share an architecture and overlap in training data. The two decoders provide complementary views: CTC, lacking an autoregressive component, isolates encoder-side behavior; RNN-T exposes how a predictor and joiner interact with it. Probing both models on the WildASR robustness benchmark with representational geometry, similarity analysis, layer-wise probing, and causal ablation, we find a surprising geometry of failure. The two encoders are near-identical at the input (CKA ≈ 0.97) but, counterintuitively, converge onto a shared subspace under heavy spectral distortion while diverging on cleaner inputs. Hallucinated utterances exhibit a coupled low-rank signature on both sides of the joiner, and skip-layer ablation localizes the recoverability boundary to a single component: removing the final encoder block raises the hallucination rate by $+97$ pp in CTC and $+71$ pp in RNN-T, while ablating any earlier layer leaves it essentially unchanged. The predictor, long suspected as the source of transducer hallucination, amplifies failure only where the encoder retains rank. Grounded text becomes irrecoverable upstream of the final encoder block.

1 Introduction

A deployed speech recognition system, fed thirty seconds of codec-distorted call audio or reverberant room noise, will sometimes emit a confident, fluent sentence about something the speaker never said. These *hallucinations* are not signaled by low intelligibility: the surface form is well-formed English, trading a wrong word for a confidently-asserted wrong word. They have been documented in encoder-decoder ASR (Wang et al., 2025b), characterised diagnostically (Chen et al., 2024; Atwany et al., 2025), and shown to concentrate under distribution shifts where every model already degrades (Glazer et al., 2026). Unlike LLM hallucination, they arise inside an acoustic encoder over a continuous signal with a fixed vocabulary, and call for an account of *where in the encoder grounded text becomes irrecoverable*.

Existing remedies sit firmly on the decoder side: ILME (Meng et al., 2021), hybrid autoregressive transducers (Variani et al., 2020), lookahead decoding (Unni et al., 2023), and adaptive LM discounting (Unni et al., 2022) all treat the predictor as the suspect. Mechanistic-interpretability work on Whisper (Glazer et al., 2026), building on layer-wise probing of self-supervised speech encoders (Pasad et al., 2021; Roll et al., 2025), has begun probing where decoded text becomes irrecoverable in AED architectures. Transducers, the dominant production architecture, have no comparable account: *at which depth and in which component does the hallucination signal actually form?*

We exploit a methodological control offered by two pretrained NVIDIA NeMo (Kuchaiev et al., 2019) checkpoints on HuggingFace: `nvidia/stt_en_conformer_ctc_large`[1] and `nvidia/stt_en_conformer_transducer_large`[2]. Both share the same Conformer encoder backbone (Gulati et al., 2020) but differ in their head: a stateless CTC layer (Graves et al., 2006) on one, an autoregressive predictor + joint network (Graves, 2012) on the other. Anything both models do on the same input is encoder-driven; anything that appears only with the predictor in the loop is decoder-side. We can therefore localise

[1] https://huggingface.co/nvidia/stt_en_conformer_ctc_large

[2] https://huggingface.co/nvidia/stt_en_conformer_transducer_large

the hallucination signal to a specific architectural depth without retraining.

Probing both models on WildASR (Tay et al., 2026) with representational geometry (ER (Roy and Vetterli, 2007), CKA (Kornblith et al., 2019)), layer-wise probes, a logit-lens-style (nostalgebraist, 2020) head re-tying, and causal skip-layer ablation, we find a coherent picture (Fig. 1):

(i) **Encoders converge under distortion, not diverge.** Near-identical at L=0 (CKA $\approx$ 0.97), they converge at the final layer under heavy spectral distortion (CKA 0.69–0.70 on far-field, reverberation) while diverging on cleaner inputs (0.49–0.56) — the opposite of the naive prediction.

(ii) **Hallucinated utterances show a coupled low-rank signature on both sides of the joiner.** Encoder ER and decoder-side ER are both depressed (RNN-T predictor ER 33–36$\rightarrow$10–25; CTC pre-softmax $-$30–55%).

(iii) **Skip-layer ablation pins the recoverability boundary to the final encoder block.** Removing it raises HER by +97 pp (CTC) and +71 pp (RNN-T); any earlier layer leaves HER essentially unchanged.

The predictor amplifies failure only where the encoder retains rank: HER reaches 86–91% on clean / accent / clipping / codec but only 6–37% on noise-gap / far-field / reverberation. Grounded text becomes irrecoverable at the final encoder block; the predictor amplifies the failure where the encoder leaves room.

Sections 2–3 describe the setup and the WildASR layer-localisation results; §4 confirms the same locus under demographic shift on L2-ARCTIC and exposes a second response mode (directional rotation); §5 discusses implications.

2 Method

2.1 Models

We use the two pretrained NVIDIA NeMo (Kuchaiev et al., 2019) checkpoints `nvidia/stt_en_conformer_ctc_large` and `nvidia/stt_en_conformer_transducer_large`, downloaded from the HuggingFace Hub. Per the NeMo model cards, both checkpoints are trained on NeMo's English ASRSET (LibriSpeech (Panayotov et al., 2015) 960h, Fisher, Switchboard-1, WSJ, National Speech Corpus, VCTK, VoxPopuli-EN (Wang et al., 2021), Europarl-ASR-EN, MLS-EN 2,000h subset, Mozilla Common Voice); the RNN-T checkpoint additionally includes a 12,000-hour subset of People's Speech, so its training set is the same as the CTC's plus that one corpus. Both share a Conformer-Large encoder (Gulati et al., 2020): hidden dimension 512, 8 attention heads, $4\times$ subsampling, 40 ms output frame stride; the full models contain $\approx$120 M parameters each (CTC 121.5 M, RNN-T 120.4 M). The CTC head is a single 1×1 convolution mapping 512 channels to a 129-class vocabulary. The transducer head consists of an LSTM-based predictor, a joint network projecting both encoder and predictor outputs to a 640-dim joint space, and a final linear layer to a 1,025-class vocabulary. The CTC encoder has 18 Conformer blocks (L=0..17) and the transducer encoder has 17 (L=0..16); cross-model comparisons that require matched layer indices (CKA, paired probe accuracies) restrict to L=0..16, while within-model claims about "the final encoder block" refer to each model's own last layer (L=17 for CTC, L=16 for RNN-T). Each block ends with its own `norm_out` LayerNorm and there is no separate post-encoder LayerNorm or out-projection, so each layer's hooked output is dimensionally and statistically compatible with the head.

2.2 Datasets

We use two complementary diagnostic benchmarks. **WildASR** (Tay et al., 2026) provides 10,058 utterances across six environmental degradation conditions (clean, clipping, phone codec, noise gap, far field, reverberation) plus a demographic shift split (accent). **L2-ARCTIC** (Zhao et al., 2018) provides 3,599 scripted utterances of read English from 24 non-native speakers across six L1 backgrounds (Arabic, Chinese, Hindi, Korean, Spanish, Vietnamese, 600 utts each). Decoding is greedy throughout; we use Whisper's English text normaliser (Radford et al., 2023) for all WER comparisons.

2.3 Hallucination flag (HER)

Atwany et al. (2025) introduce the hallucination error rate (HER) as the fraction of utterances flagged as hallucinations, using an LLM-as-judge classifier. We adopt the same rate formulation but instantiate the per-utterance flag with a deterministic, LLM-free rule for reproducibility: our flag

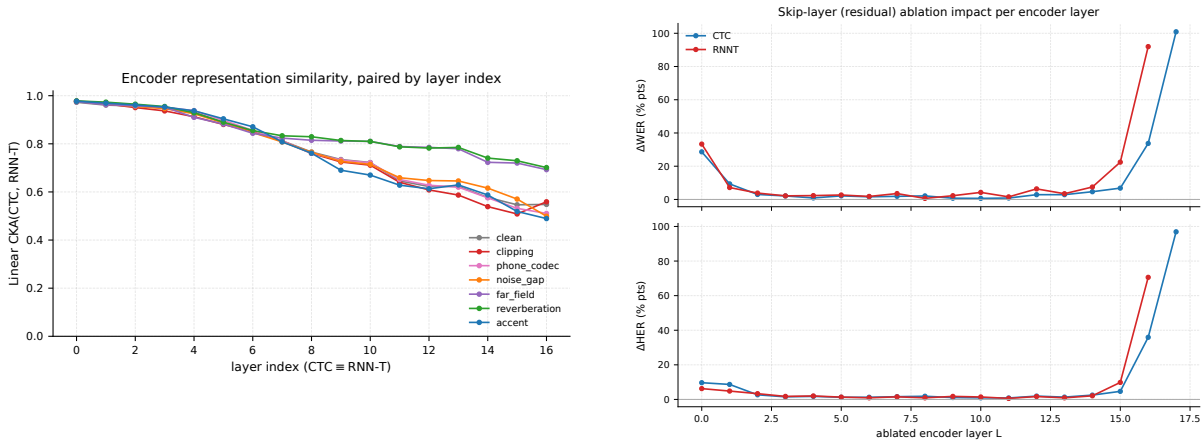

Figure 1: *Left:* per-layer linear CKA between the two encoders, by condition. Networks start identical at L=0 (CKA $\approx$ 0.97) and, counter-intuitively, *converge* under heavy spectral distortion (far-field, reverberation; CKA 0.69–0.70 at L=16) while diverging on cleaner inputs (0.49–0.56). *Right:* skip-layer ablation. Replacing each encoder block in turn with the residual identity. Ablating any block but the final one leaves HER essentially unchanged; ablating the *final* block raises it by +97 pp (CTC) and +71 pp (RNN-T) on every condition.

fires when any of (a) the insertion rate exceeds 50% of the reference token count, (b) the longest fluent 4-gram in the hypothesis has zero overlap with the reference's 4-gram set, or (c) the hypothesis length exceeds 1.5$\times$ the reference and the character error rate exceeds 0.5. This is a conservative fluent-fabrication detector; empirically all positive flags come from rule (b). HER as defined here is deliberately blind to model *collapse* (empty / repetitive output), which is failure but not hallucination.

2.4 Layer-wise probes

Following the layer-wise analysis tradition for speech representation models (Pasad et al., 2021; Roll et al., 2025), for each encoder layer of each model we train a single `nn.Linear` probe with AdamW on a held-out 80/20 split. Four probe targets: **silence** (binary, frame-level, Silero VAD (Team, 2024)), **distribution shift type** (six-way over WildASR's environmental conditions excluding accent), **grapheme** (V-class, frame-level, using the CTC head's argmax on the final-layer logits as a frame-aligned proxy), and **SNR** (regression on per-utterance speech-vs-silence energy ratio). For L2-ARCTIC we add a **six-way native-language probe** over the six L1 backgrounds, balanced via stratified subsampling.

2.5 Encoder lens

Let $h_L(x) \in \mathbb{R}^{T\times D}$ denote layer L's post-norm output and $\mathrm{head}(\cdot)$ the full post-encoder decode path (1$\times$1 conv + blank-collapse for CTC; predictor + joint + greedy decode for RNN-T). The encoder lens (nostalgebraist, 2020) substitutes h_L for the encoder's final output: $\mathrm{LensWER}(L) = \mathrm{WER}(\mathrm{head}(h_L(x)), r)$. For CTC we apply the head directly to h_L; for RNN-T we install a forward hook on the encoder (with CUDA-graph compilation disabled so the hook fires) and run the standard greedy batch decoder. The two procedures are structurally different: the CTC lens is encoder-only because its head is stateless; the RNN-T lens necessarily includes the predictor and joint in the decode path. We therefore use lens *discontinuities across* L as the comparable quantity across decoder families. At the deepest layer the lens reproduces the model's actual decode WER for both families (§3.4), confirming the substitution.

2.6 Representational metrics

Per layer, per condition, and per accent we compute three measures of representational dimensionality. **Effective rank** (Roy and Vetterli, 2007), $\mathrm{ER}(L) = \exp(-\sum_i p_i \log p_i)$, takes p_i as the normalised squared singular values of the centred activation matrix; it is the entropy of the spectrum, in nats. **Participation ratio** $\mathrm{PR}(L) = (\sum_i \sigma_i^2)^2 / \sum_i \sigma_i^4$ on the same spectrum gives an alternative effective count of dominant directions. Both ER and PR are accumulated from a streaming Gram on every frame of the dataset. The **TwoNN** intrinsic-dimension estimator (Facco et al., 2017) (implemented via scikit-dimension (Bac et al., 2021)) infers a non-linear local manifold dimension from the ratio of distances to each

frame's two nearest neighbours, independent of variance scaling. ER is the headline dimensionality measure here; PR and TwoNN are robustness checks (Appendix J), and the fact that PR tracks ER while TwoNN does not is what we read as evidence that the final-layer compression is a linear, variance-anisotropic event rather than a manifold collapse. Between matched layers of the two encoders we additionally compute **linear centred kernel alignment** (Kornblith et al., 2019) on a paired subset of 64 utterances per condition.

2.7 Causal interventions

Skip-layer ablation $\mathrm{Skip}_L : h_L \leftarrow h_{L-1}$ replaces a single encoder layer's output with its input via paired hooks, the standard layer-importance perturbation in residual networks (Veit et al., 2016). **Predictor ablation** (transducer only) replaces the predictor LSTM's output with a constant during greedy decoding in three modes: $\hat{g}_t \in \{0, \bar{g}, g_{\text{blank}}\}$. The channel-mean $\bar{g}$ is computed per-channel across the time axis within each utterance and reused at every decoding step, so the predictor's typical scale is preserved while all time-varying context is removed. Mean-replacement on the encoder collapses subsequent self-attention to uniform weights and produces layer-independent failure (App. B); skip-layer is the cleaner encoder intervention.

3 Results

3.1 Decoding and hallucination behaviour

Table 1 shows that both decoder families rank the seven WildASR conditions in the same order of WER (clean $\ll$ phone_codec $\approx$ accent $\ll$ reverberation $\ll$ far_field $\approx$ clipping $\approx$ noise_gap). They differ in how they fail: on every environmental-degradation condition the RNN-T HER is below the CTC HER, but on the demographic-shift accent condition the order flips (CTC HER 2.10% vs. RNN-T HER 2.50%). Hypothesis disagreement scales with WER on environmental shifts (66.9% on clean to 95.3% on noise gap) but drops to 40.8% on accent, the smallest disagreement among the seven conditions: when the audio is acoustically clean but speaker-shifted, the two decoders agree more than on FLEURS-clean.

3.2 Representational geometry across depth

Figure 2 shows that both encoders exhibit the same three-phase ER curve: a gentle ramp through L=0–4 (ER $\approx$ 60–80), a fast rise through L=5–13, a peak at L=15 in both models (CTC 186, RNN-T 246 on clean), and a *sharp drop from the peak to the final encoder layer* (CTC clean: 186 $\rightarrow$ 137, a 26% drop; CTC far-field: 150 $\rightarrow$ 89, 41%; RNN-T clean: 246 $\rightarrow$ 200, 18%; RNN-T far-field: 169 $\rightarrow$ 102, 39%). Across non-accent conditions the peak-to-final drop is 18–41% (CTC) and 18–44% (RNN-T). The drop is the architectural compression that hands a fixed-rank representation to the head; in CTC it is spread across the last two layers (a deeper minimum at L=16, partial recovery at L=17), in the transducer it is concentrated in the single final layer. Conditions split into two clusters from L=5 onwards: spectrally-distorted (far_field, reverberation) at the bottom, 40–60 ER points below the rest at peak, and all others above. The cluster structure is identical in both decoder families.

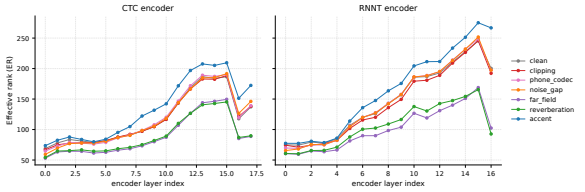

Figure 2: Effective rank per encoder layer for each WildASR condition. Both decoder families exhibit the same three-phase shape (gentle ramp, fast rise to peak, sharp drop at the final layer) and an identical two-cluster structure: spectrally-distorted conditions (far_field, reverberation) sit 40–60 ER points below the rest from L=5 onwards.

Linear CKA between matched layers of the two encoders (Fig. 1, left) shows the inverse pattern: starting from $\approx$ 0.97 at L=0, CKA declines to 0.49–0.70 at L=16, with spectrally-distorted conditions remaining *highest* (0.69 for far_field and 0.70 for reverberation at L=16, vs. 0.49–0.56 for clean, clipping, phone_codec, noise_gap, and accent). The spectral-cluster gap is robust to resampling: 100 bootstrap resamples of the 64 paired utterances per condition give 95% CIs of $[0.68, 0.71]$ for far-field and $[0.67, 0.72]$ for reverberation, separated cleanly from the upper bound of the next-highest condition ($\leq$ 0.59; full table in App. I). Heavy spectral distortion collapses both encoders onto a shared low-rank subspace, leaving less room for decoder-specific specialisation; easier conditions leave more room, and the two encoders diverge more. The ER and CKA cluster orderings are inversely related, exactly as they should be if

Condition	N	dur (min)	CTC WER	RNN-T WER	CTC HER	RNN-T HER	Disagree
clean	647	107.5	7.86	7.14	0.31	0.46	66.9
clipping	647	107.5	17.96	15.76	1.55	0.77	92.4
phone_codec	1294	215.1	9.19	7.41	0.46	0.46	73.7
noise_gap	2588	481.9	17.26	11.90	2.20	1.24	95.3
far_field	1941	405.1	16.24	14.48	2.01	0.88	90.4
reverberation	1941	403.6	14.82	13.00	2.58	1.44	84.8
accent	1000	57.9	5.65	7.16	2.10	2.50	40.8

Table 1: Per-condition WER, HER, and CTC/RNN-T hypothesis disagreement (%) on the seven WildASR splits.

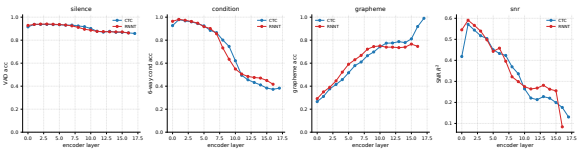

Figure 3: Linear probe accuracy per encoder layer, in both decoder families, on four targets. Silence is encoded uniformly. Distribution-shift type peaks at L=1 and decays through depth. Grapheme readability rises monotonically, with a sharp jump at the final layer (CTC: 0.811→0.919→0.990 at L=15/16/17). SNR (R^2) peaks early and decays.

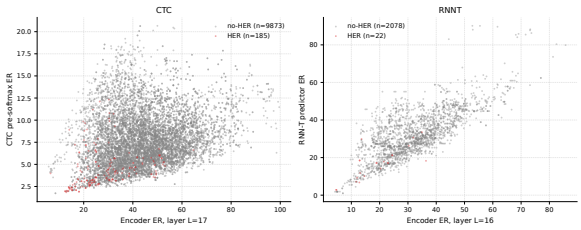

Figure 4: Per-utterance encoder ER (final-layer) versus decoder-side ER. *Left:* CTC, with the pre-softmax ER on the y-axis. *Right:* RNN-T, with the predictor LSTM-output ER on the y-axis. Hallucinated utterances (HER=True; red) cluster in the lower-left: low encoder ER *and* low decoder-side ER. The signature is coupled; either side alone is a weaker predictor.

low-rank conditions force both encoders into the same low-rank directions.

3.3 What each layer encodes (probes)

Figure 3 shows the four probe targets per layer in both decoder families. Three of the four curves point to an *inverted information hierarchy* through depth: acoustic features (condition, SNR, silence) are maximal at L=1 and decline through depth, while symbolic features (grapheme) are minimal at L=0 and rise through depth, with the crossover at approximately L=8–10 in both encoders. The 6-way condition probe peaks at L=1 (CTC 0.978, RNN-T 0.981) and decays to 0.38 (CTC) / 0.42 (RNN-T) at the final layer. SNR (R^2) peaks at L=1 (CTC 0.571, RNN-T 0.591) and decays to $\approx$ 0.13 (CTC) / 0.08 (RNN-T) at the final layer. The grapheme curve has a sharp inflection in the last two CTC layers (0.811 at L=15, 0.919 at L=16, 0.990 at L=17), confirming that the symbolic representation crystallises at the very last CTC encoder block; the RNN-T grapheme curve plateaus around 0.75 because the model produces text through its predictor + joint, not through a stateless head. Because the grapheme target is derived from CTC's own L=17 argmax, the 0.99 at L=17 is partly trivial by construction; the load-bearing observation is the *jump* 0.81→0.99 between L=15 and L=17.

3.4 Encoder lens: when does the representation become text?

We tie the frozen head of each model to each intermediate encoder layer and decode (§2). For both decoder families, lens WER stays above 100% for every layer *except the absolute final one*, where it drops sharply (CTC L=17: 7.9% on clean, 8.3% on phone_codec, 19.4% on clipping; RNN-T L=16: 9.3% on clean, 10.4% on phone_codec, 25.6% on far_field and reverberation). HER is uniformly ∼100% at every layer except the final, where it drops to the model's actual HER. At the deepest layer the lens reproduces the model's actual decode WER to within a few percentage points, confirming the substitution. This places the *recoverability boundary* $L^\star$ (smallest layer with $\mathrm{LensWER}(L^\star) < 100\%$) at the final encoder block for both decoder families. Overlaying lens WER on encoder ER per layer (Fig. 6 in App. D) shows that the two quantities undergo a *simultaneous* discontinuity at $L^\star$: ER drops by 18–41% and lens WER drops from over 100% to below 30% in one architectural step. The dimensional compression and the re-coding into vocab space are the same event.

3.5 Predictor and joint output geometry

Beyond aggregate encoder geometry, we ask whether the per-utterance hallucination signature also shows up on the decoder side. Effective rank has previously been used as a hallucination-detection signal in language models (Wang et al., 2025a); we extend the idea to ASR encoders and to the transducer's predictor LSTM. Figure 4 scatters encoder ER against decoder-side ER per utterance, coloured by HER. In both decoder families, hallucinated utterances cluster in the lower-left: their encoder ER is depressed (consistent with the spectral-distortion cluster of §3.2), and their decoder-side ER is depressed too. Concretely, the RNN-T predictor LSTM-output ER falls from a non-hallucinated baseline of 34–36 to 10–25 on hallucinated utterances (a 29–41% drop, condition-pooled), and the CTC pre-softmax ER falls 15–55% on hallucinated utterances, most steeply on reverberation ($6.8 \rightarrow 3.1$) and far-field ($6.9 \rightarrow 3.6$). The per-utterance signature of hallucination is therefore visible on both sides of the joiner: when the encoder reaches its final layer in a low-rank state, the predictor inherits that state and is also low-rank. This is what we would expect if recoverability has already been decided at the encoder and the predictor passively reflects the encoder's collapsed representation, rather than the two contributing independently. The joint pre-softmax one step further downstream is structurally rank-1 by greedy decoding (per-utt ER 1.04–1.15 on environment-degradation conditions, 1.48 on accent) and does not separate HER from non-HER (Cohen $d \in [-0.66, +0.30]$ across conditions; Appendix E), so the signature is no longer detectable past the joint.

3.6 Causal layer ablation

Skip-layer ablation across encoder depth (Fig. 1, right) shows a clear U-shape: L=0–1 hurts modestly (driven by the spectrally-distorted conditions; on far-field, ablating CTC L=0 raises WER by 72 pp), L=2–13 is essentially flat (ΔWER 1–7 pp, ΔHER $\leq$ 3 pp), and L=14–17 climbs steeply. The HER curve is concentrated almost entirely at the final 1–2 layers: skipping CTC L=17 raises HER by 94–99 pp on every condition (mean +97 pp across non-clean); skipping RNN-T L=16 raises it by 42–88 pp (mean +71 pp across non-clean). *Skipping any middle layer barely moves the hallucination rate; skipping the final layer turns nearly every utterance into a hallucination.* This is consistent with the lens result and the ER curve: the final layer is where text is produced, and removing it produces fluent garbage.

	baseline		mean ablation		
Cond	HER	WER	HER	95% CI	WER
clean	0.5	7.0	**86.5**	[82.0, 91.0]	441
accent	1.0	6.1	**86.0**	[81.0, 90.5]	479
phone_codec	0.5	7.2	**90.5**	[86.5, 94.5]	459
clipping	1.0	17.4	**90.5**	[86.5, 94.0]	376
noise_gap	0.0	12.1	**6.0**	[3.0, 9.5]	40
far_field	0.0	15.8	**32.0**	[26.0, 38.5]	119
reverberation	0.5	13.1	**37.0**	[30.5, 44.5]	104

Table 2: RNN-T predictor ablation, all values in %. Baseline (greedy) and mean-replacement of the predictor LSTM output, on the same 200-utt subsample per condition; full-split baseline WER/HER are in Table 1. Zero and blank replacements saturate at $\sim$100% HER / $\sim$3000% WER on every condition (runaway over-emission) and are omitted. The 95% CI on the mean-ablation HER comes from 2,000 bootstrap resamples; the lowest CI lower bound in the high-HER cluster (clean: 82.0%) sits 37 pp above the highest CI upper bound in the low-HER cluster (reverberation: 44.5%). The WER pattern echoes HER: mean-ablation on the high-HER cluster blows WER up to 376–479% (heavy over-emission of plausible English), while on the low-HER cluster it stays at 40–119%.

3.7 Predictor ablation (transducer only)

Zero and blank predictor replacements saturate at $\sim$100% HER and $\sim$3000% WER everywhere: with no predictor signal, the joint loses its "have I emitted X yet?" cue and over-emits non-blank tokens at every encoder frame until the max-symbols cap, producing runaway repetition. These two modes upper-bound predictor importance but are uninformative across conditions. Mean-replacement is the discriminating intervention: replacing each predictor output with its own channel mean preserves the LSTM's typical scale but removes time-varying context.

Table 2 shows the result. Conditions where the encoder retains rank (clean, accent, clipping, phone_codec) jump to 86–91% HER and 376–479% WER under mean-ablation (95% bootstrap CIs all $\geq$ 81% at the lower HER bound), a 85–90 pp rise in HER from baseline and a 22–71$\times$ WER blow-up driven by token over-emission. Conditions where the encoder has already compressed (noise_gap 6%, far_field 32%, reverberation 37% HER; 40–119% WER; HER CI upper bounds $\leq$ 44.5%) move much less. The split is

	WER		HER		
L1	CTC	RNN-T	CTC	RNN-T	Dis.
Hindi	2.4	2.3	1.3	0.7	13.2
Korean	5.0	5.2	2.8	4.0	33.0
Spanish	6.6	7.6	3.8	5.5	37.0
Arabic	7.1	6.0	3.3	2.2	38.2
Chinese	9.0	8.3	6.2	5.3	46.7
Vietnamese	18.1	16.2	18.8	15.5	62.7

Table 3: L2-ARCTIC scripted split: per-L1-background WER, HER, and CTC/RNN-T hypothesis disagreement (Dis.), all in %. 600 utts per L1 background; same Arctic prompt set across all 24 speakers.

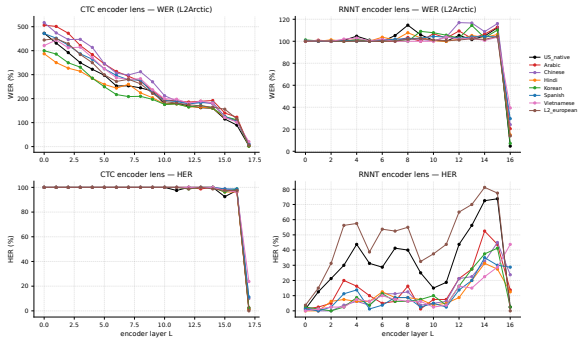

Figure 5: Encoder lens on L2-ARCTIC, per L1 background, both decoder families. Lens WER stays above 100% at every layer except the absolute final one, where it drops to the model's per-accent decode WER. The single-layer text-readiness discontinuity is invariant across accents.

well-separated: the lowest CI lower bound of the high-HER cluster (clean: 82%) sits 37 pp above the highest CI upper bound of the low-HER cluster. The mean-ablation HER ranking matches the encoder ER ranking and the CKA ranking from §3.2 exactly. *The predictor amplifies the hallucination signal where the encoder leaves room, and contributes little where the encoder has already collapsed.* This is causal confirmation of the coupled low-rank signature observed in Fig. 4: the predictor amplifies a recoverability failure that has already been decided upstream by the encoder's final-layer geometry, rather than producing it independently. Existing transducer-side mitigations (ILME (Meng et al., 2021), HAT (Variani et al., 2020), lookahead (Unni et al., 2023), AdaptLMD (Unni et al., 2022)) therefore act on the amplifier, downstream of where recoverability is lost.

4 Generalization to demographic shift

WildASR localises hallucination to the final encoder block under environmental degradation. To test whether this locus is invariant or specific to a particular shift type, we re-run the same probing, lens, and geometry analyses on L2-ARCTIC, with native-language background as the shift dimension. The acoustic conditions are clean (read English over a close-talk mic) and the speakers all read the same Arctic prompt set; the only structured source of variation is the speaker's L1.

4.1 Per-accent decoding

Table 3 shows that WER spans 2.41 (CTC on Hindi) to 18.08 (CTC on Vietnamese), a 7.5× spread, and HER spans 0.67 (RNN-T on Hindi) to 18.83 (CTC on Vietnamese), a 28× spread. Both decoder families produce the same per-accent difficulty ranking, with Vietnamese the hardest and Hindi the easiest by every metric. Hindi is in fact *easier* for both models than WildASR's clean split (CTC HER 1.33 vs 0.31 on WildASR-clean; RNN-T HER 0.67 vs 0.46), consistent with the observation that L2-ARCTIC's acoustic channel is cleaner than FLEURS.

4.2 Locus invariance and accent encoding depth

The encoder lens per accent (Fig. 5) shows that for every L1 background, lens WER stays above 100% at every layer except the final, where it drops to the model's decode WER (CTC L=17: Hindi 2.2, Korean 2.1, Arabic 6.0, Chinese 9.1, Spanish 11.4, Vietnamese 20.5; RNN-T L=16: Korean 7.2, Hindi 14.9, Arabic 20.5, Chinese 24.2, Spanish 29.7, Vietnamese 39.4). The single-layer text-readiness discontinuity is therefore not specific to environmental degradation; it occurs for every accent at the same architectural depth. The *locus* is invariant; the *fidelity at the locus* varies with per-accent difficulty.

The 8-way accent probe per layer (Appendix F) shows that accent identity peaks at L=4–8 (CTC 0.956 at L=4, 0.973 at L=8; RNN-T 0.977 at L=4), much deeper than the environmental-condition probe (which peaks at L=1, §3.3), and decays more slowly. By L=17 the CTC probe is still at 0.699 (chance $1/8 = 0.125$). Per-accent, L2-european is best preserved throughout (0.97 at L=17), US-native nearly as well (0.87), and Korean most aggressively normalised away (0.24 at L=17). Native-language identity is therefore propagated deeper into the encoder than environmental-shift identity, and partially retained even in the head-input representation.

4.3 Accent is encoded as direction, not dimensionality

Per-accent encoder effective rank is tightly clustered: at the peak layer (L=15), the spread across eight accents is $\approx$ 28 ER points for CTC (153–181) and $\approx$ 17 points for RNN-T (227–244), versus 60+ points across environmental conditions (§3.2). Per-accent ER curves overlay nearly perfectly through depth (Appendix G). Combined with the probe result (Fig. 8), this gives a sharp dissociation: accent is encoded as a *direction* in a representation of largely invariant dimensionality, while environmental degradation collapses the representation onto a lower-dimensional shared subspace. Both shifts route through the same final-layer compression to produce text, but their fingerprints inside the encoder are different.

5 Discussion

Where grounded text becomes irrecoverable. Four independent measurements converge on the same architectural event: the final encoder block (Table 4). Earlier layers show none of these effects: ablating any layer in L=2–14 leaves HER essentially unchanged ($\leq$ 3 pp mean), and lens WER stays above 100% throughout L=0–14. The final encoder block performs a sharp simultaneous compression along three axes (dimensional, representational-to-symbolic, condition-invariant); the decoder families *read off* this representation rather than constructing it.

The predictor amplifies; the encoder decides recoverability. Existing transducer-side mitigations — ILME (Meng et al., 2021), HAT (Variani et al., 2020), lookahead decoding (Unni et al., 2023), AdaptLMD (Unni et al., 2022) — target the predictor or joint at decode time, on the implicit assumption that the predictor is the source of spurious LM bias. Our predictor mean-ablation tests this directly: HER rises to 86–91% on shifts that preserve encoder rank but only to 6–37% where the encoder has already collapsed (Table 2). The predictor amplifies hallucination *where the encoder leaves room*; on collapsed inputs it has little variance to inject. Per-utterance, the hallucination signature is coupled low-rank on both sides of the joiner (Fig. 4). The existing decoder-side toolkit acts on the amplifier, not the source.

Shift typology and decoder-mode separation. The two shift types leave qualitatively different fingerprints inside the encoder. Environmental degradation collapses both encoders onto a low-rank shared subspace (CKA 0.69–0.70 on far-field/reverberation at L=16 vs. 0.49–0.56 elsewhere), where predictor neutralisation has limited additional effect. Native-language shift, in contrast, leaves effective rank tightly clustered ($\approx$ 28 ER-point spread across eight accents at L=15 vs. 60+ elsewhere) but rotates the representation: accent identity peaks at L=4–8 (vs. L=1 for environmental condition) and is partially decodable at the final layer. Both shift types route through the same recoverability boundary, but the encoder's response differs in kind: *dimensional collapse* under spectral distortion, *directional rotation* under speaker shift. Decoder family selects the failure mode; the encoder selects the locus.

Measurement	Earlier L	Final L
Encoder ER ↓	185–250 (peak)	**–18 to –44%**
Grapheme probe (CTC)	0.81 (L=15)	**0.99** (L=17)
Lens WER	>100%	**7–30%**
Skip-layer ΔHER	$\leq$ 3 pp	**+97 / +71 pp**

Table 4: Four convergent measurements locating the recoverability boundary at the final encoder block. ΔHER values are CTC L=17 / RNN-T L=16 means across non-clean conditions. See §3 and §4 for full numbers.

6 Conclusion

Hallucination in Conformer-based ASR is localised to the final encoder block. Three converging measurements — a sharp drop in encoder effective rank, the encoder lens transitioning from incoherent to near-baseline WER, and skip-layer ablation raising HER by +97 pp (CTC) and +71 pp (RNN-T) — pinpoint this recoverability boundary in both decoder families, invariant across environmental degradation and demographic (L1-background) shift. The encoder's response differs in kind by shift type: dimensional collapse under spectral distortion, directional rotation under speaker shift.

Decoder family selects what fluent output fills the gap, not where the gap opens. The predictor, long suspected as the source of transducer hallucination, amplifies the failure only where the encoder retains rank; mitigations acting only on the predictor or joint cannot move the recoverability boundary. The productive direction is upstream — on the final-layer compression itself, or on whatever encoder-internal dynamics produce it.

7 Limitations

HER is conservative and rule-based, with all positive flags coming from rule (b) (off-topic 4-gram). It is by design blind to model *collapse* (empty / repetitive output, i.e. catastrophic failure rather than fluent fabrication), so $\Delta\text{HER} = 0$ should not be read as "no impact": we report ΔWER alongside HER throughout, and the two are distinguished cleanly by the zero/blank predictor modes in Table 2. Cross-validation against an LLM-judge or detector-style hallucination metric is left for future work.

Our causal toolkit stops short of three interventions that would tighten specific claims: full activation patching across paired utterances for finer per-utterance causal claims; learned predictor replacement (e.g. shuffling predictor states across utterances) to isolate the internal-LM contribution beyond mean-ablation; and sparse autoencoders or per-head ablation to surface distributed head-level circuits beyond the attention-entropy negative (Appendix C). We use greedy decoding throughout. On the transducer, beam search at sizes 4 and 16 (App. H) yields modest WER improvements over greedy (<1 pp on clean and reverberation) while leaving HER within ± 0.5 pp; the recoverability boundary does not move. NeMo's CTC beam decoder requires an external KenLM and we do not use one. We report 100-bootstrap CIs on CKA (App. I); bootstrap CIs on ER and ΔHER are not reported in this version but effect sizes are large enough that significance is not in doubt.

8 Ethical Considerations

Our work characterizes failure modes in widely-deployed ASR systems. Two dual-use considerations bear mention. First, the layer-localized causal account we provide could in principle inform adversarial degradation of deployed transducer models though such attacks would require white-box access to the encoder, which limits their practical reach. Second, our finding that representation collapse under distribution shift drives hallucination has fairness implications for model compression: rank-based pruning calibrated on clean, high-resource speech may disproportionately degrade performance on non-native accents and acoustically degraded inputs populations that already see worse ASR performance. We encourage future compression work to evaluate on shifted distributions and accent-diverse benchmarks.

All experiments use publicly released artifacts, each used in accordance with its release terms. **Models**: the two NVIDIA NeMo Conformer-Large checkpoints are released under CC-BY-4.0 per their HuggingFace model cards; the NeMo framework itself is Apache 2.0. **Datasets**: LibriSpeech (Panayotov et al., 2015) is released under CC-BY-4.0; VoxPopuli (Wang et al., 2021) is released under CC0 1.0 (European Parliament public-domain speeches); L2-ARCTIC (Zhao et al., 2018) is distributed for non-commercial research under its original academic release terms; WildASR (Tay et al., 2026) is released on HuggingFace under the terms stated in its repository (research use). **Tooling**: `whisper-normalizer` (MIT), Silero VAD (Team, 2024) (MIT), and scikit-dimension (Bac et al., 2021) (MIT). No new data was collected, no human subjects were involved, and no personally identifying information appears in any of the datasets used; all use is non-commercial academic research consistent with the source licenses. We follow the ACL Code of Ethics.

References

Hanin Atwany, Abdul Waheed, Rita Singh, Monojit Choudhury, and Bhiksha Raj. 2025. Lost in transcription, found in distribution shift: Demystifying hallucination in speech foundation models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23181–23203, Vienna, Austria. Association for Computational Linguistics.

Jonathan Bac, Evgeny M. Mirkes, Alexander N. Gorban, Ivan Tyukin, and Andrei Zinovyev. 2021. Scikit-dimension: A python package for intrinsic dimension estimation. *Entropy*, 23(10):1368.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan ... Tao, Zhihang Fu, and Jieping Ye. 2024. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. 2017. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140.

Neta Glazer, Yael Segal-Feldman, Hilit Segev, Aviv Shamsian, Asaf Buchnick, Gill Hetz, Ethan Fetaya, Joseph Keshet, and Aviv Navon. 2026. Beyond transcription: Mechanistic interpretability in ASR. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence*.

Alex Graves. 2012. Sequence transduction with

recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Interspeech 2020*, pages 5036–5040.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3519–3529.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, Patrice Castonguay, Mariya Popova, Jocelyn Huang, and Jonathan M. Cohen. 2019. NeMo: a toolkit for building AI applications using neural modules. https://github.com/NVIDIA-NeMo/NeMo. *Preprint*, arXiv:1909.09577.

Zhong Meng, Jinyu Li, Xie Chen, Yangyang Chou, Shuo Chang, Yifan Liu, Liang Huang, and Yifan Gong. 2021. Internal language model estimation for domain-adaptive end-to-end speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5679–5683.

nostalgebraist. 2020. Interpreting GPT: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: An ASR corpus based on public domain audio books. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pages 914–921.

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

Nathan Roll, Ankita Pasad, Karthik Sridharan, and Alan W Black. 2025. Categorize early, integrate late: Decoupling frame diagnosis from sentence diagnosis in ASR. In *Interspeech*.

Olivier Roy and Martin Vetterli. 2007. The effective rank: A measure of effective dimensionality. In *Proceedings of the 15th European Signal Processing Conference*, pages 606–610.

Geeyang Tay, Wentao Ma, Jaewon Lee, Yuzhi Tang, Daniel Lee, Weisu Yin, Dongming Shen, Silin Meng, Yi Zhu, Mu Li, and Alex Smola. 2026. Back to basics: Revisiting asr in the age of voice agents. ArXiv:2603.25727.

Silero Team. 2024. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad.

Jari Unni, Anuroop Sriram, Sriram Ganapathy, and Preethi Jyothi. 2023. Improving RNN-Transducers with acoustic LookAhead. In *Proceedings of Interspeech 2023*, pages 4064–4068.

Vinit Unni, Shreya Khare, Ashish Mittal, Preethi Jyothi, Sunita Sarawagi, and Samarth Bharadwaj. 2022. Adaptive discounting of implicit language models in RNN-Transducers. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8122–8126.

Ehsan Variani, David Rybach, Cyril Allauzen, and Riley Michael. 2020. Hybrid autoregressive transducer (HAT). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6139–6143.

Andreas Veit, Michael Wilber, and Serge Belongie. 2016. Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems*, volume 29, pages 550–558.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 993–1003.

Rui Wang, Zeming Wei, Guanzhang Yue, and Meng Sun. 2025a. Revisiting hallucination detection with effective rank-based uncertainty. *arXiv preprint arXiv:2510.08389*.

Yingzhi Wang, Anas Alhmoud, Saad Alsahly, Muhammad Alqurishi, and Mirco Ravanelli. 2025b. Calm-Whisper: Reduce Whisper hallucination on non-speech by calming crazy heads down. In *Proceedings of Interspeech 2025*, pages 3414–3418.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. L2-arctic: A non-native english speech corpus. In *Proc. Interspeech*, pages 2783–2787.

A Reproducibility details

Models are loaded via NeMo's (Kuchaiev et al., 2019) `from_pretrained` (`nvidia/stt_en_conformer_ctc_large` and `nvidia/stt_en_conformer_transducer_large`). All probes are trained for 1,500 SGD steps with AdamW at lr=5×10^{-3} and weight decay 10^{-4}, with stratified class balancing capped at 8,000 frames per class. The CTC argmax used as a frame-level grapheme proxy is taken from the same model's layer-17 logits. The encoder lens for RNN-T disables NeMo's CUDA-graph greedy decoder via `change_decoding_strategy`; without this, the forward hook needed to substitute the encoder output collides with stream capture and the substitution is silently bypassed. All WER values are computed after Whisper's `EnglishTextNormalizer`. WildASR and L2-ARCTIC are loaded via the HuggingFace `datasets` library. Probe hyperparameters and the exact intervention code (skip-layer hooks, predictor-replacement context manager) are in the released repository. All experiments — activation collection, lens decoding, skip-layer and predictor ablations, and probe training — were run on a single NVIDIA H200 GPU.

B Mean-replacement ablation

Strict mean-replacement on the encoder (replacing layer L's output with the channel mean computed on clean data, broadcast across time) collapses self-attention at every subsequent layer to uniform weights and produces an essentially layer-independent failure ($\sim$100% WER for every L). We therefore present skip-layer ablation in the main paper. The mean-replacement results are provided in the supplementary parquet but do not separate layers.

C Attention entropy

Per-(layer, head) Shannon entropy of the post-softmax attention probabilities, averaged across batch and query positions, on 64 utterances per condition. Entropy declines through depth in both encoders ($\sim$4.0 nats at L=2 $\rightarrow$ $\sim$2.8 nats at L=17). Spectrally-distorted conditions carry $\sim$0.7–1.0 extra nats of entropy at every layer compared with clean, indicating more diffuse attention under acoustic corruption. We find no individual head with condition-specific outlier behaviour, which is the negative result that justifies the layer-level focus of the main paper.

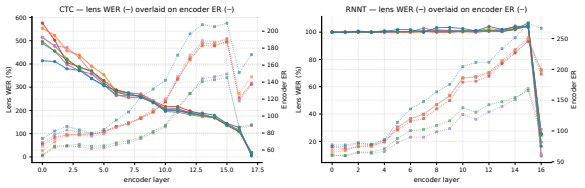

Figure 6: Lens WER (solid) and encoder ER (dotted, twin axis) per WildASR condition, both decoder families. The two quantities undergo a simultaneous discontinuity at the final encoder layer: ER drops by 18–41% and lens WER drops from over 100% to single-digit-to-low-twenties WER in one architectural step.

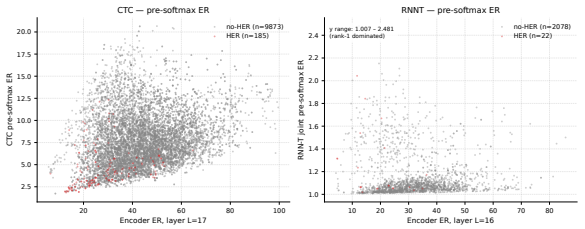

Figure 7: Per-utterance joint pre-softmax ER (RNN-T) by condition. The joint output is structurally rank-1 by greedy decoding (mean ER 1.04–1.15 on environment-degradation conditions; 1.48 on accent), and HER does not modulate it (Cohen $d \in [-0.66, +0.30]$ across conds, all underpowered). The hallucination signal does not survive compression into the joint; we therefore measure it upstream, on the encoder and the predictor LSTM (§3.5).

D Lens WER overlaid on encoder ER

Figure 6 overlays the encoder lens WER on the encoder effective rank, per layer per condition, for both decoder families. The simultaneity of the two discontinuities is the visual evidence behind the "dimensional compression and text re-coding are the same architectural event" claim in §3.4.

E Joint pre-softmax effective rank

The joint network is one step downstream of the predictor in the RNN-T inference path, and a natural place to look for an integrated hallucination signature. Figure 7 shows that the joint pre-softmax effective rank is structurally rank-1 across conditions: mean ER 1.04 (far-field) to 1.15 (phone-codec) on the six environment-degradation condi-

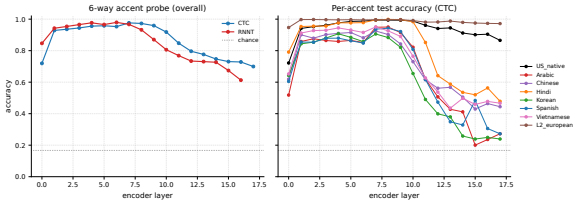

Figure 8: Linear 8-way accent probe accuracy per encoder layer (six L2-ARCTIC backgrounds plus US-native LibriSpeech (Panayotov et al., 2015) and L2-european VoxPopuli (Wang et al., 2021)). Accent identity peaks at L=4–8 (CTC 0.96–0.97; RNN-T 0.93–0.98), much deeper than the env-condition probe (which peaks at L=1), and is partially preserved into the final encoder layer (CTC 0.70 at L=17, chance $1/8$).

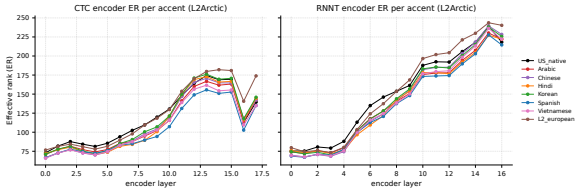

Figure 9: Encoder effective rank per layer, per accent (eight L1 backgrounds). In contrast to the environment-degradation panel (Fig. 2, where spectrally-distorted conditions sit 40–60 ER points below the rest from L=5 onwards), per-accent ER curves are tightly clustered: at the peak layer (L=15) the spread across eight accents is $\approx$ 28 ER points (CTC: 153–181) and $\approx$ 17 points (RNN-T: 227–244).

tions, and 1.48 on accent. This is a property of greedy RNN-T inference: at every decoding step the joint produces a near-binary blank-vs-argmax-token distribution, so the variance across steps lives almost entirely along that single axis. The cross-condition ranking of joint ER is identical to the cross-condition ranking of encoder ER at peak, indicating that the condition signal survives the compression; the hallucination signal does not. Per-condition Cohen d for the HER stratification spans $[-0.66, +0.30]$ and is centred on zero, with all hallucinated-utterance sample sizes $\leq$ 9 per cell. The joint therefore does not carry the per-utterance recoverability signature, which is why the main paper measures it on the encoder and on the predictor LSTM (Fig. 4).

F Accent probe per encoder layer

Figure 8 reports per-layer probe accuracy. The companion lens analysis (Fig. 5 in the main text) shows that accent does not affect the depth of text formation, only the WER at that depth.

G Per-accent effective rank

Figure 9 reports the per-layer effective rank stratified by accent, on the 8-way L2-ARCTIC + LibriSpeech + VoxPopuli accent pool. The contrast with the environment-degradation analysis (Fig. 2) is the load-bearing observation behind the "direction, not dimensionality" claim in §4. The L2-european and US-native curves sit slightly above the L2-ARCTIC L1 backgrounds throughout, reflecting that VoxPopuli and LibriSpeech are recorded under different acoustic conditions than the L2-ARCTIC studio set; the within-L2-ARCTIC spread is narrower still ($\approx$ 17 ER points across six accents at L=15 for CTC).

H RNN-T beam-search sensitivity

We re-decode the WildASR clean and reverberation splits at three beam sizes for the RNN-T (Table 5). Beam search reduces WER by under 1 pp at $b = 16$ on either condition relative to greedy, and HER stays within ± 0.5 pp; the recoverability boundary inferred from the lens is unchanged. The CTC checkpoint's beam decoder in NeMo requires an external KenLM, which we do not use; CTC greedy WER is reported for reference. We test 300 utterances per cell rather than the full split because beam decoding at $b = 16$ is approximately $16\times$ slower than greedy.

Model	Cond	b	WER (%)	HER (%)
CTC	clean	1	8.27	0.33
CTC	reverberation	1	15.28	2.00
RNN-T	clean	1	7.38	0.67
RNN-T	clean	4	7.21	0.67
RNN-T	clean	16	6.91	0.33
RNN-T	reverberation	1	13.13	1.00
RNN-T	reverberation	4	12.69	1.00
RNN-T	reverberation	16	12.24	1.33

Table 5: Beam-search sensitivity on the transducer. WER decreases monotonically with b (by 0.47 pp on clean, 0.89 pp on reverb) and HER stays within ± 0.5 pp of the greedy value; the locus does not move with decoding strategy. CTC beam-search rows omitted (require external LM).

I CKA bootstrap CIs

We resample the 64 paired utterances per condition 100 times (with replacement) and recompute linear CKA at each layer. Table 6 reports the bootstrap mean and 95% CI at L=16, the deepest layer both encoders share. The spectral-cluster

gap is well separated: far-field and reverberation CIs both sit above 0.67 while every other condition's upper bound is below 0.59.

Condition	mean	95% CI
far-field	0.693	[0.677, 0.711]
reverberation	0.699	[0.674, 0.723]
clipping	0.564	[0.539, 0.587]
clean	0.549	[0.517, 0.585]
phone-codec	0.517	[0.497, 0.543]
noise-gap	0.505	[0.490, 0.525]
accent	0.502	[0.482, 0.523]

Table 6: CKA bootstrap CIs at L=16, 100 resamples of 64 paired utts per condition. Far-field/reverberation cluster cleanly above the rest, with no CI overlap.

J Alternative dimensionality measures: PR and TwoNN

In the main paper we use effective rank (ER) as the dimensionality measure of an encoder layer's activation distribution. To check robustness to the choice of measure, we also compute two alternatives on the same streaming Gram: the **participation ratio** $\mathrm{PR} = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$ on the eigenvalues of the centred covariance, which (like ER) measures how variance is allocated across linear directions; and the **TwoNN intrinsic dimension** (Facco et al., 2017) via the scikit-dimension implementation (Bac et al., 2021), a non-linear estimator that recovers the local dimensionality of the underlying manifold of points, independent of variance scaling.

Table 7 reports all three measures at the peak (L=15) and the final encoder layer, per condition. The PR pattern tracks ER almost perfectly: peak-to-final PR drops 34–62% across non-accent conditions for both decoder families (compared with 18–44% for ER), so PR amplifies the same final-layer compression that ER reveals. The spectrally-distorted cluster gap is also present and is larger in PR than in ER (e.g. far-field PR at L=17 is 15 vs. 38–65 for the non-spectral cluster). PR therefore confirms the ER picture: there is a sharp variance-direction compression at the final encoder block, and it is sharper on spectrally-distorted conditions.

TwoNN goes the opposite way. From L=15 to the final layer, TwoNN *increases* for almost every (model, condition) cell (CTC: +17% to +36% across non-accent conditions; RNN-T: +8% to +31%). The spectrally-distorted cluster gap visible in ER and PR is also absent at the final layer for TwoNN: far-field TwoNN at L=17 (CTC) is

	ER		PR		TwoNN	
cond	peak	final	peak	final	peak	final
CTC (peak L=15, final L=17)						
clean	186	137	57	37	9.0	12.3
clipping	188	138	61	38	10.4	12.8
phone_codec	190	139	59	38	9.6	11.2
noise_gap	192	146	62	41	11.0	12.0
far_field	150	89	32	15	10.1	12.4
reverberation	145	90	33	15	11.7	11.9
accent	209	172	69	65	10.3	12.7
RNN-T (peak L=15, final L=16)						
clean	246	200	97	51	9.1	11.0
clipping	246	192	97	44	9.6	12.6
phone_codec	249	197	101	49	9.0	11.5
noise_gap	252	197	103	47	9.6	12.1
far_field	169	102	37	15	10.3	11.3
reverberation	165	93	35	13	11.1	11.9
accent	275	267	136	130	10.6	13.2

Table 7: Three dimensionality measures at the peak encoder layer (L=15) and the final encoder layer (L=17 CTC; L=16 RNN-T), per condition. ER and PR drop sharply at the final layer; TwoNN does not.

12.4, sitting near the middle of the 11.2–12.8 range across conditions, with no clear separation from clean. We interpret this as evidence that the final-layer compression is a *linear, variance-anisotropic* event rather than a manifold collapse: the encoder folds the activations into a low-rank variance subspace (ER, PR drop) but the underlying manifold of points retains its intrinsic dimensionality. This is consistent with the architectural function of the final layer, which is to produce a representation that a linear head (CTC's 1×1 conv) or a linear joint projection (RNN-T's $512 \rightarrow 640$) can read off as text: that requires anisotropic variance, not low-dimensional support.

The headline claim of the paper (the final encoder block is the layer past which grounded text becomes irrecoverable, with effects measurable by ER, lens WER, and skip-layer ablation) does not depend on which dimensionality measure one prefers: ER and PR agree, and the lens and ablation results are independent of either. TwoNN does not contradict the claim; it just measures something different.