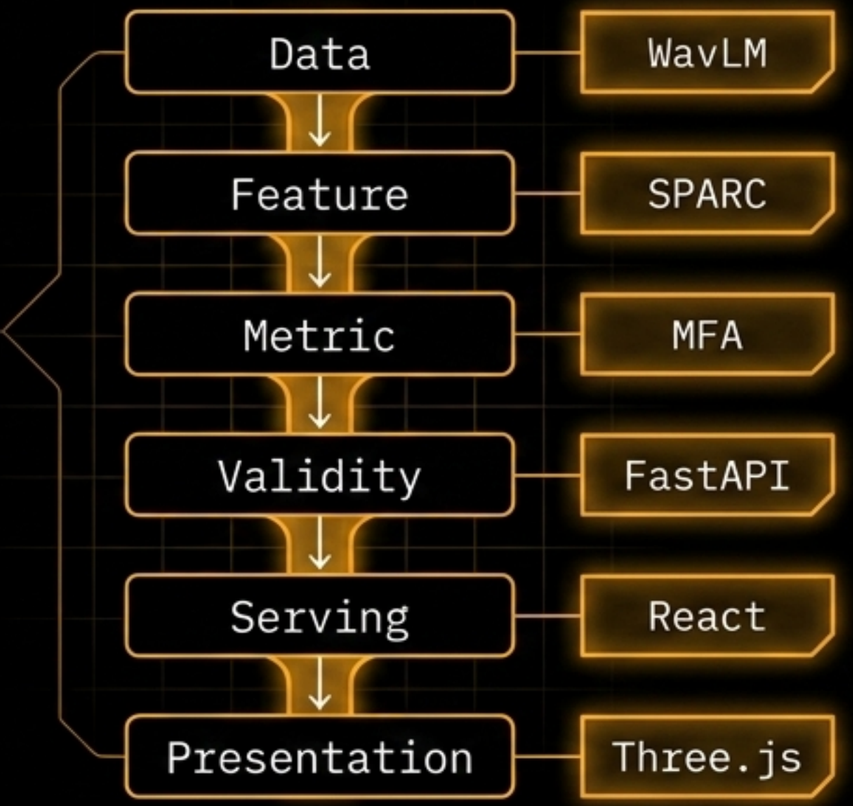
VERITAS

The instrument that puts clinical-speech AI on trial.

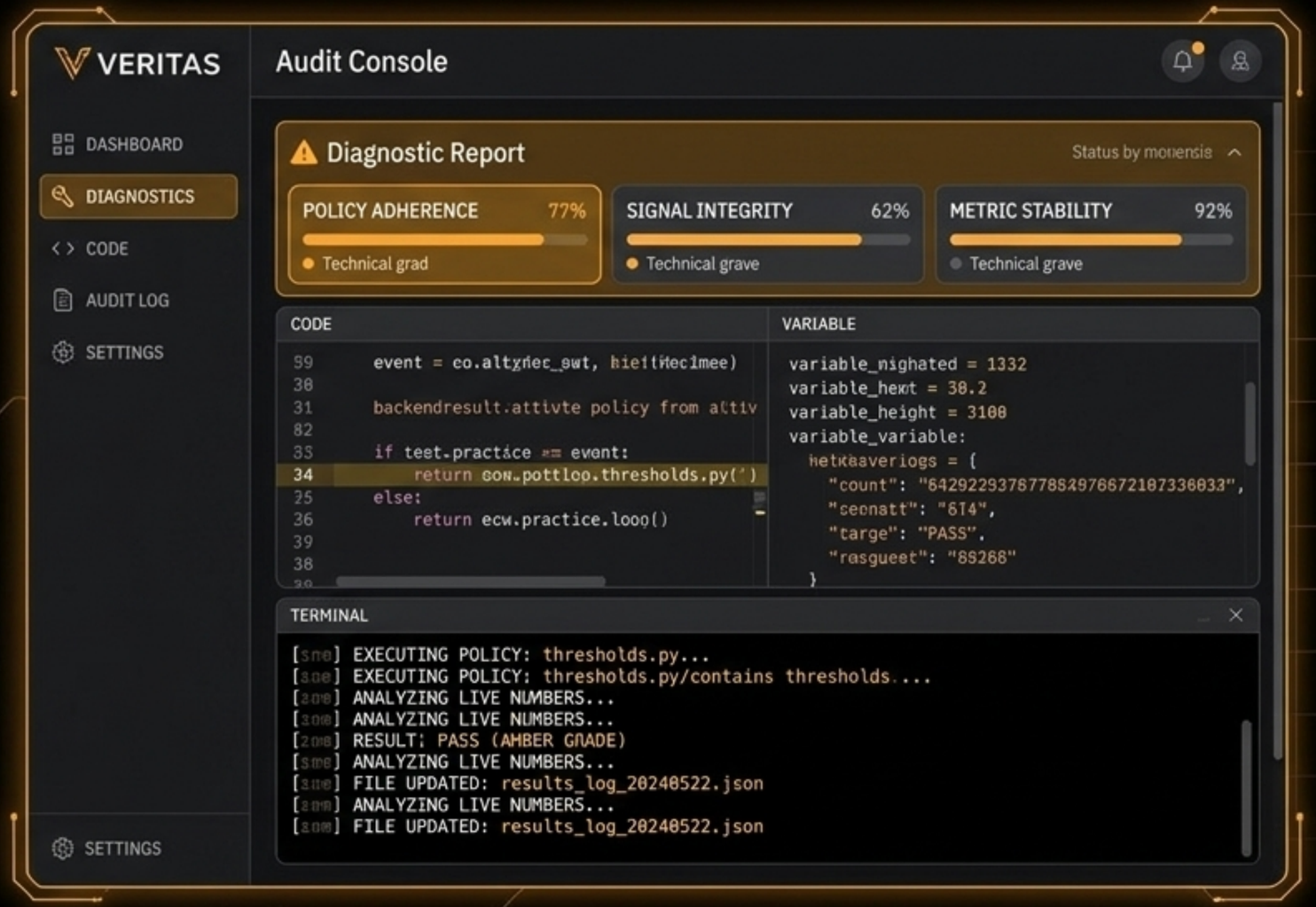
Shaurya Punj · Solo Research · 2026

The Architecture. Validated science backed by full-stack engineering.



Infrastructure built on a dedicated PyTorch/CUDA environment, Dockerized, and gated by ~57 Pytest tests.

The Live Proof. Nothing is mocked.



BACKEND MECHANICS:

- The `FastAPI` backend directly imports the **exact** `thresholds.py` used in the published papers.
- The **console** runs the **real policy** on your **live** numbers.
- **Delete** a **result file** in the backend, and the site dynamically goes blank. It is a **live reflection** of the **science**.

Methodological Rigor. No fabricated rows.

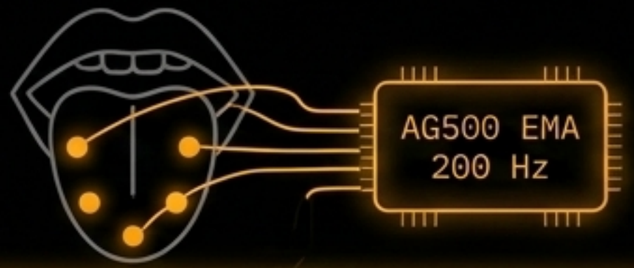
SEED 42

FROZEN WEIGHTS

SPEAKER-DISJOINT SPLITS

~50 NUMBERED SCRIPTS

~57 TESTS



Real Articulography Grounding

Validation uses real sensors on the tongue (AG500 EMA at 200 Hz), not just perceptual approximations.

Every number maps to a committed file.

A single correlation is not validation.

It’s one number that can be right for the exact wrong reason.

Everyone reports one number. Nobody checks the confounds.

ρ = 0.756

A strong apparent signal for dysarthria severity.

76% SILENCE

The signal vanishes when pausing is removed. It’s measuring silence, not speech.

AUC 0.996

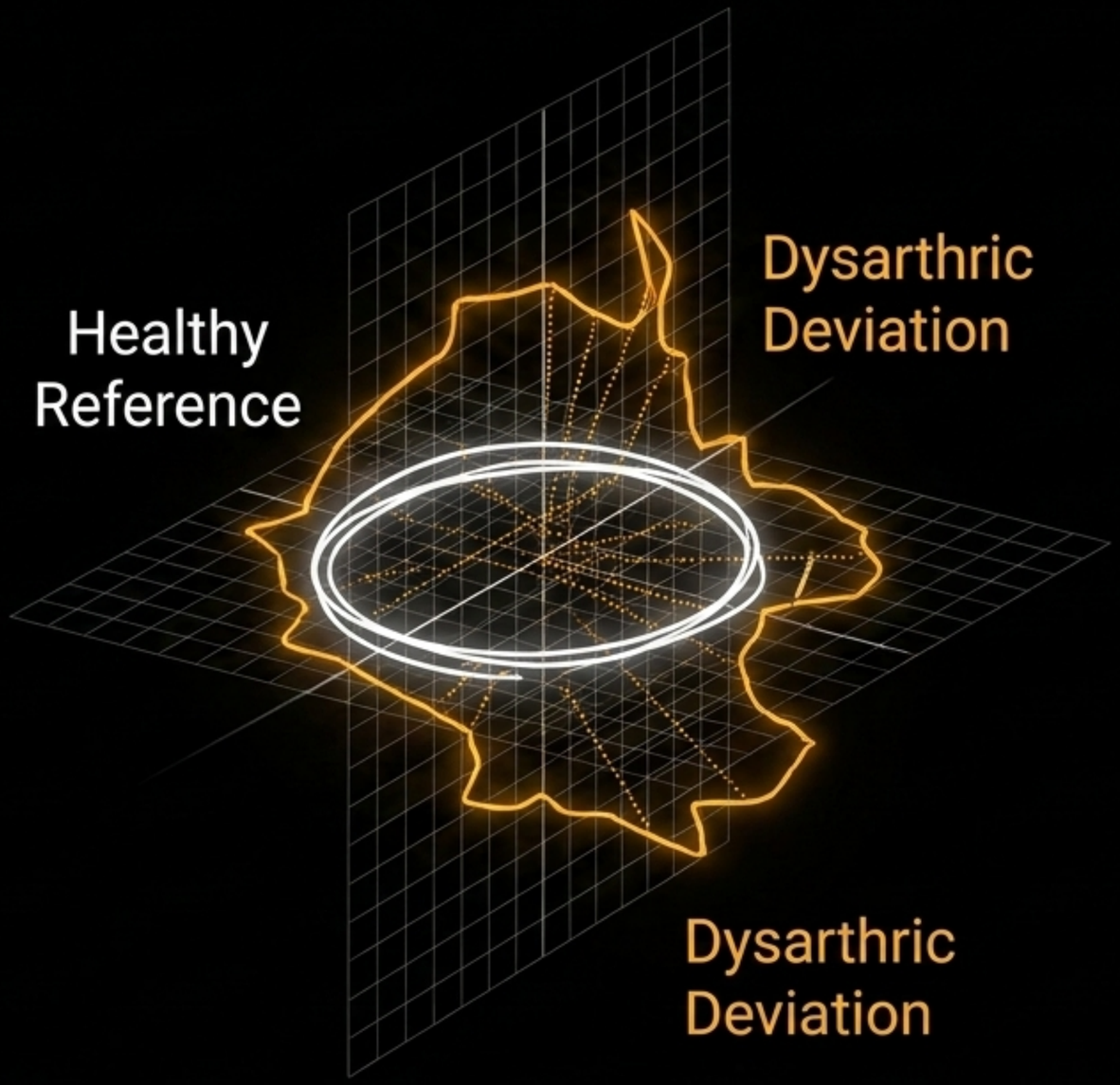
The representation perfectly encodes speaker identity. It knows who is speaking, not how.

It started as a metric of its own.

A spoken word traces a path through WavLM’s 1024-dimensional space at 50 frames/second.

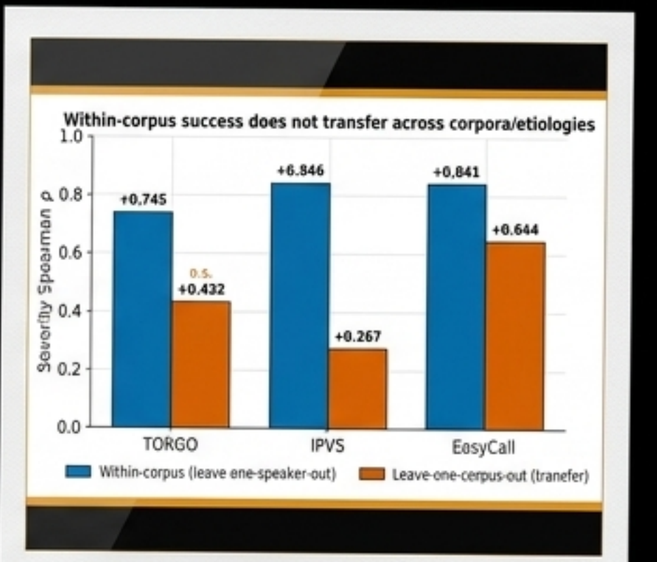
Healthy speakers trace tight paths; dysarthric speakers wander. The deviation is the severity.

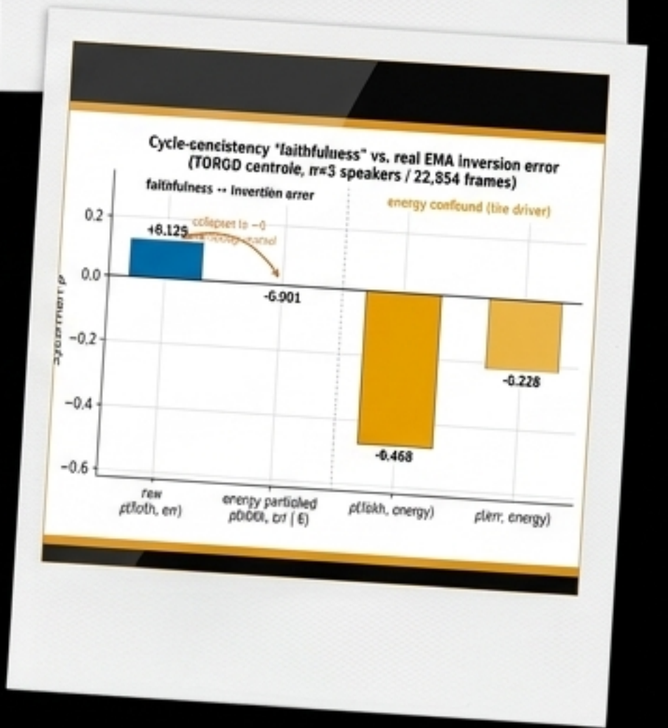
Headline Result:
Spearman ρ = 0.756
on TORGO corpus.



Then it broke. Every attempt to break it *succeeded*.

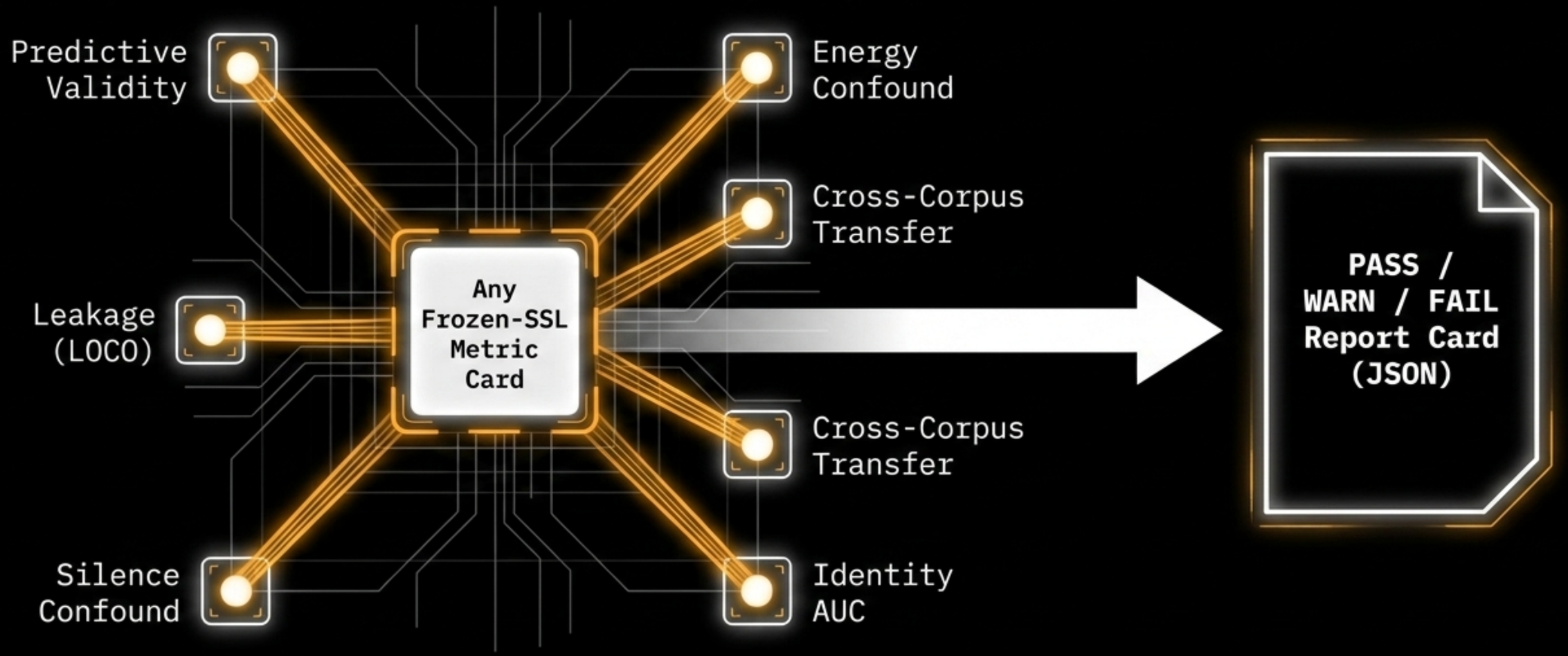
Silence	ρ 0.756 → 0.181 (n.s.) FAIL
	The signal is mostly silence.
Identity	AUC 0.996 FAIL
	It memorized the speaker.
Faithfulness	partial ρ −0.001 FAIL
	Cycle-consistency is just a loudness artifact.
Transfer	ρ 0.432 (n.s.) FAIL
	Does not survive cross-corpus transfer.





So I built the lie-detector: VERITAS-Bench.

A reusable validity instrument for any per-speaker clinical metric.



The six confound controls

Predictive Validity

Leave-one-speaker-out Spearman

`PASS: ρ ≥ .50, p < .05`

Identity Leakage

Naive vs LOCO reference inflation.

`PASS: Δ ≤ 0.05`

Silence

Retention of ρ on speech-only frames.

`PASS: retain ≥ .70 & silence < 50%`

Energy

Partial Spearman controlling for loudness.

`PASS: partial retain ≥ .70`

Cross-Corpus Transfer

Train on others, test held-out

`PASS: ρ ≥ .50, p < .05`

Identifiability

Speaker-classification AUC of features.

`PASS: AUC ≤ .70 | FAIL: > .85`

One FAIL fails the entire card.

The Report Card. *Every metric passes its headline. Every metric fails the audit.*

	ATD·L9	ATD·L1	Pooled Probe	Phono-Sub SOTA (n=3,374 reimplementation)
Predictive	PASS (.77)	PASS (.69)	PASS (.74)	PASS (.74)
Leakage	PASS (.00)	WARN (.09)	N/A	PASS (.00)
Silence	FAIL (24%)	-	WARN (100%)	WARN (100%)
Energy	PASS (96%)	-	PASS (80%)	PASS (88%)
Cross-Corpus	FAIL (n.s.)	-	FAIL (n.s.)	FAIL (n.s.)
Identity	FAIL (.996)	-	FAIL (.996)	FAIL (.999)
Overall Row	FAIL	WARN	FAIL	FAIL

*The **battery discriminates**. It certifies the SOTA's silence/energy robustness while exposing its near-perfect identity confound.*

The Five Findings of Clinical-Speech AI.

1. **The signal is silence, not articulation.**
   Removing silence drops ρ from 0.756 to 0.181. `[silence_trim.py]`
2. **The representation is the speaker.**
   Models perfectly encode identity across the board. `[AUC 0.996–0.999]`
3. **Early-layer leakage is real.**
   Naive references artificially inflate scores. `[Δ = 0.089]`
4. **Cycle-consistency is loudness.**
   Raw faithfulness is just an energy artifact. `[−0.001 over 22,854 frames]`
5. **Ground truth barely exists.**
   Usable public dysarthric EMA is dangerously scarce. `[n ≈ 2]`