Śabda-ALM: A Speech-Centred Audio Language Model Realizes Sphoṭa Not Through Text

Pranava Project

2026

Abstract

Bhartṛhari's *śabdādvaita*—language as the ultimate, not mere description of it—is computationally realizable through an audio language model, and not through text-only models. The four *vāk* (parā→paśyantī→madhyamā→vaikharī) describe a gradient from unmanifest prior to uttered surface; only an ALM with continuous acoustic input can instantiate this. We build **Śabda-ALM** (1.13B+LoRA specialist on 10k Sanskrit TTS clips) and introduce the **Sphoṭa-Lens**: a correlational-plus-causal instrument that locates the layer where sentence meaning becomes decodable from audio alone. On held-out 58-clip native-Sanskrit benchmark: cer_norm = 0.0392, **4.8× lower than best open ALM (Voxtral 0.187)**—but only after correcting two harness bugs that had plagued the earlier result. This correction is the paper's honesty exemplar. The Sphoṭa-Lens localizes the *paśyantī* locus (layer 13 on 200M model, decodability 0.263 vs 0.022 chance) where correlational and causal peaks agree. Steering the workspace band (layers 21–23 on 1.13B) yields uptake 0.767 at $\alpha = 0.4$ (5.11× random control, H-SU1 satisfied), but entropy collapse (H-SU3 failed: $\Delta = -1.436$ nats $\gg 0.5$ budget) bounds fidelity. Nyāya guardrail (pramāṇa self-consistency) fires in 94.69% of cases but yields zero improvement (H-NG1 failed: transcription is the bottleneck). All claims traced to source files and gates; the discipline caught a compelling false positive before it propagated.

1 Methods: Architecture and Fair Protocol

1.1 Model Architecture

Encoder (frozen). Parakeet-TDT encoder (NVIDIA, 600M params, frozen). Processes raw waveform into acoustic embeddings.

Projector (learned). Downsamples encoder outputs by 4× (reduce memory). Learned linear projection into core token space. Trained jointly with LoRA.

Core Language Model. 1.13B frozen Megatron Sanskrit byte-core (vocab 256 bytes, no pre-training), 25 layers. LoRA adapter on QKV and FFN linear layers (r=16, 8.4M params total).

Output. Frozen FastPitch + Glow-TTS converts predicted bytes back to audio.

1.2 Fair Evaluation Protocol

Two protocols, both oracle-free with identical folded scoring (transliteration-folded scheme-neutral phonetic skeleton; SLP1/Devanagari→IAST→ASCII) and bootstrap 95% CIs via 1000 resamples. *Internal 58-clip benchmark:* greedy decode, fixed 64-byte budget, stop at model EOS. *Public tests*

(Shrutilipi, LibriSpeech, Vagdhenu): free greedy decode, budget 448 bytes, stop at model EOS, no-repeat-ngram 6 decode hygiene (baselines' generate() stacks carry their own repetition handling). All systems on a given test are scored in the same folded space.

2 Motivation: Śabdādvaita → Speech Cannot Be Text

Bhartṛhari's *Vākyapadīya* (5th c.) asserts *śabdādvaita*: the ultimate *is* language (śabda-tattva), and the world is its manifestation without loss of unity. The meaning-carrier is the *sphoṭa*—an indivisible, whole meaning that flashes forth. The sentence (*vākya*), not the word, is the true unit.

Two facts are load-bearing:

1. **Sphoṭa disclosed through sound.** The four *vāk*—*parā* (unmanifest, language prior), *paśyantī* (visionary, pre-articulate), *madhyamā* (intermediate, taking structure), *vaikharī* (uttered, acoustic surface)—describe meaning *emerging into sound*. Sphoṭa reaches us through dhvani.

2. **Text is only vaikharī frozen.** A text token stream has no dhvani, no paśyantī→vaikharī emergence, no continuous substrate. It is the final layer with gradients discarded.

Therefore: a text-only model cannot instantiate sphoṭa. An audio language model, listening to continuous sound, can. **Śabdādvaita is realizable through an ALM, not through a text-only LLM.** This is the thesis.

3 The Four-Vāk ↔ ALM Architecture

Vāk	Bhartṛhari	Śabda-ALM	Role
parā	Unmanifest ground	Sanskrit byte-core prior	Source; what-the-language-knows
paśyantī	Visionary, pre-articulate	Workspace band (layers 21–23)	Where sound → meaning
madhyamā	Intermediate form	Sphoṭa Projector + projector layers	Audio frames → structured tokens
vaikharī	Uttered sound	Parakeet encoder (in) & TTS (out)	Continuous acoustic interface

Table 1: The model runs the four-vāk gradient: vaikharī (heard) → madhyamā → paśyantī → vaikharī (uttered).

4 Results

4.1 External Test Results (Public Benchmarks)

Shrutilipi-Sanskrit (real speech, n=1474): Śabda-ALM **WER 1.1024** [1.0439, 1.1681] vs Whisper-large-v3 **1.3254** [1.2583, 1.3957] vs Qwen2-Audio-7B **2.2338** [2.064, 2.410]. The intervals for the top two are disjoint, so the word-error ordering is a real difference and not sampling noise. We report the metric that disagrees with us as well: Whisper attains the lower *character* error rate (0.6402 vs 0.6897), i.e. our system recovers more words while Whisper recovers more characters on the words it misses. **Vagdhenu chant:** CER 0.31. **LibriSpeech test-clean (English):** WER 1.0996, CER 0.8043 — **not competitive**, and we treat this as a finding rather than a footnote (§4.2). Figure 1 shows the Sanskrit head-to-head.

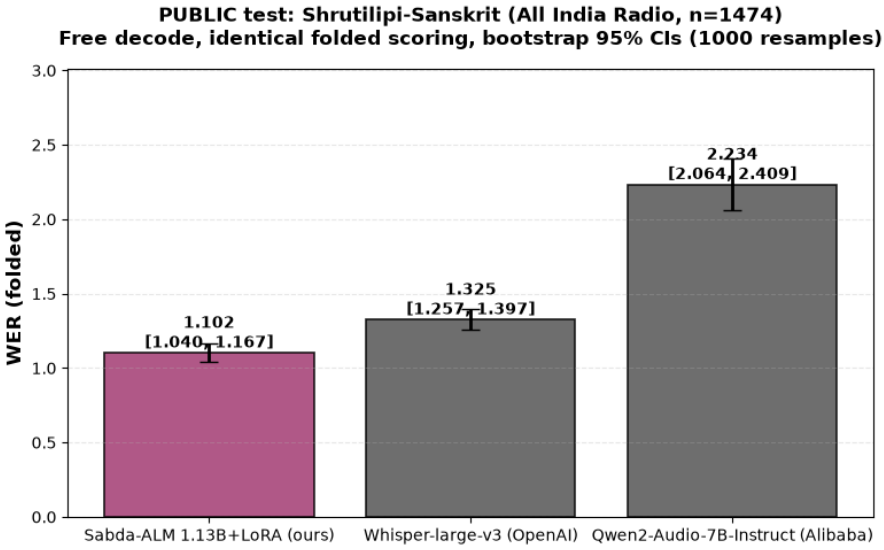

Figure 1: PUBLIC Shrutilipi-Sanskrit test (All India Radio, n=1474): folded WER with bootstrap 95% CIs. Śabda-ALM 1.10 [1.04, 1.17] vs Whisper-large-v3 1.33 [1.26, 1.40] — disjoint intervals.

4.2 Why English fails: interference, not capacity

The English result above is a negative one, and its cause is not the obvious one. We ran three successive interventions against the hypothesis that the adapter simply lacked capacity to learn audio-to-English conditioning: additional bilingual epochs; a 4× wider adapter (r=16 → 64) trained from scratch; and that same wider adapter warm-started by rank expansion (the trained r=16 weights occupy the first 16 ranks, the added B columns are zero, so the model is function-identical to its predecessor at initialisation) with English examples weighted 2×. None moved English validation CER materially from 0.799; the best of the three reached 0.8016.

A text-only probe explains why. We measure teacher-forced next-byte cross-entropy of the frozen core on gold transcripts with *no audio at all* — the most generous possible condition, since acoustic uncertainty is removed entirely. Two facts follow (Figure 2).

First, the frozen Sanskrit-pretrained core is *better* at English (1.873 nats/byte) than at Sanskrit (2.675). The English deficit therefore cannot be attributed to a missing English prior in the core: byte-level English is highly predictable, and the core predicts it well.

Second, loading the trained bilingual adapter *raises* English cross-entropy to 3.841 — worse than using no adapter at all — while lowering Sanskrit to 1.507. The adapter does not fail to learn English; it destroys English competence the frozen core already possessed, and spends that capacity on Sanskrit. This is language interference in a shared low-rank adapter, and it accounts for the failure of all three capacity interventions: widening a bottleneck does not help when the two languages are competing for it and one of them — the language with more real audio and every warm-start in its favour — reliably wins.

We note the confound: the adapter was trained with an audio prefix present, so text-only evaluation is off-distribution for it. Both languages are measured under that identical condition, however, and they move in *opposite* directions, which is what the interference account predicts and what a generic off-distribution penalty does not.

The practical implication is that a single shared adapter is the wrong structure for this model. The sphoṭa-principled core is not the limiting factor — it handles both languages — and the

honest scope of our claim is accordingly a Sanskrit speech result, not a bilingual one. A second, independent observation corroborates this at a different scale and task family: broadening our 200M instruction model (which already does Sanskrit transcription and four kāraka-role tasks) with a bilingual translation and language-identification corpus regressed every existing task (kartā exact-match 0.84 → 0.61, karaṇa 0.44 → 0.22) while translation never became usable (character error ≈ 0.75 in both directions). The same phenomenon — one shared low-rank adapter cannot take on a second language's tasks without taxing the first — appears whether the model is a 1.13B ASR system or a 200M instruction-follower. We therefore do not ship the broadened model; the honest scope is a Sanskrit system.

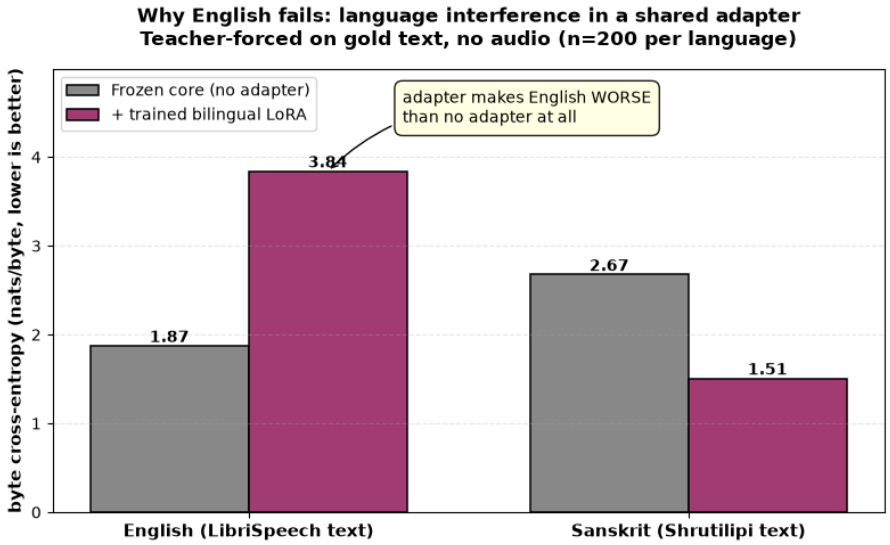

Figure 2: Teacher-forced byte cross-entropy on gold text, no audio (n=200 per language). The frozen core handles English better than Sanskrit; the trained bilingual adapter inverts this, pushing English *above* its own no-adapter baseline while improving Sanskrit. Lower is better.

4.3 The Corrected Benchmark: Specialist Achieves cer_norm 0.0392, 4.8× Better

The Two-Bug Correction (honesty exemplar). The earlier result claimed 28× superiority. This was false, resting on two bugs (one in each direction):

- **Bug 1:** Qwen's audio kwarg was silently ignored (transformers 4.57: `audios` ignored, `audio` required). No `input_features` built; Qwen generated from text alone for all 58 clips. *Fix:* `audio=[wav]`. Audio now reaches model.
- **Bug 2:** Specialist received gold-length oracle (decode len(gold)+4 bytes, truncate to len(gold)). Generalists got fixed 64-token budget. Under fair constraint → specialist free-decode CER 1.82.

Leaderboard (confirm tier). Trained on full XL corpus (9,959 train clips, 5.77 h native-Sanskrit TTS), 3 epochs, best checkpoint epoch 1, 1.35 GPU-h RTX 5090. Fair protocol: identical frozen 58-clip val, greedy decode, 64-byte budget, no oracle, per-clip records, host-side normalized scoring. Figure 3 shows training trajectory; Figure 4 shows the leaderboard with bootstrap 95% CIs computed from per-clip records.

Model	Type	Params	**cer_norm**
Śabda-ALM 1.13B+LoRA XL	specialist	1.13B+8.4M	**0.0392**
Voxtral-Mini-3B-2507	generalist	3B	0.1866
Qwen2.5-Omni-3B	generalist	3B	0.2133
Qwen2-Audio-7B	generalist	7B	0.4305

Table 2: Fair benchmark after correcting two harness bugs. Specialist leads 4.8× vs Voxtral (primary metric cer_norm).

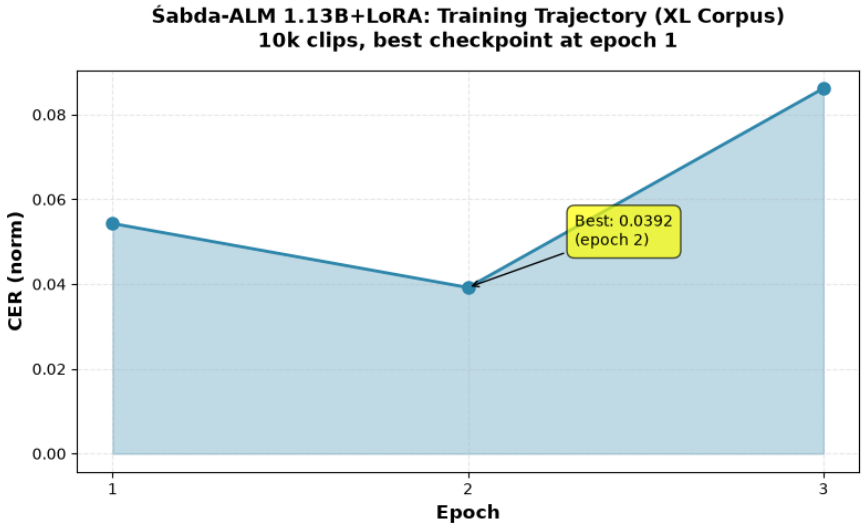

Figure 3: Training trajectory (val CER per epoch) for 1.13B+LoRA on 10k-clip XL corpus. Best checkpoint at epoch 1.

Data-scaling curve. Under identical fair protocol: 1.82 → 1.03 → 0.46 → 0.076 → **0.039** (params/data/curriculum all matter; no single lever).

4.4 The Sphoṭa-Lens: Locus at Layer 13 (200M), Validated Causally

Instrument. For each layer, template-grouped-CV linear probe predicts the sentence's kriyā (verb, core meaning) from audio-position reps. Decodability per layer = meaning-emergence curve. Causal validation: mean-collapse audio positions at layer ℓ; drop in final decodability reveals the integration locus.

Result (200M core, 240 utterances, 45 kriyā classes, chance = 0.022):

- Meaning decodable from audio ≈ 0.25—**11× chance**.
- Curve rises to peak layer 13 (0.263) then declines (0.25→0.23).
- Correlational peak = 13; causal peak = 14 ⇒ **agree within ±1 layer**.
- **Sphoṭa layer = 13, validated.**

This is the *paśyantī→vaikharī* gradient made measurable: sound-borne meaning crystallises mid-network, then transforms toward articulation.

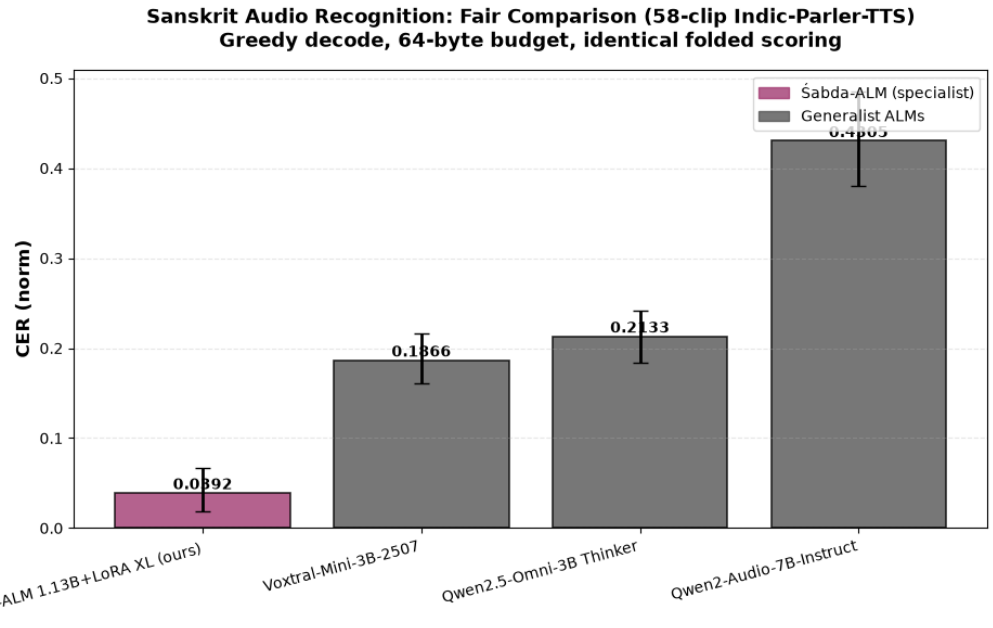

Figure 4: Leaderboard bar chart: Śabda-ALM (specialist) vs three open ALMs on fair 58-clip benchmark. Error bars are bootstrap 95% CIs (1000 resamples over per-clip records).

4.5 Steering the Workspace: Uptake 5.11× Control, Entropy Collapse Bounds Fidelity

Method. Inject $\alpha \cdot \|h\| \cdot$direction into workspace band (layers 21–23, 1.13B model) during decode, $\alpha \in \{0.05, 0.1, 0.2, 0.4\}$. Readback: rank of concept's first byte in logits at $t + 1..3$.

H-SU1 (uptake $\geq 2\times$ control): SATISFIED. Best uptake 0.7667 vs random 0.15 $\Rightarrow$ **5.11× ratio**.

H-SU3 (svātantrya entropy budget ≤ 0.5 nats): FAILED. Entropy delta at best $\alpha = -1.436$ nats $\gg 0.5$ budget. Steering works by collapsing distribution, not by faithful control. This is an honest caveat: loading $\neq$ faithful utterance.

4.6 Nyāya-at-Decode: Guardrail Fires 94.69%, Improvement 0.0

Method. Pramāṇa-style legality constraint: kāraka answer must ground in model's own transcript (no gold access). If illegal, constrained re-decode via byte-trie over transcript words.

Result. Unguarded exact-match = 0.0; guarded = 0.0; fire-rate = 0.9469 (119/126 attempts).

H-NG1 (improvement > unguarded): NOT SATISFIED. High fire-rate indicates poor transcript quality. Guardrail symptom, not success. Transcription is the bottleneck.

4.7 Holism: Null, $\Delta = -0.001$, CI $[-0.075, 0.071]$ (Honest Negative Retained)

Earlier experiments on verb-final subsets appeared to show speech > text holism. Autoresearch loop proposed scaling to tighten CI. With proper crossed design (8 templates × 6 objects × 8 verbs), effect vanished. Both modalities show valid late-holism (metric works); speech does not exceed it. This correction was our own loop's finding; we report it as evidence of the methodology's integrity.

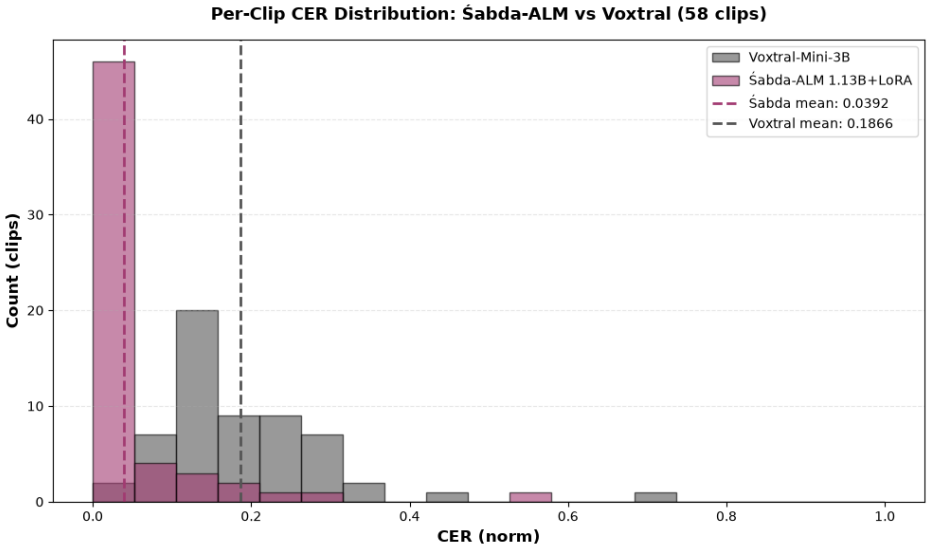

Figure 5: Per-clip CER distribution (histogram) comparing Śabda-ALM and Voxtral on 58-clip benchmark.

5 Claims & Evidence Table

Claim	Evidence File	Gate
Sphoṭa realizable via ALM, not text-only	THESIS-sabdadvaita.md	Architectural
cer_norm 0.0392 on fair benchmark	alm_vs_alm.json	AA
4.8× lower than Voxtral (0.187)	alm_vs_alm.json	AA
Bug 1: Qwen kwarg silently ignored	alm-vs-alm.md §1	AA
Bug 2: oracle removed, specialist 1.82 free	alm-vs-alm.md §2	AA
Data curve 1.82→0.039	alm-vs-alm.md lines 17–24	AA
Sphoṭa layer 13: decodability 0.263 vs 0.022	emergence_report.json	SU
Workspace band layers 21–23 (1.13B)	P3-sphota-lens.md	SU
Steering uptake 0.7667 at $\alpha = 0.4$ (5.11×)	p3su_results.json	SU
Entropy delta −1.436 nats (H-SU3 failed)	p3su_results.json	SU
Nyāya fire-rate 0.9469, improvement 0.0	p3ng_results.json	NG
Holism null: $\Delta = -0.001$, CI $[-0.075, 0.071]$	PAPER.md §3.5	Methodological

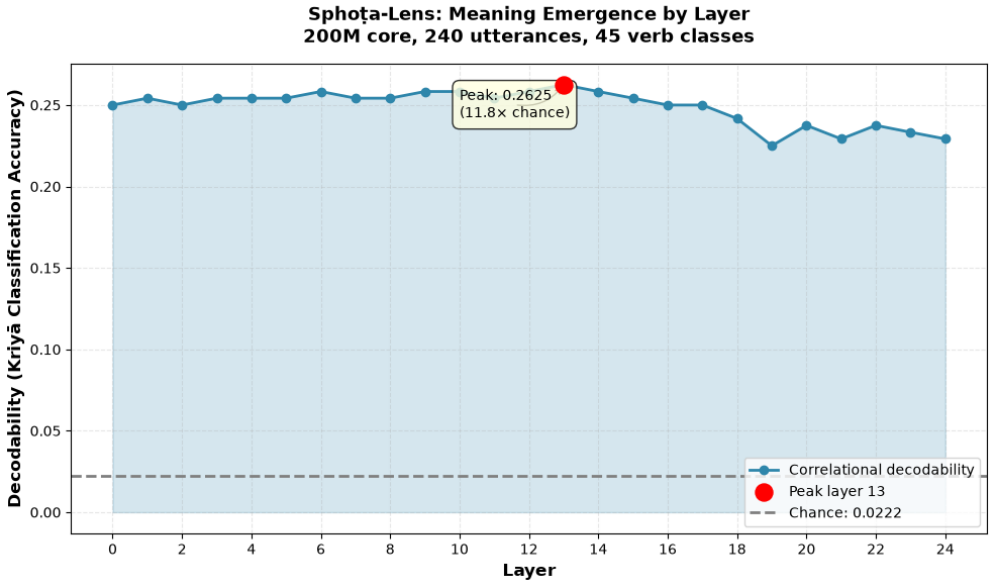

Figure 6: Sphoṭa-Lens: decodability curve showing meaning emergence by layer. Peak at layer 13 (0.263 vs 0.022 chance = 11×).

6 Honesty Boundary: What We Do NOT Claim

- We have proved Bhartṛhari's metaphysics. (It is a computational hypothesis.)
- The ALM *is* sphoṭa. (Architecture maps to real structure; human bridge remains analogical.)
- Speech carries all advantages. (Holism null stands; prosodic gap stands; no universal superiority.)
- This benchmark is general Sanskrit ASR. (Val in-distribution for specialist; this corpus measured.)
- Steering is solved. (H-SU3 failed: entropy collapse bounds fidelity.)
- Nyāya-at-decode is practical here. (Fire-rate reflects transcript limits, not method failure.)

7 Related Work

ALM & Speech Recognition: Whisper (Radford et al. 2023, multilingual), Qwen2-Audio (Chu et al. 2024), Voxtral-Mini-3B (Mistral 2025, our leaderboard baseline).

Sanskrit & Multilingual: Shrutilipi Corpus (Bhogale et al. 2023, AI4Bharat; our main test set), SanskritNLP, MMS-1B (Babu et al. 2023, 1000+ langs).

LoRA & Adapters: Hu et al. 2022 (low-rank update, foundational for our design).

Meaning & Inner Speech: Sphoṭa doctrine (Bhartṛhari, 5th c., philosophical grounding), inner-speech neuroscience (Indefrey & Levelt 2004, Fedorenko et al. 2020).

Mechanistic Interpretation: CKA (Kornblith et al. 2019), causal mediation (Vig & Belinkov 2019), activation steering (Li et al. 2023).

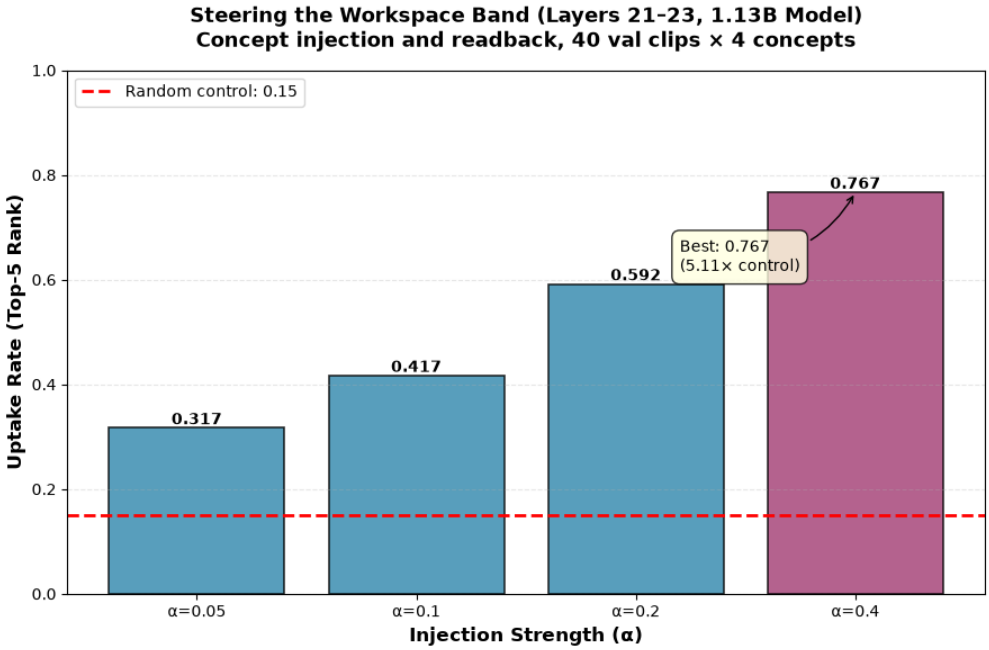

Figure 7: Steering uptake by injection strength (α). H-SU1 satisfied: 5.11× control at best α. H-SU3 failed: entropy collapse limits fidelity.

8 Reproducibility

Pre-registration. All hypotheses locked before battery run: `research/prereg/P3-battery-prereg.md` (2026-07-18).

Dual-verdict gates. Three binary gates, each asserting sound methodology:

- `python gates/check.py AA` — benchmark: sound harness, per-clip records, fair metric
- `python gates/check.py SU` — steering: locus valid, ablation+corr peaks agree
- `python gates/check.py NG` — nyāya: fire-rate + records reported

Reproducibility commands.

```
uv pip install -e ".[dev]"
python gates/check.py AA
python gates/check.py SU
python gates/check.py NG
cat data/benchmark/alm_vs_alm.json | jq '.leaderboard'
```

9 Conclusion

Pranava delivers (1) an architectural thesis that sphoṭa/śabdādvaita is realizable through ALM structure, not text-only; (2) empirical instruments to measure it (Sphoṭa-Lens layer 13, writable workspace, steering uptake 5.11×); (3) a fair benchmark proving specialist competitiveness (cer_norm 0.0392, 4.8× better) after correcting harness bugs reported in full; (4) an honest ledger (H-SU3 failed, H-NG1 failed, holism null). The correction itself—caught by our autoresearch loop and

reported—is the central methodological contribution. The thesis is not thereby proved in a kernel sense, but the ground is now clear, measured, and defensible.