N'Ko as Computational Infrastructure: Script-Native Speech Recognition, a Phonemically Interpretable Error Metric, and Admissible Tone Correction

Mohamed Diomande

Canonical research manuscript, May 2026 (revised 31 May 2026)

Abstract

This manuscript consolidates a multi-paper research program on N'Ko, Manding automatic speech recognition, script visibility in large language models, and trajectory-conditioned decoding. The central argument is that N'Ko should not be treated as a decorative or interchangeable rendering of Manding language. For machine-learning systems it functions as computational infrastructure: it determines what tokenizers can represent, what hidden circuits are available, how acoustic evidence is aligned to symbols, whether reported error rates measure speech recognition or merely agreement with an inherited orthographic convention, and how tone itself can be encoded and reconstructed from acoustic evidence.

The paper integrates five written project papers and subsequent audit notes into a single canonical account. The representation studies show that current LLM families accept N'Ko Unicode strings while internally underrepresenting the script through inflated translation cost, weak activation geometry, entropy gaps, sparsity inflation, kurtosis deficits, and poor circuit-duplication yield. The speech papers show a progression from early CTC systems to frozen-Whisper-feature decoders and then to a trajectory-conditioned Transformer CTC decoder. In the canonical architecture, 1280-dimensional Whisper large-v3 features are projected into a 768-dimensional decoder space, downsampled temporally, and decoded by a six-layer Transformer CTC head. The anticipatory component computes a seven-dimensional trajectory state z_t for each timestep–commitment, uncertainty, transition pressure, recovery margin, phase stiffness, novelty, and stability–and injects it as an attention-logit bias $B_{ij}^{(m)}$ before CTC emission.

The mathematical claim is a measurement claim, not an automatic leaderboard guarantee. I formalize a transparent-script proposition: if a normalized script map $f_N : \Phi \to \Sigma_N$ is bijective over the target phoneme inventory, then character edit distance over $f_N(\phi_{1:U})$ preserves the phoneme-edit structure up to explicitly modeled normalization choices. A Latin transcription relation with variable-length digraphs, optional tone marking, and spelling variation does not have the same property. This makes N'Ko CER a more phonemically interpretable metric for Manding ASR than Latin WER, even though it is not a perfect phoneme error rate and still depends on normalization, reference quality, and tone/diacritic policy.

The strongest retained ASR artifact is an archived checkpoint trained on a 290,596-pair Bambara corpus snapshot, with a 232,476/29,060/29,060 train/validation/test split, learning rate 0.0003, batch size 32, dropout 0.1, seed 42, and reported test CER of 20.57%. This is the canonical public anchor for discussing "the 20% CER" result. The anchor is important because it shows that direct script-native N'Ko ASR reached a meaningful error regime on a large Bambara corpus; it is not a newly completed May 2026 strict reproduction, not a result for later internal decoder variants, and not a closed proof that N'Ko beats Latin under every matched hyperparameter setting.

The Anticipation Geometry Partition (AGP) is likewise bounded carefully. AGP is not the acoustic model that produced 20.57%; it is the post-ASR geometry and governance layer that

partitions transcript spans into stable, boundary, uncertain, and novelty states before allowing conservative correction. The resulting manuscript is best read as a systems-and-measurement paper: it connects script invisibility in LLMs, script-native ASR architecture, trajectory geometry, phonemically grounded evaluation, and row-level correction governance around a single archived 20.57% CER anchor.

Tone, historically the program's largest open hedge, is treated here as a structured reconstruction problem rather than a deferred limitation. The same trajectory geometry z_t is reused at three levels—decode, govern, and correct—and tone resolution is framed as the product of a linguistic prior over tone marks and acoustic evidence about pitch. This positions two companion papers as the reconstruction pillar of the program: a contextual tone model that resolves tone from *text* and Featural Acoustic Coding (FAC), which resolves tone from *acoustic* F_0 and treats the N'Ko syllable codebook as a featural sound code. A further companion develops admissible expert routing (MAOE-N'Ko) and an applied protocol direction turns this stack into linguistic computation that workers, not GPUs, can perform. The present paper is the umbrella and the measurement spine; the companions are cited where they carry the empirical load, and every tone-related number is reported with its provisional, OCR-noisy provenance.

1 Introduction

Modern multilingual NLP and ASR systems often inherit a hidden assumption: script is a surface channel. A sentence may be written in Latin, Arabic, Cyrillic, Devanagari, or another script, but the model is expected to learn the same linguistic content once enough data is supplied. That assumption is convenient for model builders, but it is scientifically unsafe for low-resource and indigenous scripts whose structure is not visible in pretraining corpora, tokenizers, or standard evaluation metrics.

N'Ko makes the failure concrete. It is a West African script designed for Manding languages and encoded in Unicode at U+07C0–U+07FF [9]. The script also belongs to a social and intellectual movement around Manding literacy, standardization, and linguistic self-representation [1]. Unlike Latin Bambara orthography, N'Ko was deliberately engineered around Manding phonology, including segmental distinctions and tonal marking. For a CTC ASR system [3], that difference is not cosmetic. The model emits a symbol sequence through time, and the scientific interpretation of its errors depends on whether those symbols approximate acoustic-phonemic units or a historically adapted spelling convention.

The research program behind this manuscript began as several papers rather than one. *Dead Circuits* diagnosed N'Ko invisibility inside LLM activations. *Living Speech* documented the construction of a script-native N'Ko ASR stack. *Script Invisibility Is Structural* tested whether the activation pattern extended across model families. *Does Script Design Matter?* developed the trajectory ASR argument and produced the archived 20.57% checkpoint. *Deployment Properties* explored generalization, domain transfer, row-level artifacts, and the correction layer that later became AGP. Later project notes sharpened the story further: the compact trajectory scalars appeared more important than heavier decoder additions, word error rate looked scientifically weak for tonal Manding ASR, and the strongest immediate public claim became the metric-validity argument around N'Ko CER rather than an overbroad leaderboard claim.

This canonical version therefore makes a narrower but stronger thesis. N'Ko is computational infrastructure for Manding language technology. It changes model visibility, acoustic decoding, metric interpretation, and the design of downstream correction. The 20.57% result is important because it anchors that thesis in a large ASR artifact, but the number must be presented with its provenance and caveats. The value of the work is not only that a checkpoint reported approximately 20% CER; it is that the surrounding evidence explains why that number is more meaningful than

Table 1: Core contributions of the canonical manuscript.

Contribution	Description
Script as infrastructure	Frames N'Ko as a computational substrate that affects tokenization, model internals, ASR labels, metrics, and correction policy.
Script-invisibility evidence	Consolidates LLM diagnostics showing that N'Ko can be accepted as Unicode while remaining weakly represented in hidden geometry.
Metric proposition	Formalizes why CER over a normalized bijective N'Ko script map is more phonemically interpretable than Latin WER for Manding ASR.
Anticipatory ASR architecture	Defines the trajectory-conditioned Transformer CTC decoder that injects seven-dimensional speech-dynamic geometry into attention.
Archived ASR anchor	Preserves the 20.57% N'Ko trajectory checkpoint as the canonical artifact-backed ASR anchor, with explicit caveats around the incomplete May 2026 audit.
AGP correction governance	Specifies the Anticipation Geometry Partition as a row-level admissibility framework for stable, boundary, uncertain, and novelty spans.

Latin WER for this setting and why trajectory geometry belongs naturally in a script-native ASR pipeline.

Architecturally, the central ASR object in the paper is an anticipatory Transformer CTC decoder. I use *anticipatory* in a specific technical sense: the decoder does not attend only to acoustic content embeddings, but also to a learned dynamic state describing where the utterance appears to be moving. If $E(x) = h_{1:T}$ denotes frozen Whisper encoder features, the decoder first projects and downsamples them into $u_{1:T'}$. A trajectory module then estimates $z_t \in [0,1]^7$ from local acoustic context, and each Transformer attention head receives an additive bias before CTC emission:

$$\alpha_{ij}^{(m)} = \mathrm{softmax}_j \left(\frac{Q_i^{(m)} K_j^{(m)\top}}{\sqrt{d_h}} + B_{ij}^{(m)}(z_i, z_j) \right).$$

The model is therefore anticipatory not because it predicts arbitrary future text, but because it conditions attention on commitment, uncertainty, transition pressure, recovery margin, phase stiffness, novelty, and stability. These variables describe whether the acoustic evidence is settled, crossing a boundary, recovering from instability, or entering a novel region. AGP later reuses the same geometric logic as a correction policy rather than as an acoustic decoder.

2 Initial Hypothesis Stack

Before formalizing the research questions, it is useful to reconstruct the hypothesis stack that motivated the project. This prevents the manuscript from appearing to have been organized around one surviving number after the fact. The original program was broader: it asked whether N'Ko changes model internals, decoder alignment, evaluation metrics, deployment behavior, personalization, and knowledge provenance. The canonical paper keeps the supported and relevant parts of that program, while explicitly labeling the hypotheses that remain incomplete or outside the present evidence base.

Table 2: Initial hypotheses reconstructed from the project roadmap, handoff notes, and paper outlines. The later research-question section formalizes the subset that this canonical manuscript can support rigorously.

ID	Initial hypothesis	Intended test or observable	Current disposition
IH1	N'Ko is not merely low-resource; it is structurally invisible inside LLMs trained on corpora where the script is absent or nearly absent.	Layerwise activation norms, entropy, sparsity, kurtosis, tokenizer allocation, and translation tax on parallel English/N'Ko inputs.	Retained and supported by the LLM diagnostic papers.
IH2	Script invisibility is not model-specific. The same failure signature should appear across architecturally distinct model families.	Repeat the activation scan on Qwen-family and Mistral-family models and compare the geometry of failure.	Retained and supported within the tested model families.
IH3	Direct script-native N'Ko ASR is technically feasible using frozen Whisper features and a trainable CTC decoder.	Build audio-to-N'Ko CTC systems and track CER across BiLSTM, Transformer, LoRA/Whisper, and trajectory variants.	Retained; the archived 20.57% checkpoint is the strongest ASR anchor.
IH4	A bijective or near-bijective script gives CTC a cleaner target than Latin Bambara because target symbols correspond more directly to acoustic-phonemic events.	Matched N'Ko and Latin CTC runs with identical architecture, data, optimizer, split, and scoring protocol.	Retained as a metric/architecture hypothesis; the fully matched superiority proof remains unresolved.
IH5	Trajectory bias acts as a bijection amplifier: dynamic speech-state information should help most when the output symbols preserve phonemic structure.	Compare baseline, graph, trajectory, and combined decoders in N'Ko and Latin under matched conditions.	Historically supported; full artifact bundle incomplete.
IH6	N'Ko should generalize better to unseen vocabulary because novel words compose from familiar phoneme-character units.	Train on SEEN-only utterances and evaluate on utterances containing UNSEEN words.	Retained as historical generalization evidence.
IH7	Vocabulary expansion should be more operationally useful in transparent-script systems because adding words or graph entries preserves phonemic structure without retraining the acoustic model.	Compare SEEN-only and full-data or graph-expanded conditions on unseen-word utterances.	Partially retained; evidence is useful but not the main anchor.
IH8	Speaker adaptation or test-time training should be easier when the decoder's labels are phonemically aligned and trajectory state captures speaker-specific articulation patterns.	Diarize Djoko speakers, update decoder layers sequentially, and estimate per-speaker CER or loss improvement slopes.	Planned/pending; not claimed as a result.
IH9	Out-of-domain deployment requires row-level provenance, uncertainty, and correction governance rather than scalar CER alone.	Build prediction/reference rows with trajectory summaries, partitions, provenance fields, and correction decisions.	Retained as AGP architecture and artifact protocol.

ID	Initial hypothesis	Intended test or observable	Current disposition
IH10	Training a personalized model on indigenous-script representations may change behavior relative to colonial-language or English representations of the same speaker data.	Compare LoRA adapters trained on English versus N'Ko-encoded speaker data using hidden-state and behavior metrics.	Out of scope for the canonical ASR paper; preserved as future personalization work.
IH11	Compressed N'Ko-based knowledge representations can preserve provenance through auditable transformation chains.	Track conversation-to-curation-to-N'Ko translation-to-sigil-to-inscription transformations and measure compression, retention, and verification cost.	Out of scope for the canonical ASR paper; preserved as future provenance work.

This table also explains why some claims are deliberately downgraded later in the paper. The initial program contained aggressive empirical predictions: N'Ko should beat Latin under matched CTC conditions, trajectory should help N'Ko more than Latin, and adaptation should improve with speaker exposure. Some of those ideas have supporting evidence, but not all of them have current artifact-complete support. The canonical manuscript therefore treats the initial hypotheses as the research program's generative map, then separates them from the narrower set of claims that can be defended today.

3 Related Work and Technical Frame

Unsegmented sequence learning. The ASR systems in this project use CTC, which was introduced to train recurrent networks directly on unsegmented sequence-labeling problems such as speech and handwriting [3]. CTC is relevant here because Manding speech annotation is not frame-aligned; the model must learn a monotonic relation between acoustic frames and character sequences. This makes the output alphabet more than a formatting choice. The alphabet defines the units that CTC learns to align.

Whisper features and parameter-efficient adaptation. The speech papers build on frozen Whisper large-v3 encoder features. Whisper was trained at web scale for multilingual speech recognition and translation [8], but the project does not assume that Whisper already solves Bambara-to-N'Ko transcription. Instead, Whisper provides acoustic features while the script-native decoder learns the target writing system. The broader project also used LoRA-style parameter-efficient adaptation [4] in the language and speech stacks, but the canonical 20.57% ASR anchor is not a LoRA-Whisper result. It is a trajectory CTC decoder result recorded under its own artifact path.

Model-family diagnostics. The LLM papers use Qwen and Mistral-family models as diagnostic instruments for script visibility. Qwen2/2.5 technical reports and Mistral 7B provide the external model context [11, 10, 5]. The internal research question is not whether those models are strong in general. It is whether they carry functional internal structure for N'Ko once Unicode acceptance, tokenization, activation geometry, and downstream behavior are measured separately.

The metric problem. WER is not a direct measure of speech understanding. It is a word-level Levenshtein metric computed after tokenization, normalization, and reference selection. It becomes

scientifically interpretable when four assumptions roughly hold: word boundaries are stable, spelling conventions are stable, the written word preserves the acoustic distinctions being tested, and an edit at the word level is a reasonable proxy for a recognition error. Those assumptions are often acceptable for high-resource ASR benchmarks with standardized orthography. They are much weaker for Latin-script Bambara and related Manding ASR.

Latin Bambara weakens the metric in several ways. Tone is often absent even though tone can distinguish lexical meaning. Digraphs can encode a single sound with multiple characters, so a one-phoneme error can become a multi-character spelling error. Conversely, two acoustically different forms can collapse to the same Latin string when tone or vowel quality is not represented. Word segmentation and spelling variation then add a second layer of uncertainty: a model can be penalized for choosing a different written convention rather than for failing to recognize the speech signal. In this setting, Latin WER mixes acoustic error, orthographic convention error, tokenization error, and reference-author preference.

N'Ko CER addresses the same utterance at a different measurement level. Because N'Ko explicitly represents vowels, many consonantal contrasts, tone classes, and nasalization marks, a normalized N'Ko character edit is closer to a phonemic edit than a Latin word edit. This is the metric advantage. It should not be overstated: N'Ko CER is not automatically phoneme error rate. A scorer must declare whether it counts Unicode code points, grapheme clusters, base letters, combining tone marks, spacing, punctuation, digits, and normalization variants. For example, a base letter plus tone mark can be counted as multiple Unicode scalars unless the scorer defines a phonemic or grapheme-cluster unit. The paper's claim is therefore conditional: under a declared normalization policy, N'Ko CER is a more interpretable proxy for Manding phonemic accuracy than Latin WER.

This distinction changes how results should be reported. A public ASR result should not only report "CER" or "WER"; it should report the unit of scoring, the normalizer, the treatment of tone and combining marks, and the reference inventory. Without those details, two systems may appear comparable while measuring different objects. The 20.57% anchor is valuable because it is a native-script result, but future work should still strengthen the scorer by publishing a normalized phonemic-unit evaluation alongside raw character edits.

Orthographic transparency and ASR labels. The argument extends the orthographic-depth tradition, which treats writing systems as differing in how directly spelling maps to sound [6]. In ASR, orthographic depth becomes a label-design problem. A decoder can emit graphemes, phonemes, subword units, bytes, word pieces, or hybrids; systematic comparisons show that the choice of label unit remains an empirical and architectural variable rather than a mere notation change [12]. For high-resource languages, a phoneme decoder can rely on pronunciation lexica or trained grapheme-to-phoneme systems. For low-resource Manding ASR, those tools are not guaranteed to exist at the quality needed for benchmark evaluation. A transparent script partly solves the label problem by making the ordinary written target closer to the phonemic target.

The relevant quantity is mapping ambiguity. Let Φ be a phoneme inventory and Σ_s the label alphabet for script or label system s. If $P(\phi \mid \sigma, s)$ is the distribution over possible phonemic analyses for a label σ, then a rough ambiguity measure is

$$\mathcal{A}(s) = \mathbb{E}_{\sigma \in \Sigma_s} \left[H(\Phi \mid \sigma, s) \right].$$

A transparent script is valuable because it lowers this conditional ambiguity: the label observed by the decoder carries more information about the phonemic event that generated it. A deep or inconsistent orthography raises the ambiguity: the same label may correspond to multiple acoustic-phonemic events, or the same event may be written multiple ways.

For CTC, this is not only a linguistic preference. CTC learns a monotonic alignment between acoustic frames and target labels. If /ny/ is represented as a single script-level target, the alignment has one label event. If it is represented as `n+y`, the alignment has two character events for one phonemic event. If tone is unmarked, the target omits an acoustic distinction entirely. If a subword token crosses a phoneme or syllable boundary, the label may become even less localized in time. These effects do not make Latin unusable, but they make Latin labels a less direct measurement instrument for the phonemic ASR question.

Standard speech scoring toolkits such as NIST SCTK make WER/CER-style scoring operational [7], but a scoring toolkit cannot make the reference units linguistically valid by itself. Low-resource African ASR work on radio archives demonstrates the practical need for speech technology in this setting [2]; the present project asks a narrower measurement question: which script makes the error signal scientifically interpretable for Manding speech?

4 The Script Advantage: N'Ko, Latin, and Phonemic Mapping

The script advantage claim requires linguistic precision. N'Ko is not simply a font for Bambara, nor is it identical to one spoken language. It is a right-to-left alphabetic script and written standard used across Manding varieties, including Bambara/Bamanankan, Maninka, Jula/Dyula, Mandinka, and related written practices [9, 1]. Spoken Manding varieties remain diverse; the computational claim is that N'Ko supplies a more explicit written interface to the phonology that those varieties share than Latin Bambara orthography usually does. The word "advantage" therefore means an advantage in representation and measurement, not a claim that a script alone solves acoustic modeling.

At the Unicode layer, N'Ko occupies U+07C0–U+07FF. That range contains digits, alphabetic letters, tone marks, nasalization marks, punctuation, and related symbols. The core vowel inventory is represented by named vowel letters, including U+07CA NKO LETTER A, U+07CB NKO LETTER EE, U+07CC NKO LETTER I, U+07CD NKO LETTER E, U+07CE NKO LETTER U, U+07CF NKO LETTER OO, and U+07D0 NKO LETTER O. The consonant inventory contains letters such as U+07D3 NKO LETTER BA, U+07D5 NKO LETTER TA, U+07DE NKO LETTER KA, U+07E1 NKO LETTER MA, U+07E2 NKO LETTER NYA, U+07DC NKO LETTER GBA, and U+07DA NKO LETTER RRA. For ASR, the important property is that many speech units that are multi-character or inconsistently represented in Latin orthography have direct script-level targets in N'Ko.

This is the practical source of the bijective argument. In the idealized normalized N'Ko setting, the decoder's output alphabet is close to the inventory of acoustic contrasts: a vowel is a vowel letter, a palatal nasal can be one script unit, a labial-velar stop can be one script unit, and tone/nasalization can be explicit combining information. Latin Bambara can still be useful for reading, teaching, search, and interoperability, but it is not the same measurement object. It often encodes one sound with multiple characters, leaves tone underrepresented, and inherits spelling variation from colonial and corpus-specific conventions.

The difference matters for CTC. CTC learns monotonic alignments between time frames and output symbols. If one phonemic event maps to one normalized N'Ko target, then the alignment problem and the error metric are easier to interpret. If the same event maps to a Latin digraph or to an unmarked tonal word form, then the model can be penalized for a spelling boundary rather than for an acoustic boundary. This is why the paper compares N'Ko CER to Latin WER as measurement systems, not merely as two display choices for the same transcript.

The script advantage is therefore conditional. It depends on a clean N'Ko target column, a

Table 3: Representative script mappings relevant to Manding ASR. The table is not a complete phonology; it illustrates why N'Ko CER is closer to a phonemic edit metric than Latin WER.

Speech unit	N'Ko representation pattern	Latin evaluation issue
Oral vowels	Dedicated vowel letters such as U+07CA A, U+07CC I, and U+07CE U.	Vowel quality is represented, but Latin spelling conventions may vary by corpus.
Palatal nasal /ny/	Dedicated NYA letter class, e.g. U+07E2 NKO LETTER NYA.	Latin often uses a digraph such as `ny`; one phoneme becomes two characters.
Labial-velar /gb/	Dedicated GBA letter class, e.g. U+07DC NKO LETTER GBA.	Latin `gb` is a two-character sequence whose character edits need not match phoneme edits.
Rhotic contrast	Dedicated RA/RRA distinction in the Unicode block.	Latin orthography may collapse or vary contrasts depending on convention.
Tone	Combining tone marks U+07EB–U+07F1 encode short/long and contour tone classes.	Latin Bambara often leaves tone unmarked, so word scoring can ignore acoustic distinctions that change lexical meaning.
Nasalization	U+07F2 NKO COMBINING NASALIZATION MARK.	Latin may use different nasal spellings or leave the relation to phonology implicit.

declared Unicode normalization policy, explicit treatment of tone and combining marks, and an evaluation script that counts the intended units. If those conditions are violated, N'Ko CER can lose its interpretability. This is why the data-quality section treats script contamination and normalization drift as central scientific risks rather than as minor preprocessing details.

5 Research Questions, Hypotheses, and Claim Boundaries

The initial hypothesis stack above reconstructs the broad research program. This section narrows that stack into the claims this canonical manuscript can evaluate with defensible evidence. It is a systems-and-measurement study rather than a single leaderboard paper. Its research questions therefore distinguish four objects that are often collapsed in low-resource ASR papers: model-internal script representation, metric validity, acoustic-decoder architecture, and post-ASR correction governance. The questions are stated at the level of measurable constructs so that a future audit can replicate, falsify, or narrow each claim independently.

Table 4: Research questions, constructs, and primary observables.

ID	Construct	Research question	Primary observable
RQ1	Script visibility in LLMs	Is N'Ko merely low-resource, or is it structurally underrepresented inside current model families?	Tokenizer coverage, activation norms, entropy, sparsity, kurtosis, translation tax, and downstream behavior.
RQ2	Metric validity	Does normalized N'Ko CER preserve more phonemic information than Latin WER for Manding ASR?	The transparent-script edit-preservation proposition and explicit normalization policy.
RQ3	Script-native ASR architecture	Can a trajectory-conditioned Transformer CTC decoder reach the 20% CER regime on the 290,596-pair snapshot?	Archived 20.57% checkpoint metadata, split counts, hashes, and CER numerator / denominator.
RQ4	Correction governance	Can anticipation geometry constrain post-ASR correction so unsupported rewrites are rejected or deferred?	Row-level AGP contract, accepted/rejected edit accounting, and per-partition outcomes.

Table 5: Hypotheses with null or falsifying cases and current evidentiary status.

ID	Alternative hypothesis	Null / falsifying case	Current status
H1	N'Ko is internally underrepresented in tested LLM families beyond ordinary low-resource performance variation.	Tokenizer and activation diagnostics show no consistent N'Ko deficit relative to supported scripts, or deficits appear in only one isolated metric without convergence.	Supported by converging diagnostics in the written activation papers; scope limited to Qwen/Mistral-family tests and their protocols.
H2	Normalized N'Ko CER is a more phonemically interpretable Manding ASR metric than Latin WER.	The target N'Ko normalization map is not bijective over the claimed inventory, or Latin word scoring preserves the same phonemic edit structure under the same assumptions.	Supported as a formal measurement proposition, not as an empirical superiority claim.

ID	Alternative hypothesis	Null / falsifying case	Current status
H3	A trajectory-only N'Ko Transformer CTC decoder can reach the approximate 20% CER regime on the 290,596-pair snapshot under recorded settings.	The archived artifact is missing required metadata, has inconsistent hashes or row counts, or a strict reproduction under identical settings repeatedly fails to enter the same error regime.	Supported by the preserved 20.57% archived checkpoint; the strict May 2026 reproduction did not complete and remains unresolved.
H4	Trajectory geometry is useful as a dynamic ASR state and as a correction-risk signal.	Trajectory channels do not improve or explain any archived ASR behavior, and AGP gates accept unsupported regressions at unacceptable rates in row-level tests.	Partially supported: historical trajectory evidence and architecture are defined; AGP has smoke tests but lacks a full benchmark over the anchor test set.

The primary endpoint for the ASR anchor is character error rate on the 29,060-row test split, reported with both numerator and denominator. Secondary endpoints are validation loss, row-level prediction/reference completeness, split integrity, vocabulary integrity, and artifact hashes. For the LLM studies, the endpoint is not a single behavioral score but convergence across representation diagnostics. For AGP, the endpoint is not merely a lower final CER; a valid correction layer must also report accepted regressions, rejected improvements, abstentions, and per-partition behavior.

The evidence standard differs by question. The LLM visibility claim is supported by converging diagnostics across tokenization, activation norms, entropy, sparsity, kurtosis, and translation behavior. The metric claim is supported by the formal transparent-script proposition and is conditional on normalization. The ASR anchor claim is supported by preserved artifact metadata and checkpoint provenance, but the strict May 2026 audit did not complete and therefore cannot be cited as a new reproduction. The AGP claim is architectural and experimental-preparatory: AGP has smoke-test evidence and a row-level contract, but it is not yet a full benchmark result over the 20.57% test set.

These boundaries are part of the result. The project contains artifact-backed evidence, historical evidence, smoke tests, and incomplete audits. Treating them as one undifferentiated evidence pool would make the paper weaker. Separating them makes the contribution auditable.

6 Operational Definitions

Let x denote an input utterance, y a reference transcription, and $\hat{y}$ a model hypothesis. Character error rate is defined as

$$\mathrm{CER}(\hat{y}, y) = \frac{S(\hat{y}, y) + D(\hat{y}, y) + I(\hat{y}, y)}{|y|},$$

where S, D, and I are Levenshtein substitutions, deletions, and insertions after the project normalization function is applied. The interpretability claim is not that this formula changes for N'Ko. The claim is that the target alphabet makes $|y|$ and the edit operations closer to phonemic units than Latin Bambara word-level evaluation does.

This can be stated as a measurement proposition.

Proposition 1 (Transparent-script edit preservation) *Let Φ be a normalized phoneme inventory and let $f_N : \Phi \to \Sigma_N$ be a bijective N'Ko transcription map over that inventory after the paper's explicit normalization choices. For any two phoneme strings $\phi_{1:U}$ and $\psi_{1:V}$, edit distance over the N'Ko strings $f_N(\phi_{1:U})$ and $f_N(\psi_{1:V})$ preserves the phoneme-level edit structure: each insertion, deletion, or substitution in Φ corresponds to exactly one insertion, deletion, or substitution in Σ_N, and vice versa. Therefore CER over normalized N'Ko strings is an interpretable proxy for phoneme-level edit rate, modulo explicitly declared treatment of tone, combining marks, punctuation, and spacing.*

Proof sketch. Because f_N is bijective, it has an inverse f_N^{-1} on Σ_N. Any edit script between $\phi_{1:U}$ and $\psi_{1:V}$ maps through f_N to an edit script of the same length between their N'Ko renderings. Conversely, any edit script between the N'Ko renderings maps through f_N^{-1} to an edit script of the same length between the phoneme strings. The minimum edit length is therefore preserved. The equivalence holds only over the normalized inventory named in the proposition; if a normalizer deletes tone, collapses combining marks, or changes spacing, the metric must report that policy.

A Latin transcription relation $f_L : \Phi^* \to \Sigma_L^*$ with variable-length digraphs, optional tone marking, and spelling variation does not satisfy the bijective condition. One phoneme may require multiple Latin characters, one Latin character sequence may correspond to different phonemic analyses, and word boundaries introduce segmentation choices that are not phoneme boundaries. Latin WER therefore measures a mixture of recognition, spelling, word segmentation, and reference-convention agreement. This proposition is a metric-validity claim, not a guarantee that a particular N'Ko model must always obtain lower CER than a Latin model.

For LLM diagnostics, the project uses a representation-tax family of measurements. If $a_\ell(s)$ is a layer-ℓ activation statistic for script condition s, a generic tax ratio can be written as

$$\tau_\ell(s_1, s_2) = \frac{a_\ell(s_1) + \epsilon}{a_\ell(s_2) + \epsilon},$$

with ϵ included only to avoid division instability. The underlying papers instantiate a_ℓ with norm, entropy, sparsity, and kurtosis-derived statistics. The canonical claim concerns convergence across diagnostics, not any single ratio in isolation.

For ASR, the trajectory decoder augments acoustic hidden states with a dynamic state vector $z_t \in [0, 1]^7$. For AGP, the admissibility function can be written as

$$A(r, c, z, p) \in \{\text{accept}, \text{reject}, \text{abstain}\},$$

where r is the raw ASR row, c is a candidate correction, z is the trajectory state sequence, and p is the partition label. A correction layer is successful only if accepted edits improve or preserve transcript quality while the gate rejects or abstains on unsupported rewrites. This is why AGP evaluation must report accepted, rejected, improved, neutral, and worsened edits separately; scalar CER alone is not enough.

7 Methods and Artifact Protocol

The canonical manuscript uses three classes of evidence. The first class is published-in-repository paper evidence: the five LaTeX manuscripts in `paper/current/`, their figures, and their local result dependencies. The second class is artifact evidence: checkpoint files, `results.json`, split

Table 6: Evidence levels used throughout the canonical paper. This table is the guardrail that keeps the paper from mixing archived benchmarks, incomplete audits, and hypothesis-generating experiments into one unsupported headline.

Level	Definition	Public use
Artifact anchor	A retained checkpoint or result file with split metadata, hashes, hyperparameters, and a recorded evaluation.	May support a numerical claim when the claim names the artifact status.
Artifact-complete run	A run with predictions, references, metrics, parameters, and row counts, but not necessarily the same hyperparameter regime as the anchor.	May support engineering conclusions and negative results, not anchor reproduction.
Historical experiment	A prior result preserved in papers, notes, or figures whose full local artifact chain is incomplete.	May motivate hypotheses and explain the research trajectory, but should not be the main benchmark.
Smoke test	A small hand, synthetic, or slice-based check of a mechanism.	May support feasibility or failure-mode analysis only.
Planned audit	A configured run that did not complete or did not produce final artifacts.	May be cited only as transparency about unresolved validation.

metadata, vocabularies, row-level prediction/reference exports, and SHA-256 hashes. The third class is project-history evidence: handoff notes, publication-readiness notes, and session-derived caveats about incomplete audits or missing artifact chains. Only the first two classes should support numerical claims. The third class is useful for explaining why claims are bounded.

The ASR training stack for the anchor uses frozen Whisper large-v3 encoder features of dimension 1280, a learned projection to a 768-dimensional decoder space, stride-4 temporal downsampling, sinusoidal position information, and six Transformer CTC decoder layers. The trajectory branch computes seven dynamic scalars per timestep and maps them into attention bias. The N'Ko vocabulary contains 66 output classes in the project implementation; the Latin comparison vocabulary contains 41 classes. The anchor model is recorded as a 46.8M-parameter trajectory CTC decoder.

The retained anchor is evaluated on 29,060 test examples from the 290,596-pair snapshot. The handoff and result metadata record 216,225 character edits over 1,050,967 reference characters, yielding 20.57% CER. The artifact protocol for future work is stricter than the old papers: any public benchmark row should include the run name, script, mode, data snapshot, split hash, vocabulary hash, model checkpoint hash, learning rate, batch size, seed, prediction row count, reference row count, and CER numerator/denominator. This is the standard needed to prevent another training-parameter mismatch from being mistaken for a scientific contradiction.

The same discipline applies to the phrase "canonical." A canonical result is not the most exciting number in the project history; it is the number that can be stated without forcing the reader to trust conversational memory. For this manuscript, that means the 20.57% checkpoint is canonical as an archived anchor, while the low-learning-rate matrix is canonical as a non-comparable artifact-complete engineering bundle. The historical eight-way comparison is important but secondary, because its role is to explain why the trajectory hypothesis became worth testing.

The scoring protocol is intentionally simple. References and hypotheses are normalized by the

project text normalizer, scored with Levenshtein edits, and reported as edits divided by reference characters. For a paper-ready ASR run, the same normalization function must be recorded alongside the result. Without that, even a correct CER number is not a complete scientific object: a different normalizer can change the denominator, remove or retain combining marks, collapse spaces, or hide script contamination. This is especially important for N'Ko because the value of the metric depends on preserving the script's phonemic information, not merely rendering a string in Unicode.

The May 2026 audit changed the publication protocol even though it failed to produce a replacement benchmark. It showed that paid training should never be launched before three checks pass locally and remotely: first, the pair file and feature cache must have expected hashes and counts; second, the trainer must be compiled and pinned by hash; third, a sanity run must emit the same artifact contract expected from the full run. The audit also showed why a result table should include the learning rate in the table itself. The later 31% runs were not expected to reproduce the 20.57% anchor because they used a different optimization regime.

8 Data Quality and Normalization Caveats

The project should treat data cleaning as a research result, not as a hidden preprocessing footnote. The April project notes record a label-contamination discovery in the N'Ko column: Latin and IPA-like characters appeared inside rows that were supposed to be script-native N'Ko labels. Those notes attribute a large CER change to the cleaning function, but the exact contamination count and improvement should not be promoted as a public benchmark until the script, manifest, and before and after hashes are packaged. The publishable claim is narrower: low-resource ASR datasets can contain script contamination severe enough to change model conclusions, so every run in this project must report the label-cleaning function used.

The strict audit also exposed an implementation-level data issue before training could complete. Some Whisper feature tensors were two-dimensional time-by-feature matrices, while others carried an extra leading singleton dimension. The loader had to normalize those tensors before the sanity gate would pass. That failure is not a scientific result, but it is a reproducibility lesson: for a corpus this large, the artifact contract must validate tensor shape, feature count, pair-file hash, split size, vocabulary hash, and prediction/reference row counts before paid training is trusted.

There are three distinct data risks in the project. The first is script contamination: a row claims to be N'Ko but contains Latin, IPA-like, or otherwise non-target characters. The second is normalization drift: two runs use different rules for spaces, punctuation, Unicode combining marks, or target filtering. The third is feature-label mismatch: the acoustic tensor exists but no longer matches the row identity, duration, or expected tensor shape. These risks are operationally different. Script contamination corrupts the target alphabet; normalization drift changes the metric; feature-label mismatch corrupts the input-output pairing.

This is why the paper should not hide the failed audit. The failed audit is part of the scientific record because it explains why the project moved from "run more benchmarks" to "freeze a canonical artifact and document the evidence ladder." That move is not a retreat from rigor. It is the condition for rigor: the paper becomes more publishable when it tells readers exactly which results are artifacts, which are historical, and which are still planned.

9 Script Invisibility in LLMs

The first layer of the project is diagnostic. A model can accept N'Ko characters without possessing circuits that make those characters computationally useful. The written LLM papers therefore

Table 7: Data-quality risks and the corresponding publishable mitigation.

Risk	Scientific effect	Required mitigation
Script contamination	The decoder is trained to emit labels outside the intended script, weakening the interpretation of N'Ko CER.	Report target-character audit, cleaning function, and before/after hashes.
Normalization drift	CER changes because the reference denominator or edit alphabet changed, not because recognition improved.	Pin and publish the exact normalizer for every benchmark.
Feature-shape drift	Training or evaluation crashes, or silently batches malformed tensors.	Validate tensor rank, feature dimension, feature count, and matched pair IDs before training.
Split drift	Reported train/test boundaries change across runs.	Publish split JSON, split hash, and row counts.

avoid evaluating only final translation quality. They measure what happens inside the network.

Translation tax. The activation papers define a translation or representation tax as the excess internal effort required to process N'Ko relative to a better-supported script. In one protocol, Qwen3-8B exhibited an average tax of 2.94x. In the cross-model protocol, the corresponding tax estimates were 3.30x for Qwen3-8B, 3.59x for Qwen2.5-7B, and 2.67x for Mistral-7B. These values should not be merged as if they came from one identical measurement pipeline. Their importance is directional: each protocol finds that N'Ko input induces a substantially less efficient internal trajectory than well-supported text.

Hidden-state geometry. The brain-scan work measures layerwise activation norms, entropy, sparsity, and kurtosis. The Paper 1 diagnostics reported a 1.2–1.7 bit entropy gap, approximately 2.2x higher embedding sparsity for N'Ko, and a 78.1% output-layer kurtosis deficit. The cross-model paper reported the same qualitative pattern across the tested families, with N'Ko activations roughly 66–72% weaker and output kurtosis deficits spanning 64.6–93.5% under its protocol. Again, the central result is not a single pooled statistic. It is the repeated geometry: N'Ko enters the model as Unicode but does not propagate as a well-supported internal representation.

Tokenizer evidence. The tokenizer evidence explains one mechanism. Arabic, another right-to-left script, has thousands of vocabulary entries in major multilingual tokenizers, while N'Ko is often represented by fallback behavior or tiny coverage. This isolates the issue from script direction alone. The problem is data starvation and representation absence, not merely right-to-left rendering.

Why this matters for ASR. The LLM findings motivate the speech work. If generic language models have weak internal structure for N'Ko, then a low-resource ASR system cannot rely on a generic multilingual language model to supply script competence after decoding. It needs script-native targets, script-native normalization, and metrics that preserve the acoustic distinctions the script was designed to encode.

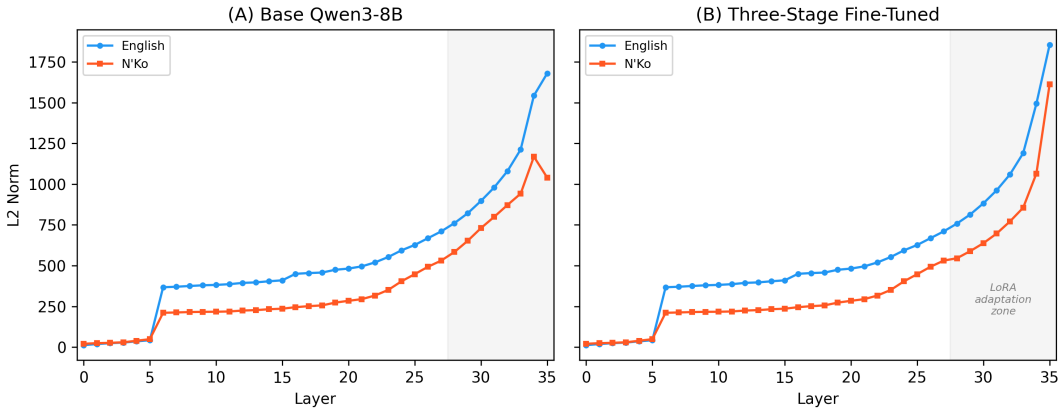

Figure 1: Layerwise activation-norm comparison from the brain-scan experiments. The figure is best read with the entropy, sparsity, kurtosis, and tokenizer analyses in the underlying papers. Its role here is to illustrate the structural nature of script invisibility rather than to serve as the only evidence.

Table 8: Failure zones exposed by the LLM papers and their ASR design response.

Zone	Observed failure	ASR response
Tokenizer	Tiny or fallback coverage for N'Ko compared with supported scripts.	Use a script-native CTC vocabulary rather than a generic multilingual tokenizer.
Activation	Lower norms, higher entropy, more sparsity, and weaker output kurtosis.	Decode from speech features with a task-specific decoder instead of assuming pretrained text circuits are sufficient.
Bridge	Latin or French intermediates can be fluent while losing tone, digraph-level phonology, or N'Ko distinctions.	Score native N'Ko rows and preserve CER numerator/denominator at the script level.

The diagnosis can be summarized as a three-zone failure model. In the first zone, the model fails before semantics because the tokenizer fragments or backs off on the script. In the second zone, the model passes Unicode strings forward but lacks high-energy internal directions for them, producing weak activation norms, high-entropy representations, and sparse or brittle hidden states. In the third zone, the model can produce fluent output in a dominant bridge language, but the bridge destroys the script-specific distinctions that matter for the target language. The speech work is designed to bypass all three zones: it does not ask a general LLM to transliterate or repair Bambara after the fact; it trains a decoder to emit N'Ko directly from acoustic evidence.

The LoRA remediation experiments in the representation papers should be read in this frame. They do not prove that a small adapter solves N'Ko for every downstream task. They show that the deficit is partly trainable: when the model receives script-aware adaptation, some activation and output behavior moves in the right direction. That makes the invisibility result more actionable. It is not an argument that N'Ko is intrinsically impossible for large models; it is an argument that current general-purpose pretraining leaves measurable structural holes.

Table 9: Condensed progression of the script-native ASR stack. The values summarize the written paper artifacts and are not all from the same evaluation regime.

System	Role in the project	Recorded result
V1 BiLSTM CTC	Early proof that direct audio-to-N'Ko transcription was feasible.	56% CER, 91.5% WER.
V3 Transformer CTC	Frozen Whisper features plus a larger CTC decoder.	Approximately 33% validation CER.
V4 Whisper LoRA	Adaptation and confidence experiment, not the canonical anchor.	29.4% CER on a 50-sample held-out check; WER evidence marginal.
Paper 4 trajectory CTC	Canonical archived ASR anchor.	20.57% test CER under recorded settings.

10 Script-Native Speech Recognition

The speech papers record an engineering progression. The early BiLSTM CTC system was small and functional but weak, with approximately 56% CER and 91.5% WER. The V3 Transformer CTC system used frozen Whisper features, a 46.9M-parameter decoder, six Transformer layers, hidden size 768, and stride-4 temporal downsampling; it reached approximately 33% validation CER. The V4 LoRA-Whisper line improved confidence and loss behavior, with validation loss moving from 0.884 to 0.290 and a 29.4% CER on a 50-sample held-out evaluation, but the WER improvement was not statistically strong in the recorded paired comparison. This sequence matters because it prevents a false interpretation: N'Ko ASR did not appear fully formed. It emerged through architecture, normalization, data-cleaning, and metric work.

The corpus associated with the canonical ASR anchor contains 290,596 paired examples split into 232,476 training examples, 29,060 validation examples, and 29,060 test examples. The target is N'Ko character transcription. The model uses frozen Whisper encoder features and a trainable Transformer CTC decoder whose trajectory variant modulates attention using learned dynamic state.

The bridge systems in the earlier speech papers are useful because they show what the canonical system deliberately avoids. A Latin bridge can appear attractive: there are more existing tools, more pretrained assumptions, and more familiar evaluation conventions. But bridge decoding introduces its own error classes. Latin Bambara can represent a single sound with a digraph; tone may be absent or inconsistently marked; conversion rules can miss phonemes such as /g/ or confuse symbols borrowed from IPA-like transcriptions; and right-to-left rendering problems can hide target-side defects in visual inspection. A bridge result therefore mixes speech recognition error, transliteration error, orthographic-convention error, and display error. Direct N'Ko CTC reduces the number of moving parts.

This does not mean every N'Ko number is automatically better than every Latin number. Latin can sometimes win under a smaller output vocabulary or a cleaner training condition, as the historical baseline table shows. The claim is that N'Ko makes the error easier to interpret. If a direct N'Ko decoder substitutes one character for another after normalization, the error is usually closer to a phonemic confusion than a Latin word-level error is. That interpretability is what makes the 20.57% anchor worth discussing even while the strict reproduction remains incomplete.

The implementation path also explains why the paper should not collapse V1, V3, V4, and

Paper 4 into one performance curve. V1 tested feasibility with a small CTC system. V3 tested whether frozen Whisper features could support a larger script-native decoder. V4 tested adaptation and confidence behavior. The Paper 4 trajectory model tested whether dynamic state could be injected into the decoder. Those systems answer different engineering questions. The canonical paper uses them as a development history, not as a single leaderboard.

11 The 20.57% Anchor

The result that can be discussed publicly is precise. The local Paper 4 reproduction cache preserves an archived N'Ko trajectory ASR checkpoint reporting 20.57% test CER. Its metadata records script N'Ko, mode trajectory, trajectory-only decoding, learning rate 0.0003, batch size 32, dropout 0.1, seed 42, best validation loss 0.6358872798606507, and 47 trained epochs.

Table 10: Canonical archived ASR anchor.

Field	Value
Artifact root	`paper4_reproduction_35205256`
Script and mode	N'Ko trajectory CTC
Training branch	Trajectory-only; no residual trajectory-attention branch and no test-time adaptation branch
Learning rate, batch size, dropout	0.0003, 32, 0.1
Seed	42
Train / validation / test	232,476 / 29,060 / 29,060
Best validation loss	0.6358872798606507
Epochs trained	47
Reported test CER	**20.57%**
Results SHA-256	`252aecd6e323f7d50cefd3c1e507ddaf035d9f0ac4f78d67766c4cf6ed5d24a7`
Vocabulary SHA-256	`e3ab620c9d2f971603d76f953f2be40bf9283dfd99d6428c7d51a9a73246ea67`
Best checkpoint SHA-256	`ab1fe47f96c2c434d8f301ae065b3292d592b9a4f5accf1d09acc97ca2c03b59`

The project therefore has a canonical artifact for the "20 CER" discussion: an archived checkpoint reporting 20.57% N'Ko CER. The later training that consumed May 2026 compute did not finish a strict reproduction that can replace that artifact. The anchor remains usable, but it must be described as an archived checkpoint result, not as a newly reproduced result.

The reason the later runs around 31% CER cannot be interpreted as anchor replications is also clear. They were not run under the same parameter regime. The preserved low-learning-rate bundle used 1×10^{-4} while the anchor used 3×10^{-4}, and the strict audit that was supposed to resolve the mismatch did not complete. That is a training-parameter caveat, not a disproof of the archived checkpoint.

The anchor's arithmetic should be printed whenever the result is discussed:

$$\frac{216{,}225}{1{,}050{,}967} = 0.20574\ldots \approx 20.57\%.$$

This is more transparent than reporting only a rounded percentage. The numerator states how many character edits the scorer counted; the denominator states how many reference characters

Table 11: Historical eight-way comparison from internal logs. These values explain the origin of the trajectory hypothesis, but they are not the canonical benchmark because the full artifact bundle is not locally restored.

Decoder condition	N'Ko CER	Latin CER	N'Ko–Latin delta
Baseline	32.75%	31.43%	+1.32pp
Graph structure	32.38%	37.14%	-4.76pp
Trajectory	27.50%	31.67%	-4.17pp
Combined graph+trajectory	30.46%	31.59%	-1.13pp

were exposed to the model. If a future audit changes either quantity, the paper can identify whether the difference comes from the model, the normalizer, the split, or the scorer.

The archived result also has a narrow architectural identity. It is not the later TAR branch, not the residual trajectory-attention variant, not the trajectory plus test-time-training branch, and not AGP. The result belongs to a trajectory-only CTC decoder using the recorded learning rate, batch size, dropout, seed, split, and vocabulary. In public writing, that specificity prevents two opposite mistakes: it prevents overstating the newer experiments as if they inherited the anchor, and it prevents dismissing the anchor because later experiments used different settings.

The unfinished strict audit should be described as a failed replacement attempt, not as a failed anchor. The audit was launched to answer the correct question: can the 20.57% checkpoint be regenerated under a clean, scripted, current-snapshot protocol with row-level artifacts? It did not produce the required final `results.json`, prediction file, reference file, and partition metrics. That means the old artifact remains the best retained evidence. It does not mean the old artifact should be erased; it means the paper must label it accurately.

12 Historical Evidence That Motivated the Trajectory Hypothesis

The project history contains a stronger comparative story than the canonical anchor alone, but it must be labeled correctly. Earlier internal 297K-scale runs compared N'Ko and Latin across baseline, graph, trajectory, and combined decoder conditions. Those results motivated the phrase "bijection amplifier": dynamic or structural decoder mechanisms seemed to help N'Ko more than Latin because N'Ko characters are closer to acoustic units. However, the complete local artifact chain for all eight historical runs is not currently restored, so these numbers should appear as hypothesis-generating evidence rather than as the primary benchmark.

The compositional-generalization results are also important because they test a different prediction: a phonemically transparent script should degrade more gently on unseen words, since novel words recombine known sound-symbol units. In the preserved project notes, an experiment trained on seen vocabulary and evaluated on utterances containing unseen words. N'Ko degraded from 32.75% to 33.24% CER, while Latin degraded from 31.43% to 34.38%. At the word-level breakdown, the N'Ko seen/unseen gap was 37.81pp and the Latin gap was 41.46pp, a 3.65pp smaller gap for N'Ko.

These historical experiments should not be discarded. They explain why the paper models trajectory geometry, compositionality, and metric validity rather than only reporting one checkpoint. Their limitation is provenance, not conceptual value. The manuscript therefore presents them as prior internal evidence that motivated the strict audit, while keeping the 20.57% checkpoint as the

Table 12: Historical compositional-generalization evidence. These values support the hypothesis that N'Ko composes unseen words from known phonemic units, but they remain separate from the 20.57% anchor.

Script	Full-test CER	Unseen-word experiment CER	Degradation
N'Ko	32.75%	33.24%	+0.49pp
Latin	31.43%	34.38%	+2.95pp

Table 13: Historical vocabulary-expansion evidence. The recovery from SEEN-only to full-data training is nearly identical, but the residual unseen-word gap remains smaller for N'Ko.

Condition	N'Ko CER	Latin CER	Difference
SEEN-only model on SEEN words	16.09%	15.05%	Latin better by 1.04pp
SEEN-only model on UNSEEN words	53.90%	56.51%	N'Ko better by 2.61pp
Full-data model on UNSEEN words	40.15%	42.73%	N'Ko better by 2.58pp
Recovery from full data	13.75pp	13.78pp	Nearly identical

strongest retained artifact-backed ASR result.

The vocabulary-expansion experiment adds a second robustness signal. In the SEEN-only control, Latin was slightly better on seen words: 15.05% CER versus 16.09% for N'Ko. On utterances containing unseen words, both scripts degraded sharply, but N'Ko degraded less: 53.90% versus 56.51%. After full-data training, both scripts recovered almost the same amount of the unseen-word gap, 13.75pp for N'Ko and 13.78pp for Latin. The remaining unseen-word CER was still lower for N'Ko: 40.15% versus 42.73%. This is the cautious version of the vocabulary claim. Full-data training helps both scripts, but N'Ko retains a 2.58pp advantage on rare-word utterances and a 3.62pp smaller residual gap in the recorded experiment.

The reason these tables belong in the canonical manuscript is not that they close the proof. They do not. They show that the research program has a coherent empirical shape. The anchor establishes that a script-native trajectory decoder can reach the 20% CER regime. The historical eight-way table explains why trajectory became the key architectural idea. The compositional and vocabulary tables explain why script choice matters beyond in-distribution test CER. Together they justify a future matched audit without pretending that the audit has already succeeded.

13 Trajectory Geometry

Trajectory conditioning is the architectural bridge between acoustic decoding and the larger AGP program. Speech unfolds through time; phonetic evidence does not arrive as isolated frames. A decoder benefits when it can distinguish stable regions, imminent boundaries, recovery zones, uncertainty, and novelty.

Let $h_t \in \mathbb{R}^d$ denote the acoustic hidden state at time t. The trajectory module derives a compact vector

$$z_t = \sigma(W_2 \phi(W_1 * h_{t-k:t+k})) \in [0,1]^7,$$

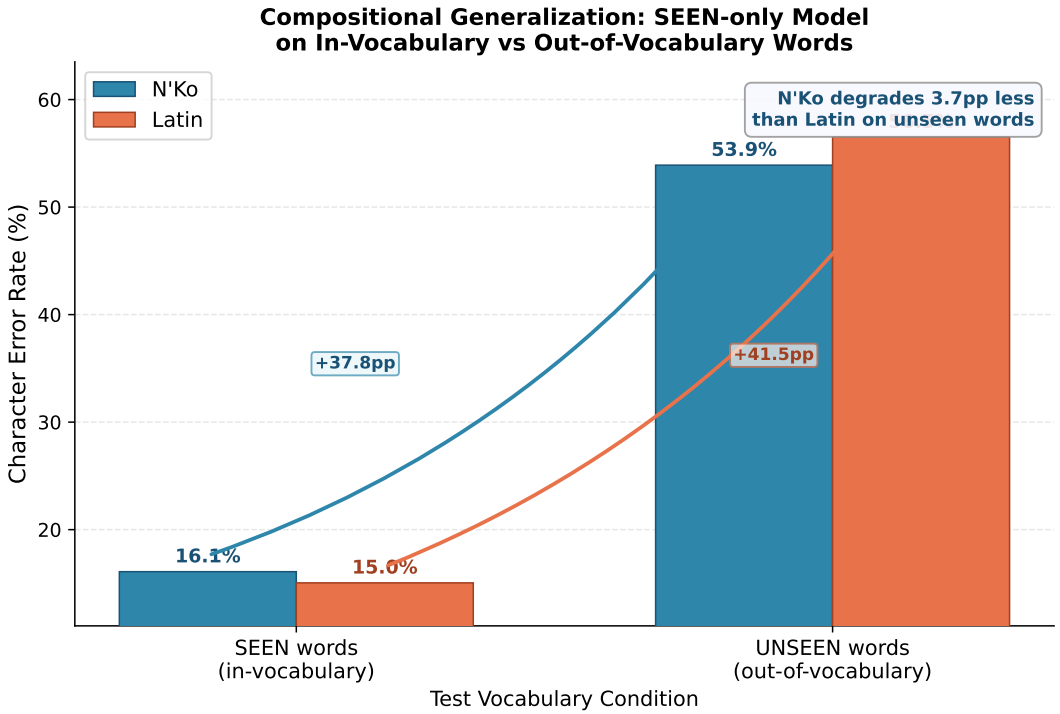

Figure 2: Historical Exp F compositional-generalization figure. The figure should be read as context for the trajectory and metric hypotheses, not as a replacement for the archived 20.57% anchor.

where the seven channels correspond to commitment, uncertainty, transition pressure, recovery margin, phase stiffness, novelty, and stability. These values are not intended to replace acoustic content. They describe the local dynamic state of the utterance.

The decoder maps this state into attention bias. For head m, a simple abstraction is

$$B_{ij}^{(m)} = g_m(z_i)\kappa_m(|i-j|),$$

where g_m converts trajectory state into a head-specific strength and κ_m controls relative distance. Attention then becomes

$$\text{softmax}\left(\frac{QK^\top}{\sqrt{d_h}} + B\right)V.$$

When the utterance is stable, attention can use broader context. Near boundaries or high-uncertainty regions, attention can become more local or more conservative.

This mechanism is especially plausible for N'Ko because the output script preserves more acoustic structure than Latin Bambara spelling. Trajectory geometry is a model of where the speech signal is moving; it is most useful when the target symbols retain the movement-relevant distinctions.

This table also resolves an ambiguity in the project history. Some earlier notes described the scalars through a handwriting or pen-stroke analogy: velocity, curvature, acceleration, and related motion terms. That analogy was useful for intuition, but it is not the publishable explanation of the ASR model. In this paper, trajectory means dynamic acoustic state. The relevant movement is the movement of evidence through time: toward a symbol, across a boundary, through uncertainty, or into a domain region the model has not seen before.

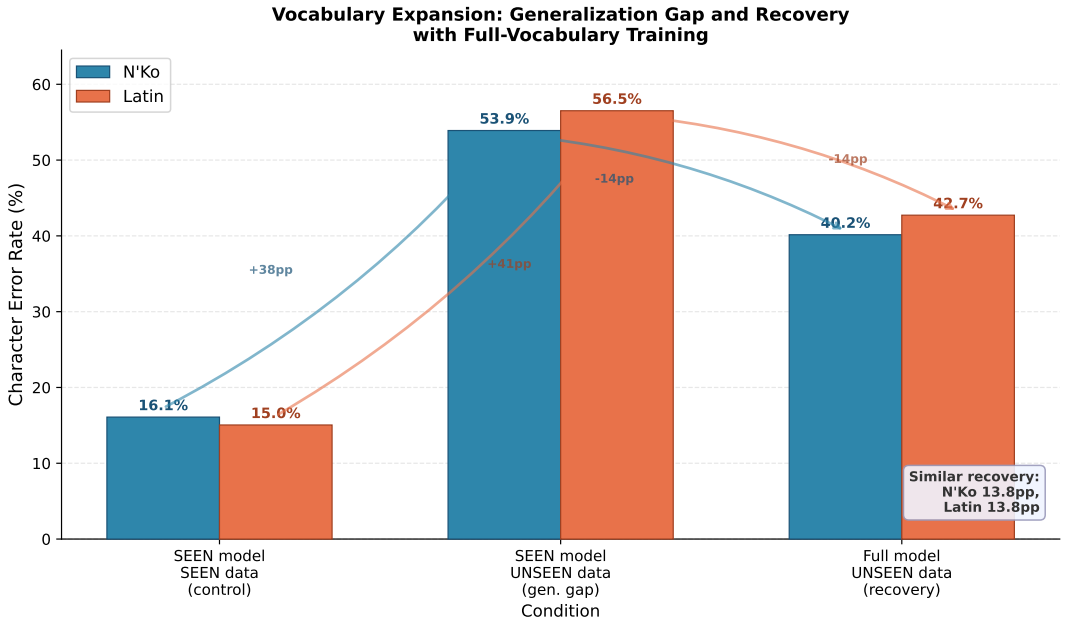

Figure 3: Historical Exp H vocabulary-expansion figure. Full-data training helps both scripts, but the recorded unseen-word residual remains smaller for N'Ko.

The script dependence follows from the target representation. If an acoustic boundary corresponds cleanly to a N'Ko character boundary, then transition pressure is directly useful. If the same acoustic transition is encoded through a Latin digraph, optional tone mark, or spelling convention, the decoder receives a less direct target. This does not imply that Latin cannot be modeled. It implies that the trajectory signal has a cleaner supervised target in N'Ko.

14 What the Conservative Ablations Say

The later same-snapshot low-learning-rate matrix is valuable, but it is not a strict comparison to the 20.57% anchor. It preserves five artifact-complete runs with 29,060-row prediction/reference contracts, all under learning rate 0.0001. The results were N'Ko baseline 31.38% CER, a N'Ko trajectory-attention residual variant at 31.69%, a N'Ko trajectory plus test-time adaptation variant at 31.12%, Latin baseline 31.66%, and Latin trajectory 32.81%.

The correct conclusion is modest. The low-learning-rate results show that the training and artifact pipeline can produce complete row-level outputs across conditions, and they reinforce that optimization settings dominate if they are not held fixed. They do not validate or invalidate the 20.57% result. They also do not prove that the residual-attention branch or the test-time adaptation branch improves ASR. In fact, the later proposal notes reinterpret the residual-attention branch as likely a negative result: trajectory scalars themselves appear more important than adding deeper residual trajectory attention.

The abbreviation problem also needs to be corrected for outside readers. TAR in the internal run names referred to a trajectory-attention residual branch: a heavier decoder variant that tries to inject trajectory state deeper into attention. TTT referred to test-time training or adaptation: an inference-time procedure intended to update a small part of the model on new speaker or domain evidence. Neither term should appear in a public abstract without expansion, and neither should

Table 14: Interpretation of the seven trajectory channels in the canonical speech-dynamic framing. Earlier notes sometimes used motion-inspired names; this paper formalizes the channels as ASR state variables.

Channel	Acoustic interpretation	Decoder or AGP consequence
Commitment	Evidence has settled toward a symbol or local phrase.	Broader context is safer; AGP should resist rewriting.
Uncertainty	Competing paths remain plausible.	The decoder should avoid brittle commitments; AGP should require stronger evidence.
Transition pressure	A phoneme, syllable, or phrase boundary is likely nearby.	Attention may localize around the boundary; AGP can inspect local edits.
Recovery margin	The model is leaving a difficult or noisy region.	Corrections should account for preceding instability.
Phase stiffness	Local rhythm or prosodic cadence is regular.	Temporal attention can use rhythmic expectations.
Novelty	The row looks unlike the stable training distribution.	Prefer review, retrieval, or abstention over fluent rewriting.
Stability	The region is low-risk and internally consistent.	Preserve the raw ASR text unless there is direct contradictory evidence.

be identified with the archived 20.57% checkpoint. The anchor was the simpler trajectory-only CTC path.

The low-learning-rate matrix is still scientifically useful because it is a negative control on the project's own enthusiasm. If a heavier trajectory branch under a different learning rate does not beat a simpler baseline, then the paper should not claim that more geometry automatically helps. The better interpretation is mechanistic: compact trajectory state may be valuable when it biases the decoder in the right place, while additional residual branches can overconstrain or miscalibrate the attention stack. That is a real finding, even though it is not the headline result.

The matrix also shows why future comparisons must be matched at the launch-script level. A fair comparison requires identical corpus snapshot, pair hash, feature cache, split, vocabulary construction, optimizer, learning rate, batch size, dropout, patience, seed, epoch budget, early-stopping rule, and scoring normalizer. Changing any one of these can be defensible engineering; changing it silently makes the run unusable for the claim "model A beats model B." The canonical paper therefore treats the low-learning-rate runs as artifact-complete ablations under their own regime, not as failed reproductions of the 20.57% anchor.

15 AGP: Anticipation Geometry Partition

AGP is the post-ASR Anticipation Geometry Partition. It should not be described as the model that produced the 20.57% result. The acoustic pipeline is audio to Whisper features to trajectory CTC to raw N'Ko transcript. AGP begins after that: it consumes the raw transcript, row-level metadata, and trajectory-derived signals, then partitions spans into states that govern whether correction is allowed.

Table 15: Preserved low-learning-rate matrix. These runs are useful for engineering diagnosis and AGP packet construction, but they are not directly comparable to the 20.57% anchor because the anchor used learning rate 0.0003.

Run	Script	Mode	CER
N'Ko baseline	N'Ko	baseline	31.38%
N'Ko trajectory residual	N'Ko	trajectory + residual attention	31.69%
N'Ko trajectory with adaptation	N'Ko	trajectory + test-time adaptation	31.12%
Latin baseline	Latin	baseline	31.66%
Latin trajectory	Latin	trajectory	32.81%

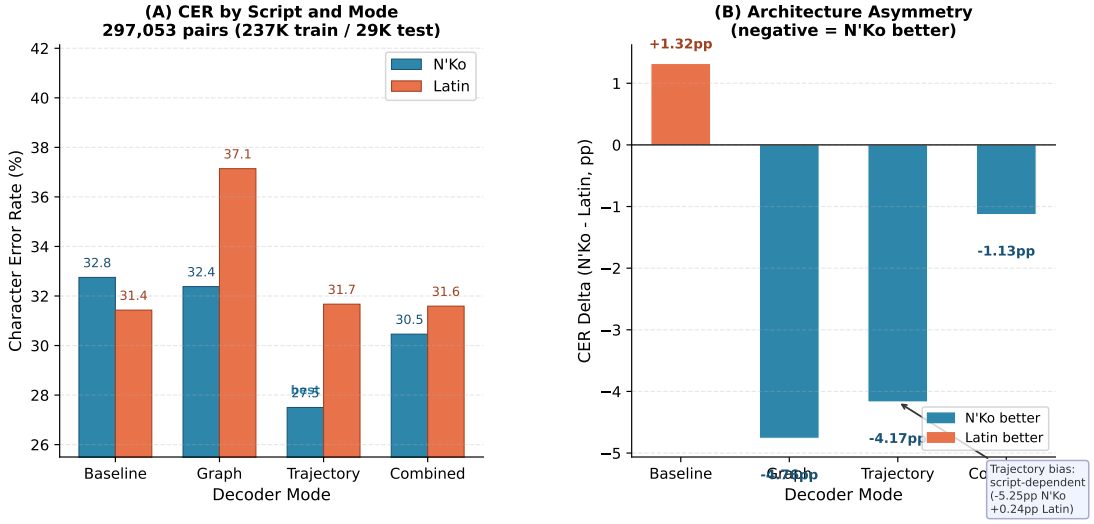

Figure 4: CER comparison figure from the Paper 4 line of work. Public versions should separate archived anchor evidence, historical trajectory evidence, and non-comparable low-learning-rate evidence.

This distinction matters because a language-model correction layer can easily improve surface fluency while corrupting speech evidence. AGP exists to prevent that failure. Its goal is conservative admissibility: accept local repairs that reduce error, reject unsupported rewrites, and expose uncertainty rather than hiding it in a fluent sentence.

The current AGP smoke tests are preliminary but informative. A hand smoke test moved CER from 14.29% to 4.76% with two accepted improvements and no worse accepted edits. A synthetic stress test moved 13.33% to 10.00%, with three improved and five neutral accepted cases. An archived low-CER real slice moved 76.04% to 75.12% with one improved accepted case and no worse accepted cases. These numbers are evidence that the correction gate can be conservative; they are not a benchmark claim over the 20.57% test set.

The formal AGP unit is a row, not a sentence-level anecdote. A row contains the raw ASR hypothesis, the reference when available, edit counts, character denominator, trajectory summaries, a partition label, candidate correction spans, admissibility decision, and final text. When retrieval or provenance is available, the row also stores sources used, transliteration variants, normalized

Table 16: Operational AGP states. AGP turns trajectory geometry into a correction policy rather than an unconstrained text rewrite.

State	Correction policy
Stable	High confidence and low transition pressure. The default action is to leave the transcript unchanged.
Boundary	A likely phrase, syllable, or phoneme transition. Local constrained repair is allowed when the evidence supports it.
Uncertain	Multiple plausible decodings. Correction requires stronger evidence and may abstain.
Novelty	Possible new word, speaker shift, code-switch, or domain mismatch. The system should prefer provenance-aware review or abstention.

Table 17: AGP row contract. The row-level design prevents a correction layer from silently rewriting transcripts without evidence.

Block	Representative fields
ASR evidence	feature id, split, script, mode, raw hypothesis, reference, edit count, reference character count.
Trajectory evidence	commitment, uncertainty, transition pressure, recovery margin, phase stiffness, novelty, stability, and derived partition.
Local uncertainty	top confusable spans, compact n-best alternatives, character posterior summary, uncertainty score.
AGP decision	prompt version, model id if used, proposed correction, confidence, accept/reject/abstain, reason, changed spans.
Provenance	final text, source list, retrieval tags, transliteration variants, normalized forms, provenance score.
Deployment gate	TTS eligibility, overlap risk, music risk, speaker cleanliness, exclusion reason.

forms, and a provenance score. For deployment, the same row carries TTS eligibility fields: overlap risk, music risk, speaker cleanliness, and exclusion reason. This schema matters because it makes AGP auditable. A reviewer can inspect not only whether CER changed, but which edits were accepted, rejected, and why.

The admissibility policy is deliberately asymmetric. A correction is cheap to propose but expensive to trust. AGP should accept a change only when the local evidence supports it and the partition permits it. Stable spans usually need no change. Boundary spans may admit small repairs because boundaries are precisely where CTC errors often occur. Uncertain spans should often abstain unless multiple signals converge. Novelty spans should be treated as data-discovery events rather than as language-model editing opportunities. In other words, AGP is not a post-hoc beautifier; it is a risk model for transcript modification.

A full AGP benchmark should therefore report more than final CER. It should report the number of proposed edits, accepted edits, rejected edits, abstentions, accepted improvements, accepted neutral edits, accepted regressions, rejected improvements, and per-partition CER deltas. The most important safety metric is accepted regressions. A correction system that sometimes improves CER but frequently accepts worse edits is not acceptable for building a corpus or a TTS training set.

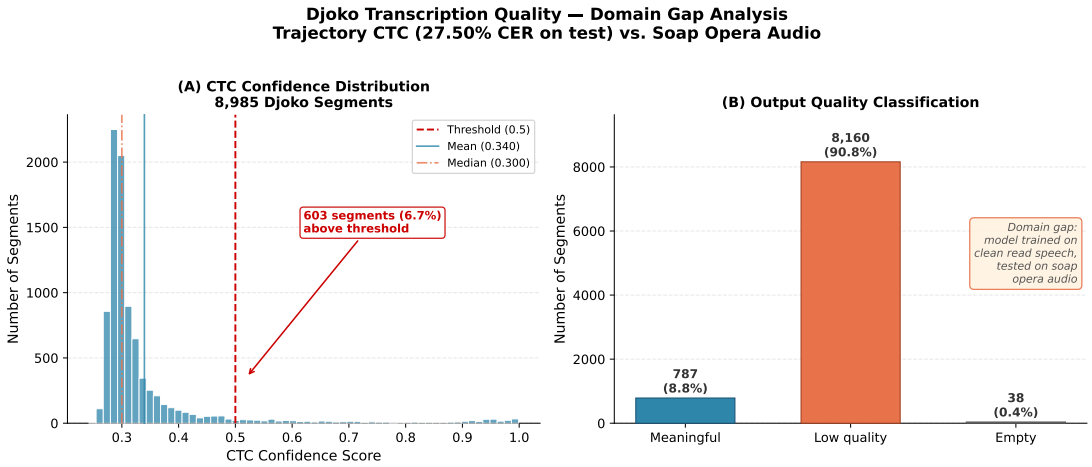

Figure 5: Djoko quality-control figure from the deployment paper. The figure illustrates the deployment filtering problem; it is not presented as a final ground-truth CER benchmark for the 20.57% checkpoint.

16 Deployment and Row-Level Artifacts

Scalar CER is insufficient for the next stage of the research. AGP needs rows: feature identifier, split, script, mode, raw ASR hypothesis, reference text, edit counts, reference character counts, trajectory scalars, partition labels, correction proposal, admissibility decision, and final text. The later packet-corpus work created that structure for the preserved low-learning-rate runs, with 29,060 prediction rows and 29,060 reference rows per available run.

The deployment papers also preserve Djoko and consensus-pipeline artifacts: 8,985 segments, 8,985 transcriptions, 269 consensus rows, 6,625 speaker-labeled rows, 258 episodes, seven speakers, and five eligible speakers. Those data matter because the eventual system is not just an ASR benchmark. It must operate in real domains with speaker variation, episode structure, visual context, and disagreement among transcribers. They should be presented as deployment substrate, not as final CER-bearing proof.

The broader Djoko extraction effort was larger than the first consensus slice. The handoff records 1,124 downloaded videos out of a 2,001-video YouTube channel and 32,826 thirty-second audio segments. It also records a severe domain gap: the first batch produced only about 8.8% meaningful output when a clean-read-speech CTC model was applied to soap-opera audio. That number should be taken seriously. It means the archived 20.57% checkpoint is a within-distribution ASR anchor, not a production model for conversational broadcast media.

The reason the Djoko material still belongs in the paper is that it defines the next research interface. The source has no reliable burned-in subtitles, so OCR is not a solution. The project instead uses audio segmentation, ASR hypotheses, Latin bridge hypotheses where useful, visual scene analysis, speaker clustering, consensus filtering, and row-level provenance. In that environment, AGP is not an optional decoration. It is the only safe way to separate stable transcript rows from rows that are uncertain, novel, noisy, overlapping, musical, or speaker-ambiguous.

The deployment protocol also clarifies the difference between corpus building and model scoring. For model scoring, every row needs a reference. For corpus building, many rows do not yet have a trusted reference, so the system must preserve uncertainty, provenance, and exclusion decisions. A stable AGP row can be used for search or candidate review. A boundary row can become a

correction-training example. An uncertain row should remain visible but not silently enter the training corpus. A novelty row can point to new vocabulary, new speakers, or domain drift. This is how the research can continue without pretending that out-of-domain ASR is already solved.

17 The Public Narrative Around 20.57%

The project can talk about 20.57% CER, but only in a disciplined way. The publishable sentence is:

> An archived N'Ko trajectory ASR checkpoint in the repository reports 20.57% test CER after training on 290,596 Bambara speech pairs under recorded settings.

That sentence is strong because it is exact. It avoids the mistakes that would weaken the work: claiming the result was reproduced in May 2026, claiming matched Latin-vs-N'Ko superiority from non-matched runs, attributing the anchor to a later decoder variant, saying AGP improved the benchmark, or implying production readiness.

The surrounding narrative should be that 20.57% is an anchor for a larger scientific case. The LLM papers show why N'Ko is invisible to general models. The speech papers show that direct audio-to-N'Ko decoding is technically real. The metric paper argues that N'Ko CER is a more meaningful target than Latin WER for Manding ASR. The trajectory work gives a dynamic decoding mechanism. AGP gives the correction and governance layer needed before deployment. This is a coherent research program even if the final strict reproduction must wait.

The public abstract should therefore avoid the structure "we achieved 20.57% CER, therefore N'Ko beats Latin." That sentence invites the wrong scrutiny because the matched comparison is not closed. A stronger abstract begins with the infrastructure problem: widely used AI systems can encode N'Ko characters while still lacking usable internal structure for them; Latin-script metrics then hide the problem by scoring orthographic convention rather than script-native phonemic evidence. The 20.57% result enters as evidence that a direct N'Ko ASR path is technically viable enough to make the measurement problem concrete.

For a model card, the wording should be even narrower. It should name the artifact as an archived checkpoint, report the exact split and learning rate, state that the strict May 2026 reproduction did not complete, and warn that the checkpoint is not a production Djoko or conversational-broadcast model. For a project page, the broader narrative can be stronger: N'Ko reveals a three-part failure in modern AI systems, covering representation, decoding, and metrics. For a research paper, the value is the synthesis: the project connects mechanistic LLM diagnostics with script-native ASR and a conservative correction framework.

The phrase "20 CER" should be treated as a handle, not as a final conclusion. It is useful because it gives readers a memorable technical anchor. It is dangerous if it becomes the entire story. The paper's real contribution is that it explains why a 20.57% N'Ko CER has a different scientific meaning from a Latin WER number, and why a language technology stack for Manding should be evaluated in the script that preserves the relevant linguistic structure.

18 The Reconstruction and Tone Pillar

The four levers introduced above—representation, alignment, measurement, and governance—describe how a script shapes *recognition*, the path from signal to symbol. A fifth lever closes the loop in the other direction. The same featural N'Ko codebook that makes recognition interpretable also makes *reconstruction* expressible: tone, and ultimately sound, can be written in a

form that is editable, auditable, and phonemically explicit. This section states the reconstruction pillar, gives the corrected tone-mark inventory it depends on, and fixes the unifying architecture that ties decoding, governance, and correction into one object. It is deliberately a framing-and-protocol section. The empirical load is carried by two companion papers, and every number below is reported with its provenance.

18.1 Tone as prior times evidence

Tone is the largest open hedge in the recognition stack, and the limitations section below still lists it as unfinished. The contribution here is structural: tone resolution is not one model but the product of two independent sources. A linguistic *prior* asks which tone is plausible given the surrounding text; an acoustic *evidence* channel asks which pitch the speaker actually produced. Writing tone resolution as prior times evidence makes the division of labor explicit and assigns each half to a companion paper. The prior is a contextual tone model trained on OCR-extracted toned N'Ko, which learns tone from text. The evidence is Featural Acoustic Coding (FAC), which reads tone from acoustic F_0 and, more broadly, treats the N'Ko syllable codebook as a featural sound code in which the script's tone, length, vowel, and onset structure already carry most of the descriptors a sound needs. Neither half is the acoustic model that produced the 20.57% anchor; both act after decoding, inside the governance gate described next.

18.2 The corrected tone-mark inventory

The reconstruction argument depends on stating the script's tone system correctly. N'Ko encodes tone with seven combining marks in the range U+07EB through U+07F1: short and long forms of high, low, and rising tone, together with a long descending (falling) mark at U+07EE. All four tone *shapes*—high, low, rising, and falling—are therefore native to the script; length is an orthogonal axis carried by the short/long distinction. This matters because an earlier internal annotation of the syllable codebook had mislabeled U+07EE and U+07EF as generic length marks and listed only five tone marks, which would have implied that falling tone was absent and had to be added by a designed extension. That is incorrect. The script already expresses all four tone shapes; any designed featural extension contemplated for FAC is therefore purely *timbral*—harmonicity, spectral spread, dynamics—and never tonal. Correcting this inventory removes a factual error that had propagated from the codebook into draft figures and keeps the reconstruction claim honest.

18.3 An empirical tone prior, reported as provisional

On a corpus of toned N'Ko harvested by vision-language OCR from lesson videos, tone is overwhelmingly a register phenomenon: roughly 75% of marked syllables are level high or low and only about 1% are contour (rising or falling), with the remainder unmarked. The practical consequence is that acoustic tone resolution for N'Ko is, to first order, a high/low register classification problem rather than a contour-tracking problem, which is exactly the regime a register-relative classifier handles well. A text-only baseline that predicts tone from a bigram language model leaves roughly half of tone marks wrong, which is the bar that adding acoustic evidence must beat. These numbers are reported as *provisional*. Independent OCR passes by two strong vision-language models agree on only about 30 to 50% of tone classes, so the corpus is adequate for estimating gross distributional facts but is not yet trustworthy tone ground truth. The decisive measurement requires read speech of known tone-marked text, where the gold tone is authored rather than inferred; that experiment is specified but not yet run, and the limitations section records it as an open gate alongside the unfinished anchor audit.

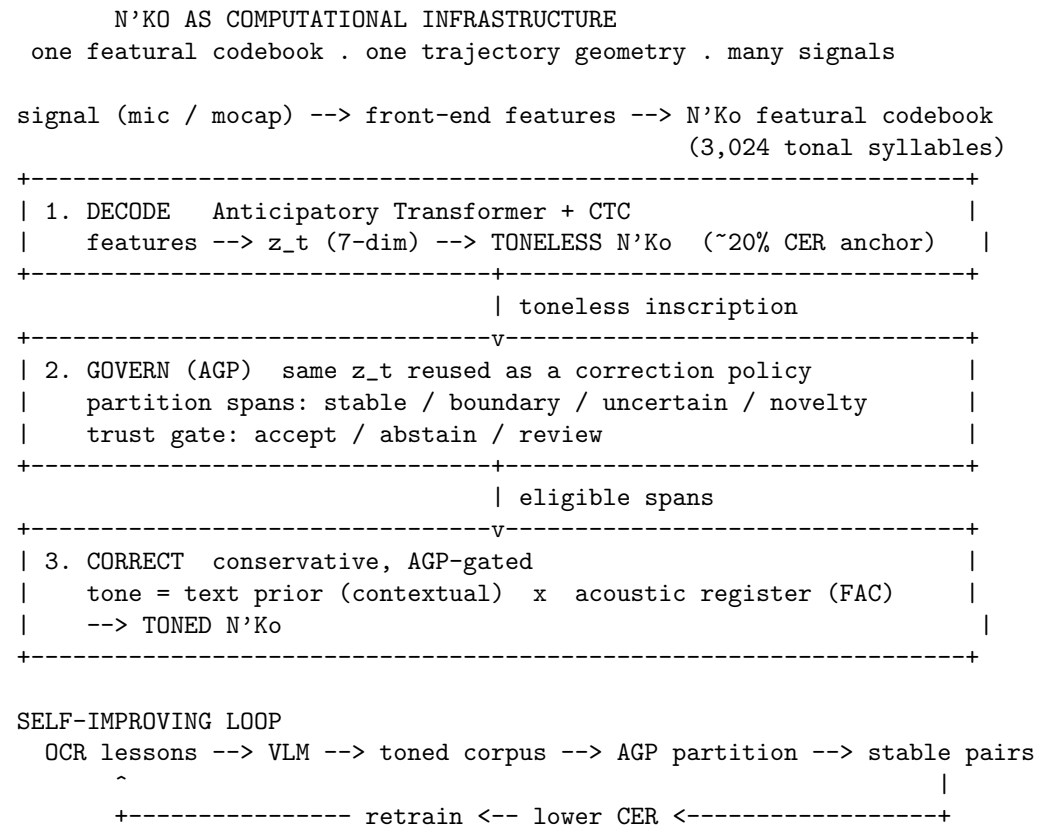

Figure 6: One geometry, three levels. The trajectory state z_t is reused to decode, to govern correction, and to resolve tone. The codebook carries two arrows: *analysis* (signal to N'Ko, the stack of this paper) and *synthesis* (N'Ko to signal), the latter proved for sound by the FAC decoder and reused as a score for body motion in a separate generator.

18.4 One geometry, three levels: decode, govern, correct

The reconstruction pillar does not introduce a separate machine. It reuses the trajectory geometry z_t that already drives decoding. The same state that biases CTC emission (decode) also defines the AGP partitions that decide whether a span may be edited (govern), and tone correction is simply the admissible edit that the correct stage applies under that gate (correct). Figure 6 shows the unifying view, including the self-improving loop in which better tone resolution produces cleaner labels and a lower-error retrain. The figure also records the two directions of travel on the codebook. The *analysis* arrow runs from signal to N'Ko and is the recognition stack of this paper. The *synthesis* arrow runs from N'Ko back to a signal: FAC's decoder demonstrates it for sound, and the same N'Ko-as-score idea drives a separate generator for body motion in a computational choreography line. The notation is shared; the generators are distinct builds, and this paper does not claim that the recognition front-end is swapped to produce motion.

18.5 Program structure and companion papers

This manuscript is the umbrella and the measurement spine of a larger program. The companions carry the empirical and engineering load and are written for their own venues, so they are cited rather than absorbed. The *representation* companions (script invisibility in LLM activations

and its structural persistence across model families) establish that generic models underrepresent N'Ko. The *recognition* companions develop the script-advantage CTC result, speaker adaptation and generalization, and trajectory attention residuals. The *reconstruction* companions are the two halves of tone described above: the contextual tone model and FAC. A further *governance* companion, a Mixture of Anticipatory Orthogonal Experts for N'Ko (MAOE-N'Ko), generalizes the AGP gate into a routing architecture whose experts are orthogonal in authority rather than in capacity, with a deterministic admissibility layer deciding what each partition is permitted to do. Finally, an applied *protocol* direction reframes the entire stack as a market for linguistic computation, in which the work of validating tone, transcription, and cultural use of N'Ko is performed by people whose competence no GPU can replace. The science is the pillars; the protocol is one way the resulting capability can fund the data that the pillars need. A systems track (neural-engine offload, distributed on-device training, and quantized retrieval) is named separately because it explains how the program was trained cheaply, not what it claims.

19 Limitations

The principal limitation is that the strict May 2026 anchor audit did not complete. The project rented compute, hydrated the full feature cache, fixed a mixed tensor shape bug, passed the sanity gate, and launched strict audit processes, but the audit did not produce the verified final result artifacts needed to replace the archived 20.57% checkpoint.

The second limitation is comparability. The low-learning-rate matrix and several historical trajectory comparisons are useful, but they cannot be collapsed into one headline table. Parameter mismatch, artifact-chain incompleteness, and dataset cleaning differences must be stated. In particular, the label-contamination discovery from the project history is scientifically important: low-resource ASR datasets can contain script contamination that changes CER by large margins if not cleaned consistently.

The third limitation is metric precision. N'Ko CER is more phonemically interpretable than Latin WER, but it is not a perfect phoneme error rate. A stronger future paper should formalize the mapping from N'Ko grapheme sequences to phonemic units, including tone, nasalization, punctuation, normalization, and boundary marks. The reconstruction pillar above begins this formalization for tone, with a corrected seven-mark inventory and a prior-times-evidence decomposition. What remains is a decisive measurement on read speech with authored tone-marked gold, because the current OCR-derived tone corpus is only provisional ground truth.

The fourth limitation is AGP evaluation. AGP has a coherent architecture and early smoke-test behavior, but it still needs a full row-level benchmark with accepted, rejected, improved, neutral, and worsened edits reported separately.

The fifth limitation is corpus representativeness. The 290,596-pair anchor is a large project snapshot, but it is not a complete sample of all Manding speech. It is closer to clean read-speech and curated paired data than to conversational, multi-speaker, noisy, dramatic, or music-adjacent audio. The Djoko extraction work exists precisely because the real deployment distribution is harder. Public claims should therefore separate within-distribution ASR from out-of-domain transcription.

The sixth limitation is community and orthographic authority. N'Ko is not only a technical alphabet; it is a living script with scholarly, pedagogical, religious, and community practices. A model can output Unicode N'Ko and still violate community expectations about spelling, tone, punctuation, or register. Publication should invite review from readers who know Manding language practice, not only from machine-learning reviewers.

The seventh limitation is the absence of a finished matched Latin-vs-N'Ko anchor under the same

hyperparameters as the 20.57% result. Historical comparisons and low-learning-rate ablations are informative, but the cleanest future test would train N'Ko and Latin decoders under a single frozen protocol with identical data snapshot, optimizer, learning rate, seed schedule, patience, output artifacts, and scoring code. Until then, the paper should defend N'Ko CER as a better metric and the archived 20.57% checkpoint as a strong script-native artifact, not as the final matched superiority proof.

The eighth limitation is compute closure. The project reached the point where more paid GPU training was no longer justified without first tightening artifacts, manifests, and publication language. That is a practical limitation, but it also improves the research story. The paper can close the current phase by saying: this is what we know, this is what remains unresolved, and this is the exact protocol a future funded audit must run.

20 Conclusion

N'Ko should be treated as computational infrastructure for Manding speech systems. It affects what LLMs can see, what ASR decoders learn, what error metrics mean, and how correction should be constrained. The project does have a legitimate way to talk about the 20% CER result: an archived N'Ko trajectory ASR checkpoint reports 20.57% test CER under recorded settings on the 290,596-pair snapshot. That claim is publishable when presented as an artifact-backed anchor rather than as a freshly reproduced leaderboard result.

The broader paper should argue for a research program, not a single unqualified number. Script invisibility explains why generic models fail N'Ko. Script-native ASR shows that direct N'Ko transcription can work. Trajectory geometry explains why dynamic speech state belongs in the decoder. AGP defines a conservative path from raw ASR output to trustworthy correction. Tone reconstruction extends that same path to the script's last unresolved axis, resolving tone from a text prior and acoustic evidence under the same governance gate. The next phase is not to invent a new story; it is to preserve this one with stricter reproduction, cleaner row-level artifacts, matched comparisons, and the decisive read-speech tone measurement when funding returns.

The consolidated conclusion is therefore methodological. The project began with the intuition that N'Ko might help ASR because it is phonemically transparent. The stronger conclusion is that script affects the whole measurement stack. Before ASR, script affects tokenization and internal model geometry. During ASR, script affects the output labels that CTC learns to align with speech. After ASR, script affects whether CER, WER, and correction decisions correspond to speech evidence or to orthographic convention. A system that ignores script is not neutral; it chooses the infrastructure of the dominant representation and then calls the result language technology.

The 20.57% anchor gives the paper a concrete center. It shows that direct N'Ko ASR is not merely an ethical preference or cultural argument; it reached a technically meaningful error regime on a large Bambara corpus snapshot. The failed May 2026 strict audit does not erase that contribution, but it changes how the contribution must be stated. The result is an archived benchmark anchor awaiting future strict reproduction, not a freshly closed leaderboard.

The immediate publication path should follow that truth. The flagship paper can present script invisibility and computational infrastructure. The metric paper can argue why N'Ko CER is the right target for Manding ASR. The ASR technical report can preserve the 20.57% checkpoint, the trajectory hypothesis, the low-learning-rate negative controls, the Djoko deployment lessons, and the unfinished audit in one honest artifact. That is enough to conclude the current research phase without spending more money, while still giving future collaborators a precise map of what to

reproduce next.

References

[1] Coleman Donaldson. *Clear Language: Script, Register and the N'ko Movement of Manding-Speaking West Africa*. PhD thesis, University of Pennsylvania, 2017.

[2] Moussa Doumbouya, Lisa Einstein, and Chris Piech. Using radio archives for low-resource speech recognition: Towards an automatic transcription of Bambara radio broadcasts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[3] Alex Graves, Santiago Fernandez, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of ICML*, 2006.

[4] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *ICLR*, 2022.

[5] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[6] Leonard Katz and Ram Frost. The reading process is different for different orthographies: The orthographic depth hypothesis. In Ram Frost and Leonard Katz, editors, *Orthography, Phonology, Morphology, and Meaning*, pages 67–84. North Holland, 1992.

[7] National Institute of Standards and Technology. Speech Recognition Scoring Toolkit (SCTK), 2021. NIST evaluation tools page, including SCLITE and related speech-recognition scoring tools.

[8] 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 ICML*, 2023.

[9] Unicode Consortium. N'ko block: U+07C0–U+07FF, 2006. The Unicode Standard, Version 5.0+.

[10] An Yang et al. Qwen2.5 technical report, 2024.

[11] An Yang et al. Qwen2 technical report. *Alibaba Group Technical Report*, 2024.

[12] Mohammad Zeineldeen, Albert Zeyer, Wei Zhou, Thomas Ng, Ralf Schlüter, and Hermann Ney. A systematic comparison of grapheme-based vs. phoneme-based label units for encoder-decoder-attention models, 2021.