Script-Native ASR for N'Ko: Anticipatory Transformer CTC Decoding and the 20.57% CER Anchor

Mohamed Diomande

Final paper series, May 2026

Abstract

This paper preserves the technical ASR center of the N'Ko research program: an archived script-native trajectory checkpoint reporting 20.57% character error rate on a 290,596-pair Bambara corpus snapshot. The model uses frozen Whisper large-v3 acoustic features, a trainable Transformer CTC decoder, and a compact trajectory state that biases attention with speech-dynamic information. The result is the strongest retained ASR artifact in the project and is the correct way to discuss the phrase "20 CER" publicly.

The paper gives the architecture, training regime, artifact contract, and claim boundaries. Input audio is encoded by Whisper large-v3 [3]; 1280-dimensional features are projected to a 768-dimensional decoder space, temporally downsampled, and decoded by a six-layer Transformer CTC head [1]. A trajectory module estimates a seven-dimensional state for each timestep: commitment, uncertainty, transition pressure, recovery margin, phase stiffness, novelty, and stability. This state produces an additive attention-logit bias before CTC emission, giving the decoder an anticipatory geometry over speech dynamics.

The archived anchor was trained on 290,596 paired examples split into 232,476 training rows, 29,060 validation rows, and 29,060 test rows, with learning rate 0.0003, batch size 32, dropout 0.1, seed 42, and best validation loss 0.6358872798606507. The reported test CER is 20.57%, computed as 216,225 edits over 1,050,967 reference characters. It is an archived checkpoint result with preserved metadata. Later low-learning-rate runs around 31% CER are not comparable to the anchor because they used a different learning-rate regime. The conclusion is therefore bounded: direct N'Ko ASR reached a meaningful error regime under recorded settings, and the artifact should not be silently replaced by non-comparable runs.

1 Introduction

The first two papers in this series establish the premise. General language models can be weak processors of N'Ko even when they accept the Unicode string, and Latin WER is not a sufficient metric for script-native Manding speech recognition. This paper asks the next question: what did the ASR system actually achieve, and how should that achievement be stated without overstating it?

The answer is the 20.57% anchor. The project retains an archived N'Ko trajectory CTC checkpoint trained on a 290,596-pair corpus snapshot. It reports 20.57% test CER with explicit scorer arithmetic. This is the number that can be discussed publicly, but only with provenance: later ablations used a different learning-rate regime and should not be merged into the same claim. The paper therefore treats the result as an artifact-backed anchor rather than as a loose leaderboard slogan.

The scientific contribution is not only the number. It is the architecture and the measurement stack around the number. The model decodes directly into N'Ko, not through Latin. It uses CTC,

Table 1: Research questions for the script-native ASR anchor.

ID	Question	Required evidence
RQ1	What architecture produced the retained 20.57% CER anchor?	Frozen acoustic encoder, CTC decoder, trajectory-state definition, and training metadata.
RQ2	What makes the anchor inspectable?	Row counts, split sizes, scorer numerator, denominator, hashes, and artifact paths.
RQ3	Which later runs are non-comparable?	Hyperparameter table showing learning rate, architecture branch, and artifact differences.
RQ4	What should be preserved for future work?	Row exports, partition metrics, feature/pair hashes, and the exact scoring contract.

Table 2: Allowed and disallowed public claims about the ASR result.

Allowed	Disallowed
An archived N'Ko trajectory CTC checkpoint reports 20.57% test CER.	A later variant with different hyperparameters is the same result.
The score was recorded under lr=0.0003, batch size 32, dropout 0.1, seed 42.	The later lr=0.0001 matrix directly refutes or replaces the anchor.
The result supports direct script-native N'Ko ASR as a serious research path.	The result proves universal N'Ko superiority over Latin under all settings.
The anchor should be reported with its scorer arithmetic and metadata.	AGP, TAR, or TTT produced the 20.57% number.

whose labels are script-native characters rather than words. It adds a compact trajectory state to attention, making speech dynamics visible to the decoder. The surrounding artifacts record splits, row counts, hashes, and scorer denominators. This is what makes the result useful for research even when the public story is kept concise.

2 Research Questions and Claim Boundaries

The paper is organized around three questions: what architecture produced the retained number, what metadata makes the number interpretable, and which later branches should not be merged into the same claim. This keeps the paper from turning the 20.57% result into an overclaim while avoiding a defensive narrative.

Table 3: Condensed development path. Values are from different regimes and should not be plotted as one homogeneous leaderboard.

System	Role	Recorded outcome
V1 BiLSTM CTC	Feasibility: direct audio-to-N'Ko output.	Approximately 56% CER.
V3 Transformer CTC	Frozen Whisper features plus six-layer Transformer decoder.	Approximately 33% validation CER in the development regime.
V4 Whisper LoRA	Acoustic adaptation and confidence experiment.	Loss and confidence improved; not the canonical anchor.
Trajectory CTC	Script-native trajectory decoder at larger corpus scale.	20.57% archived test CER.

3 Background

3.1 CTC and script-native labels

Connectionist Temporal Classification solves sequence alignment without frame-level labels [1]. For an input sequence $x_{1:T}$ and target label sequence $y_{1:U}$, CTC marginalizes over all framewise paths π that collapse to y:

$$\mathcal{L}_{\mathrm{CTC}}(x, y) = -\log \sum_{\pi \in \mathcal{B}^{-1}(y)} \prod_{t=1}^{T} p(\pi_t \mid x).$$

The label inventory matters. If labels are N'Ko characters, the model learns a direct relation between acoustic evidence and N'Ko script units. If labels are Latin words or Latin characters, the model learns a different output relation that may hide tone, digraph boundaries, or orthographic convention.

3.2 Whisper features as acoustic substrate

The project uses Whisper large-v3 as a frozen acoustic encoder [3]. This is pragmatic. Training an acoustic encoder from scratch for Bambara is unrealistic under the available data and compute. Frozen Whisper features give the decoder access to strong multilingual acoustic representations while keeping the script-specific learning problem in the CTC head.

3.3 Development path

The final anchor should be understood against the development sequence. Early BiLSTM CTC systems established feasibility but were weak. Transformer CTC decoders on frozen Whisper features improved validation CER into the low-30s in earlier development regimes. LoRA experiments on Whisper improved confidence and loss behavior but were not the canonical benchmark [2]. The archived anchor belongs to a later trajectory-conditioned CTC line trained at larger scale.

Table 4: Seven trajectory channels.

Channel	Intended meaning
Commitment	Evidence that the current local acoustic-symbol state is settled.
Uncertainty	Local ambiguity in the acoustic or decoder state.
Transition pressure	Evidence that the utterance is crossing a phoneme, syllable, or phrase boundary.
Recovery margin	Room to recover after an unstable local decision.
Phase stiffness	Resistance to abrupt changes in the local trajectory.
Novelty	Evidence of unseen word, unusual speaker behavior, or domain shift.
Stability	Persistence of a coherent local decoding path.

4 Architecture

4.1 Encoder and decoder

Let a be an audio segment. Whisper large-v3 produces acoustic features

$$H = E_{\text{Whisper}}(a) \in \mathbb{R}^{T \times 1280}.$$

The decoder first projects these features:

$$U_0 = HW_p + b_p, \qquad W_p \in \mathbb{R}^{1280 \times 768}.$$

A temporal downsampling module reduces the effective sequence length to T':

$$U = \text{Downsample}(U_0) \in \mathbb{R}^{T' \times 768}.$$

The downsampled sequence passes through six Transformer blocks and a CTC output projection over the normalized N'Ko character vocabulary plus blank.

4.2 Trajectory state

The anticipatory component computes a compact state:

$$z_t = \sigma(g_\theta(U_{t-k:t+k})) \in [0,1]^7.$$

The seven channels are defined operationally, not metaphysically:

For attention head m, the trajectory state defines an additive bias:

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

The model is called anticipatory because attention is conditioned not only on content similarity but also on local speech dynamics. A frame near a boundary or uncertainty region should not be treated the same as a stable frame deep inside a settled character span.

Table 5: Ablation logic for the ASR architecture family.

Mechanism	Hypothesis tested	What a negative result means
Script-native CTC	Direct N'Ko labels reduce label ambiguity compared with a Latin bridge.	The model or data may still be insufficient; it does not refute N'Ko metric validity.
Compact trajectory state	Local speech dynamics help boundary and uncertainty handling.	The state may be misplaced, undertrained, or unnecessary for a given regime.
TAR branch	Deeper trajectory residuals improve attention decisions.	More geometry can overconstrain or destabilize training.
TTT branch	Inference-time adaptation improves difficult rows.	Test-time updates may add variance or fail without a calibrated objective.

4.3 Why trajectory belongs with N'Ko

The trajectory hypothesis is that compact dynamic state is most useful when output labels preserve acoustic structure. N'Ko characters are closer to Manding phonemic units than Latin word tokens or many Latin character sequences. If a trajectory channel detects transition pressure, a N'Ko CTC decoder can use it at a boundary where a single script unit often corresponds to a sound unit. In a Latin digraph regime, the same acoustic event may need multiple written characters, so the boundary relation is less direct.

This hypothesis is not the same as claiming every trajectory variant improves every run. Later low-learning-rate experiments suggest the heavier trajectory-attention residual branch can underperform. The retained claim is narrower: the archived 20.57% checkpoint is associated with the simpler trajectory-conditioned CTC path, and the historical comparisons motivated trajectory as a script-sensitive mechanism.

4.4 Ablation logic

The architecture family has three separable mechanisms. The baseline mechanism is direct script-native CTC: the model emits normalized N'Ko characters instead of Latin words or post-converted strings. The trajectory mechanism adds a compact state that summarizes local speech dynamics before or during attention. The heavier TAR mechanism, short for trajectory-attention residual, injects trajectory information as a deeper residual branch. TTT, short for test-time training or test-time adaptation, changes the inference procedure rather than simply changing the decoder's forward pass.

5 Training Regime and Artifact Anchor

The canonical anchor metadata is explicit.

Table 6: Canonical archived N'Ko ASR anchor.

Field	Value
Corpus snapshot	290,596 paired examples
Train / validation / test	232,476 / 29,060 / 29,060
Script and mode	N'Ko trajectory CTC
Learning rate	0.0003
Batch size	32
Dropout	0.1
Seed	42
Best validation loss	0.6358872798606507
Epochs trained	47
Reported test CER	**20.57%**
CER arithmetic	216,225 edits / 1,050,967 reference characters
Results SHA-256	`252aecd6e323f7d50cefd3c1e507ddaf035d9f0ac4f78d67766c4cf6ed5d24a7`
Vocabulary SHA-256	`e3ab620c9d2f971603d76f953f2be40bf9283dfd99d6428c7d51a9a73246ea67`
Best checkpoint SHA-256	`ab1fe47f96c2c434d8f301ae065b3292d592b9a4f5accf1d09acc97ca2c03b59`

The anchor should be reported with its arithmetic:

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

This protects the claim from rounding ambiguity. If a future scorer review changes the numerator or denominator, the difference can be localized to scorer behavior, normalization, split composition, model output, or reference material.

6 Provenance Protocol

The anchor is useful because it carries enough metadata to be inspected rather than only repeated as a rounded percentage. The provenance protocol verifies the pair file hash, row count, feature count, feature tensor shapes, vocabulary hash, split sizes, and output directories. It also preserves the prediction and reference rows needed to recompute the score.

Table 7: Required provenance stages for the anchor.

Stage	Required check
Data identity	Pair file hash, row count 290,596, feature tensor count 290,596, and feature-shape normalization.
Split identity	Train, validation, and test row identities; expected counts 232,476, 29,060, and 29,060.
Vocabulary identity	Script-native vocabulary file and SHA-256 `e3ab620c9d2f971603d76f953f2be40bf9283dfd99d6428c7d51a9a73246ea67`.

Table 8: Artifact contract for the 20.57% anchor.

Artifact	Purpose
`results.json`	Stores scalar metrics, hyperparameters, row counts, hashes, and artifact paths.
`test_predictions.jsonl`	Provides one hypothesis per test row for independent rescoring.
`test_references.jsonl`	Provides the aligned reference rows and denominator source.
`test_metrics_by_partition.json`	Supports AGP or error-partition analysis without repeating inference.
`split.json`	Preserves exact row membership for train, validation, and test.
`vocab.json`	Defines output labels and lets readers verify script-native decoding.
`best.pt` and `final.pt`	Preserve the selected checkpoint and the terminal training state.
Logs and launch scripts	Record command line, hardware, dependency versions, and guardrail checks.

Stage	Required check
Training identity	Learning rate 0.0003, batch size 32, dropout 0.1, seed 42, patience, optimizer, maximum epochs, and stopping rule.
Runtime identity	Trainer hash, library versions, GPU type, feature cache path, and exact launch command.
Evaluation identity	CER normalizer, prediction export, reference export, partition metrics, edit numerator, and reference denominator.
Publication identity	Results JSON, best checkpoint, final checkpoint, logs, split, vocabulary, and hashes copied into a durable bundle.

7 Artifact Contract

The main research artifact is not only a checkpoint. A checkpoint without row-level predictions cannot defend the metric. A row export without a split file cannot defend the data boundary. A metric without a normalizer cannot defend the scorer. The durable anchor bundle should therefore contain the following artifacts.

8 Historical Comparative Evidence

The project history contains an eight-way internal comparison across baseline, graph, trajectory, and combined conditions in N'Ko and Latin. These runs explain why the trajectory hypothesis became central. They are not the canonical benchmark, because the full local artifact bundle for all eight historical runs is not restored.

The historical pattern is scientifically useful because it suggests that dynamic or structural mechanisms may help N'Ko more when the output labels preserve phonemic structure. But it

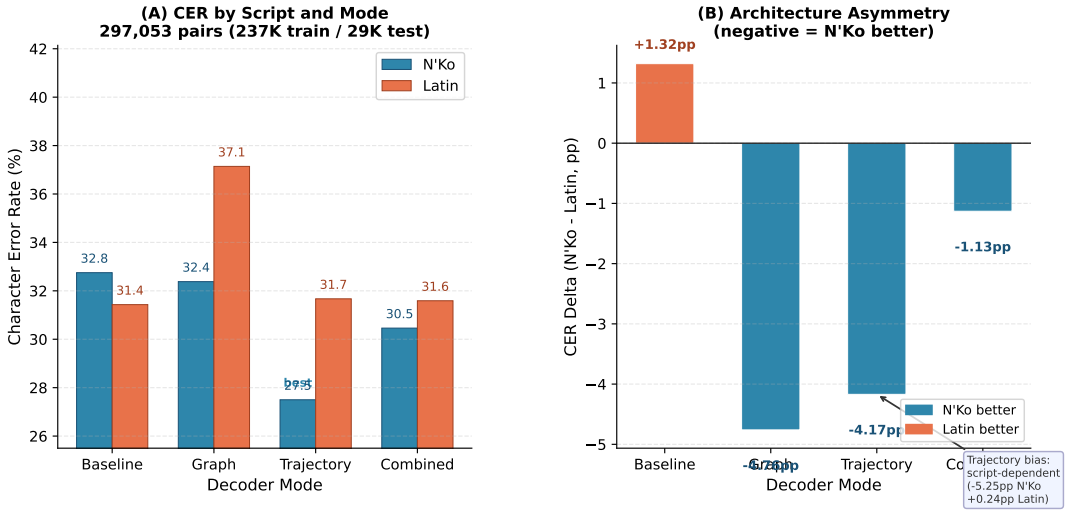

Figure 1: CER comparison context from the ASR experiments. The final paper series separates archived anchor evidence from historical comparisons and non-comparable low-learning-rate ablations.

Table 9: Historical eight-way comparison. These values are hypothesis-generating context, not the primary benchmark.

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
Graph + trajectory	30.46%	31.59%	-1.13pp

should not be promoted as the final matched proof. Artifact status determines claim strength.

9 Non-Comparable Later Runs

The project also preserved a low-learning-rate matrix around 31% CER. Those runs are valuable engineering evidence, but they are not anchor replacements because the learning rate differed. The main lesson is comparability: a change from lr=0.0003 to lr=0.0001 changes the training regime, so the result should be read as a separate condition rather than as a direct contradiction.

9.1 Operational lessons

The later execution work exposed practical issues that matter for future work: feature hydration required substantial disk, feature tensors appeared in mixed shapes and needed loader normalization, and guardrails had to distinguish pre-hydration disk capacity from post-hydration free space. These are engineering lessons about running the pipeline, not reasons to bury the archived anchor.

This table is important because it prevents separate mechanisms from being merged into one story. TAR means trajectory-attention residual: a heavier branch that injects trajectory state deeper into attention. TTT means test-time training or adaptation: an inference-time update procedure. Neither is the archived 20.57% result. The fact that the heavier variants did not

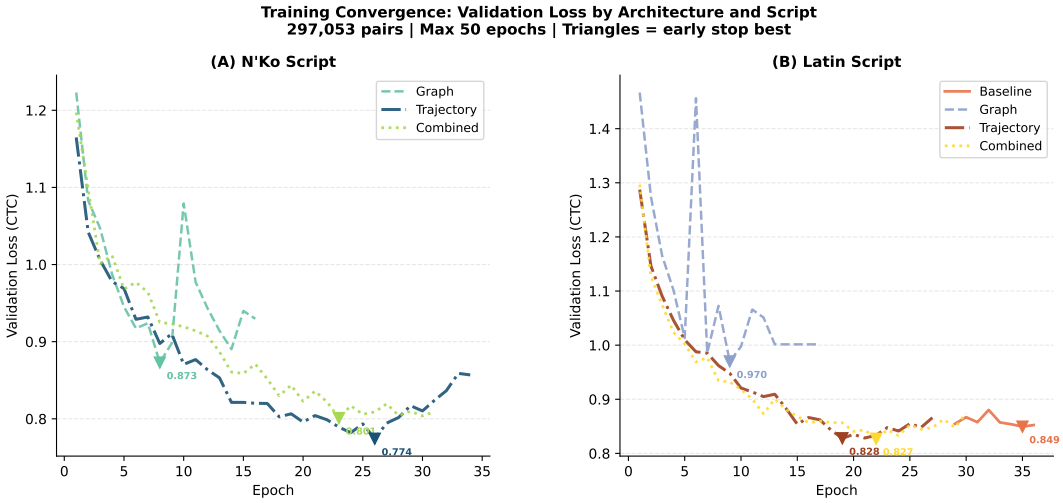

Figure 2: Loss-curve context from the ASR line of work. Loss curves are useful for training diagnosis but do not by themselves establish matched script superiority.

Table 10: Low-learning-rate ablation context. These runs used lr=0.0001 and should not be compared directly against the lr=0.0003 anchor as if only architecture changed.

Run	Script	Mode	CER
N'Ko baseline	N'Ko	baseline	31.38%
N'Ko TAR	N'Ko	trajectory-attention residual	31.69%
N'Ko trajectory TTT	N'Ko	trajectory + test-time training branch	31.12%
Latin baseline	Latin	baseline	31.66%
Latin trajectory	Latin	trajectory	32.81%

obviously improve the low-learning-rate matrix is a useful mechanistic warning: compact trajectory state may be helpful in the right position, but more geometry is not automatically better.

10 Data Scale

Data scale matters. Earlier development used smaller corpora for architecture search and feasibility. The anchor uses the larger 290,596-pair snapshot. The paper should not mix results from 37-hour or 37K-row development runs with the 290K-row anchor without labeling them. The development runs answer engineering questions; the anchor answers the retained benchmark question.

11 Publication Language

The correct public sentence is:

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

That sentence is strong because it is specific. It names the script, architecture family, metric, corpus scale, and artifact status. The unsafe sentence is:

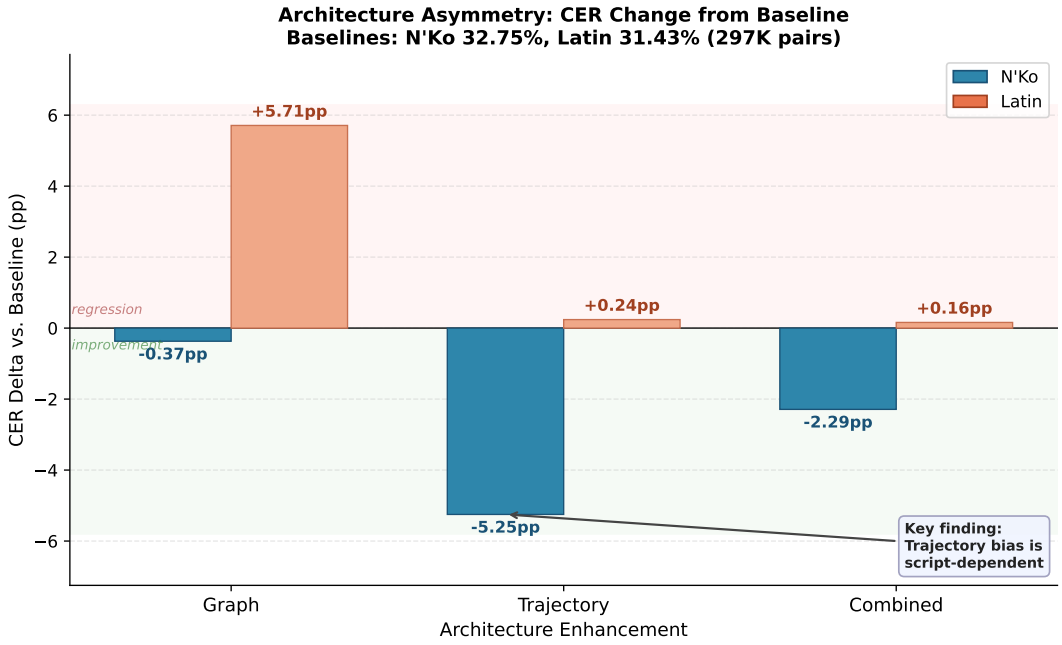

Figure 3: Delta figure from the ASR experiments. The public interpretation should keep historical and canonical evidence separate.

> The later TAR, TTT, or AGP branch produced 20.57% CER and proves N'Ko always beats Latin.

That sentence is false or at least unsupported by the retained artifact chain.

12 Model Card and Intended Use

The retained anchor should be accompanied by a minimal model card. The intended use is research on script-native N'Ko ASR for Manding speech under controlled evaluation. The intended metric is normalized N'Ko CER with explicit edit counts and denominators. The system is not intended as a final legal, medical, emergency, or fully automated subtitle system. It is also not intended to replace community orthographic authority. The correct deployment path is benchmarked ASR followed by row-level governance, review, and domain-specific validation.

13 Limitations

The main limitation is scope. The archived checkpoint is a retained benchmark anchor, not a deployment guarantee. Future work should preserve the exact lr=0.0003 anchor contract with the frozen split, pair hash, vocabulary hash, feature validation, row exports, and partition metrics.

The second limitation is matched comparison. The historical N'Ko/Latin tables are informative, but artifact status and hyperparameter mismatch prevent them from closing a universal superiority claim. The next fair comparison needs identical data, split, feature cache, optimizer, learning rate, patience, seed schedule, normalizer, scorer, and artifact export.

The third limitation is deployment. The anchor is a within-distribution ASR result. It is not a finished conversational-broadcast model, a Djoko production model, or a guarantee of subtitle-

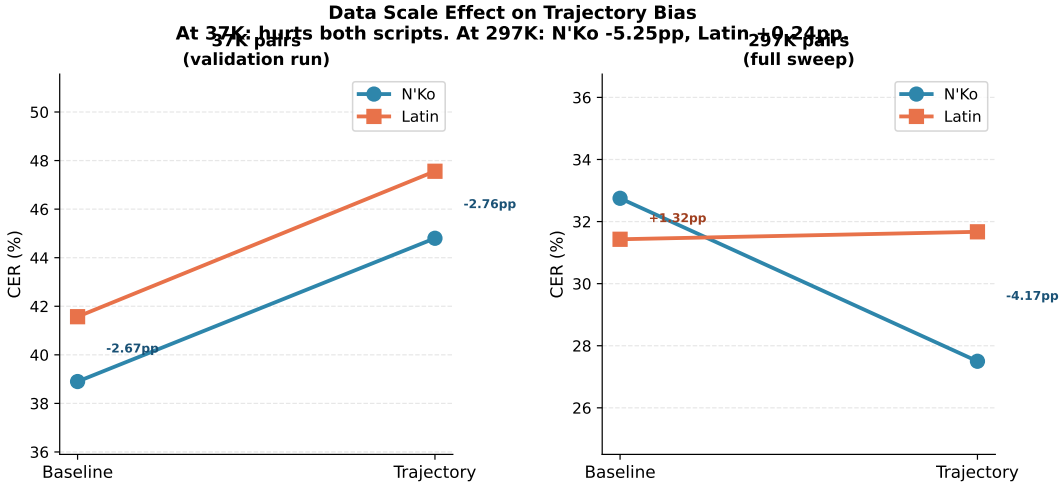

Figure 4: Data-scale context from the ASR experiments. Scaling changes both model behavior and claim strength, which is why artifact provenance is central.

quality transcription. Out-of-domain use requires the AGP governance layer described in the fourth paper.

14 Conclusion

The 20.57% result is usable, but only if stated with discipline. It is an archived N'Ko trajectory CTC checkpoint, trained under recorded settings on a 290,596-pair snapshot, with explicit scorer arithmetic. It is not the later TAR branch, not the TTT branch, and not AGP.

That bounded claim is still important. It shows that direct script-native N'Ko ASR reached a meaningful error regime on a large Bambara corpus. Combined with the metric argument from the previous paper, it gives the project a concrete scientific anchor: Manding ASR should be evaluated in a script that preserves the linguistic structure the system is supposed to recognize.

References

[1] 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.

[2] 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.

[3] 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.

Table 11: Minimal model-card fields for the N'Ko ASR anchor.

Field	Statement
Primary task	Direct acoustic-to-N'Ko character transcription for Manding speech.
Primary metric	Normalized N'Ko CER with prediction/reference row exports.
Known strengths	Script-native output, explicit scorer arithmetic, large retained corpus snapshot.
Known limitations	Limited out-of-domain evidence and no universal Latin comparison claim.
Unsafe uses	High-stakes transcription, unreviewed corpus expansion, or automatic normalization of uncertain N'Ko text.
Required governance	AGP-style row contracts, correction admissibility, and human review for uncertain or novel rows.