Anticipation Geometry Partition: Row-Level Governance for Script-Native N'Ko ASR Deployment

Mohamed Diomande

Final paper series, May 2026

Abstract

This paper defines the deployment layer of the N'Ko ASR project: Anticipation Geometry Partition (AGP). AGP is not the acoustic model that produced the archived 20.57% CER anchor. It begins after ASR. Its role is to convert trajectory and uncertainty signals into row-level decisions about correction, provenance, corpus admission, and deployment eligibility. The motivation is simple: a scalar CER number is not enough to build a trustworthy transcript corpus or a production speech system. A model can make local mistakes, a language model can propose fluent but false corrections, and out-of-domain data can contain music, overlapping speakers, dialect drift, visual context, and uncertain references.

AGP partitions transcript spans into stable, boundary, uncertain, and novelty states. Stable spans should usually remain unchanged. Boundary spans may admit local repairs when acoustic and textual evidence agree. Uncertain spans require stronger evidence or abstention. Novelty spans should be treated as data-discovery events rather than as opportunities for language-model rewriting. The formal unit is a row containing ASR hypothesis, reference when available, edit counts, trajectory summaries, partition labels, candidate correction, admissibility decision, provenance metadata, and deployment gates.

The paper also consolidates the deployment evidence that motivated AGP. Historical compositional generalization and vocabulary expansion experiments suggest that N'Ko degrades less than Latin on unseen-word utterances and retains a smaller residual gap after full-data training. Djoko soap-opera extraction created a much harder deployment substrate: 1,124 downloaded videos out of a 2,001-video channel, 32,826 audio segments, an initial 8,985-segment batch, 269 high-confidence consensus rows at a strict threshold, 6,625 speaker-labeled rows, 258 episodes, seven speakers, and five eligible speakers. The domain gap was severe; the first batch produced only about 8.8% meaningful output under a clean-read-speech model. The conclusion is that AGP is not optional polish. It is the governance structure needed before script-native ASR output can safely become search data, correction data, subtitle candidates, or TTS training material.

1 Introduction

The first three papers in this series establish the representation, metric, and ASR anchor. This final paper asks what happens after the model emits text. That question matters because ASR deployment is not simply a sequence of benchmark runs. A deployed system ingests new speakers, new domains, background music, overlapping dialogue, dialect variation, channel noise, spelling variation, and missing references. CTC-style ASR [2] and frozen Whisper features [3] can produce useful transcripts, but a scalar CER cannot decide whether a particular row should enter a corpus, be corrected by a language model, be routed to human review, or be excluded from TTS training.

Table 1: Research questions for AGP.

ID	Question	Evidence required
RQ1	Can trajectory summaries partition rows into useful uncertainty states?	Per-state row counts, CER deltas, and accepted-regression rates.
RQ2	Can correction be made conservative rather than merely fluent?	Accepted, rejected, and abstained edits with improvement/regression labels.
RQ3	Can out-of-domain data be converted into inspectable review material?	Source metadata, speaker metadata, confidence gates, and exclusion reasons.
RQ4	Can corpus and TTS eligibility be separated from search eligibility?	Deployment gates that distinguish review, search, correction, TTS, and exclusion.

AGP, the Anticipation Geometry Partition, is the project response. It turns the trajectory geometry introduced in the ASR paper into a post-ASR governance layer. The acoustic model produces a raw transcript and trajectory summaries. AGP assigns states to rows and spans, evaluates proposed corrections, records provenance, and decides what the row is safe for. It is deliberately conservative: improving surface fluency is not enough. The correction must be admissible under local evidence.

The paper has two goals. The first is formal: define AGP states, row contracts, and admissibility. The second is practical: document the deployment substrates and historical robustness experiments that made AGP necessary. These include unseen vocabulary experiments, vocabulary expansion, and Djoko out-of-domain extraction. None of these should be confused with the archived 20.57% anchor. They answer a different question: how does a script-native ASR project move from a benchmark number to a governed corpus-building system?

2 Research Questions and System Boundary

AGP is evaluated as a governance layer, not as an acoustic model. Its research questions are therefore about decision quality, provenance, correction safety, and deployment eligibility. A paper about AGP must not be scored only by asking whether the final transcript looks more fluent. Fluency can hide unsupported rewrites. AGP is successful when it admits improvements, rejects or abstains from risky changes, and preserves enough evidence for later review.

3 From Trajectory Geometry to Governance

The ASR model computes a seven-dimensional trajectory state:

$$z_t = (c_t, u_t, p_t, r_t, s_t, n_t, q_t) \in [0, 1]^7,$$

where the channels correspond to commitment, uncertainty, transition pressure, recovery margin, phase stiffness, novelty, and stability. During decoding, these states can bias attention. After

Table 2: System boundary: what AGP is and is not.

AGP is	AGP is not
A post-ASR row-governance layer.	The acoustic model that produced 20.57% CER.
A correction-admissibility policy.	A guarantee that all accepted text is perfect.
A provenance and deployment gate.	A replacement for human N'Ko orthographic authority.
A way to preserve uncertain evidence.	A license to train on every extracted row.

Table 3: Operational AGP states.

State	Governance interpretation
Stable	High commitment and stability, low uncertainty and transition pressure. Default action: preserve the transcript.
Boundary	Elevated transition pressure near a likely phoneme, syllable, word, or phrase boundary. Default action: allow constrained local repair.
Uncertain	High ambiguity or low confidence without a clear boundary explanation. Default action: require stronger evidence or abstain.
Novelty	Possible unseen word, speaker shift, domain shift, code-switch, or out-of-training-distribution pattern. Default action: preserve provenance and route to review or data discovery.

decoding, they can be summarized over spans and rows:

$$\bar{z}_{a:b} = \frac{1}{b-a+1} \sum_{t=a}^{b} z_t.$$

AGP uses such summaries to classify local transcript regions.

AGP's central design choice is that novelty is not treated as error by default. In low-resource ASR, novelty may be the most valuable part of the row: a new word, a speaker-specific pronunciation, or a domain-specific expression. A generic language-model correction layer might normalize it away. AGP should instead retain the evidence and mark the risk.

4 Pipeline Formalization

The AGP pipeline is a sequence of evidence-preserving transformations:

$$a_i \xrightarrow{\text{ASR}} \hat{y}_i \xrightarrow{\text{partition}} p_i \xrightarrow{\text{candidate}} c_i \xrightarrow{\text{admissibility}} d_i \xrightarrow{\text{gate}} G_i,$$

where a_i is an audio segment, $\hat{y}_i$ is the raw N'Ko hypothesis, p_i is the partition state, c_i is an optional correction candidate, d_i is the accept/reject/abstain decision, and G_i is the deployment gate. If a reference y_i exists, the row can also carry edit statistics

$$e_i = \text{EditDistance}(\hat{y}_i, y_i), \qquad e'_i = \text{EditDistance}(c_i, y_i).$$

Table 4: AGP row lifecycle.

Stage	Primary question	Output
ASR	What did the acoustic model emit?	Raw N'Ko hypothesis and confidence signals.
Partition	What kind of uncertainty is present?	Stable, boundary, uncertain, or novelty state.
Correction	Is there a local candidate change?	Candidate string and changed-span metadata.
Admissibility	Is the candidate defensible?	Accept, reject, or abstain with reason code.
Deployment gate	What may this row be used for?	Search, review, correction, TTS, or exclusion status.

The critical rule is that no transformation deletes the previous state. The raw ASR hypothesis, candidate correction, decision, and final gated text remain distinct fields. This prevents the project from losing the evidence needed to review whether correction helped or merely rewrote the row.

5 Row Contract

The formal unit is not a paragraph or a final transcript. It is a row. Each row should be inspectable from acoustic source to final decision.

Table 5: AGP row contract.

Block	Representative fields
Identity	row id, feature id, audio source id, segment start/end, split, script, mode, model id.
ASR evidence	raw hypothesis, reference when available, edit count, reference character count, character denominator, posterior summary.
Trajectory evidence	commitment, uncertainty, transition pressure, recovery margin, phase stiffness, novelty, stability, derived AGP state.
Correction candidate	proposed correction, source of proposal, changed spans, local confidence, prompt or rule version.
Admissibility	accept, reject, or abstain; reason code; partition state; evidence thresholds; regression flag when reference exists.
Provenance	retrieval sources, transliteration variants, normalized forms, speaker metadata, episode metadata, provenance score.
Deployment gate	search eligibility, corpus-training eligibility, TTS eligibility, overlap risk, music risk, speaker cleanliness, exclusion reason.

This contract prevents the common failure mode in which a correction layer silently rewrites a transcript and only the final string is preserved. For scientific use, the row must show what changed, why, and under which evidence.

Table 6: Admissibility policy by AGP state.

State	Default decision	Accept only if
Stable	Reject or preserve	Correction fixes a clear normalization artifact and does not alter acoustic content.
Boundary	Consider local accept	Candidate is small, local, and supported by boundary evidence or reference-aligned confusion patterns.
Uncertain	Abstain	Multiple independent signals converge and no high-risk rewrite is introduced.
Novelty	Abstain or route to review	Candidate is provenance-backed and does not erase a possible new lexical item or speaker feature.

6 Admissibility

Let r be a row, c a candidate correction, z the trajectory summary, and p the provenance bundle. AGP defines an admissibility function:

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

The function is asymmetric. A correction is accepted only when evidence supports it and the partition state permits it. Rejection is appropriate when the candidate changes a stable span without evidence, conflicts with the acoustic hypothesis, or reduces provenance. Abstention is appropriate when the evidence is insufficient, especially for uncertain and novelty spans.

The metric that matters most for AGP safety is not total accepted edits. It is accepted regressions. A correction system that improves some rows while frequently accepting worse edits is not safe for corpus construction. A full AGP benchmark should report proposed edits, accepted edits, rejected edits, abstentions, accepted improvements, accepted neutral edits, accepted regressions, rejected improvements, and per-partition CER deltas.

7 Partition Scoring

AGP should be evaluated with a decision matrix rather than a single aggregate CER. For row i, define $\Delta_i = e_i - e'_i$, where positive values mean the candidate improved the row. An accepted correction is an *accepted improvement* if $\Delta_i > 0$, an *accepted neutral edit* if $\Delta_i = 0$, and an *accepted regression* if $\Delta_i < 0$. Rejected candidates can be labeled the same way when a reference is available, which exposes whether the gate is too conservative. Rows without references should still carry provenance and partition labels but should not be counted as metric improvements.

8 Smoke Tests

The current AGP tests are preliminary but useful for defining expected behavior. They should not be presented as a final benchmark over the archived 20.57% test set.

Table 7: AGP benchmark accounting.

Quantity	Interpretation
Proposed edits	How often the correction layer tries to change ASR output.
Accepted edits	How often AGP permits a change.
Accepted improvements	Corrections that reduce edit distance when references exist.
Accepted regressions	Corrections that make the row worse; primary safety failure.
Rejected improvements	Useful edits blocked by the gate; primary conservatism cost.
Abstentions	Cases where evidence is insufficient for a safe decision.
Per-state CER delta	Whether stable, boundary, uncertain, and novelty rows behave differently.

Table 8: AGP smoke-test outcomes. These tests evaluate conservative correction behavior, not ASR model quality.

Test	CER movement	Correction behavior
Hand smoke test	14.29% to 4.76%	Two accepted improvements; no worse accepted edits.
Synthetic stress test	13.33% to 10.00%	Three improved and five neutral accepted cases.
Archived real slice	76.04% to 75.12%	One improved accepted case; no worse accepted cases.

The important pattern is conservatism. AGP is allowed to abstain. In fact, it should abstain often in uncertain or novelty regions. Its purpose is not to maximize the number of edited rows; its purpose is to admit only defensible changes.

9 Correction Benchmark Design

A full benchmark should contain at least four partitions. The first is a stable clean-read partition where corrections should be rare. The second is a boundary partition containing likely local segmentation, elision, or character-boundary errors. The third is an uncertain partition with low confidence, overlapping speech, or noisy acoustics. The fourth is a novelty partition containing unseen names, domain-specific words, dialectal variants, or speaker-specific forms. Each partition should be scored separately because the safe action differs by state.

10 Compositional Generalization

The deployment story also includes historical robustness experiments. In a compositional generalization setup, models trained on seen vocabulary were evaluated on utterances containing unseen words. The hypothesis was that N'Ko should degrade less because unseen words can be composed

Table 9: Target benchmark design for AGP correction.

Partition	Expected gate behavior	Required report
Stable	Preserve most rows; accept only normalization repairs.	False rewrite rate and accepted-regression rate.
Boundary	Permit small local fixes.	Local edit precision and CER delta.
Uncertain	Abstain unless evidence converges.	Abstention rate and unsafe-accept rate.
Novelty	Route to review or data discovery.	Novelty preservation and human-review yield.

Table 10: Historical compositional-generalization evidence.

Script	Seen-word CER	Unseen-word CER	Gap
N'Ko	16.09%	53.90%	+37.81pp
Latin	15.05%	56.51%	+41.46pp

from known phoneme-character units.

N'Ko has a 3.65 percentage-point smaller generalization gap in this historical experiment. The result is consistent with the script-advantage hypothesis: a transparent script makes unseen words more compositional. It is not a substitute for a future matched evaluation.

11 Vocabulary Expansion

The vocabulary-expansion experiment asks whether full-data training recovers unseen-word performance and whether any script advantage remains.

This result matters for deployment because vocabulary expansion is a corpus-building problem. A system should improve as new rows become trustworthy. AGP provides the filtering mechanism for deciding which rows are trustworthy enough to enter future training or review loops.

12 Djoko Domain Transfer

The Djoko deployment effort is the harshest test environment in the project. The source is a Bambara-language soap-opera domain rather than clean read speech. The data include music, overlapping speech, varying speakers, likely dialect variation, episode structure, and no reliable subtitle track. It is not the same distribution as the 20.57% anchor.

The extraction record includes 1,124 downloaded videos from a 2,001-video channel and 32,826 thirty-second audio segments. A first batch included 8,985 segments and 8,985 ASR transcriptions. Consensus filtering at a strict threshold produced 269 high-confidence rows; speaker and episode processing produced 6,625 speaker-labeled rows, 258 episodes, seven speakers, and five eligible speakers. The first batch had only about 8.8% meaningful output under a clean-read-speech model, showing a severe domain gap.

The practical conclusion is that deployment needs gates. A row can be useful for search, candidate review, correction training, TTS training, or none of these. The decision depends on

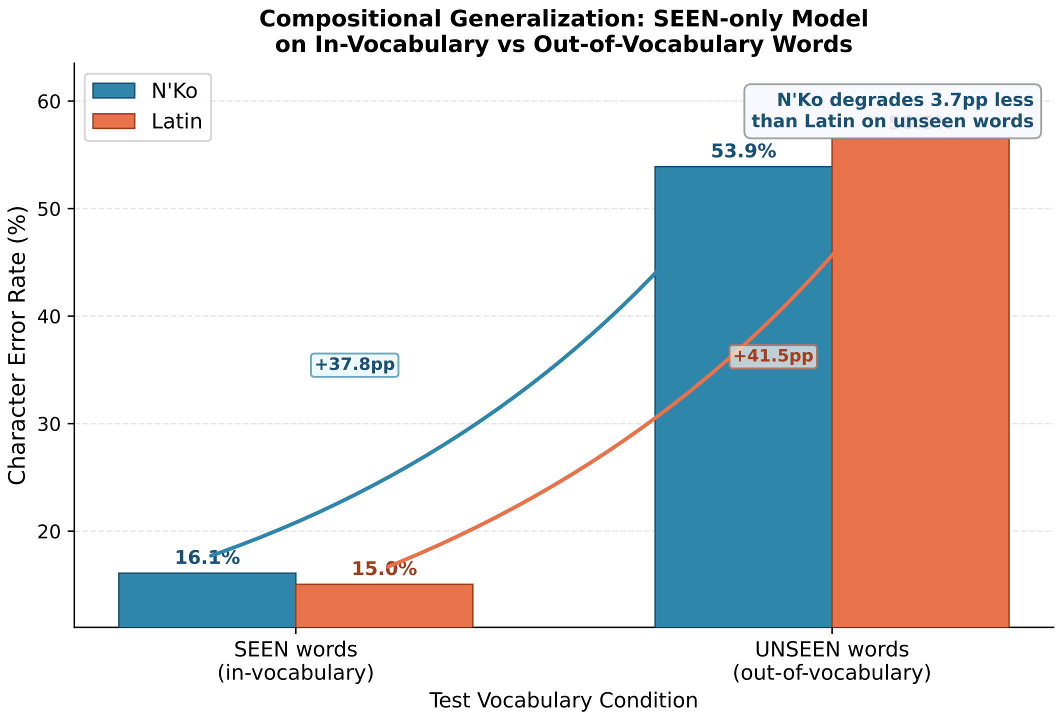

Figure 1: Experiment F compositional-generalization figure. The result is used as deployment motivation, not as the canonical 20.57% anchor.

evidence, not on the existence of an ASR string.

13 Failure Mode Taxonomy

The deployment data expose failure modes that are invisible in clean benchmark summaries. Music can create confident nonsense. Overlapping speakers can produce a hybrid transcript that should not be assigned to either speaker. Dialect drift can look like an error to a model trained on a narrower reference distribution. Code-switching can make a purely script-native gate overconfident. Episode metadata can leak duplicates into training and evaluation if not handled carefully.

14 AGP and TTS Eligibility

TTS training introduces a stricter requirement than search. A search index may tolerate uncertain transcript candidates if provenance is retained. TTS training should not ingest rows with overlapping speakers, music, severe noise, unclear speaker identity, or untrusted text. AGP therefore carries deployment gates:

$$G(r) = \{\text{search}, \text{review}, \text{correction}, \text{TTS}, \text{exclude}\}.$$

The same row can be eligible for review but excluded from TTS. This separation is essential for low-resource projects, where the temptation to use every available row is strong.

Table 11: Historical vocabulary-expansion evidence.

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

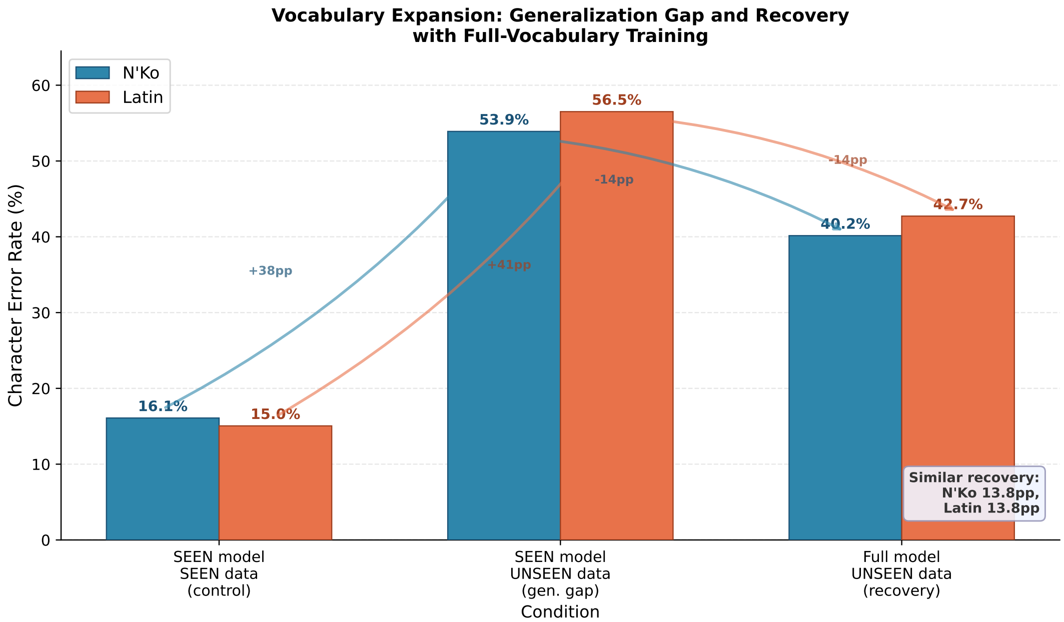

Figure 2: Experiment H vocabulary-expansion figure. Full data helps both scripts, but the residual unseen-word gap remains smaller for N'Ko in the recorded experiment.

15 Human Review and Community Authority

AGP is a machine gate, not an orthographic authority. Human review remains necessary for tone policy, dialectal variants, register, names, religious or formulaic phrases, and acceptable spelling variation. The row contract is designed so a reviewer can see the raw audio source, raw ASR hypothesis, candidate correction, partition state, and reason code. That makes review cheaper and more accountable: the reviewer is not asked to trust the model, only to adjudicate a visible evidence bundle.

16 Data Lifecycle

Rows should move through a staged lifecycle. A raw row may enter search if it is useful for retrieval and clearly marked as machine-generated. A reviewed row may enter the correction set if its evidence and human decision are preserved. A cleaner single-speaker row may enter TTS only after stricter audio and speaker gates. Rows with music, overlap, weak provenance, or novelty

Table 12: Djoko deployment substrate.

Artifact	Count
Channel videos available	2,001
Videos downloaded	1,124
Audio segments extracted	32,826
First-batch segments/transcriptions	8,985
Strict high-confidence consensus rows	269
Speaker-labeled rows	6,625
Episodes represented	258
Speakers detected in speaker table	7
Eligible speakers under gate	5
Approximate meaningful-output rate in first batch	8.8%

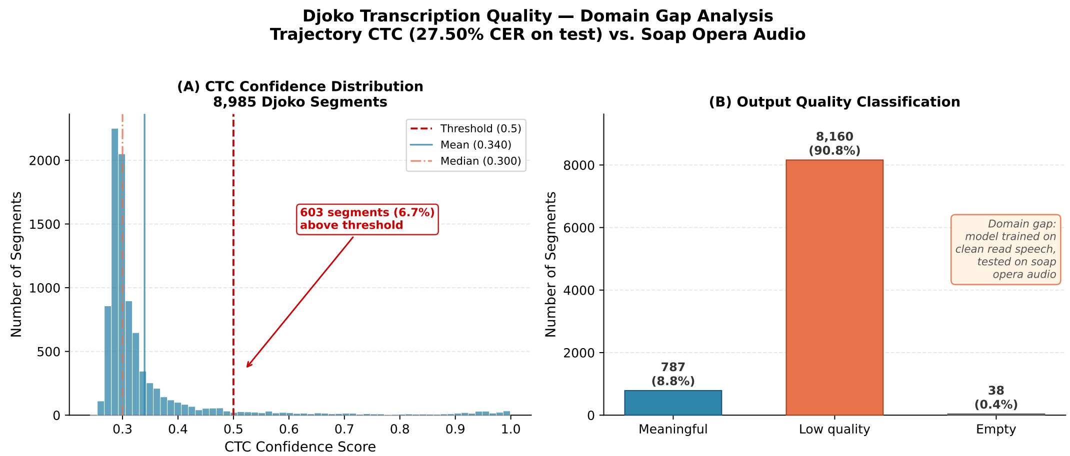

Figure 3: Djoko quality-control figure. The figure illustrates why row-level filtering is necessary before using out-of-domain ASR output as training data or subtitle material.

uncertainty should remain in review or exclusion states until additional evidence is available. This staged lifecycle lets the project stop spending compute while still preserving the value of extracted data for later review, correction, and deployment work.

17 Limitations

AGP is a downstream governance architecture with smoke tests and deployment substrates. It should be evaluated separately from the 20.57% ASR anchor: the ASR number measures decoding quality, while AGP measures whether transcript rows can be safely searched, reviewed, corrected, retained, or excluded. A complete AGP evaluation would run on the prediction/reference row set and report accepted improvements, accepted regressions, rejected improvements, and abstentions by partition.

The historical ExpF and ExpH results are useful but not a substitute for a new matched evaluation. They should be presented as deployment motivation. The Djoko data are also not ground-truth CER evidence; they are out-of-domain corpus-building evidence. The severe 8.8%

Table 13: Deployment failure modes and AGP response.

Failure mode	Risk	AGP response
Background music	ASR may hallucinate speech-like text.	Music-risk flag and TTS exclusion.
Overlapping speakers	Transcript cannot be assigned cleanly.	Speaker-overlap flag and review-only gate.
Dialect or register shift	Valid speech may be misclassified as error.	Novelty state and community review.
Named entities	Language model may normalize away new lexical items.	Preserve raw hypothesis and require provenance.
Duplicate episodes	Train/test leakage or inflated coverage.	Episode-level provenance and deduplication.
Low-confidence consensus	False agreement across weak models.	Threshold review and abstention by default.

meaningful-output rate should be presented honestly because it defines the real deployment difficulty.

Finally, AGP needs community review. A conservative correction gate can prevent many machine-learning errors, but orthographic authority, tone policy, punctuation, register, and acceptable spelling variation require Manding and N'Ko expertise, not only model confidence. This is especially important because N'Ko is a living scholarly and community script, not merely a model output alphabet [1, 4].

18 Conclusion

AGP is the bridge from benchmark ASR to governed deployment. It does not produce the archived 20.57% CER result, and it should not be described that way. Its role is downstream: preserve row-level evidence, partition uncertainty, constrain correction, and decide which outputs can safely become corpus material.

The broader conclusion of the final paper series is that script-native ASR is an infrastructure problem. Representation, metrics, decoding, correction, and deployment gates all interact. Without AGP, a project can obtain a promising CER number and still build an unsafe corpus. With AGP, the project has a path to stop spending compute now while preserving the exact evidence needed for later correction, review, and deployment.

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] 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.

[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.

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