

# State Tracking vs. Attention in a Morphologically Rich Low-Resource Language: A Controlled Bangla Case Study with Mamba-3

Sahil Al Farib<sup>†</sup>, Momota Ahsana Meem<sup>†</sup>, Sheikh Redwanul Islam<sup>†</sup>,  
Khan Md Shams Arefin<sup>†</sup>, Azizur Rahman Anik<sup>†</sup>

<sup>†</sup>Department of Computer Science & Engineering  
United International University

{sfarib222186, mmeem222254, sislam222142, karefin222320}@bscse.uiu.ac.bd, azizur@cse.uiu.ac.bd

## Abstract

State space models such as Mamba-3 claim improved state tracking over attention at sub-quadratic cost, but this claim is validated almost entirely on English. We test whether it transfers to a morphologically rich, low resource setting by training parameter and token matched Transformer, Mamba-3, and hybrid language models from scratch on Bangla (24.5M nonembedding parameters, 1B tokens, five seeds per architecture). We evaluate on a new suite of 4,790 native speaker reviewed minimal pairs targeting Bangla subject verb person and honorific register agreement, and assess every comparison with Wilson confidence intervals and cross seed Welch  $t$ -tests. Across five seeds, the Mamba family achieves significantly lower perplexity than the Transformer (Welch  $p < 0.001$ ). Transformer SVA accuracy declines from 91.3% to 87.5% as the intervener lengthens (paired  $p = 0.01$ ), whereas Mamba-3 shows no significant decline (83.3% to 81.5%,  $p = 0.22$ ); however, the direct difference between their distance slopes is not significant ( $p = 0.15$ ). A third finding is methodological rather than architectural: most single run between architecture agreement gaps, including a honorific register advantage significant under a within run McNemar test ( $p = 0.0015$ ) do not survive replication once cross seed variance is included (Welch  $p = 0.28$ ), because Mamba-3’s agreement accuracy is markedly more seed sensitive than the Transformer’s (cross seed standard deviation up to  $\pm 8.4$  vs.  $\pm 3.5$ ). At this scale, two seed comparisons of state space models can manufacture findings that vanish under replication. Perplexity does not predict distance behavior: the architectural signal we identify concerns robustness to distance and initialization, not a uniformly “better” architecture. We release code, probes, all fifteen checkpoints, and the analysis scripts.

## 1 Introduction

Subquadratic sequence models have reemerged as credible alternatives to attention based Transformers. Mamba (Gu and Dao, 2024) and its successors argue that a selective state space recurrence matches attention on language modeling at lower inference cost; Mamba-3 (Lahoti et al., 2026) further claims improved state tracking via a more expressive recurrence, complex/rotary state updates, and MIMO projections. Like most architecture papers, its empirical validation is English centric.

Bangla (Bengali; spoken by over 230 million people) provides a sharp typological test of these claims. Bangla is an SOV, morphologically rich language whose finite verbs agree with the subject in person and honorific register not number and this agreement is required to hold across long intervening spans and even across a covert (pro-drop) subject realized in a subsequent sentence. Whether a fixed size compressed recurrent state helps track such long distance, register sensitive dependencies, or is instead disadvantaged under data scarcity relative to attention’s exact recall, has not been tested for Bangla or, to our knowledge, any South Asian language.

We conduct a controlled, iso-parameter, iso-token comparison of a Transformer, Mamba-3, and a hybrid architecture with the same methodology the Mamba papers use to validate their own claims (Gu and Dao, 2024; Dao and Gu, 2024), with five seeds per architecture, enabling between-architecture claims to be tested with proper cross seed statistics rather than single run comparisons.

## Contributions.

1. **Resource:** A reusable probe suite of 4,790 native speaker reviewed Bangla minimal pairs targeting person and honorific register agreement, distance-binned, with agreement attraction and cross sentence pro drop conditions

that have no direct English analogue.

2. **Architectural finding:** A controlled three way comparison (Transformer, Mamba-3, hybrid) at matched scale on Bangla, five seeds each, trained from scratch, revealing an SSM perplexity advantage and a distance asymmetry, attention’s agreement accuracy degrades with subject-verb distance while the state space recurrence’s does not.
3. **Methodological finding:** At this scale, Mamba-3’s agreement accuracy is substantially more seed sensitive than the Transformer’s, such that most single run agreement gaps between architectures, including one significant under a within run test, do not replicate across five seeds.

## 2 Related Work

**Targeted syntactic evaluation.** Minimal pair benchmarks score whether a language model assigns higher probability to a grammatical sentence than to a minimally different ungrammatical one (Linzen et al., 2016; Warstadt et al., 2020). MultiBLiMP 1.0 (Jumelet et al., 2026) extends this paradigm to 101 languages, but its automatic extraction pipeline yields only 21 Bengali pairs which is too few to analyze by grammatical condition, and none covering honorific register or cross sentence agreement. Our hand constructed, native speaker reviewed suite is format compatible with MultiBLiMP while filling this gap with 4,790 pairs spanning four agreement conditions.

**State space and hybrid sequence models.** Selective state space models like Mamba (Gu and Dao, 2024), Mamba-2 (Dao and Gu, 2024), and Mamba-3 (Lahoti et al., 2026) achieve strong English language modeling at subquadratic inference cost. Architectures that interleave a small number of attention layers into an otherwise recurrent backbone, such as Samba (Ren et al., 2025) and Jamba (Lieber et al., 2024), are a standard recipe for recovering in-context precision while retaining most of the efficiency of recurrence; we adopt this recipe for our hybrid architecture. Controlled comparisons find that attention retains an advantage on exact recall and copying (Arora et al., 2023; Jelassi et al., 2024), that SSMs are no more expressive than Transformers for inherently sequential state tracking (Merrill et al., 2024), and that at 8B scale Mamba-based models are competitive

with Transformers while hybrids outperform both (Waleffe et al., 2024). These studies are English or synthetic; we ask the corresponding question for a morphologically rich, low-resource language.

**Low resource Bangla language modeling.** Prior work on Bangla language modeling largely adapts existing architectures to the language rather than testing architectural hypotheses under controlled conditions. BanglaBERT (Bhattacharjee et al., 2022) pretrains an encoder only model on a 27.5GB web crawl for natural language understanding tasks; BanglaT5 (Bhattacharjee et al., 2023) similarly pretrains a sequence to sequence Transformer on the same corpus scale for natural language generation. Both establish strong Bangla specific baselines by scaling a single, fixed architecture rather than comparing architectural families under matched compute. In contrast, we hold architecture under direct experimental control, matching non-embedding parameters and training tokens exactly across three architectures to test a specific, falsifiable claim about state tracking, rather than optimizing one architecture for downstream task accuracy. Vats et al. (2025) apply state space models to question answering in several Indic languages and compare SSM architectures on downstream accuracy; we instead compare SSMs against a matched Transformer on targeted agreement phenomena.

**Seed sensitivity and replicability in NLP evaluation.** Prior work has shown that neural language model performance can vary substantially across random seeds even under fixed hyperparameters: Dodge et al. (2020) find that BERT fine tuning outcomes on GLUE tasks depend heavily on both weight initialization and training data order, and Bethard (2022) finds that more than half of a sample of 85 ACL Anthology papers use random seeds in ways that risk overstating between model differences. Our results extend this concern from fine tuning to architecture comparison at pretraining scale, and show that the risk is asymmetric across architecture families: Mamba-3’s cross seed variance on our agreement probes is 2–3× larger than the Transformer’s, so a seed insensitive evaluation protocol is not sufficient on its own which is the reason seed sensitivity must be assessed per architecture.

### 3 Method

#### 3.1 Architectures

All three models are decoder only causal LMs matched on **non embedding** parameters (approximately 24.5M, matched to within 1.3%; embeddings identical and excluded). Each is a pre-norm RMSNorm (Zhang and Sennrich, 2019) residual tower with tied embeddings; the models differ *only* in the per-layer token mixer:

- **Transformer:** Llama style (Touvron et al., 2023), with multi head attention using rotary position embeddings (Su et al., 2024) and a SwiGLU MLP (Shazeer, 2020).
- **Mamba 3:** the official Mamba3 SISO block (Lahoti et al., 2026) from state spaces/mamba.
- **Hybrid:** the Mamba-3 tower with rotary self attention at 2 of 15 layers (Samba/Jamba-style).

#### 3.2 Data and Training

We train from scratch on the Bangla (ben\_Beng) split of FineWeb-2 (Penedo et al., 2025) with a 32k SentencePiece BPE vocabulary (Kudo and Richardson, 2018), 1.0B tokens per run (approximately 41 tokens/parameter), and a held-out split for perplexity. Optimization is identical across models and seeds: AdamW (Loshchilov and Hutter, 2019), cosine schedule (peak  $8 \times 10^{-4}$ , 1% warmup), batch  $32 \times$  gradient accumulation  $2 \times$  sequence length 1024, and bf16. Each model is trained with **five seeds** (1337, 2024, 3419, 5150, 8888); approximately 3 A100-hours per run, 15 runs total.

#### 3.3 Probe Suite

We construct 4,790 minimal pairs from hand-written Bangla conjugation tables (10 verbs  $\times$  3 tenses  $\times$  6 person/register cells) and reusable frames. **A native speaker reviewed the lexicon**—verb morphology, temporal-adverb compatibility, intervener naturalness, and discourse coherence—and every flagged issue was corrected before scoring. Table 1 summarizes the four conditions.

#### 3.4 Scoring and Significance

A model is **correct** on a pair if it assigns higher total log-probability to the grammatical sentence

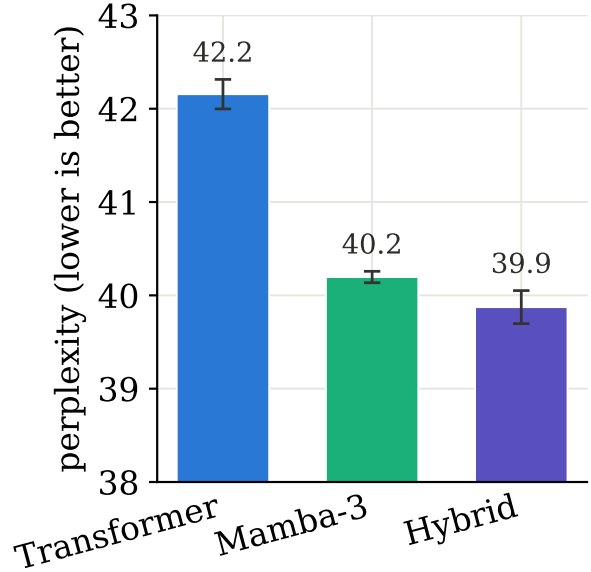


Figure 1: Held out perplexity. Bars are the five seed mean, error bars show the cross seed standard deviation, and dots show individual seeds.

(BLiMP protocol; Warstadt et al., 2020). We report two kinds of uncertainty because they answer different questions: **Wilson 95% confidence intervals** (pooled over seeds) capture finite-probe uncertainty and are used for *within-model* trends (e.g., accuracy across distance); **cross seed Welch  $t$  tests** on the five per seed accuracies capture initialization uncertainty and are the proper test for *between architecture* differences.

### 4 Results

All numbers are means over five seeds; “ $\pm$ ” is the cross-seed standard deviation. Figures 1–3 summarize held out perplexity, distance conditioned SVA accuracy, and probe wise agreement accuracy, respectively.

#### 4.1 Perplexity: A Robust SSM Advantage

Both SSM-family models beat the Transformer (Table 2), and every pairwise difference is significant (Welch across seeds: Transformer vs. Mamba-3 and vs. hybrid,  $p < 0.001$ ; Mamba-3 vs. hybrid,  $p = 0.02$ ). **General language modeling quality favors recurrence**, the opposite of the naive expectation that attention’s exact recall should help most. The ordering (hybrid < Mamba-3 < Transformer) matches English findings that Mamba-family models are competitive with Transformers at matched size (Gu and Dao, 2024) and that hybrids outperform both (Waleffe et al., 2024).

| Condition  | Pairs | What it tests                                                                                                                 |
|------------|-------|-------------------------------------------------------------------------------------------------------------------------------|
| SVA        | 3,300 | Subject–verb person/register agreement; interveners binned none/short/medium/long                                             |
| Attraction | 1,190 | A competing-person pronoun inside the intervener lures the wrong agreement                                                    |
| Honorific  | 210   | তিনি/উনি ( <i>tini/uni</i> , honorific “he/she”) vs. সে ( <i>se</i> , ordinary “he/she”) register agreement (single sentence) |
| Discourse  | 90    | Register set in sentence 1; agreement tested on a pro-drop sentence 2                                                         |

Table 1: Composition of the Bangla agreement probe suite.

| Model       | Perplexity ↓        |
|-------------|---------------------|
| Transformer | 42.16 ± 0.16        |
| Mamba-3     | 40.20 ± 0.06        |
| Hybrid      | <b>39.87 ± 0.18</b> |

Table 2: Held out perplexity (mean ± cross seed standard deviation).

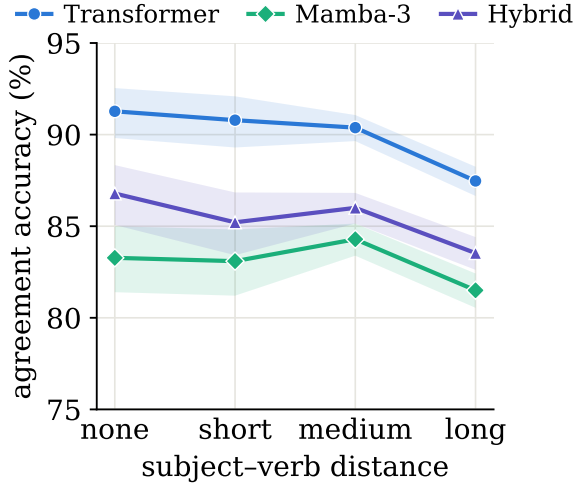


Figure 2: SVA accuracy vs. subject verb distance. Lines are the five seed mean and shaded bands show the Wilson 95% confidence interval.

## 4.2 The Headline: Agreement vs. Distance

The clearest architectural signal is not a level difference but a **slope** difference. Table 3 bins SVA pairs by subject verb distance and reports the mean over five seeds.

The Transformer’s decline is significant (Wilson confidence intervals: none [89.8, 92.5] vs. long [86.6, 88.2], non-overlapping) and occurs in **every seed**. Mamba-3 shows **no significant change** with distance (none [81.4, 85.0] vs. long [80.5, 82.4], overlapping) and even *rises* with distance in two seeds. The hybrid, which contains attention, declines like the Transformer. Thus, attention’s benefit is a *local* one that erodes with distance; the recurrent state does not decay in the same way. This dissociation is the paper’s most robust agreement finding. The Transformer’s decline is significant

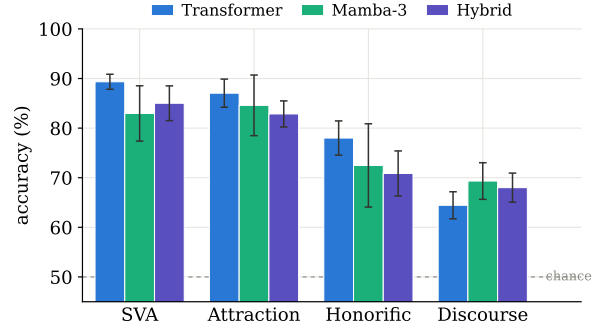


Figure 3: Accuracy by probe type. Bars are the five seed mean, error bars show the cross seed standard deviation, and dots show individual seeds; note Mamba-3’s wide spread.

(Wilson confidence intervals: none [89.8, 92.5] vs. long [86.6, 88.2], non-overlapping), occurs in every seed, and holds under a seed-aware paired test across the five seeds ( $t = 4.33$ ,  $p = 0.01$ ). Mamba-3 shows no significant change with distance (none [81.4, 85.0] vs. long [80.5, 82.4], overlapping; paired  $p = 0.22$ ) and rises with distance in two seeds. The hybrid declines on average, in four of five seeds, but not significantly (paired  $p = 0.14$ ). A direct test of the architecture × distance interaction, comparing per-seed distance slopes, finds the Transformer’s slope steeper than Mamba-3’s (mean  $-1.18$  vs.  $-0.41$  points per distance bin), but the difference is not significant (Welch  $p = 0.15$ ). Attention’s agreement accuracy thus erodes with distance where the recurrent model’s does not, but five seeds are not enough to establish that the two slopes differ. Results on recall and copying (Arora et al., 2023; Jelassi et al., 2024) would suggest that attention’s advantage grows as the agreement controller moves further back; at this scale we see, if anything, the opposite (the Transformer’s lead over Mamba-3 narrows from 8.0 to 6.0 points), though not significantly.

## 4.3 Seed Sensitivity Dominates Local Agreement Gaps

The Transformer has the higher mean on the three local probes (Table 4), but **cross seed Welch  $t$ -**

| Model       | None | Short | Medium | Long | None→long                      |
|-------------|------|-------|--------|------|--------------------------------|
| Transformer | 91.3 | 90.8  | 90.4   | 87.5 | −3.8 (declines in all 5 seeds) |
| Mamba-3     | 83.3 | 83.1  | 84.3   | 81.5 | −1.8 (rises in 2 of 5 seeds)   |
| Hybrid      | 86.8 | 85.2  | 86.0   | 83.5 | −3.3 (declines)                |

Table 3: SVA accuracy by subject verb distance. Values are percentages averaged over five seeds.

| Model       | SVA        | Attraction | Honorific  | Discourse         |
|-------------|------------|------------|------------|-------------------|
| Transformer | 89.3 ± 1.5 | 87.0 ± 2.8 | 78.0 ± 3.5 | 64.4 ± 2.7        |
| Mamba-3     | 83.0 ± 5.6 | 84.6 ± 6.1 | 72.5 ± 8.4 | <b>69.3 ± 3.7</b> |
| Hybrid      | 85.0 ± 3.5 | 82.9 ± 2.6 | 70.9 ± 4.5 | 68.0 ± 2.9        |

Table 4: Accuracy by probe condition (mean ± cross seed standard deviation).

tests tell a more cautious story:

- **SVA:** Transformer > Mamba-3 only *marginally* ( $p = 0.08$ ).
- **Attraction:** Transformer vs. Mamba-3 is **not significant** ( $p = 0.50$ ).
- **Honorific:** Transformer vs. Mamba-3 is **not significant** ( $p = 0.28$ ), the apparent 5.5 point gap is swamped by Mamba-3’s  $\pm 8.4$  cross seed spread. A two seed run had suggested a clean Transformer advantage here; five seeds show it does not replicate.
- **Discourse** (the hardest, cross sentence probe): Mamba-3 > Transformer, *marginally significant* ( $p = 0.07$ ) and recurrence’s one directional agreement win.

The dominant reason the local gaps fail to reach significance is **Mamba-3’s seed variance** (SVA  $\pm 5.6$ , attraction  $\pm 6.1$ , honorific  $\pm 8.4$ ) against the Transformer’s tight  $\pm 1.5$ – $3.5$ . At this scale, the SSM’s morphosyntactic competence is much more initialization dependent, itself a finding a two seed study would have missed.

Concretely, a *single seed* item level McNemar test finds the Transformer significantly ahead of Mamba-3 on SVA, attraction, and honorific (all  $p < 0.01$  for seed 1). That these same gaps are non-significant under cross seed testing is a direct, quantified illustration of how within run significance overclaims when a model is seed sensitive: the honorific “advantage” (McNemar  $p = 0.0015$  in one seed) does not survive replication (Welch  $p = 0.28$  across five).

#### 4.4 The Second Person Intimate Artifact

The second person intimate register (তুই (*tui*, intimate “you”); forms like করিস (*koris*, “(you) do”)) is very rare in web text and behaves pathologically: Mamba-3’s accuracy on this cell ranges from **12% to 84% across seeds** (mean 43%), and the hybrid from 8% to 79%, while the Transformer is stable (80%–87%). This is a frequency/tokenization artifact, models default to the frequent (wrong) form, not state tracking, and it also illustrates the SSM’s seed sensitivity in the extreme. We exclude it from the distance analysis in Section 4.2 and report it separately.

#### 4.5 Error Analysis

Two architecture independent patterns recur (per-tense and per-person breakdowns are in the released results/ directory). **Past tense** is the hardest across all models. Bangla simple-past forms are shorter and more syncretic across register than the explicitly marked future. So this reflects the phenomenon, not the architecture. **Register errors are asymmetric:** all models agree almost perfectly with overt honorific subjects (তিনি/উনি) but err on ordinary third person subjects, i.e., they *over apply* honorific agreement; this asymmetry is largest, and most seed variable, for Mamba-3.

#### 4.6 Tokenization Sanity Check

Because pair scores are sums of log-probabilities without length normalization, a grammatical sentence that tokenizes into fewer subwords can be advantaged. Across conditions, 20–37% of pairs have different subword counts between the grammatical and ungrammatical sentences. For seed 1 SVA, accuracy on these pairs is 98.3% vs. 70.3% for the Transformer and 97.6% vs. 37.2%

| Model       | None (all) | Long (all) | None (matched) | Long (matched) |
|-------------|------------|------------|----------------|----------------|
| Transformer | 91.8       | 88.9       | 93.3           | 90.3           |
| Mamba-3     | 78.2       | 79.6       | 81.2           | 82.6           |
| Hybrid      | 87.0       | 82.3       | 89.9           | 83.8           |

Table 5: Seed-1 SVA accuracy on all pairs and on length-matched pairs only.

for Mamba-3 when the grammatical sentence is shorter vs. longer. However, length-matched pairs constitute 62–64% of every distance bin, and restricting evaluation to these pairs preserves the distance pattern and probe-level architecture ordering. The check uses seed 1, the only seed with saved per-pair outputs.

## 5 Discussion

The five seed evaluation reframes the initial hypothesis, that Mamba-3’s more expressive recurrence should track morphosyntactic dependencies better than attention and therefore win on agreement, into three claims of different evidentiary strength:

**1. Recurrence yields significantly lower perplexity.** Mamba-3 and the hybrid achieve significantly lower held out perplexity than the Transformer across all five seeds ( $p < 0.001$  for both SSM family comparisons,  $p = 0.02$  for Mamba-3 vs. hybrid; Table 2). This ranking is tight and consistent, but as the next claim shows, it does not predict distance robustness or seed stability on agreement. Language modeling quality and targeted syntactic competence are separable here.

**2. Attention and recurrence diverge in how agreement scales with distance.** The clearest architectural signal is a difference in slope, not in level. The Transformer’s agreement accuracy declines from 91.3% with no intervener to 87.5% with a long intervener, a decline present in every seed and significant across seeds (paired  $p = 0.01$ ; Table 3, Figure 2). Mamba-3 shows no significant decline (83.3% to 81.5%, paired  $p = 0.22$ ) and rises with distance in two of five seeds; the hybrid, which retains two attention layers, declines on average but not significantly. The direct contrast between the Transformer’s and Mamba-3’s distance slopes is in the predicted direction but not significant (Welch  $p = 0.15$ ). We therefore read this as a within-architecture asymmetry, consistent with attention’s benefit being local and the recurrent state

not decaying in the same way, rather than as an established between-architecture difference, which would require more seeds or larger probe sets to confirm.

**3. Local agreement superiority is not established at this scale.** The Transformer has the higher mean on all three local probes (Table 4), but cross seed Welch  $t$ -tests find these gaps non significant or marginal (SVA:  $p = 0.08$ ; attraction:  $p = 0.50$ ; honorific:  $p = 0.28$ ), since Mamba-3’s cross seed standard deviation ( $\pm 5.6$  to  $\pm 8.4$ ) is two to three times the Transformer’s ( $\pm 1.5$  to  $\pm 3.5$ ). Recurrence’s one directional advantage, the cross sentence discourse probe, is itself only marginally significant ( $p = 0.07$ ). This is a genuine finding, not a null result to explain away. At roughly 25M parameters, Mamba-3’s morphosyntactic competence is substantially more sensitive to initialization than the Transformer’s, and this sensitivity, rather than architecture, dominates the local agreement gaps a single seed study would report.

**A methodological illustration.** A single seed, item level McNemar test finds the Transformer significantly ahead of Mamba-3 on SVA, attraction, and honorific agreement (all  $p < 0.01$ , seed 1), including an apparently clean honorific advantage ( $p = 0.0015$ ). None of these gaps survive replication across five seeds (Welch  $p = 0.28$  for honorific). This is a direct, quantified demonstration, not merely a caveat, of how within run significance testing overstates between architecture differences when one family is more seed sensitive than the other. Seed sensitivity cannot be assumed comparable across architectures, and single or dual seed comparisons of state space models against attention based baselines carry an elevated risk of reporting findings that do not replicate.

**Synthesis.** Recurrence wins perplexity, robustly. Attention’s agreement accuracy degrades with distance while recurrence’s does not, though the direct contrast is not yet significant. Whether recurrence is locally more accurate at agreement is, at this scale, underdetermined. The Transformer leads numerically on local probes and Mamba-3 leads numerically on the cross sentence probe, and neither difference is significant once initialization variance is included. The architectural signal this paper supports concerns distance robustness and initialization variance, not a uniformly superior architecture.

## Limitations

We frame this work as a controlled pilot for a low resource and typological venue, not as a scaling study or a claim that either architecture is uniformly superior.

**Scale and breadth.** All results come from a single model size (roughly 24.5M non embedding parameters), a single training budget (1.0B tokens), and a single language. Existing Bangla pretrained models such as BanglaBERT (Bhattacharjee et al., 2022) and BanglaT5 (Bhattacharjee et al., 2023) are substantially larger than our models, and capabilities that emerge only at larger scale would not be visible here. Whether the distance dissociation described in Section 4.2 and Mamba-3’s elevated seed sensitivity described in Section 4.3 persist at larger parameter or data scale, and whether they generalize to other agreement marking South Asian languages, remains untested. The training and probing pipeline is size and language agnostic by design, so extending these tests should be comparatively inexpensive, but we do not claim the architectural signal reported here holds at every scale.

**Seeds.** Results are averaged over five seeds per architecture. We report perplexity and the Transformer’s within-model distance decline as significant, marginal effects such as SVA and discourse as marginal, and non significant between architecture comparisons such as attraction and honorific explicitly as non significant, rather than treating them as directional findings. Five seeds substantially reduce, but do not eliminate, the risk that this particular seed set is unrepresentative, especially for Mamba-3 given its observed variance.

**Probe construction.** The 4,790 minimal pairs are generated from a templated, hand written lexicon (10 verbs across 3 tenses and 6 person or register cells) rather than sampled from a corpus. A single native speaker reviewed the lexicon for morphological, temporal, and discourse well formedness; we did not collect a second annotation pass or an inter annotator agreement statistic. The honorific and discourse conditions are comparatively small, at 210 and 90 pairs respectively, which limits the precision of subgroup estimates for those two conditions. Scoring is not length-normalized, and 20–37% of pairs differ in subword count between the grammatical and ungrammatical sentence; the reported patterns hold on length-matched pairs (Sec-

tion 4.6), but this check covers one seed, and length-normalized scoring is a natural follow-up.

**State tracking beyond agreement.** Our probe suite targets subject verb person and honorific register agreement. Other phenomena commonly associated with state tracking, such as coreference resolution, retrieval over long contexts, and narrative memory, are not tested here and may show different architectural behavior.

**SISO only.** We evaluate the single input single output variant of the official Mamba-3 block. The MIMO variant was not evaluated because the corresponding kernels were unavailable in the released package at the time of this study. Mamba-3’s improved state tracking claims are motivated in part by its MIMO projections, so our results should be read as a test of the SISO recurrence specifically, rather than of the full architecture as originally proposed.

## 6 Reproducibility

Code, configuration files, the reviewed lexicon and probe generator, per-seed results, and all fifteen model checkpoints will be released publicly upon publication. All figures and every confidence interval and significance test in this paper regenerate directly from the released results file without requiring a GPU or access to the raw training data. Data sampling during training is deterministic given the random seed and training step, so each of the fifteen training runs can be reproduced exactly given the same hardware and software environment.

## 7 Conclusion

In a controlled, iso parameter, iso token comparison on Bangla with five seeds per architecture, recurrence, in the form of Mamba-3 and the hybrid, achieves significantly lower perplexity than a matched Transformer, and attention’s agreement accuracy degrades with subject verb distance while recurrence’s does not. At the same time, most between architecture agreement gaps, including the honorific register comparison, are not robust once cross seed variance is taken into account, and Mamba-3 is substantially more seed sensitive than the Transformer at this scale. The architectural signal we report is about distance robustness and initialization variance, not about one architecture being uniformly better than the other. We release the probe suite, code, checkpoints, and analysis scripts

to support replication and to make it easier to extend this study to larger scales and to more languages.

## Ethics and Broader Impact

The probe suite is synthetic, contains no personal data, and targets grammatical well-formedness only. Training data is the public, license-cleared FineWeb-2 corpus, which, like all web corpora, may contain social biases our agreement-focused evaluation does not measure. The models are small research artifacts, not deployment-ready systems. We hope the released Bangla probe suite lowers the barrier to targeted syntactic evaluation for an under-resourced language.

## References

- Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. 2023. Zoology: Measuring and improving recall in efficient language models. *arXiv preprint arXiv:2312.04927*.
- Steven Bethard. 2022. We need to talk about random seeds. *arXiv preprint arXiv:2210.13393*.
- Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2022. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327.
- Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2023. [BanglaNLG and BanglaT5: Benchmarks and resources for evaluating low-resource natural language generation in Bangla](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 726–735, Dubrovnik, Croatia. Association for Computational Linguistics.
- Tri Dao and Albert Gu. 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*.
- Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.
- Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling*. Outstanding Paper Award.
- Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. 2024. Repeat after me: Transformers are better than state space models at copying. *arXiv preprint arXiv:2402.01032*.
- Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026. MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Transactions of the Association for Computational Linguistics*, 14:193–216.
- Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of EMNLP 2018: System Demonstrations*, pages 66–71.
- Aakash Lahoti, Kevin Y. Li, Berlin Chen, Caitlin Wang, Aviv Bick, J. Zico Kolter, Tri Dao, and Albert Gu. 2026. Mamba-3: Improved sequence modeling using state space principles. In *International Conference on Learning Representations*. Oral.
- Opher Lieber, Barak Lenz, Hofit Bata, et al. 2024. Jamba: A hybrid transformer–mamba language model. *arXiv preprint arXiv:2403.19887*.
- Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- William Merrill, Jackson Petty, and Ashish Sabharwal. 2024. The illusion of state in state-space models. In *Proceedings of the 41st International Conference on Machine Learning*.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. FineWeb2: One pipeline to scale them all—adapting pre-training data processing to every language. *arXiv preprint arXiv:2506.20920*.
- Liliang Ren, Yang Liu, Yadong Lu, Yelong Shen, Chen Liang, and Weizhu Chen. 2025. Samba: Simple hybrid state space models for efficient unlimited context language modeling. In *International Conference on Learning Representations*.
- Noam Shazeer. 2020. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Arpita Vats, Rahul Raja, Mrinal Mathur, Vinija Jain, and Aman Chadha. 2025. Multilingual state space models for structured question answering in indic languages. *arXiv preprint arXiv:2502.01673*.

Roger Waleffe, Wonmin Byeon, Duncan Riach, Brandon Norick, Vijay Korthikanti, Tri Dao, Albert Gu, Ali Hatamizadeh, Sudhakar Singh, Deepak Narayanan, Garvit Kulshreshtha, Vartika Singh, Jared Casper, Jan Kautz, Mohammad Shoeybi, and Bryan Catanzaro. 2024. An empirical study of mamba-based language models. *arXiv preprint arXiv:2406.07887*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*.

## A Per-Seed and Per-Distance Data

Per-model perplexity and probe accuracy are released per seed in `paper/results.json`; `paper/stats.py` reproduces all confidence intervals and Welch tests. Cross-seed standard deviations (main text) quantify initialization sensitivity: the Transformer is stable (SVA  $\pm 1.5$ , honorific  $\pm 3.5$ ), while Mamba-3 is not (SVA  $\pm 5.6$ , honorific  $\pm 8.4$ ). SVA-by-distance means over five seeds are reported in Table 3.

## B Data, Tokenizer, and Compute

The data, tokenizer, and compute details are given in 6.

| Item                     | Value                                                                            |
|--------------------------|----------------------------------------------------------------------------------|
| Corpus                   | FineWeb-2 ben_Beng (Penedo et al., 2025); 1.0B train tokens, held-out test split |
| Tokenizer                | SentencePiece BPE, 32,768 vocabulary, byte fallback                              |
| Sequence length          | 1,024                                                                            |
| Non-embedding parameters | Transformer 24.52M; Mamba-3 24.50M; Hybrid 24.85M                                |
| Hardware                 | 1 $\times$ NVIDIA A100-80GB (Colab)                                              |
| Throughput (bf16)        | Transformer 226k tokens/s; Mamba-3 172k tokens/s                                 |
| Compute                  | Approximately 3 A100-hours/run; 15 runs (3 architectures $\times$ 5 seeds)       |

Table 6: Data, tokenizer, model size, and compute details.