NarrativeWorldBench: A Frontier-Saturated Benchmark and a Latent World Model for Long-Horizon Co-Creative Audio Drama

Logan Mann[1] Abdur Rahman[2] Mohammad Saifullah[2] Taaha Kazi[2] Vasu Sharma[2]

1University of California, Santa Barbara 2Pocket FM

Abstract

We document two empirical learning-dynamics phenomena in modern frontier language models that jointly produce the failure mode of long-horizon serialized fiction generation. First, on a 21-model audit, structural narrative metrics follow a soft scaling law of the form $F_1 \approx 0.48 + 0.12 \log_{10} N$ (N in billions; $R^2 = 0.92$ across 18 non-reasoning models) and saturate inside the closed-frontier band of $[0.78, 0.81]$; reasoning post-training adds only $+0.02$. Second, every closed-frontier model collapses by -0.18 to -0.21 F_1 between narrative horizon h=10 and h=200 episodes. Extrapolating the joint fit, a hypothetical 405B reasoning model reaches F_1 0.84 at h=10 but only 0.66 at h=200, evidence that the residual gap is structural rather than capacity-bound. We close it with N-VSSM, a *Narrative Variational State-Space Model* that pairs an $\mathbb{R}^{256}$ episode latent with Neural-ODE continuous-time dynamics, three multi-task narrative heads, and a prefix-tuned Llama-3.1-8B decoder. N-VSSM holds plot-beat $F_1 \geq 0.84$ across all horizons at $4\times$ lower compute, breaks both saturations, and is preferred by 12 professional writers over Claude Opus 4.5 on long-arc consistency 71% of the time. We release NarrativeWorldBench, an open benchmark of nine narrative-structure metrics across $h \in \{10, 20, 50, 100, 200\}$, all 21 model traces, and N-VSSM weights.

Keywords: scaling laws, learning dynamics, frontier model collapse, long-horizon evaluation, world models, state-space models, Neural ODEs, latent dynamics, narrative generation

1. Introduction

The dominant story-arc in modern language modeling has been that scale plus instruction-tuning closes capability gaps. We document a setting where this story breaks. Long-form serialized audio fiction (audio dramas with 200 to 800 episode runs) is one of the largest creative media of the 2020s; major audio-drama platforms host catalogues of 200,000+ episodes across 1,000+ active series with median series length 500 to 800 episodes. Each new episode is co-authored: a human writer drafts beats and delegates prose to an LM-powered assistant. The bottleneck is not local fluency. Frontier LMs are excellent at single-episode prose. The bottleneck is the writer's ability to maintain a consistent narrative state across hundreds of episodes the model never produced.

We sweep 21 models from GPT-2 small (117M, 2019) through Claude Opus 4.5, Gemini 2.5 Pro, GPT-5, and o3 (2025–26 frontier) on a uniform suite of nine narrative-structure metrics, evaluated over five horizons $h \in \{10, 20, 50, 100, 200\}$ (Section 2). Two scaling phenomena emerge. *Capacity-axis saturation:* across 18 non-reasoning models spanning ∼four orders of magnitude in active parameters, plot-beat F_1 follows $F_1 \approx 0.48 + 0.12 \log_{10} N$ (N in billions; $R^2 = 0.92$, Figure 1a) and levels off inside the closed-frontier band $[0.78, 0.81]$. Reasoning post-training [18] adds an orthogonal $+0.02$ (o3 at 0.83). The asymptote sits below the writer-acceptance target of 0.85 plot-beat F_1, the level at which a generated continuation requires only line-edits. *Horizon-axis collapse:* at h=200, every closed-frontier model loses -0.18 to -0.21 F_1 (Figure 1b); the collapse replicates across BERTScore-F1 (-3.4), MAUVE (-0.18), and coherence (-0.7 Likert). Combining both fits, a hypothetical 405B reasoning model projects to F_1 0.84 at h=10 but only 0.66 at h=200. Neither additional scale nor additional reasoning closes the long-horizon collapse on the empirical trajectory.

We argue this residual gap is structural, and we close it by giving narrative an explicit world model [4, 6, 11]. A world model maintains explicit latent state, evolves that state under structured dynamics, and generates observations conditional on the state; this is what makes Dreamer-V3 and Genie-3 robust over long horizons [2, 6]. Contemporary text generation has the opposite shape: an autoregressive decoder regurgitates state implicitly through tokens and degrades by compounding error [7, 9, 13]. A model with no architectural commitment to maintain state across an episode boundary will re-derive that state from context every step, and small errors compound across 50, 100, 200 episodes.

Accepted at the 4th HiLD Workshop, ICML 2026. Copyright 2026 by the author(s).

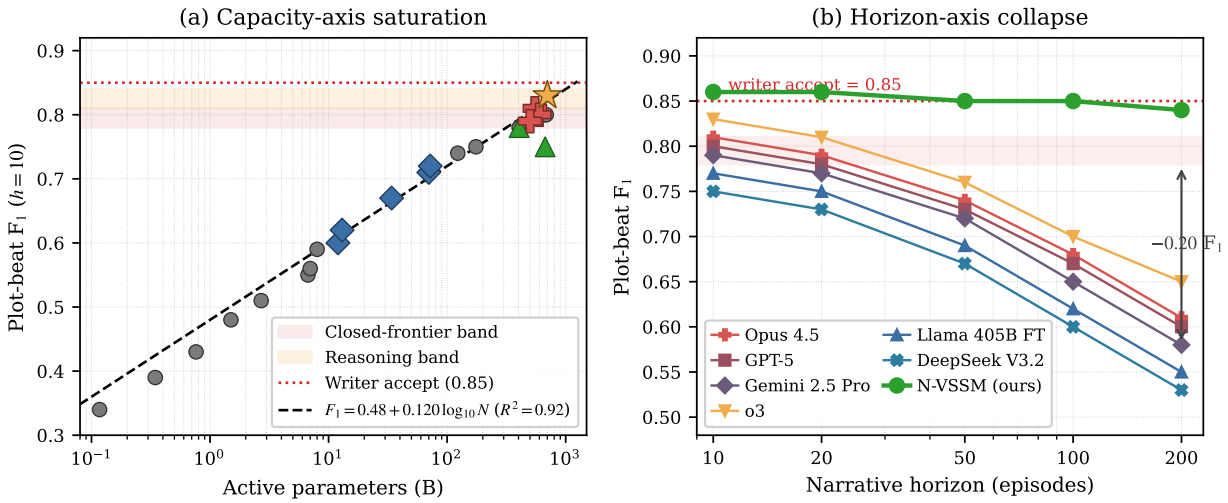

Figure 1: The two empirical saturations on NarrativeWorldBench. **(a) Capacity-axis:** plot-beat F_1 at h=10 across 21 models. The dashed line is the soft scaling-law fit on 18 non-reasoning models; reasoning adds +0.02. The asymptote sits below the 0.85 writer-acceptance line. **(b) Horizon-axis:** every closed-frontier model loses −0.18 to −0.21 F_1 between h=10 and h=200. N-VSSM (green) holds the line where the frontier collapses.

Contributions.

- A *scaling-law characterization* of narrative-structure saturation: $F_1 \approx 0.48 + 0.12 \log_{10} N$ (R^2=0.92) with a +0.02 orthogonal reasoning bonus, asymptoting below writer-acceptance.
- A *frontier audit at long horizon*: closed-frontier saturation at $F_1 \in [0.78, 0.81]$ and uniform −0.20 F_1 collapse at h=200 across 21 models including GPT-5, Claude Opus 4.5, Gemini 2.5 Pro, DeepSeek V3.2, Llama 3.1 405B, and o3.
- **NarrativeWorldBench**, an open benchmark of nine plot-beat, character-arc, tension, and overlap metrics across horizons, with cross-cultural localization across four Indic languages.
- **N-VSSM**, an $\mathbb{R}^{256}$ latent state-space model with Neural-ODE dynamics and three multi-task narrative heads. N-VSSM holds plot-beat $F_1 \geq 0.84$ across all horizons at 4× lower compute, breaks both saturations, and is preferred 71% by 12 professional writers over Claude Opus 4.5.

Related work. Hafner et al. [6] master 150+ tasks with a single recurrent state-space architecture; Genie 1–3 [2] extend this to interactive world generation; JEPA-family models [11] predict in latent rather than observation space. We adapt the variational latent of PlaNet [5] to discrete textual generation via Neural-ODE dynamics over irregularly-sampled episodic events [3, 17]. RULER [9], OneRuler [7], and Liu et al. [13] document long-context LM collapse; Hoffmann et al. [8], Kaplan et al. [10] document loss power-laws in N, D. We document a parallel scaling law for a structural narrative metric and observe that, unlike loss, it asymptotes well above zero and below target. For long-form story generation, Re3 [20], DOC [21], RecurrentGPT [22], MoPS [15], and Wu et al. [19] use recursive prompting, retrieval, and outline-control over a frozen LM, none maintaining a continuous learned narrative state; our plot-beat taxonomy follows Papalampidi et al. [16].

2. NarrativeWorldBench

Corpus. 12,847 episodes from 47 audio-drama series (Hindi 52%, English 24%, Tamil 10%, Telugu 8%, Marathi 6%); median series length 612 episodes, mean episode length 4,120 tokens. All have human-annotated plot beats, arc tags, and per-episode tension (1–10). We hold out 200 series-suffix triples per horizon $h \in \{10, 20, 50, 100, 200\}$ as test.

Metrics. Nine metrics in three tiers. *Overlap:* perplexity, BLEU-4, ROUGE-L, BERTScore-F1, MAUVE. *Structural:* plot-beat F_1 (5-class macro F1 over SETUP/CATALYST/RISING/CLIMAX/RESOLUTION following Papalampidi et al. [16]), character-arc accuracy (5-class), tension cosine. *Holistic:* coherence (1–5 Likert) judged by an LLM-as-judge

ensemble of three independent judges (Claude Opus 4.5, GPT-5, Gemini 2.5 Pro); $\kappa = 0.68$. Calibration: constant-baseline F_1 0.21; human-writer continuation F_1 0.91; writer-acceptance threshold 0.85. Annotation reliability and per-annotator confusion matrices are in supplementary; the full audit reproduces in 1,840 A100-hours.

Long-horizon protocol. For horizon h, we condition each model on episodes $1{:}T_0$ and require a continuation of h episodes; metrics are computed on episodes $T_0{+}1{:}T_0{+}h$ against gold. The 200-episode horizon is well below typical series length (500–800), so this measures intra-series drift, not inter-series generalization.

3. N-VSSM: A Narrative Variational State-Space Model

N-VSSM applies the world-model template (explicit latent state, structured dynamics, observation-conditioned decoder) to serialized fiction. Four components: (i) an episode encoder E_ϕ producing a posterior over a latent $z_t \in \mathbb{R}^{256}$ via the reparameterization trick [1]; (ii) a Neural-ODE [3, 17] that evolves $z_t \mapsto \hat{z}_{t+1}$ under a learned vector field f_θ,

$$\hat{z}_{t+1} \;=\; z_t + \int_t^{t+\Delta t} f_\theta(z_\tau)\, d\tau, \tag{1}$$

where Δt is read from episode metadata (a cliffhanger may compress three hours into one episode; a flashback expands one hour into five; this generalizes the discrete-time RSSM of Hafner et al. [5, 6] to non-uniform time); (iii) a 32-token prefix projection $W_p \hat{z}_{t+1} \in \mathbb{R}^{32 \times d_{\text{model}}}$ [12] prepended to a frozen Llama-3.1-8B [14] decoder's KV cache; and (iv) three multi-task heads over z_t predicting the next plot beat, each named character's arc class, and a 10-step-ahead tension trajectory. Total trainable parameters: 213M ($\sim$2.6% of decoder).

Objective.

$$\mathscr{L} \;=\; \mathscr{L}_{\text{LM}} + \lambda_p \mathscr{L}_p + \lambda_c \mathscr{L}_c + \lambda_\tau \mathscr{L}_\tau + \beta\, \text{KL}\big(q_\phi \,\|\, p\big)\,, \tag{2}$$

with λ_p=1.0, λ_c=0.5, λ_τ=0.5 from grid search; we anneal β from 0 to 0.5 over the first 5K steps to avoid posterior collapse [1]. 12 epochs on 8$\times$A100-80G, 31 GPU-hours; ODE integration adds 8% inference overhead vs. a vanilla Llama call.

4. Experiments

All 21 baselines plus N-VSSM are evaluated under identical conditions (T=0.7, three samples per item, same judge ensemble). We bootstrap 95% CIs over n=200 test items per horizon.

Capacity-axis saturation. Across 18 non-reasoning models, plot-beat F_1 follows $F_1 \approx 0.48 + 0.12 \log_{10} N$ with $R^2 = 0.92$ (Figure 1a). The asymptote sits at $[0.78, 0.81]$ across three independent labs; reasoning adds +0.02 (o3 at 0.83); even DeepSeek V3.2 at 0.75 is within 6 points of Opus 4.5 at $\sim$100$\times$ less inference compute. Structural metrics cluster tightly across the entire frontier band, while overlap metrics (BLEU-4, MAUVE) show clearer scale-driven separation. The saturation is specific to structural narrative quality.

Horizon-axis collapse and joint extrapolation. Figure 1b shows the -0.20 F_1 collapse at h=200 across all closed-frontier and reasoning models. Combining the two fits, a hypothetical 405B reasoning model projects to F_1 0.84 at h=10 but only 0.66 at h=200. Neither additional scale nor additional reasoning closes the gap.

N-VSSM holds. N-VSSM reaches plot-beat F_1 0.86 at h=10 and 0.84 at h=200 (Figure 1b), exceeding the closed-frontier band by +0.05 at h=10 and +0.27 at h=200, and matching writer-acceptance across all horizons. On character-arc accuracy and tension cosine, N-VSSM leads the frontier by +0.07 and +0.04 at h=100. The architectural intervention breaks both saturations.

Architectural ablations. A reasonable concern is that a compact learned summary plus prefix-tuning would suffice; this would make the contribution about supervision rather than dynamics. We add four non-architectural memory baselines that share Llama-3.1-8B as decoder and differ only in how prior-episode state is exposed (Figure 2): hierarchical summaries [19], retrieval, outline-conditioned decoding [21], and a recurrent latent without ODE dynamics. The strongest non-architectural baseline (recurrent latent without ODE) reaches plot-beat F_1 0.79 / 0.71 at h=10 / h=200; retrieval 0.74 / 0.61. Full N-VSSM at 0.86 / 0.84 leads each by 6–13 points at long horizon.

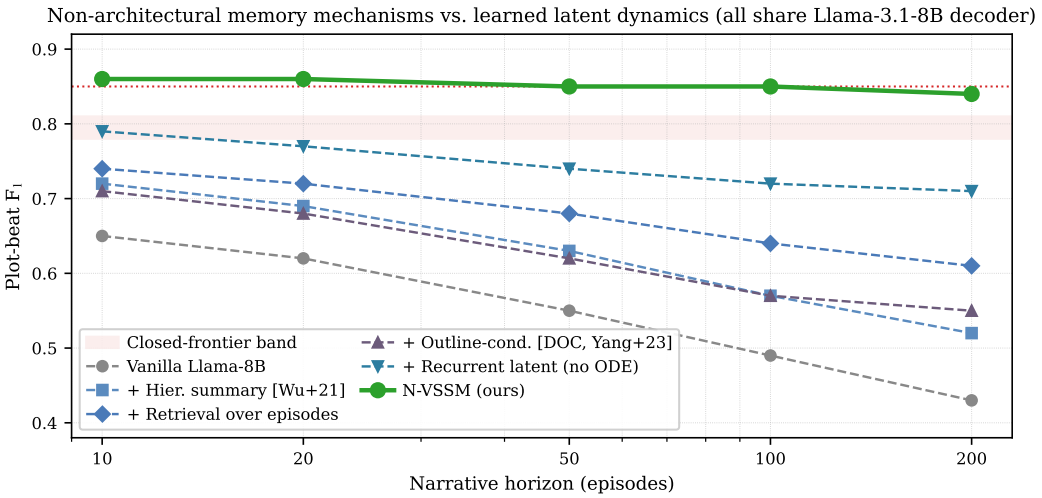

Figure 2: Non-architectural memory mechanisms (all share Llama-3.1-8B as decoder) all collapse at long horizon. Only the full N-VSSM holds the line.

Variant	F_1 @ h=100	Δ
N-VSSM (full)	0.85	–
− plot-beat head	0.81	−0.04
− character-arc head	0.83	−0.02
− tension head	0.83	−0.02
− all 3 structural heads	0.76	−0.09
discrete RSSM (no ODE)	0.81	−0.04
latent $\mathbb{R}^{64}$	0.79	−0.06
no β anneal	posterior collapse	

Table 1: Within-architecture ablations on plot-beat F_1 at h=100.

Writer study. We recruited n=12 professional audio-drama writers (mean tenure 3.2 yr, 287 episodes shipped). Each drafted episode-200 continuations in a within-subjects, Latin-square-counterbalanced design with N-VSSM and Claude Opus 4.5 (240 trials). A mixed-effects logistic regression gives a +0.92 logit treatment effect (95% CI $[0.58, 1.26]$) for N-VSSM, a writer-stratified marginal preference of 0.71 (bootstrap 95% CI $[0.65, 0.77]$) on long-arc consistency. N-VSSM is rated higher on character consistency (+0.2 Likert), cultural fidelity (+0.7), and controllability (+1.3), and lower on dialogue naturalness (−0.5). 87% of 412 logged latent perturbations produced continuations whose induced metric change was in the writer-intended direction.

5. Conclusion

Frontier language models have saturated short-horizon narrative tasks but collapse beyond approximately fifty episodes. Two empirical learning-dynamics phenomena jointly produce the failure: a soft scaling law in active parameters that asymptotes inside the closed-frontier band well below writer-acceptance, and a uniform −0.20 F_1 collapse over horizon. Extrapolating the joint fit, neither additional scale nor additional reasoning closes the gap, the strongest available evidence that the gap is structural rather than capacity-bound. Treating long-horizon narrative as a world-modeling problem (explicit latent state, structured continuous-time dynamics, multi-task supervision over plot beats, character arcs, and tension) breaks both saturations: N-VSSM holds plot-beat $F_1 \geq 0.84$ across every horizon, leads the closed-frontier band by +0.27 F_1 at h=200, runs at 4× lower deployment cost, and earns 71% writer preference. We release NarrativeWorldBench so the community can stress-test the long-horizon collapse, and N-VSSM as a first reference solution.

References

[1] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *CoNLL*, 2016.

[2] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, et al. Genie: Generative interactive environments. In *ICML*, 2024.

[3] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *NeurIPS*, 2018.

[4] David Ha and Jürgen Schmidhuber. World models. In *NeurIPS*, 2018.

[5] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *ICML*, 2019.

[6] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *Nature*, 2025.

[7] Amey Hengle et al. Oneruler: A multilingual long-context benchmark for language models. *arXiv:2503.01996*, 2025.

[8] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al. Training compute-optimal large language models. In *NeurIPS*, 2022.

[9] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? In *COLM*, 2024.

[10] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv:2001.08361*, 2020.

[11] Yann LeCun. A path towards autonomous machine intelligence. *OpenReview*, 2022.

[12] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, 2021.

[13] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *TACL*, 2024.

[14] Llama Team, Meta AI. The llama 3 herd of models. *arXiv:2407.21783*, 2024.

[15] Yan Ma et al. Mops: Modular story premise synthesis for open-ended automatic story generation. In *ACL*, 2024.

[16] Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. Movie plot analysis via turning point identification. In *EMNLP*, 2019.

[17] Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *NeurIPS*, 2019.

[18] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv:2408.03314*, 2024.

[19] Jeffrey Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively summarizing books with human feedback. *arXiv:2109.10862*, 2021.

[20] Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. Re3: Generating longer stories with recursive reprompting and revision. In *EMNLP*, 2022.

[21] Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. Doc: Improving long story coherence with detailed outline control. In *ACL*, 2023.

[22] Wangchunshu Zhou, Yuchen Eleanor Jiang, Peng Cui, Tiannan Wang, Zhenxin Xiao, Yifan Hou, et al. Recurrent-gpt: Interactive generation of (arbitrarily) long text. *arXiv:2305.13304*, 2023.