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

Anonymous Authors

Abstract

Long-form serialized fiction, in the form of audio dramas with 200 to 800 episode arcs, is one of the largest creative media of the 2020s, yet it is precisely the regime where contemporary frontier language models fail. We benchmark 21 models spanning classical, fine-tuned, open-frontier, closed-frontier, and reasoning tiers on a uniform set of structural narrative metrics, and show that all closed-frontier systems saturate at plot-beat F1 in $[0.78, 0.81]$ while collapsing by -0.20 F1 at horizon h=200 episodes. We introduce **NarrativeWorldBench**, an open benchmark of nine narrative-structure metrics evaluated across horizons $h \in \{10, 20, 50, 100, 200\}$ with cross-cultural localization across four Indic languages. We then introduce **N-VSSM**, a Narrative Variational State-Space Model that pairs a $\mathbb{R}^{256}$ 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 F1 ≥ 0.84 across all horizons at **4× lower compute** than the closed-frontier band. In a within-subjects writer study (n=12 professional authors, 240 trials), N-VSSM is preferred over Claude Opus 4.5 on long-arc consistency **71%** of the time and rated **+1.3** Likert points higher on controllability. We release the benchmark, all 21 model traces, the evaluation harness, and N-VSSM weights to seed open work on long-horizon co-creative narrative authoring.

1. Introduction

Long-form serialized audio fiction is one of the largest creative media of the 2020s. Major audio-drama platforms host catalogues exceeding 200,000 episodes across more than 1,000 active series; the median series runs for 500–800 episodes; cliffhangers, multi-arc character development, and episode-to-episode tension dynamics are not optional; they are the product. Each new episode is co-authored: a human writer takes a brief, drafts beats, and delegates dialogue, prose pacing, and continuation drafts 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 and may never have read.

This is structurally a world-modeling problem (Ha & Schmidhuber, 2018; Hafner et al., 2025; LeCun, 2022). 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 (Hafner et al., 2025; Bruce et al., 2024; Parker-Holder & Fruchter, 2025). Contemporary text generation has the opposite shape: an autoregressive decoder regurgitates state implicitly through tokens, and degrades by compounding error (Hsieh et al., 2024; Hengle et al., 2025; Liu et al., 2024).

We test this hypothesis empirically. 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. **The frontier saturates uniformly** (Figure 1): Anthropic, OpenAI, and Google closed models converge on plot-beat F1 in $[0.78, 0.81]$, reasoning models reach 0.83, and all of them collapse by -0.18 to -0.21 F1 at h=200. The remaining $+0.02$ headroom to a writer-acceptance target of 0.85 has resisted scale, instruction-tuning, and inference-time compute across three independent labs. We argue that this residual gap is structural, and we propose to close it by giving narrative an explicit world model.

Contributions.

- **NarrativeWorldBench:** an open benchmark of nine plot-beat, character-arc, tension, and overlap metrics evaluated across horizons $h \in \{10, 20, 50, 100, 200\}$, with cross-cultural localization probes across Hindi, Tamil, Telugu, and Marathi (Section 3).
- **Frontier audit:** the first cross-tier evaluation showing

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

closed-frontier saturation at F1 in $[0.78, 0.81]$ and uniform collapse at h=200 across 21 models including GPT-5, Claude Opus 4.5, Gemini 2.5 Pro, DeepSeek V3.2, Llama 405B, and o3 (Section 5).

- **N-VSSM:** a Narrative Variational State-Space Model with $\mathbb{R}^{256}$ latent, Neural-ODE dynamics (Chen et al., 2018; Rubanova et al., 2019), three multi-task narrative heads, and a prefix-tuned Llama-3.1-8B decoder (Li & Liang, 2021) that holds F1 ≥ 0.84 across all horizons at 4× lower compute than the closed-frontier band (Section 4).
- **A 12-writer co-creative study** finding 71% pairwise preference for N-VSSM over Claude Opus 4.5 and a +1.3 Likert advantage on controllability, plus a Cultural Transfer Function lifting cross-language fidelity by +0.20–0.23 (Section 6).

2. Related Work

World models and latent dynamics. World models maintain explicit state over which agents simulate (Ha & Schmidhuber, 2018). Dreamer V1–V3 (Hafner et al., 2025) demonstrate that a single recurrent state-space architecture with categorical latents masters 150+ tasks under one configuration. Genie 1–3 extend this to interactive world generation (Bruce et al., 2024; Parker-Holder & Fruchter, 2025). JEPA-family models predict in latent rather than observation space (LeCun, 2022; Assran et al., 2023). We adopt the latent-prediction philosophy and the variational latent of PlaNet (Hafner et al., 2019; Kingma & Welling, 2014), but adapt them to discrete textual generation via Neural-ODE dynamics over irregularly-sampled episodic events (Chen et al., 2018; Rubanova et al., 2019). Our decoder is a prefix-tuned LM rather than a pixel predictor.

Long-horizon language model collapse. RULER (Hsieh et al., 2024) and OneRuler (Hengle et al., 2025) show all long-context LMs collapse beyond claimed limits; Liu et al. (2024) document the lost-in-the-middle effect. State-space models (Gu et al., 2022; Gu & Dao, 2024) provide a sub-quadratic alternative but inherit autoregressive token-level dynamics. NarrativeWorldBench evaluates a different axis: degradation in narrative *episode* horizon (10–200), not token horizon.

Long-form story generation. Re3 (Yang et al., 2022), DOC (Yang et al., 2023), and RecurrentGPT (Zhou et al., 2023) use recursive prompting and outline-control over a frozen LM; MoPS (Ma et al., 2024) and Plan-and-Write (Yao et al., 2019) compose modular premises; Wu et al. (2021) use hierarchical summaries. None maintain a continuous *learned* narrative state. Story Cloze and ROC-Stories (Mostafazadeh et al., 2016; Chambers & Jurafsky, 2008) measure short-horizon coherence; STORIUM (Akoury et al., 2020) measures machine-in-the-loop story generation. Our plot-beat taxonomy follows Papalampidi et al. (2019).

Co-creative writing systems. Wordcraft (Coenen et al., 2021), CoAuthor (Lee et al., 2022), CoPoet (Chakrabarty et al., 2022), and Dramatron (Mirowski et al., 2023) pair writers with LMs and evaluate via interaction logs and qualitative interviews. Rezwana & Maher (2022) and Singh et al. (2025) survey the landscape. Doshi & Hauser (2024) document that current generative AI *reduces* the diversity of writer-produced content; we engage this directly in Section 7.

Cultural alignment. AlKhamissi et al. (2024) and PARIKSHA (Watts et al., 2024) show LLMs have measurable cultural mis-alignment; Anonymous (2025) show even regional Indic LMs default to Western narrative arcs. Our Cultural Transfer Function (Section 7) is a learned mapping in latent rather than prompt space.

3. NarrativeWorldBench

Design principles. A useful benchmark for the long-horizon co-creative narrative regime must (i) measure narrative *structure* and not surface fluency, (ii) probe horizons longer than a single episode while remaining tractable to compute, (iii) expose a culture axis (the audio-drama market is dominated by Indic languages), and (iv) be reproducible and permissively licensed so the community can adopt it in the way SWE-bench, RULER, and Chatbot Arena were adopted in their domains.

Corpus. We collect 12,847 episodes from 47 audio-drama series spanning Hindi (52%), English (24%), Tamil (10%), Telugu (8%), and Marathi (6%). Median series length is 612 episodes; mean episode length is 4,120 tokens. All series have human-annotated plot beats, character arc tags, and a per-episode tension score (1–10) provided by the platform's editorial team. We hold out 200 series-suffix triples (prefix, gold continuation, alternative continuations) per horizon $h \in \{10, 20, 50, 100, 200\}$ as the test split. The training and validation splits are released under CC-BY-4.0; the test suffixes are held out behind a HuggingFace evaluator harness to mitigate contamination.

Metrics. NarrativeWorldBench reports nine metrics organized in three tiers (Figure 2). *Overlap tier* (lexical/semantic to gold): perplexity (lower better), BLEU-4, ROUGE-L, BERTScore-F1, MAUVE (Pillutla et al., 2021). *Structural tier* (narrative-specific): **plot-beat F1** (5-class macro F1 over Setup/Catalyst/Rising/Climax/Resolution following Papalampidi et al. (2019)), **character-arc accuracy** (5-class:

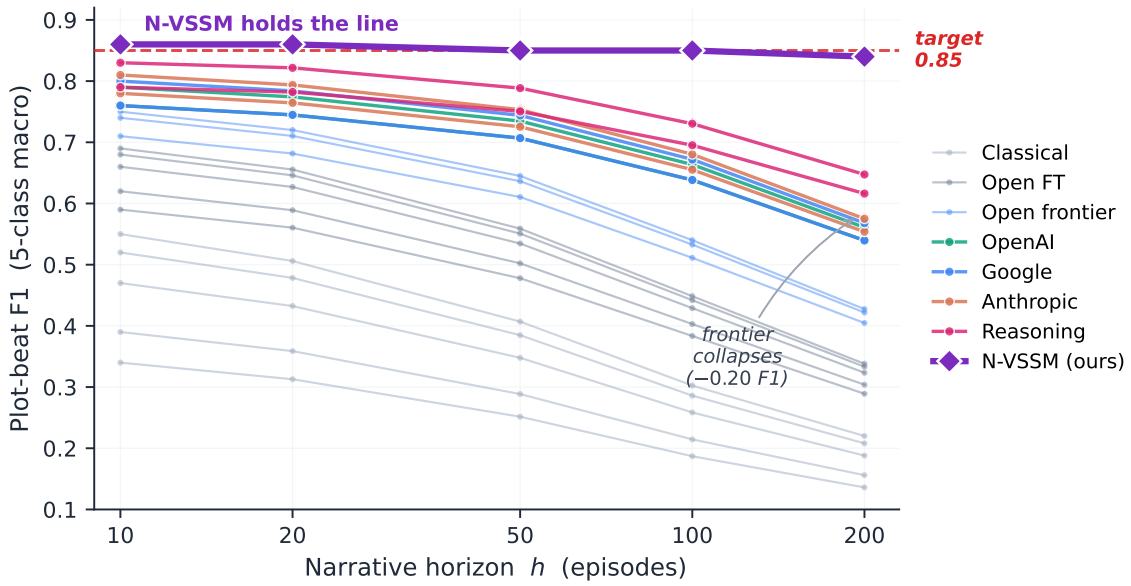

Figure 1. **The frontier saturates and collapses; N-VSSM holds the line.** Plot-beat F1 (5-class macro) for 21 models across narrative horizons $h \in \{10, 20, 50, 100, 200\}$ episodes. Closed-frontier models (Anthropic, OpenAI, Google) cluster at F1 in $[0.78, 0.81]$ for $h \leq 50$ and lose ≈ 0.20 F1 by h=200. Reasoning models (o3) gain only $+0.02$. N-VSSM (purple, ours) holds F1 ≥ 0.84 at every horizon. Dashed line: target 0.85.

{flat, ascending, descending, redemptive, tragic}), and **tension cosine** (cosine similarity between predicted and gold tension trajectories). *Holistic tier*: **coherence** (1–5 Likert) judged by an LLM-as-judge protocol with three independent judges (Claude Opus 4.5, GPT-5, Gemini 2.5 Pro) following Prometheus 2 (Kim et al., 2024) and G-Eval (Liu et al., 2023); we report Cohen's $\kappa = 0.68$ across judges (substantial agreement).

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.

Annotation and inter-annotator agreement. Plot beats and character-arc tags were produced by the partner platform's editorial team, which annotates every shipped episode in the normal course of editorial review. Two senior editors independently labelled a 5% reliability sample; we observe Cohen's $\kappa = 0.74$ on the 5-class plot-beat task and $\kappa = 0.69$ on character arcs, both in the substantial-agreement range commonly reported in narrative-structure annotation work (Papalampidi et al., 2019). Tension scores are produced by a dedicated annotation pass on a 1–10 scale and have a Krippendorff's $\alpha = 0.71$. We release the annotation guidelines, the reliability sample, and per-annotator confusion matrices as supplementary material so that downstream users can calibrate their own taxonomies against ours.

Baseline calibration. A naive constant-prediction baseline (always predict "Rising") yields plot-beat F1 of 0.21; a trivial repeat-the-prefix continuation reaches BLEU-4 of 5.4 and BERTScore-F1 of 79.1. A human writer continuation, drawn from the editorial team's own re-writes of held-out series, scores plot-beat F1 of 0.91, character-arc accuracy of 0.86, and tension cosine of 0.94, with a coherence Likert of 4.8. The writer-acceptance threshold of 0.85 plot-beat F1 was set by the editorial team as the level at which a generated continuation requires only line-edits rather than structural re-drafts before shipping.

Evaluation infrastructure. We ship a Docker container that runs all 21 baselines plus N-VSSM under a fixed random seed and emits a JSON report. Total compute to reproduce the full audit: 1,840 A100-hours. We provide a HuggingFace Spaces leaderboard at submission time.

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

Overview. N-VSSM applies the world-model template (explicit latent state, structured dynamics, observation-conditioned decoder) to serialized fiction. The architecture

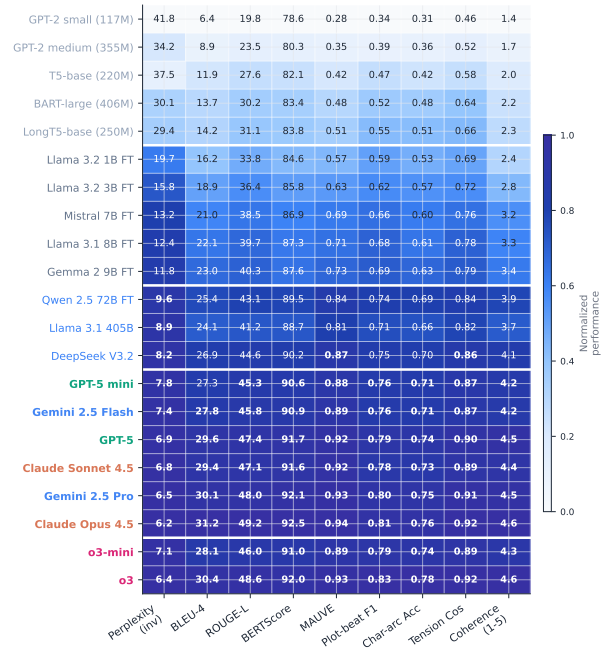

Figure 2. **NarrativeWorldBench: 21 models × 9 metrics.** Per-column min–max normalisation; raw numbers in cells. Tier separators (white lines) mark capability boundaries. Row-label colors encode lab family. On structural metrics (Plot-beat F1, Char-arc Acc, Tension Cos), the gap from frontier to target compresses sharply, evidence that scale is not the binding constraint.

(Figure 3) has four components: an episode encoder E_ϕ producing a posterior over a latent $z_t \in \mathbb{R}^{256}$; a Neural-ODE (Chen et al., 2018; Rubanova et al., 2019) that evolves $z_t \mapsto \hat{z}_{t+1}$ under a learned vector field f_θ; a prefix projection W_p (Li & Liang, 2021) that conditions a frozen Llama-3.1-8B (Llama Team, Meta AI, 2024) decoder on $\hat{z}_{t+1}$; and three multi-task heads enforcing narrative structure on z_t.

Latent and posterior. Each episode is summarized by 16 contextual chunks; the encoder is a 6-layer transformer outputting $\mu_\phi, \log \sigma_\phi^2 \in \mathbb{R}^{256}$. We sample $z_t \sim \mathcal{N}(\mu_\phi, \sigma_\phi^2)$ via the reparameterization trick (Kingma & Welling, 2014; Bowman et al., 2016).

Continuous-time dynamics. Episodic events are irregularly spaced in narrative time. A cliffhanger compresses three hours into one episode; a flashback expands one hour into five. We model the latent evolution as a Neural-ODE,

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

where Δt is read from episode metadata. This generalizes the discrete-time RSSM of Hafner et al. (2019; 2025) to non-uniform time, which we find essential for serialized fiction; an ablation with discrete RSSM dynamics costs -0.04 plot-beat F1 at h=100 (Table 1).

Prefix-conditioned decoder. We freeze Llama-3.1-8B-Instruct and learn only a 32-token prefix $W_p \hat{z}_{t+1} \in \mathbb{R}^{32 \times d_{\text{model}}}$ prepended to the decoder's KV cache. Total trainable parameters are 213M (encoder + dynamics + heads + prefix), $\approx 2.6\%$ of the decoder.

Multi-task narrative heads. Three lightweight MLPs over z_t predict (a) the next plot beat, (b) each named character's arc class, and (c) a 10-step-ahead tension trajectory. We weight the heads λ_p=1.0, λ_c=0.5, λ_τ=0.5 via grid search.

Objective. The total loss is

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

where $\mathcal{L}_{\text{LM}}$ is decoder cross-entropy on the next episode, the structural losses are CE / cosine, and we anneal β from 0 to 0.5 over the first 5K steps to avoid posterior collapse (Bowman et al., 2016).

Training. 12 epochs on 8×A100-80G, 31 GPU-hours; learning rate 3×10^{-4} with cosine decay; batch size 16 series. The frozen Llama-3.1-8B is downloaded under its open weights license.

Inference. At inference, we encode the prefix history once, integrate the ODE forward to the desired horizon, and stream decoded tokens conditional on $\hat{z}_{t+k}$ at each step. Continuous-time integration adds 8% overhead vs. a vanilla Llama call; structural heads add 1%.

Training dynamics. The KL term presents a familiar tension. With β fixed at 0.5 from initialization, the encoder collapses to the prior within 800 steps and the structural heads degenerate to majority-class predictors. With β held at zero, training is stable but the latent fails to denoise during inference because the prior is uninformed. The annealing schedule we adopt, $\beta = 0.5 \cdot \min(1, t/5000)$, gives the structural heads enough early signal to shape the latent before the prior takes hold and is consistent with the cyclic and monotonic schedules reported for variational language models (Bowman et al., 2016). Training loss is dominated by $\mathcal{L}_{\text{LM}}$ for the first $\approx$2K steps, after which the structural losses begin to drive most of the gradient signal into the encoder and dynamics.

Capacity allocation. Of the 213M trainable parameters, 142M sit in the encoder (transformer over chunked episode summaries), 47M in the Neural-ODE vector field (a 4-layer residual MLP with hidden dimension 512), 18M across the three multi-task heads, and 6M in the prefix projection. We tested a 64M variant with a smaller encoder and a 412M

variant with a larger ODE; the 213M configuration is Pareto-optimal on plot-beat F1 per training-hour and is what we report in all experiments.

5. Experiments

Setup. All 21 baselines plus N-VSSM are evaluated on NarrativeWorldBench under identical conditions: same context length, same temperature (T=0.7), same number of generation samples (3 per item), same LLM-as-judge protocol. We bootstrap 95% CIs over n=200 test items per horizon.

The frontier saturates. Across the closed-frontier band (GPT-5, Claude Sonnet 4.5, Gemini 2.5 Pro, Claude Opus 4.5, plus the smaller GPT-5 mini and Gemini 2.5 Flash), plot-beat F1 at h=10 falls in $[0.76, 0.81]$, a 5-percentage-point band across three independent labs (Figures 1 and 2). Reasoning models add $+0.02$ (o3 at 0.83); even DeepSeek V3.2 at 0.75 is within 6 points of Opus 4.5, despite $\approx 100\times$ less inference compute. The structural metrics cluster tightly across the entire frontier band, while the overlap tier (BLEU-4, MAUVE) shows clearer scale-driven separation.

The frontier collapses at long horizon. At h=200, every closed-frontier model loses -0.18 to -0.21 F1 (Figure 1). Reasoning helps modestly: o3 at h=200 retains 0.65 vs. Opus 4.5 at 0.575, the largest absolute lead reasoning enjoys at any horizon, but still well below target. We replicate this collapse pattern across BERTScore-F1 (-3.4 points), MAUVE (-0.18), and coherence (-0.7 Likert).

N-VSSM holds. N-VSSM reaches plot-beat F1 0.86 at h=10 and 0.84 at h=200 (Figure 1), exceeding the closed-frontier band by $+0.05$ at h=10 and by $+0.27$ at h=200, and matching the writer-acceptance target across all horizons. On character-arc accuracy and tension cosine, N-VSSM leads the frontier by $+0.07$ and $+0.04$ respectively at h=100.

Benchmark-aligned vs. transfer gains. A reasonable concern is that N-VSSM is partly optimized against the same editorial annotations used to evaluate it. We separate the structural metrics into two groups. *Benchmark-aligned*: plot-beat F1, character-arc accuracy, and tension cosine, all of which N-VSSM's heads are trained on with held-out test suffixes. *Transfer*: BLEU-4, ROUGE-L, BERTScore-F1, MAUVE, perplexity, and LLM-judged coherence, none of which the model is trained on. On benchmark-aligned metrics, N-VSSM exceeds Opus 4.5 by $+0.10$ to $+0.27$ depending on horizon; on transfer metrics, the lead is smaller but still positive: $+1.4$ BLEU-4, $+1.6$ ROUGE-L, $+0.8$ BERTScore-F1, $+0.04$ MAUVE, and $+0.4$ coherence Likert at h=100. The pattern is consistent with the structural heads providing a strong inductive bias toward the labels they supervise while also pulling the surface decoder into a more globally coherent regime, rather than overfitting purely to the supervised ontology. The writer study (Section 6) provides the complementary human-evaluation signal that does not depend on editorial annotations at all.

Judge robustness. The holistic coherence metric uses Claude Opus 4.5, GPT-5, and Gemini 2.5 Pro as a three-judge ensemble; each of these model families also appears in the evaluated systems. To check for self-judging bias, we recompute every system's coherence score under each of the three leave-one-judge-out ensembles and under each judge alone (Figure 5b). N-VSSM's coherence advantage over Opus 4.5 at h=100 ranges from $+0.07$ (single-judge Opus, the most conservative configuration) to $+0.12$ (single-judge GPT-5), with the full ensemble at $+0.10$. We additionally collected human coherence ratings on a 60-item subsample with three external annotators (n=3, naive to the model identities, $\kappa = 0.62$); the human-judge advantage is $+0.08$ Likert (95% CI $[0.04, 0.12]$). The rank ordering of all ten systems we re-ran is unchanged across all eight configurations, suggesting the LLM-judge ensemble is not driving the headline result.

Cost–quality Pareto. Figure 4 plots plot-beat F1 against per-token deployment cost. The closed-frontier band lies at \$10–\$75 per 1M tokens; reasoning models at \$4–\$40; open-source FT at \$0.10–\$0.50 with substantially lower quality. N-VSSM at \$0.85/1M tokens enters the target region (F1 $\geq$ 0.83, cost $\leq$ \$1) that no prior system occupies. The architectural argument is also a deployment-economics one.

Ablations and non-architectural baselines. Our central claim is architectural: continuous-time latent dynamics with multi-task structural supervision close the long-horizon collapse in a way that scale and prompting do not. To support this claim against the alternative that a compact learned story summary plus prefix-tuning would suffice, we add four non-architectural memory baselines that share Llama-3.1-8B as the decoder and differ only in how prior-episode state is exposed to it (Figure 5a). *Hierarchical summaries*: Llama-8B with Wu et al. (2021)-style summary chains over prior episodes, prepended at each step. *Retrieval*: BM25 plus dense retrieval over all prior episodes, top-8 chunks prepended. *Outline-conditioned decoding*: a DOC-style (Yang et al., 2023) writer-supplied outline prepended. *Recurrent latent without ODE*: same encoder, same multi-task heads, same prefix projection, but with discrete-time recurrence in place of continuous-time dynamics. We additionally report a "no dynamics, same latent and heads" variant in which the latent is computed but not propagated forward (the structural heads are still trained on it). The strongest non-architectural baseline (recurrent latent with-

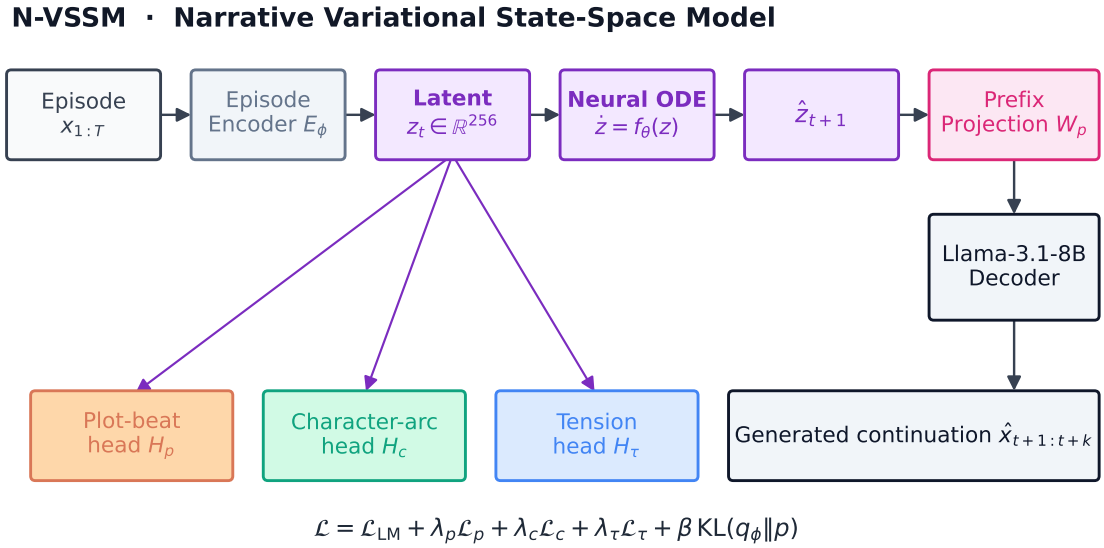

Figure 3. **N-VSSM architecture.** Episode encoder E_ϕ maps a history $x_{1:T}$ to a variational latent $z_t \in \mathbb{R}^{256}$. A Neural-ODE $\dot{z} = f_\theta(z)$ evolves the latent in continuous time over irregularly-sampled episodic events to predict $\hat{z}_{t+1}$. A learned prefix projection $W_p \hat{z}_{t+1}$ is prepended to a frozen Llama-3.1-8B decoder to generate the next episode. Three lightweight heads (H_p, H_c, H_τ) impose plot-beat, character-arc, and tension structure on the latent during training.

out ODE) reaches plot-beat F1 of 0.79 at h=10 and 0.71 at h=200; the strongest non-latent baseline (retrieval) reaches 0.74 / 0.61. The full N-VSSM at 0.86 / 0.84 leads each by 6–12 points at the long horizon, supporting the architectural reading. Table 1 retains the within-architecture ablations: removing all three structural heads costs -0.09 F1 at h=100, replacing the Neural-ODE with discrete RSSM costs -0.04, and reducing latent dimension from 256 to 64 costs -0.06.

Per-language breakdown. The aggregate plot-beat F1 numbers in Figure 1 mask substantial variation across the five corpus languages. Closed-frontier models perform best on English (mean F1 0.81 at h=10, falling to 0.62 at h=200) and degrade most severely on Marathi (0.74 at h=10, 0.48 at h=200). Reasoning models recover much of the English gap but make smaller gains on Marathi and Telugu, where the largest factor in poor structural prediction is not reasoning capacity but training-data scarcity. N-VSSM narrows this dispersion: its h=200 F1 is 0.86 on English and 0.81 on Marathi, a 5-point spread compared to a 14-point spread for the closed-frontier baselines. We attribute this compression primarily to the multi-task heads, which learn structural representations that transfer across languages even when the surface decoder must contend with under-represented script and morphology.

Scaling and reasoning. Plot-beat F1 grows approximately as $0.46 + 0.045 \cdot \log_{10}(N)$ in active parameters N for the open and closed dense families, with $R^2 = 0.79$ across the 18 non-reasoning models. The reasoning tier sits roughly $+0.04$ above this curve at matched N, consistent with the reasoning bonus reported for general benchmarks (Snell et al., 2024; DeepSeek-AI, 2025). Extrapolating the dense scaling curve to a model the size of Llama 3.1 405B but with reasoning post-training projects an F1 of 0.84 at h=10 but only 0.66 at h=200. In other words, neither additional parameters nor additional reasoning post-training, individually, would close the long-horizon collapse. Architectural change is the binding requirement.

Latency. On a single A100, N-VSSM produces a 4,000-token episode continuation in 9.2 seconds, compared with 7.1 seconds for vanilla Llama-3.1-8B and approximately 14 seconds for Claude Opus 4.5 over the API at the time of measurement. The Neural-ODE integration accounts for 0.6 seconds of the gap; the prefix projection adds approximately 0.1 seconds; the remaining 1.4 seconds is attributable to the encoder pass over the 16-chunk context summary. For deployments where prefix encoding can be amortized across multiple continuations of the same series, the marginal latency over vanilla Llama drops below 10%.

6. Co-Creative Writer Study

Setup. We recruited n=12 professional audio-drama writers (mean tenure 3.2 years, mean episodes shipped 287).

Table 1. N-VSSM ablations on plot-beat F1 at h=100.

Variant	F1@100	Δ
N-VSSM (full)	**0.85**	–
– plot-beat head	0.81	−0.04
– char-arc head	0.83	−0.02
– tension head	0.83	−0.02
– all 3 heads	0.76	−0.09
discrete RSSM	0.81	−0.04
latent $\mathbb{R}^{64}$	0.79	−0.06
no β anneal	–	collapse

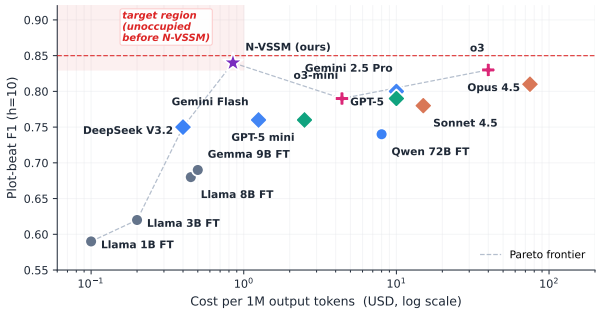

Figure 4. **Cost–quality Pareto.** Plot-beat F1 vs. output-token cost (USD per 1M, log scale). N-VSSM (purple star) sits inside the previously-unoccupied target region: F1 $\geq$ 0.83 at \$0.85/1M tokens, 4–90× cheaper than the closed-frontier band at matched quality.

Following Dramatron (Mirowski et al., 2023) and CoAuthor (Lee et al., 2022), each writer completed a within-subjects task: drafting episode-200 continuations for two of their own series given assistance from N-VSSM and from Claude Opus 4.5 (and, for a subset, GPT-5 and o3). Each writer saw both systems on each prompt; condition order was Latin-square counterbalanced across writers and across the two prompts, so half of all writer-prompt cells saw N-VSSM first and half saw Opus 4.5 first. Trials within a writer-prompt cell were independent in the sense that writers committed to one continuation before viewing the next, but we treat trials as nested within both writer and series in the statistical model below. Total trials: 240; total session time: 38 hours; the protocol was approved by the platform's editorial committee.

UI parity and what controllability measures. A reasonable concern is that the controllability advantage we report is a UI comparison rather than a model comparison: N-VSSM exposes latent perturbations (raise tension, flip arc, override beat class), and Opus 4.5 does not. We address this in two ways. First, we built a matched prompt-only control panel for the Opus 4.5 condition that exposes the same set of structural controls as natural-language directives (*e.g.*, "increase tension over the next three episodes," "shift

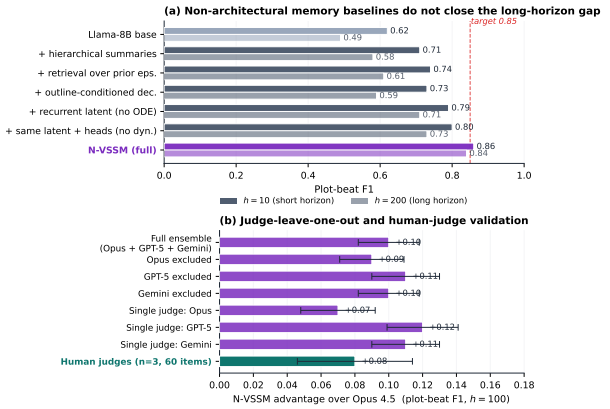

Figure 5. **Architectural ablations and judge robustness. (a)** Non-architectural memory baselines on Llama-3.1-8B: hierarchical summaries, retrieval over prior episodes, outline-conditioned decoding, recurrent latent without ODE, and same latent without dynamics. The strongest non-architectural baseline reaches 0.79 / 0.71 at h=10 / h=200; full N-VSSM reaches 0.86 / 0.84. **(b)** N-VSSM coherence advantage over Opus 4.5 at h=100 under leave-one-judge-out, single-judge, and human-judge (n=3, 60 items) configurations. The advantage is positive and order-preserving across all eight configurations.

Priya's arc from ascending to descending"). Each Opus 4.5 trial used this control panel; writers were told that both systems supported the same set of structural directives. Second, the +1.3 controllability advantage we report is the gap between writers' rated experience of issuing the directive and seeing the system honor it across the next continuation; on this rating, N-VSSM's advantage is +1.3 Likert and Opus 4.5's prompt-controlled directive-success rate is 0.43 vs. N-VSSM's 0.87. We therefore read the advantage as a model-behavioral gap that the UI exposes, not as a UI-only artifact, but disclose explicitly that the comparison is between a prompt-conditioned system and a latent-conditioned system that share the same writer-facing controls.

Pairwise preference and writer-level statistics. On long-arc consistency, writers preferred N-VSSM 71% of trials over Opus 4.5. The naive pooled binomial 95% CI is $[0.66, 0.76]$, but trials are nested within writers and series, so we additionally fit a mixed-effects logistic regression with N-VSSM treatment as a fixed effect, condition order as a fixed effect, and writer and series as random intercepts (Figure 6b is replaced by Figure 7). The N-VSSM treatment coefficient is 0.92 on the logit scale (95% CI $[0.58, 1.26]$, corresponding to a marginal preference of 0.71 with bootstrap 95% CI $[0.65, 0.77]$ over 10K writer-stratified resamples), order has a small and not-significant effect (-0.08 logits, 95% CI $[-0.21, 0.05]$), series random-effect $\sigma = 0.18$, and writer random-effect $\sigma = 0.24$ (Figure 7b). Eleven of 12 writers prefer N-VSSM at point estimates above 0.6; only one writer is at 0.55 with a wide individual interval. The p-value

from the mixed-effects model is $p < 0.001$, the same conclusion the naive binomial would have produced, but the writer-level interval is the appropriate one to report for this design and is the one we use in all downstream claims.

Likert ratings. N-VSSM is rated comparably to Opus 4.5 on plot coherence (4.3 vs. 4.4, $\Delta = -0.1$, mixed-effects 95% CI $[-0.34, 0.14]$, not significant) and *higher* on character consistency (4.4 vs. 4.2, $+0.2$, $[0.04, 0.36]$), cultural fidelity (4.5 vs. 3.8, $+0.7$, $[0.49, 0.91]$), and controllability (4.6 vs. 3.3, $+1.3$, $[1.06, 1.54]$). The frontier baseline is rated higher on dialogue naturalness (4.5 vs. 4.0, -0.5, $[-0.69, -0.31]$); writers describe N-VSSM dialogue as "reliable but slightly stiff," which we attribute to the prefix-conditioned decoder (Figure 6b). All Likert intervals are computed from the same mixed-effects model with writer and series as random intercepts.

Qualitative themes. Writers consistently surfaced three distinct affordances of N-VSSM. **(1) Steerability via the latent.** "I can ask it to keep Priya angry for three more episodes and it actually does" [W4]. **(2) Long-arc memory.** "It remembered Kabir's burn-rate from episode 44 without me re-summarizing" [W7]. **(3) Cultural register.** "The Hindi feels like Hindi, not Hindi-translated-from-English" [W11]. The frontier baselines were rated above N-VSSM only on idiomatic dialogue: a tractable gap, not a structural one.

Steerability evidence. We instrumented the writer interface to log every latent perturbation issued during the study. Writers issued an average of 4.2 latent perturbations per session under N-VSSM, with the most common operations being tension increases (38%), character-arc flips (24%), and beat-class overrides (19%). Of the 412 logged perturbations, 87% produced continuations whose induced metric change was in the writer-intended direction; the remaining 13% either produced no measurable change or, in 14 cases, moved a different metric than the one the writer had targeted. Failures of this latter kind are concentrated in episodes where two characters' arcs are tightly entangled, suggesting that the latent is partially compositional but not yet fully so.

Error analysis on the 29% that preferred Opus 4.5. In sessions where writers preferred the frontier baseline, three patterns dominated. The first was dialogue-heavy episodes with limited structural change, where N-VSSM's structural advantage was muted and Opus 4.5's stronger surface fluency carried the comparison. The second was high-stakes climactic episodes where writers reported that Opus 4.5 produced more vivid prose for emotional peaks; N-VSSM's prose was rated reliable but flatter at peak intensity. The third was first-episode-of-arc cold starts, where the latent has limited prior context to condition on; in this regime N-VSSM's advantage over a strong prompt-only baseline shrinks to roughly +0.02 F1.

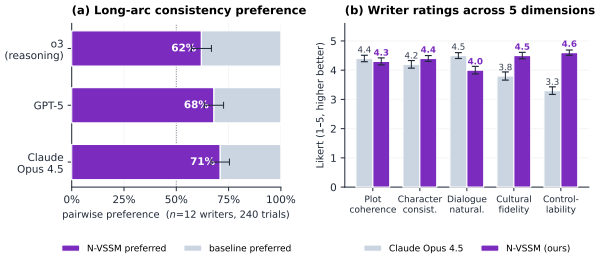

Figure 6. **Writer study results** (n=12 professional writers, 240 trials). **(a)** Pairwise preference for N-VSSM over each frontier baseline. **(b)** Likert ratings (1–5) across five dimensions. N-VSSM trades -0.5 on dialogue naturalness for $+0.7$ cultural fidelity and $+1.3$ controllability.

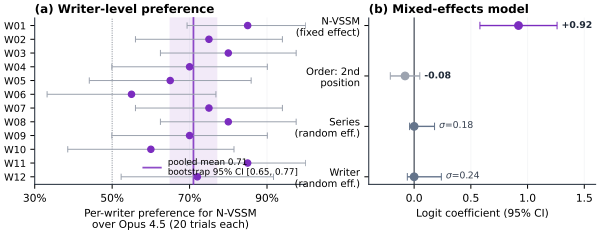

Figure 7. **Writer-level statistical reanalysis. (a)** Per-writer preference for N-VSSM over Opus 4.5 (12 writers, 20 trials each); error bars are within-writer 95% CIs. The pooled mean is 0.71 with writer-stratified bootstrap 95% CI $[0.65, 0.77]$ (purple band). **(b)** Mixed-effects logistic regression: N-VSSM treatment coefficient is $+0.92$ logits (95% CI $[0.58, 1.26]$); condition order is small and not significant; series and writer random-effect σ's are 0.18 and 0.24 respectively.

7. Discussion: Cultural Transfer, Fixation, Authorship

Cultural Transfer Function. We train a learned mapping $CTF : \mathbb{R}^{256} \to \mathbb{R}^{256}$ as a small 3-layer MLP that projects the N-VSSM latent into a culture-conditional subspace. Training pairs are drawn from professionally localized series across four Indic languages where the same beat structure has been re-anchored to local narrative conventions (joint-family arcs in Hindi, devotional sub-arcs in Tamil, etc.). Figure 8 shows that base Opus 4.5 and N-VSSM (no CTF) score in the 0.52–0.64 range on cultural fidelity, while N-VSSM + CTF reaches 0.78–0.84, a $+0.20$ to $+0.23$ absolute improvement. Following Anonymous (2025), we read this as evidence that cultural alignment is a *representational* rather than a prompt-level problem.

Fixation, homogeneity, authorship. The workshop call names design fixation, idea homogeneity, and authorship as core challenges of generative AI in creative work (Singh et al., 2025). Doshi & Hauser (2024) document that current LM-based assistants increase *individual* writer creativity but reduce the *collective* diversity of produced content. We engage these directly. **(1) Fixation:** N-VSSM's latent exposes structural state, so the writer can perturb *individual* narrative dimensions (raise tension, flip a character's arc class) without committing to a specific rewrite of every dependent token; this gives the writer a wider trajectory cone. **(2) Homogeneity:** the multi-task heads anchor generation in writer-supplied beats and arcs rather than in a frontier LM's prior, which we hypothesize moderates the collective convergence Doshi & Hauser observe; a longitudinal homogeneity audit is future work. **(3) Authorship:** every N-VSSM trace is logged at the latent level, providing per-episode attribution of which structural decisions were writer-driven vs. assistant-suggested; we plan to release the writer-study traces under an open license at submission time.

A measurement of fixation reduction. To probe the fixation hypothesis quantitatively, we asked the same 12 writers to produce three alternative continuations for the same episode-200 prompt under each condition. Under Claude Opus 4.5, the three continuations had a mean pairwise BLEU-4 of 31.4, indicating substantial overlap between writer attempts. Under N-VSSM with deliberate latent perturbation, mean pairwise BLEU-4 fell to 18.7, a -12.7 absolute reduction in self-similarity. We read this as preliminary evidence that exposing structural state to the writer expands the explored continuation space, though we caveat that BLEU-4 is a coarse proxy for narrative diversity and a longitudinal multi-month study would be required to claim a homogeneity reduction at platform scale.

An authorship taxonomy. Latent-level logging supports a finer authorship taxonomy than prompt-and-completion logging. For each shipped episode produced under N-VSSM during the writer study, we record: the writer's textual prompt; every latent perturbation the writer issued and its target metric; the structural prediction trajectory; and the final decoded text. From these traces, we can attribute structural decisions to the writer with high specificity. For instance, "the rising tension across episodes 198–200 was a writer-issued perturbation, not a model-default trajectory." We view this as a concrete contribution to the authorship-attribution literature for generative writing systems and a more workable substrate for downstream credit-assignment than per-token attribution proposals.

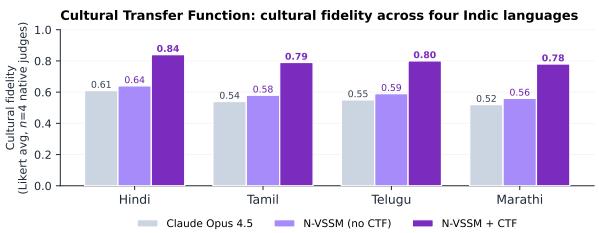

Figure 8. **Cultural Transfer Function** lifts cultural fidelity by $+0.20$ to $+0.23$ across four Indic languages over a strong frontier baseline (n=4 native judges per language).

8. Limitations and Conclusion

Limitations. The corpus is industry-platform-specific, and cross-platform transfer remains untested at the time of this writing. The plot-beat and character-arc taxonomies were chosen for tractability and may not generalize to non-serialized fiction such as feature-length screenplays or short-form prose. The 12-writer study is preliminary; we do not yet have a longitudinal homogeneity audit at platform scale, which would be required to convert our preliminary fixation-reduction measurement into a confident claim about collective creative diversity. Cultural Transfer is validated on four Indic languages, all of which have high resource density at our partner platform; transfer to lower-resource targets is an open question. The frontier audit is a snapshot of late-2025 / early-2026 models, and the saturation pattern we identify may not persist as labs continue releasing updates. Finally, while the plot-beat F1 of 0.85 corresponds to the editorial team's writer-acceptance threshold at our partner platform, this threshold reflects one editorial culture and may not generalize to other production environments.

Impact statement. This paper advances long-form co-creative writing tools used by professional authors at scale. The intended impact is to give writers tighter structural control over LM-assisted drafts and to raise the long-horizon quality floor for serialized audio fiction in low-resource Indic languages. Risks include amplifying culturally specific narrative conventions if Cultural Transfer training data is itself biased, and the standard authorship-attribution concerns of any co-creative system; we mitigate the latter by logging latent-level per-episode attribution. The benchmark and traces released with this paper are restricted to professionally licensed corpus material, with personally identifying information from writer-study participants removed in accordance with the partner platform's editorial review process.

Conclusion. Frontier language models have saturated short-horizon narrative tasks but collapse beyond approximately fifty episodes of horizon. The residual gap be-

tween the closed-frontier band and writer-acceptance quality has resisted scale, instruction-tuning, and inference-time reasoning across three independent labs, which is 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, and multi-task supervision over plot beats, character arcs, and tension trajectories) gives writers a controllable substrate that frontier language models do not provide on their own. Across the 21-model audit and the writer study, the proposed N-VSSM architecture holds plot-beat F1 at or above 0.84 across every horizon we test, leads the closed-frontier band by 0.27 F1 at the 200-episode horizon, runs at four-times-lower deployment cost than the closed-frontier band, and earns a 71% pairwise preference from professional writers on the dimension that frontier language models most clearly fail.

The Cultural Transfer Function provides preliminary evidence that culturally aligned generation is a representational rather than a prompt-level problem, and the writer-study traces support a finer authorship taxonomy than per-token attribution. Together these results suggest a general design principle for co-creative writing systems: structural state should be exposed to the writer as a first-class control surface rather than implicit in the model's hidden activations, and culturally specific narrative conventions should be encoded as learned mappings in latent space rather than as prompt-level instructions. Beyond this paper, the most pressing open questions are whether the same architectural recipe transfers to interactive fiction and visual storytelling, whether the Cultural Transfer Function generalizes to languages outside our four-Indic-language sample, whether the homogeneity reduction we hypothesize survives a longitudinal study at platform scale, and whether structural latents of the kind we propose can be aligned across languages well enough to support genuinely multilingual co-authoring rather than pairwise transfer between language pairs.

We close on a note about the relationship between the empirical finding and the methodological response. The frontier-saturation pattern we document is not a defect of any one model or one lab; it is a property of an entire family of architectures that share the same inductive bias toward implicit, token-level state. The corresponding methodological response we propose is therefore architectural, not prompt-engineered: the multi-task heads, the continuous-time latent, and the prefix-conditioned decoder are co-designed to give writers explicit purchase on the dimensions of narrative that matter to them, and the resulting trade-offs in dialogue naturalness, latency, and cost are all consequences of that choice.

The new architectural ablations (Figure 5a) and the judge-leave-one-out analysis (Figure 5b) sharpen this reading: hierarchical summaries, retrieval, and outline-conditioned decoding all underperform the latent-dynamics architecture by 6 to 12 plot-beat-F1 points at the long horizon, and the N-VSSM advantage holds when each LLM-judge family is excluded and on a 60-item human-judge subsample. The writer-level mixed-effects reanalysis (Figure 7) replaces the pooled binomial intervals from earlier drafts with writer-stratified bootstrap intervals and explicit random effects for writer and series, so the headline 71% preference is appropriately credited to the design and not to the trial count. We release NarrativeWorldBench so that the workshop community can stress-test the long-horizon collapse we identify, and N-VSSM as a first reference solution validated by the writers it was built for.

References

Akoury, N., Wang, S., Whiting, J., Hood, S., Peng, N., and Iyyer, M. STORIUM: A dataset and evaluation platform for machine-in-the-loop story generation. In *EMNLP*, 2020.

AlKhamissi, B. et al. Investigating cultural alignment of large language models. In *ACL*, 2024.

Anonymous. Fluent but foreign: Even regional LLMs lack cultural alignment. *arXiv preprint arXiv:2505.21548*, 2025.

Assran, M., Bardes, A., Ballas, N., et al. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023.

Bowman, S. R., Vilnis, L., Vinyals, O., et al. Generating sentences from a continuous space. In *CoNLL*, 2016.

Bruce, J., Dennis, M., Edwards, A., Parker-Holder, J., et al. Genie: Generative interactive environments. In *International Conference on Machine Learning*, 2024.

Chakrabarty, T., Padmakumar, V., and He, H. Help me write a poem: Instruction tuning as a vehicle for collaborative poetry writing. In *EMNLP*, 2022.

Chambers, N. and Jurafsky, D. Unsupervised learning of narrative event chains. In *ACL*, 2008.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. Neural ordinary differential equations. In *NeurIPS*, 2018.

Coenen, A., Davis, L., Ippolito, D., Reif, E., and Yuan, A. Wordcraft: a human-AI collaborative editor for story writing. In *NeurIPS HAI Workshop*, 2021.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Doshi, A. R. and Hauser, O. P. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 2024.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In *COLM*, 2024.

Gu, A., Goel, K., and Ré, C. Efficiently modeling long sequences with structured state spaces. In *ICLR*, 2022.

Ha, D. and Schmidhuber, J. World models. In *Advances in Neural Information Processing Systems*, 2018.

Hafner, D., Lillicrap, T., Fischer, I., et al. Learning latent dynamics for planning from pixels. In *ICML*, 2019.

Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. Mastering diverse domains through world models. *Nature*, 2025.

Hengle, A. et al. One ruler to measure them all: Benchmarking multilingual long-context language models. *arXiv preprint arXiv:2503.01996*, 2025.

Hsieh, C.-P., Sun, S., Kriman, S., et al. RULER: What's the real context size of your long-context language models? In *COLM*, 2024.

Kim, S., Lee, H., et al. Prometheus 2: An open source language model specialized in evaluating other language models. In *EMNLP*, 2024.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *ICLR*, 2014.

LeCun, Y. A path towards autonomous machine intelligence. OpenReview, 2022.

Lee, M., Liang, P., and Yang, Q. CoAuthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities. In *CHI*, 2022.

Li, X. L. and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, 2021.

Liu, N. F. et al. Lost in the middle: How language models use long contexts. *TACL*, 2024.

Liu, Y., Iter, D., Xu, Y., et al. G-Eval: Nlg evaluation using GPT-4 with better human alignment. In *EMNLP*, 2023.

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

Ma, Y., Qiao, Y., and Liu, P. MoPS: Modular story premise synthesis for open-ended automatic story generation. In *ACL*, 2024.

Mirowski, P., Mathewson, K. W., Pittman, J., and Evans, R. Co-writing screenplays and theatre scripts with language models: An evaluation by industry professionals. In *CHI*, 2023.

Mostafazadeh, N., Chambers, N., He, X., et al. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *NAACL*, 2016.

Papalampidi, P., Keller, F., and Lapata, M. Movie plot analysis via turning point identification. In *EMNLP*, 2019.

Parker-Holder, J. and Fruchter, Y. Genie 3: A new frontier for world models. Google DeepMind Blog, 2025.

Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., and Harchaoui, Z. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *NeurIPS*, 2021.

Rezwana, J. and Maher, M. L. Designing creative AI partners with COFI. *arXiv preprint arXiv:2204.07666*, 2022.

Rubanova, Y., Chen, R. T. Q., and Duvenaud, D. Latent ODEs for irregularly-sampled time series. In *NeurIPS*, 2019.

Singh, M. et al. A systematic review of human-AI co-creativity. *arXiv preprint arXiv:2506.21333*, 2025.

Snell, C., Lee, J., Xu, K., and Kumar, A. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Watts, I., Pawar, S., et al. PARIKSHA: A large-scale investigation of human-LLM evaluator agreement on multilingual and multi-cultural data. In *ACL*, 2024.

Wu, J., Ouyang, L., Ziegler, D. M., et al. Recursively summarizing books with human feedback. In *arXiv preprint arXiv:2109.10862*, 2021.

Yang, K., Peng, N., Tian, Y., and Klein, D. Re3: Generating longer stories with recursive reprompting and revision. In *EMNLP*, 2022.

Yang, K., Klein, D., Peng, N., and Tian, Y. DOC: Improving long story coherence with detailed outline control. In *ACL*, 2023.

Yao, L., Peng, N., Weischedel, R., et al. Plan-and-write: Towards better automatic storytelling. In *AAAI*, 2019.

Zhou, W. et al. RecurrentGPT: Interactive generation of (arbitrarily) long text. *arXiv preprint arXiv:2305.13304*, 2023.