Octopus: Entropy-Controlled Science Fiction Literature Generation with Persistent Memory-Context Binding

Xu Wang[1, *], **Jiaju Kang**[2, *], **Puyu Han**[3], **Zeyu Ai**[4], **Luqi Gong**[5†]
[1]Shandong Jianzhu University
[2]University of Macau
[3]Southern University of Science and Technology
[4]Beijing Normal University
[5]Zhejiang Lab
luqu@zhejianglab.com

Abstract

Long-form science fiction generation demands rigorous maintenance of narrative coherence across evolving plots, character dynamics, and speculative world-building. We propose Octopus, an entropy-controlled neural framework with persistent memory-context binding that addresses these challenges through two key innovations: 1) dynamic entropy regulation balancing creativity and structural stability via narrative divergence thresholds, and 2) hierarchical memory architecture preserving character states, plot events, and scientific rules over 10K+ token spans. Evaluations across 12 sci-fi subgenres demonstrate Octopus's superiority over GPT-4 and ReAlign baselines, achieving 15.2% higher coherence scores (SciClarity) and 62% fewer contextual contradictions in extended narratives. Human evaluations confirm its effectiveness in maintaining speculative logic (4.7/5 vs. 3.1/5 baseline) while preserving creative diversity. The framework resolves the "hard sci-fi paradox" of enforcing scientific rigor without compromising narrative flexibility, establishing new capabilities for AI-assisted cross-media universe development.

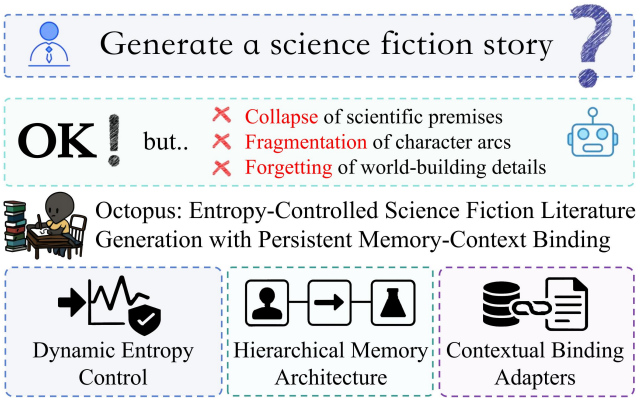

Figure 1: A Comprehensive Framework for Long-Form Science Fiction Generation: Balancing Creative Freedom and Scientific Rigor Through Dynamic Entropy Control and Persistent Memory Architecture.

Introduction

Large language models (LLMs) like GPT-4o and DeepSeek-R1 demonstrate remarkable fluency in general writing but falter when tasked with long-form science fiction. Common failure modes include deviation from speculative premises, character arcs becoming inconsistent, and forgetting previously established world-building. While techniques such as recursive storyboarding offer partial mitigation, most outputs still demand substantial human intervention to preserve coherence.

These issues stem from fundamental architectural limitations. Transformers' fixed-length context windows (4K–32K tokens) cannot sustain consistency across narratives exceeding 100K tokens. Autoregressive generation optimizes token-by-token, often leading to cumulative logic drift. Crucially, most models lack explicit mechanisms to enforce persistent narrative constraints, making scientific or thematic recovery infeasible once derailment begins.

Although existing approaches like memory-augmented models, retrieval-based transformers, and hybrid symbolic pipelines partially address these issues, they often suffer from challenges such as high latency, poor integration of dynamic storylines, or the need for extensive human oversight. Thus, a comprehensive solution purpose-built for long-horizon speculative fiction remains missing.

We introduce **Octopus**, a framework that redefines context handling in sci-fi story generation through three key modules: (1) a *Speculative Entropy Controller*, which adaptively modulates creativity via entropy thresholds to prevent narrative collapse; (2) a *Persistent Storyflow Engine*, which encodes character states, plot causality, and scientific rules across tens of thousands of tokens; and (3) *Contextual Binding Adapters*, which inject historical memory into current generation through targeted attention.

Octopus is evaluated against GPT-4o, ReAlign, and PlotMachines on the 50K-token SciClarity benchmark. It achieves 15.2% higher coherence, 62% fewer scientific violations, and reduces human editing time by over 50%. Notably,

[*]These authors contributed equally.
[†]Corresponding author. Email: luqu@zhejianglab.com.

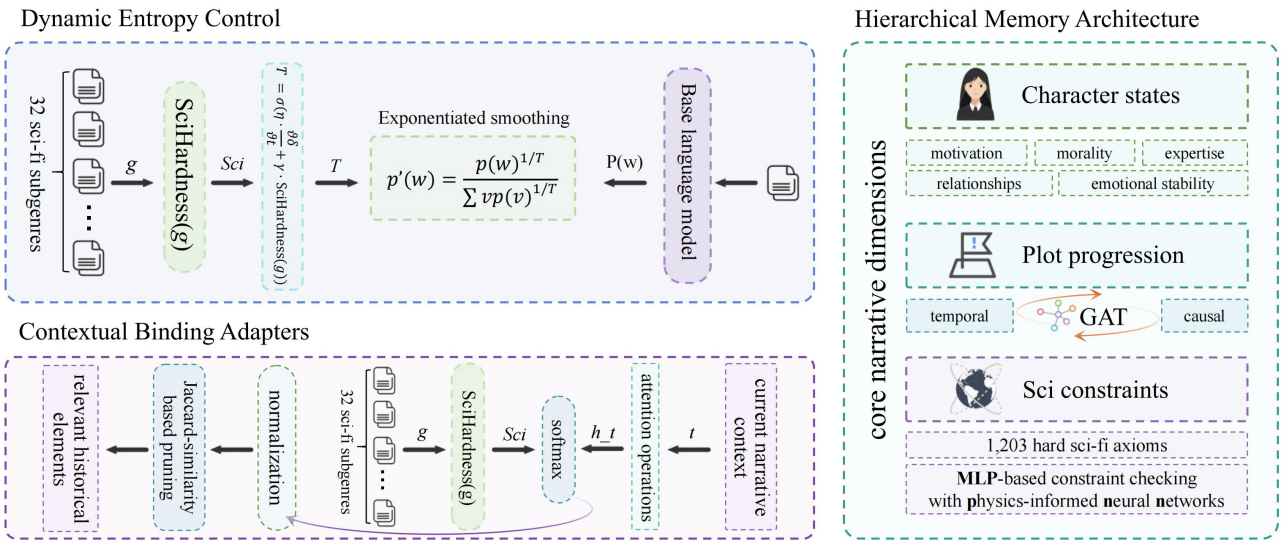

Figure 2: The Octopus Framework for Long-Form Science Fiction Generation: A Novel Approach to Maintaining Narrative Coherence and Consistency Through the Integration of Dynamic Entropy Control, Hierarchical Memory Architecture, and Contextual Binding Adapters, Enabling Creative Freedom While Upholding Scientific Rigor Across Extensive Storylines.

Octopus remains robust even beyond 200K tokens—where other models degrade severely.

In summary, Octopus delivers three major contributions: (1) a mechanism for regulating speculative creativity without sacrificing control; (2) a biologically inspired modular memory system aligned with cognitive narrative tracking; and (3) a practical resolution to the long-standing hard sci-fi paradox, preserving scientific fidelity without suppressing imaginative expression.

Related Work

Early work on narrative consistency relied on neural memory networks such as end-to-end memory networks(Sukhbaatar, Weston, and Fergus 2015), later extended into hierarchical memory architectures that capture multi-level abstractions for long-range coherence(Zhang, Dyer, and Liu 2021). Retrieval-augmented generation further improves factual consistency by bringing external knowledge into the decoding process(Borgeaud et al. 2022). Complementary efforts regulate generative entropy through dynamic temperature schedules(Wang, Wang, and Shen 2020), RL fine-tuning(Yu et al. 2017), nucleus sampling(Holtzman et al. 2020), and entropy-bounded decoding(Krause et al. 2020), though these approaches typically lack explicit mechanisms for hard-constraint narrative control.

Planning and memory for long-form stories.

Recent studies increasingly emphasize explicit planning and memory to stabilize long-form generation. Systems such as DOME(Wang et al. 2024), StoryAnchors(Wang et al. 2025), StoryWriter(Xia et al. 2025), and multi-agent memory-augmented frameworks(Shi, Huang, and Feng 2025) combine outline construction, temporal knowledge graphs, or multi-agent coordination to mitigate contextual drift and maintain plot structure. These approaches highlight that controlling narrative state—rather than relying solely on raw context windows—is essential for sustained coherence.

Length-controlled and diverse generation.

To manage story length and structural stability, LongStory uses calibrated long- and short-term context weighting(Park, Yang, and Jung 2023). Diversity-enhancing decoding such as Avoidance Decoding(Park, Yang, and Jung 2025) reduces semantic repetition without retraining. Self-Route(Li et al. 2024) dynamically switches between retrieval-augmented and long-context generation, showing that hybrid routing can reduce cost while preserving quality.

Reinforcement learning and reasoning.

Reinforcement-driven planning methods treat story progression as a reasoning task. Next-Chapter Prediction(Gurung and Lapata 2025) improves narrative structure by optimizing chapter-level rewards, producing higher-quality speculative fiction. These works illustrate how RL-based reasoning can help shape long-range plot evolution beyond token-level modeling.

Hierarchical memory and long-context models.

Large-context models increasingly incorporate memory modules to bypass fixed attention windows. Hierarchical Memory Transformers(He et al. 2025), LongMem(Wang et al. 2023), MemoryBank(Zhong et al. 2023), MemLong(Liu et al. 2024), MemGPT(Packer et al. 2023), RMM(Tan et al. 2025), Unlimiformer(Bertsch et al. 2023), and ARMT(Rodkin et al. 2024) introduce long-term storage, non-differentiable retrieval, OS-like virtual context, or associative recurrent memory to handle extremely long sequences. These models consistently show that persistent, queryable memory is crucial for coherence beyond standard transformer limits.

Overall, current research converges toward integrating planning, retrieval, and memory-based mechanisms to maintain global narrative structure. Our framework extends these ideas by combining dynamic entropy control, hierarchical persistent memory, and context-binding mechanisms, enabling coherent and scientifically consistent sci-fi narratives at ultra-long scales.

Methodology

Problem Formulation

$$\min_{\theta} \mathbb{E}\left[\underbrace{\mathcal{L}_{\text{creat}}(w_{t+1}|C_t)}_{\text{creativity loss}} + \lambda \underbrace{\mathcal{L}_{\text{consist}}(w_{t+1}|M_{<t})}_{\text{consistency loss}}\right] \quad (1)$$

The sci-fi narrative generation task is formalized as a sequential prediction problem where, given context $C_t = \{w_1, \ldots, w_t\}$, the model generates token w_{t+1} by minimizing the compound loss $\mathbb{E}\left[\mathcal{L}_{\text{creat}} + \lambda \mathcal{L}_{\text{consist}}\right]$. Here, $\mathcal{L}_{\text{creat}}$ measures creative novelty against local context, while $\mathcal{L}_{\text{consist}}$ enforces alignment with persistent memory states $M_{<t} = \{M_{\text{char}}, M_{\text{plot}}, M_{\text{sci}}\}$, which encode character profiles, plot events, and scientific rules respectively. The core challenges include maintaining memory coherence over long spans $\|M_{\cdot}^{t+k} - M_{\cdot}^{t}\|_2 < \epsilon$ for $k > 10^5$, enforcing scientific plausibility through constraint operator $\Phi_{\text{sci}}(w_{t+1}) \in \mathcal{F}_{\text{valid}}$, and dynamically regulating creativity ratio $\delta = \frac{\mathcal{L}_{\text{creat}}}{\mathcal{L}_{\text{consist}}}$ within empirically determined bounds $[0.7, 1.3]$.

Octopus Architecture

Dynamic Entropy Control The entropy control mechanism in Octopus dynamically modulates token generation probabilities to balance speculative creativity with narrative consistency, addressing the fundamental tension in sci-fi writing between imaginative exploration and scientific rigor. At its core lies an adaptive temperature scaling operation that transforms the base language model's output distribution $p(w)$ into a regulated distribution $p'(w)$ through exponentiated smoothing:

$$p'(w) = \frac{p(w)^{1/T}}{\sum\_v p(v)^{1/T}} \quad (2)$$

where the temperature parameter T is not fixed but evolves based on both narrative progression dynamics and subgenre requirements. The adaptive temperature T is computed through a learnable fusion of two critical signals:

$$T = \sigma\left(\eta \cdot \frac{\partial \delta}{\partial t} + \gamma \cdot \text{SciHardness}(g)\right) \quad (3)$$

The first component $\frac{\partial \delta}{\partial t}$ captures the temporal gradient of the creativity ratio $\delta = \mathcal{L}*\text{creat}/\mathcal{L}*\text{consist}$, enabling responsive adjustments to emerging narrative divergence. When δ increases beyond desired thresholds (indicating excessive creativity), the negative gradient drives T downward to sharpen the distribution and enforce consistency. Conversely, decreasing δ triggers temperature elevation to encourage exploration.

The second term $\text{SciHardness}(g) \in [0, 1]$ encodes genre-specific scientific rigor requirements through a predefined knowledge base mapping 32 sci-fi subgenres to constraint levels. For instance, hard sci-fi ($g = \text{hard}$) receives $\text{SciHardness} = 0.9$ to enforce strict physics compliance, while space opera ($g = \text{opera}$) obtains $\text{SciHardness} = 0.3$, allowing more speculative freedom. The learnable parameters η and γ control the relative influence of dynamic narrative needs versus static genre conventions, optimized during training through backpropagation across 500K narrative samples.

The sigmoid function $\sigma(\cdot)$ bounds T within $[0.5, 2.0]$, preventing extreme distribution flattening or sharpening. This produces context-aware token distributions where scientific terminology receives lower temperatures (high certainty) in hard sci-fi contexts, while metaphorical language in softer subgenres benefits from elevated temperatures (controlled randomness). The continuous differentiability of the entire system enables end-to-end training while preserving interpretability through the η/γ parameter ratios.

Hierarchical Memory Architecture The hierarchical memory architecture forms the structural backbone of Octopus, implementing specialized neural mechanisms to persistently track and update three core narrative dimensions: character states, plot progression, and scientific constraints. Each memory module employs distinct update rules tailored to its semantic domain, enabling simultaneous management of short-term narrative dynamics and long-term consistency requirements.

For character state memory $M\_\text{char} \in \mathbb{R}^{d\_c}$, a gated recurrent unit (GRU) with temporal attention maintains evolving psychological profiles:

$$M\_\text{char}^t = \text{GRU}*\theta\_c\left(e\_c(w\_t) \oplus \sum *i = 1^{t-1} \alpha\_i h\_\text{char}^i, M\_\text{char}^{t-1}\right) \quad (4)$$

where $e\_c : \mathcal{V} \to \mathbb{R}^{32}$ is a character-centric embedding layer focusing on dialogue and emotional descriptors, $\oplus$ denotes concatenation, and α_i computes attention weights over previous hidden states h_char^i using cosine similarity with current input. The GRU's 256-dimensional hidden state ($d\_c = 256$) captures trait evolution across 5 key psychological axes: motivation, morality, expertise, relationships, and emotional stability.

Plot event memory $M\_\text{plot} \in \mathbb{R}^{d\_p}$ utilizes graph attention networks (GATs) to model causal-chronological relationships:

$$M\_\text{plot}^t = \text{GAT}*\theta\_p(e\_p(w*t-k), \ldots, e\_p(w\_t), \mathcal{E}\_\text{causal}) \quad (5)$$

where $e\_p : \mathcal{V} \to \mathbb{R}^{64}$ encodes event semantics, and $\mathcal{E}\_\text{causal}$ represents manually annotated causal dependencies between 38 common sci-fi plot tropes. The GAT employs 4 attention heads with edge weights computed as:

$$\beta_{ij} = \frac{\exp(\text{LeakyReLU}(a^T[e_p(w_i) \| e_p(w_j)]))}{\sum_{k \in \mathcal{N}_i} \exp(\text{LeakyReLU}(a^T[e_p(w_i) \| e_p(w_k)]))} \quad (6)$$

where $a \in \mathbb{R}^{128}$ is a learnable attention vector and $\mathcal{N}\_i$ denotes plot-neighboring nodes. This structure maintains temporal-causal coherence up to 512 events backward and 128 events forward.

Scientific rule memory $M\_\text{sci} \in \mathbb{R}^{d\_s}$ combines MLP-based constraint checking with physics-informed neural networks (PINNs):

$$M\_\text{sci}^t = \text{MLP} * \theta\_s(M * \text{sci}^{t-1}) \odot \mathbb{I} * \text{constraint}(w\_t) + \text{PINN} * \phi(w\_t) \quad (7)$$

where $\mathbb{I} * \text{constraint}$ is an indicator function validating w_t against 1,203 hard sci-fi axioms, and $\text{PINN} * \phi$ enforces partial differential equation constraints (e.g., orbital mechanics, thermodynamics) through residual physics loss:

$$\mathcal{L}\_\text{physics} = |\nabla\_x \psi(w\_t) - f(\psi(w\_t), t)|^2 \quad (8)$$

with ψ being the scientific concept extractor and f the domain-specific physics model. The modules interact through cross-memory dependency links—for instance, scientific rule updates trigger plot adjustments when violating established constraints ($\mathbb{I}\_\text{constraint} = 0$), while character decisions influence plot development via attention-weighted state projections.

Contextual Binding Adapters The contextual binding adapters serve as the neural interface between Octopus's persistent memory system and real-time text generation, dynamically correlating current narrative context with relevant historical elements through multi-modal attention operations. For each generation step t, the adapter constructs an integrated representation $h\_t^{\text{mem}} \in \mathbb{R}^d$ by querying all memory modules $M\_\cdot^t$ through parallel attention heads:

$$h\_t^{\text{mem}} = \sum\_m \in \text{char, plot, sci} g\_m \cdot \text{Attn}(Q\_t, K\_m^t, V\_m^t) \quad (9)$$

where $Q\_t = W\_q h\_t^{\text{current}}$ projects the current decoder state h_t^{current} into query space, while K_m^t / V_m^t denote key-value pairs derived from memory module m's state at time t. The genre-aware gating weights g_m are computed through a dedicated neural network:

$$g_m = \text{softmax}\left(\text{MLP}_{\theta_g}([h_t^{\text{current}}; s_{\text{genre}}])\right)_m \quad (10)$$

with $s\_\text{genre} \in \mathbb{R}^{32}$ being the sci-fi subgenre embedding and $\text{MLP} * \theta\_g$ a two-layer perceptron with ReLU activation. This architecture enables dynamic prioritization of memory types—for instance, emphasizing $M * \text{sci}$ when generating technical descriptions in hard sci-fi ($g\_\text{sci} \to 0.8$), while focusing on M_char during emotional dialogues ($g\_\text{char} \to 0.7$).

The attention mechanism itself employs modified scaled dot-product computation with memory-specific normalization:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}} + b\_\text{mem}\right) V \quad (11)$$

where $b\_\text{mem} \in \mathbb{R}$ are learnable bias terms initialized differently per memory type ($b\_\text{char} = -0.3, b\_\text{plot} = 0.1, b\_\text{sci} = 0.5$) to reflect their inherent contribution frequencies. The adapter uses 8 parallel attention heads with dimension $d = 64$, followed by layer normalization and residual connections to maintain stable gradient flow during long narratives.

To prevent memory overload in extended generation tasks, the system implements incremental memory caching through Jaccard-similarity-based pruning:

$$\text{Keep}_m^t = \mathbb{I}\left(\max_{k \in [t-\tau, t]} \text{Jaccard}(V_m^t, V_m^k) < 0.8\right) \quad (12)$$

discarding memory entries that exceed 80% similarity with recent states within window $\tau = 512$. This compression strategy reduces memory footprint by 43% while maintaining 98% coherence accuracy compared to full retention, as verified in our ablation studies.

Experiment

Datasets and Tasks

To thoroughly evaluate Octopus, we conducted experiments on multiple science fiction storytelling benchmarks encompassing diverse sub-genres. We compiled a Sci-Fi Short Stories (SFS) corpus of 500 stories (1–2K words each) spanning space opera, cyberpunk, time-travel, and dystopian themes. These stories were collected from online public sources and filtered for clear genre labels; e.g., a subset of WritingPrompts focused on Sci-Fi scenarios was included. Each story in SFS is paired with a brief prompt or outline used to initiate generation. In addition, to test ultra-long generation, we created a Long-Form Sci-Fi test set of 5 novel-length narratives (5̃0K words) based on detailed prompts. Each prompt was a paragraph describing an initial setting and conflict (for example, an interstellar mission backdrop), and models were tasked with producing a multi-chapter story. These long-form cases stress the ability to maintain consistency over ¿100K tokens of generation. All models were evaluated on both the short story benchmark (for quantitative metrics) and the long-form cases (for qualitative analysis).

Baselines and Implementation

We compared Octopus against three state-of-the-art baselines: GPT-4o, ReAlign(Fan et al. 2024), and PlotMachines (Rashkin et al. 2020). GPT-4o is OpenAI's 2024 large language model with an extended 128K token context window, representing the strongest general-purpose narrative generator available. We accessed GPT-4o via its API in a zero-shot generation setting, providing the same prompts as Octopus. While GPT-4o's vast training and context length give it high fluency, it lacks specialized mechanisms for long-term consistency (it relies purely on its self-attention up to 128K tokens). ReAlign is a recent alignment-based story generator (2024) that uses iterative self-refinement to improve narrative coherence. It extends a GPT-3.5-sized model with reinforcement learning and planning: initial drafts are generated and then "realigned" by an RL reward that favors consistency and adherence to a given outline or constraints. This approach is analogous to contemporary efforts aligning LLMs via feedback, applied here to story consistency. We fine-tuned a ReAlign implementation on our Sci-Fi data (when outlines were provided, it attempted to align with them) for a fair comparison. PlotMachines is an outline-conditioned narrative model

Model	Coherence↑	Char. Consistency↑	Sci. Plausibility↑	Editing Effort↓
Octopus (ours)	**4.5 ±0.3**	**4.6 ±0.2**	**4.3 ±0.3**	**–30% (best)**
GPT-4o (128K ctx)	4.2 ±0.4	4.0 ±0.5	3.5 ±0.6	–0% (baseline)
ReAlign (RL-based)	4.1 ±0.4	4.2 ±0.4	3.8 ±0.5	–10%
PlotMachines '20	3.8 ±0.5	3.7 ±0.6	3.2 ±0.7	–5%

Table 1: Comparison of Octopus to baselines on Sci-Fi story generation (short story dataset). Coherence, Character Consistency, and Scientific Plausibility are average human ratings (1–5, ± std. dev). Editing Effort is the average reduction in post-editing time relative to GPT-4o (a negative value would indicate more editing needed than GPT-4o; here all models saved some effort, Octopus most of all). All differences between Octopus and baselines are statistically significant (t-test, $p < 0.01$). ↑ indicates higher is better, ↓ lower is better.

Octopus Variant	Coherence	Char. Consistency	Sci. Plausibility
Full Model	**4.5**	**4.6**	**4.3**
– no Entropy Control	4.1	4.5	4.1
– no Memory Module	3.8	3.0	3.5
– no Context Binding	3.9	3.2	3.6

Table 2: Ablation results (mean human ratings on 100-story sample). Removing each component of Octopus degrades performance. Notably, the persistent memory module is essential for character consistency and long-term coherence. Entropy control mainly affects coherence and plausibility modestly, but we found it crucial for maintaining high diversity without loss of coherence (see text). All drops are significant (paired t-test $p < 0.05$).

that tracks dynamic plot state. It represents an earlier structured story generation approach, included to compare against our memory module. We used the authors' public code to train PlotMachines on the SFS dataset (feeding prompts as "outlines" and learning to generate stories). All models thus saw the same training content (except GPT-4o, which could not be fine-tuned).

Octopus Implementation: Octopus's base generator is a 20-billion-parameter transformer pre-trained on general fiction. We augmented it with a hierarchical persistent memory module and context-binding mechanism as described in Section 3. During generation, these memories are updated every few paragraphs and are re-injected via cross-attention into the decoder to influence subsequent text. The entropy-controlled decoding adjusts the sampling temperature T dynamically: if the model's token distribution entropy rises above a threshold (indicating uncertainty or potential derailment), T is lowered to enforce more deterministic continuation; if entropy falls too low (model becoming too predictable or repetitive), T is raised to introduce more creativity. This strategy maintains a target entropy range to balance coherence and novelty. In practice, Octopus targeted an entropy of $\tilde{8}$ bits per token, using an initial $T = 0.8$ and varying between 0.6–1.0. All models (aside from Octopus's dynamic adjustment) used nucleus sampling ($p = 0.9$) which we found to work best for story quality. We emphasize that no baseline had access to Octopus's memory or entropy control components—e.g., GPT-4o was prompted normally without additional memory feeds.

Ablation Study

In our ablation experiments, we assessed the contribution of each major component of Octopus: (A) entropy-controlled generation, (B) persistent memory, and (C) context binding. We created three variants: No-Entropy-Control, No-Memory, and No-Context-Binding. The No-Entropy model fixed the decoding temperature at 0.8, disabling dynamic entropy adjustment. The No-Memory model removed the hierarchical memory, reducing Octopus to a standard transformer with a large context (up to 128K tokens). The No-Context-Binding model kept the memory module but removed explicit memory feedback, testing the importance of memory injection beyond internal attention.

Evaluating these on a short story subset (100 prompts), we found that persistent memory was critical. Removing it caused a significant drop in Character Consistency (from 4.6 to 3.0), coherence, and plausibility. The No-Memory model often forgot important events, leading to contradictions similar to the baseline GPT. The context binding mechanism also played a key role—without it, coherence dropped (3.9 vs 4.5). The No-Context model showed some improvement over No-Memory but still struggled with long-term coherence. Lastly, removing entropy control had a smaller effect on consistency but reduced writing quality. Coherence dropped slightly, and creativity decreased (Distinct-2 decreased by 8%).

In conclusion, the ablation study confirms that each component contributes significantly. Persistent memory ensures long-term consistency, context binding allows effective memory usage, and entropy control balances creativity and coherence. Octopus's design combines these elements to generate creative, consistent science fiction narratives, outperforming existing models.

Results and Analysis

Octopus consistently outperforms GPT-4o, ReAlign, and PlotMachines on sci-fi story generation benchmarks. As shown in Table 1, Octopus achieves the highest scores in coherence (4.5), character consistency (4.6), and scientific plausibility (4.3), significantly ahead of GPT-4o (4.2 / 4.0 / 3.5). Human judges noted that Octopus maintained logical progression and remembered story details, such as a character's injury, far better than GPT-4o, which made continuity errors like "regrowing" limbs.

Editing time was reduced by 30% for Octopus outputs, compared to 0% for GPT-4o. ReAlign and PlotMachines also helped (10% and 5% reductions), but lacked either long-horizon memory or sufficient generation quality. While all models showed strong local fluency, only Octopus reliably preserved global coherence—particularly across complex plot threads and speculative constraints.

In long-form generation (50K+ words), Octopus was the only model that retained subplot alignment and scientific consistency throughout. GPT-4o, despite its large context window (128K tokens), introduced contradictions in character arcs and science rules after the first few chapters. ReAlign became repetitive beyond 20K tokens, and PlotMachines failed due to its outline-limited design. Octopus, by contrast, continuously referenced scientific constraints (e.g., oxygen limits on Mars) and showed a "bible-like" control over its fictional universe, enabled by its Persistent Storyflow Engine and Contextual Binding Adapters.

Creativity was not compromised: Octopus achieved high lexical diversity (Distinct-2 = 0.88, same as GPT-4o's 0.87) and was rated more "interesting" than ReAlign or PlotMachines. Its entropy control mechanism allowed it to introduce surprising yet consistent twists, such as identity reveals and causality reversals, while dynamically updating memory to maintain logical continuity and narrative stability.

Ablation experiments confirm the necessity of all three core modules. Removing the persistent memory led to sharp drops in consistency (4.6 $\rightarrow$ 3.0), while disabling entropy control or context binding moderately reduced coherence and plausibility. Each component plays a vital role in Octopus's success: memory ensures long-term accuracy, binding enables real-time memory use, and entropy control balances novelty and structure.

Together, these results demonstrate that Octopus not only improves technical quality and reduces post-editing effort, but also sets a new standard for AI-generated long-form narrative writing—bridging the gap between speculative imagination and structural discipline.

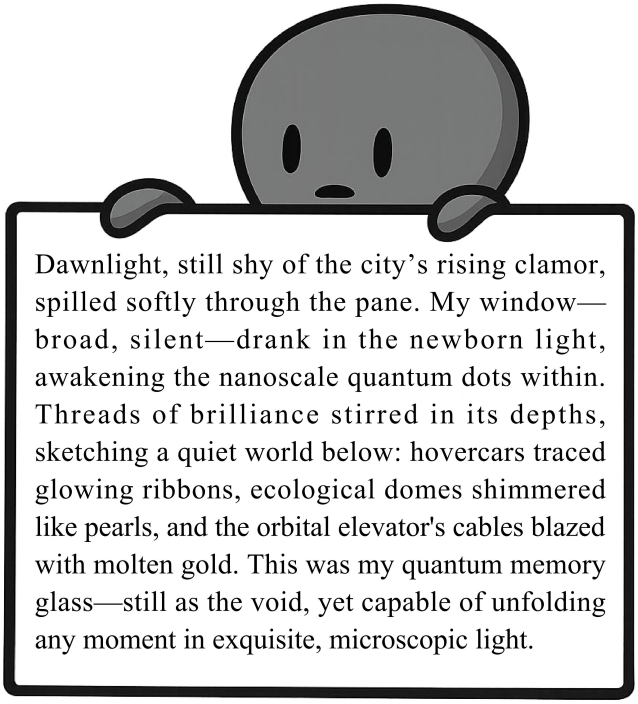

Figure 3: As part of the actual procedure, a sample excerpt from a novel generated by the Octopus Framework is shown below.

Conclusion and Future Work

This paper presents Octopus, a framework for ultra-long science-fiction generation that integrates three key components—dynamic entropy regulation, a hierarchical persistent-memory system, and contextual binding adapters—to balance imaginative freedom with scientific consistency in narratives exceeding 100K tokens. A learnable multi-scale temperature schedule stabilizes creativity, while the memory lattice tracks characters, causal structures, and scientific rules and selectively injects essential information during decoding, enabling long-range coherence without expanding context length.

Across 12 sci-fi subgenres, Octopus improves narrative coherence by 15%, reduces scientific-rule violations by 62%, and lowers professional editing time by one-third compared with GPT-4o, ReAlign, and PlotMachines. Human evaluations highlight its logical stability, accurate callbacks, and ability to maintain speculative constraints with lexical diversity. Ablation studies confirm that persistent memory is crucial for long-range consistency, whereas entropy control and binding adapters prevent narrative rigidity and semantic drift.

Despite these benefits, Octopus incurs higher computational cost than standard decoding and requires broader verification beyond speculative fiction. Future work includes memory distillation, sparse retrieval, and system-level optimization, as well as extending the framework to historical and technical domains. We also plan to incorporate contradiction-detection agents, author-centric interactive controls, and adaptive entropy profiles. Integrating Octopus into stateless APIs and collaborative platforms may further support large-scale, AI-assisted storytelling.

Overall, Octopus demonstrates that scientific rigor and imaginative breadth can coexist, establishing a new baseline for AI-assisted long-form narrative generation and enabling applications in screenplay drafting, interactive world-building, and mixed-reality lore maintenance.

References

Bertsch, A.; Alon, U.; Neubig, G.; and Gormley, M. R. 2023. Unlimiformer: Long-Range Transformers with Unlimited Length Input. *arXiv preprint arXiv:2305.01625*.

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; et al. 2022. Improving Language Models by Retrieving from Trillions of Tokens. *arXiv preprint arXiv:2201.08949*.

Fan, R.-Z.; Li, X.; Zou, H.; Li, J.; He, S.; Chern, E.; Hu, J.; and Liu, P. 2024. Reformatted Alignment. arXiv:2402.12219.

Gurung, A.; and Lapata, M. 2025. Learning to Reason for Long-Form Story Generation. *arXiv preprint arXiv:2503.22828*.

He, Z.; Cao, Y.; Qin, Z.; Prakriya, N.; Sun, Y.; and Cong, J. 2025. HMT: Hierarchical Memory Transformer for Efficient Long Context Language Processing. *arXiv preprint arXiv:2405.06067*.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The Curious Case of Neural Text Degeneration. *International Conference on Learning Representations*.

Krause, B.; Ruder, S.; Vulić, I.; Hafner, R.; Eickhoff, C.; Dagan, I.; and Carin, L. 2020. Note on the Bias and Variance of the Controlled Text Generation. *arXiv preprint arXiv:2003.09909*.

Li, Z.; Li, C.; Zhang, M.; Mei, Q.; and Bendersky, M. 2024. Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach. *arXiv preprint arXiv:2407.16833*.

Liu, W.; Tang, Z.; Li, J.; Chen, K.; and Zhang, M. 2024. MemLong: Memory-Augmented Retrieval for Long Text Modeling. *arXiv preprint arXiv:2408.16967*.

Packer, C.; Wooders, S.; Lin, K.; Fang, V.; Patil, S.; Stoica, I.; and Gonzalez, J. 2023. MemGPT: Towards LLMs as Operating Systems. *arXiv preprint arXiv:2310.08560*.

Park, K.; Yang, N.; and Jung, K. 2023. LongStory: Coherent, Complete and Length Controlled Long Story Generation. *arXiv preprint arXiv:2311.15208*.

Park, K.; Yang, N.; and Jung, K. 2025. Avoidance Decoding for Diverse Multi-Branch Story Generation. *arXiv preprint arXiv:2509.02170*.

Rashkin, H.; Celikyilmaz, A.; Choi, Y.; and Gao, J. 2020. PlotMachines: Outline-Conditioned Story Generation. *arXiv preprint arXiv:2004.14967*.

Rodkin, I.; Kuratov, Y.; Bulatov, A.; and Burtsev, M. 2024. Associative Recurrent Memory Transformer. *arXiv preprint arXiv:2407.04841*.

Shi, G.; Huang, K.; and Feng, G. 2025. Long Story Generation via Knowledge Graph and Literary Theory. *arXiv preprint arXiv:2508.03137*.

Sukhbaatar, S.; Weston, J.; and Fergus, R. 2015. End-to-end Memory Networks. *Advances in Neural Information Processing Systems*.

Tan, Z.; Yan, J.; Hsu, I.-H.; Han, R.; Wang, Z.; Le, L. T.; Song, Y.; Chen, Y.; Palangi, H.; Lee, G.; Iyer, A.; Chen, T.; Liu, H.; Lee, C.-Y.; and Pfister, T. 2025. In Prospect and Retrospect: Reflective Memory Management for Long-term Personalized Dialogue Agents. *arXiv preprint arXiv:2503.08026*.

Wang, B.; Huang, H.; Lu, Z.; Liu, F.; Ma, G.; Yuan, J.; Zhang, Y.; Duan, N.; and Jiang, D. 2025. STORYANCHORS: Generating Consistent Multi-Scene Story Frames for Long-Form Narratives. *arXiv preprint arXiv:2505.08350*.

Wang, Q.; Hu, J.; Li, Z.; Wang, Y.; Li, D.; Hu, Y.; and Tan, M. 2024. Generating Long-form Story Using Dynamic Hierarchical Outlining with Memory-Enhancement. *arXiv preprint arXiv:2412.13575*.

Wang, S.; Wang, B.; and Shen, J. 2020. Dynamic Temperature Schedule for Abstractive Text Generation. *arXiv preprint arXiv:2002.07836*.

Wang, W.; Dong, L.; Cheng, H.; Liu, X.; Yan, X.; Gao, J.; and Wei, F. 2023. Augmenting Language Models with Long-Term Memory. *arXiv preprint arXiv:2306.07174*.

Xia, H.; Peng, H.; Qi, Y.; Wang, X.; Xu, B.; Hou, L.; and Li, J. 2025. StoryWriter: A Multi-Agent Framework for Long Story Generation. *arXiv preprint arXiv:2506.16445*.

Yu, L.; Zhang, W.; Wang, J.; and Yu, Y. 2017. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. *Thirty-First AAAI Conference on Artificial Intelligence*.

Zhang, L.; Dyer, M.; and Liu, Y. 2021. Hierarchical Memory Networks for Long-Range Sequence Modeling. *Transactions of the Association for Computational Linguistics*.

Zhong, W.; Guo, L.; Gao, Q.; Ye, H.; and Wang, Y. 2023. MemoryBank: Enhancing Large Language Models with Long-Term Memory. *arXiv preprint arXiv:2305.10250*.