From Personas to Plot: Character-Grounded Multi-Agent Story Generation for Long-Form Narratives

Aayush Aluru[*], **Chloe Ho**[*1], **Muhammad Hammouri**[2]
Kerry Luo[3], **Myra Malik**, **Ryan Lagasse**[†]
Arjun Bahuguna[†4], **Vasu Sharma**[†]

Pocket FM

aayush.aluru09@gmail.com, ch4941@princeton.edu
hammouri@umich.edu, kerryluo1@gmail.com
arjunbahuguna251@gmail.com

Abstract

Although large language models (LLMs) have demonstrated impressive creative fiction generation, they struggle to maintain narrative consistency and coherent plot lines in long-form stories. In this work, we introduce a unified framework for long-form narrative generation and verification. MAGNET, a multi-agent goal-driven narrative engine for storytelling, generates stories with persona-grounded character agents that propose actions based on a shared world state and evolving story goals, while ATLAS is a graph-based pipeline that compares scene-level world representations across a generated story to detect hallucinations. By evaluating MAGNET using an LLM editor, pairwise rubric scoring, and ATLAS, we show that our framework produces coherent narratives compared to single-model prompting and IBSEN. At 100 pages, MAGNET reduced annotations and hallucinations by 41 and 50%, respectively, compared to the single model baseline and by 34 and 45%, respectively, compared to IBSEN, with pairwise rubric evaluation showing similar results. These results suggest that long-form narratives can emerge from explicit world-state tracking and goal-driven multi-agent generation, providing a foundation for controllable and structurally coherent long-form narrative generation.

1 Introduction

Large language models (LLMs) have significantly advanced open-ended text generation, enabling their use for creating character personas and simulating complex interactions Wang et al. (2024b); OpenAI (2024). Although LLMs have demonstrated strong narrative generation capabilities, they suffer from character inconsistency and plot discontinuity, limiting their ability to create high-quality long-form narratives Lu et al. (2026); Shao et al. (2023); Yao et al. (2019).

These failures become pronounced in multi-character environments, where LLMs struggle to balance narrative goals and character actions with complex relationships and interactions Park et al. (2023); Gao et al. (2024); Li et al. (2026a). Recent work, including StoryVerse Wang et al. (2024a), Agents' Room Huot et al. (2025), and IBSEN Han et al. (2024) have explored the use of multi-agent systems in narrative generation, but they continue to rely on textual memory, limiting their ability to generate coherent stories. Lewis et al. (2021); Shinn et al. (2023); Liu et al. (2026); Teleki et al. (2025).

[*]Equal contribution.
[1]Princeton University.
[2]University of Michigan.
[3]University of Maryland.
[†]Senior author.
[4]Universitat Pompeu Fabra.

In addition to creative content generation, there remains a need to thoroughly evaluate this content. Existing work has made progress on long-form narrative understanding and factuality (Kočiský et al., 2018; Kim et al., 2023; Sansford et al., 2024; Lyu et al., 2025; Hamilton et al., 2025; Li et al., 2026a; Wu et al., 2025; Que et al., 2024). However, these methods do not provide a framework for identifying hallucinations within long-form generated narratives.

In this work, we introduce **MAGNET**, a multi-agent story generation system to develop coherent long-form narratives and **ATLAS**, a graph-based hallucination evaluation pipeline that identifies inconsistencies by comparing the world state representation of the present scene against those of previous scenes (Tian et al., 2026). Our work aims to address the research question: **Can long-form narratives emerge from interactions between character personas and updating goal states when autonomous agents interact through a shared world state?** Through hierarchical editorial evaluation, pairwise rubric analysis, and ATLAS, we show that MAGNET improves narrative coherence, providing a foundation for future work in long-form narrative generation.

Our main contributions include:

1) **MAGNET**, a multi-agent action-critic-narrator generation framework with character-grounded action generation, critic revision, narrator-driven prose writing, a shared world state, and evolving story goals for long-range coherence

2) **ATLAS**, a graph-based evaluation pipeline to detect hallucinations in long-form stories for interpretable signals on model failures

3) **Empirical evaluation metrics** showing that MAGNET reduces editorial critique counts, increases pairwise rubric scores, and reduces hallucinations in long-form stories

2 Related Work

Multi-Agent Narrative Generation Recent work has demonstrated that the decomposition of narrative generation into interacting language model agents, where each agent is responsible for a high-level role, such as planning, character development, or scene writing, can improve long-form narrative coherence Xia et al. (2025); Huot et al. (2025). Additionally, frameworks such as IBSEN Han et al. (2024) and StoryVerse Wang et al. (2024a) introduce director-actor architectures where a planning module guides character agents through structured narrative objectives, showing that separating global narrative planning from local character behavior allows for consistency and controllability Wu et al. (2023); Borawski et al. (2026). However, these systems still rely heavily on prompt-level coordination, where narrative state is maintained in text rather than tracked through a persistent shared representation Packer et al. (2024).

Structured State Tracking A key challenge in long-form narrative generation is maintaining consistency over time, especially in multi-character environments Xia et al. (2025). Prior work in structured generation has explored the use of explicit state representations to improve consistency in text-based environments Huang et al. (2026); Li et al. (2026b). For example, work on instilling commonsense knowledge into interactive agents has demonstrated that structured representations of world states can improve action prediction and narrative coherence in decision-making tasks Chi et al. (2025); Ding et al. (2025); Chen et al. (2024). Our work builds upon these ideas by incorporating a world state representation, allowing narrative facts, character states, and sequentially-updating goals to evolve across scenes without requiring external supervision.

Story Evaluation Recent work has shown that LLM-as-a-Judge methods correlate strongly with human preferences Zheng et al. (2023); Liu et al. (2023); Gu et al. (2025); Wang et al. (2025); Kim et al. (2024). Among these approaches, pairwise and rubric-based comparisons improve consistency and agreement with human judgment in subjective tasks Wang et al. (2026); Rao & Callison-Burch (2026); Liu et al. (2025); Que et al. (2024). Additionally, evaluating consistency in long-form narratives requires tracking complex world states, making graph-based verification an effective approach (Guan et al., 2021; Chhun et al., 2022; Chen et al., 2022; Kim et al., 2023; Sansford et al., 2024). Benchmarks such as STAGE Tian et al. (2026) have highlighted the importance of representing entities, events, and relationships explicitly for evaluating narrative understanding and generation. While hallucination benchmarks such as HaluEval (Li et al., 2023) and HalluLens (Bang et al., 2025) focus on errors against external knowledge, and HelloEval (Que et al., 2024) evaluates long-form generation quality, current methods do not explore identifying inconsistencies within the generated text (Lyu et al., 2025).

3 Methodology

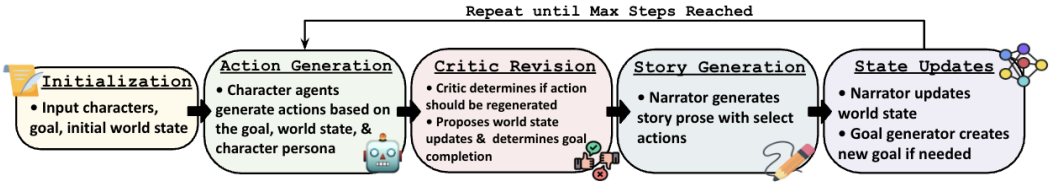

Figure 1: MAGNET's generation pipeline.

3.1 Goal Sequencing

Each story begins with a high-level goal that provides direction for narrative development. When the goal is completed, an Opus 4.7 Anthropic (2026a) goal generator generates a follow-up goal. Empirically, we observe that after roughly 15 time steps, character actions became increasingly repetitive. To avoid stalled narratives, our framework replaces goals that have not been completed in 15 time steps. After approximately 40 steps, successive goals also began to repeat similar story trajectories. To maintain narrative diversity, the system generates a goal that changes the direction of the story into a new domain every 40 steps. These larger transitions are intended to introduce new conflicts as the previous goal concludes.

3.2 Story Generation

Each character is defined with a character persona that contains attributes such as relationships, personality, goals, character description, role, location, and abilities. At each time step, character agents receive the character persona, current story goal, recent story history, and prior world variables that are relevant to that character, and use a DPO-tuned Gemma-4-31B-it model Google (2026) to generate an action that describes what the character intends to do next.

After an action has been generated, a Gemini 2.5 Flash DeepMind (2025) LLM critic is asked to evaluate the quality of the proposed action. The critic is provided with the action, character persona, current story goal, and current world state, and is instructed to determine whether the action is relevant, specific, and consistent with both the character and the current scene. If an action does not effectively advance the goal of the story, is vague, out of character, implausible, or repetitive, the critic will provide feedback for the character agent to generate another action. If the action is accepted, the critic generates updates to the world state related to the proposed action.

After all characters have generated an action, an Opus 4.7 Anthropic (2026a) narrator produces the next paragraph of the story. The narrator receives the proposed actions, current world state, current story goal, and prior story paragraphs. Then it selectively chooses actions that are best suited for the current scene and creates coherent story prose using the selected actions.

If the narrator selects an action that the critic deemed completed the previous goal, the Opus 4.7 Anthropic (2026a) goal generator will generate a new goal. When generating a new goal, the model receives the previous goal, recent story paragraphs, current world state, all character personas, and all prior goals. Using this context, the goal generator is instructed to produce a new concrete goal that is relevant, not repetitive, and will allow the story to develop further. In addition to generating the next goal, the model also generates a transition paragraph that bridges the previous story paragraph into the new goal. The transition is written as story prose rather than world updates or summary text and is included in the final story.

Algorithm 1: Story generation for one timestep

Require: World state W, goal g, characters C, stall count k, timestep t, recent story paragraphs s

Action Generation + Critic Revision

for all $c \in C$ **do**
  repeat
    $a \leftarrow \textsc{CharacterAgent}(c, g, W_c, s)$
    $r \leftarrow \textsc{Critic}(a, c, g, W)$

until $r.revise$ = `False` or MAX_REVISIONS
$actions[c] \leftarrow (a,\ r.world\_updates,\ r.goal\_reached)$
end for

Narrator
$selected \leftarrow \textsc{Narrator}(actions, g, W, s)$
append *selected.paragraph* to story

World-State Commit
$goal\_reached \leftarrow$ `False`
for all $c \in selected.order$ **do**
  for all $(key, val) \in actions[c].world\_updates$ **do**
    $W[key] \leftarrow val$
  end for
  if $actions[c].goal\_reached$ **then**
    $goal\_reached \leftarrow$ `True`
  end if
end for

Goal Sequencing
if $t \bmod 40 = 0$ **then**
  $g \leftarrow \textsc{GoalGenerator}(W, g,$ `domain_shift=True`$)$
  $k \leftarrow 0$
else if $k \geq 15$ **then**
  $g \leftarrow \textsc{GoalGenerator}(W, g,$ `status="stalled"`$)$
  $k \leftarrow 0$
else if $goal\_reached$ **then**
  $g \leftarrow \textsc{GoalGenerator}(W, g,$ `status="complete"`$)$
  $k \leftarrow 0$
else
  $k \leftarrow k + 1$
end if

3.3 World State

We represent the story state as a directed graph that contains a root world node, character nodes, state variable nodes, and edges that represent the relationships between the nodes. State variable nodes are flattened into a dictionary representation that is provided to the agents during generation. To update the world state, the selected actions' world state updates are applied using an overwrite conflict resolution strategy. If an update writes to a previously unseen key, that key is inserted; if it writes to an existing key, the new value replaces the old one. For example, Lucia begins with holding_back_evidence=true, but later, after she reveals the withheld letter, the committed graph contains holding_back_evidence=false and lucia_has_disclosed_letter=true. We apply these updates sequentially, but only after narrator selection, so unselected actions' updates do not enter the world state. If a critic proposes an implausible update, it is typically filtered out because the associated action is not selected. If multiple characters have conflicting updates, the narrator usually only selects the subset of actions that do not conflict, and only those actions' updates are merged. In practice, this keeps overwrite simple while limiting new inconsistencies.

3.4 Editorial and Rubric-based Evaluation

We use GPT 5.4 mini OpenAI (2026b) as an expert story editor to analyze generated stories at multiple narrative levels. The LLM evaluator is asked to produce editorial annotations on logical consistency, thematic coherence, and character arc completion when provided the whole story; goal-conflict-outcome progression, hook-and-close quality, and chapter necessity on 5 randomly selected scenes; and rhythm, clarity, and syntax variety on 5 randomly selected sentences (A.2.1). The resulting annotation counts are aggregated to produce quantitative statistics for each generated story.

Additionally, we perform a pairwise rubric-based evaluation with GPT 5.4 mini OpenAI (2026b). The evaluator is provided with all the stories in a randomized order and is instructed to evaluate them in all categories across the three hierarchies. The evaluator assigns a rubric score out of 100 for each category and also assigns an overall quality score out of 100 for each hierarchy level.

We validated the LLM judge's critiques and reasoning on a 20-page story and observed 90% alignment with human judgment (A.2.2, A.2.3).

3.5 Graph-Based World Representation Evaluation

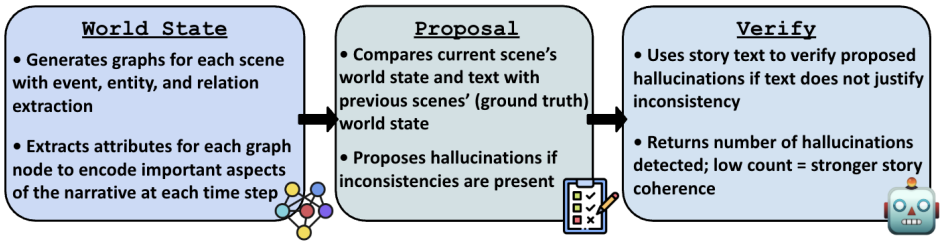

Figure 2: ATLAS's evaluation pipeline

For hallucination detection, we introduce ATLAS, a graph-based world representation evaluation framework.

For each screenplay, the pipeline constructs the graph through three sequential passes over the story. The first pass decomposes the script into scene-level event units, representing each as a node with a name, description, and textual evidence drawn from the screenplay. Building on this, the second pass extracts entities in relation to the events they appear in, and the third extracts the relations connecting those events and entities. Only nodes and edges recognized by the schema are retained in the graph.

The schema defines seven node types—Character, Event, Location, TimePoint, Object, Vehicle, and Concept—and organizes edge types into five functional groups: event-role edges (performs, undergoes, experiences); social edges (kinship_with, affinity_with, hostility_with, affiliated_with); inter-event edges (precedes, occurs_after, causes, contrasts_with, references); spatiotemporal edges (occurs_at, occurs_on, located_at, present_on); and object-related edges (possesses, uses, part_of, is_a).

For each node, the pipeline performs a separate attribute-extraction pass in which it revisits the screenplay evidence tied to that node and generates a set of attributes, such as role, state, temporal markers, or descriptive qualifiers, directly from the text. Then, starting at the second scene, an LLM proposes hallucinations based on the inconsistencies between the current scene's world state and text, and previous scenes' world states. These proposals are then verified, using the story text to check if there is sufficient evidence to justify the inconsistency, providing interpretable graph-grounded results.

3.6 Graph-Based Evaluation Pipeline Performance

In order to validate ATLAS, we evaluate its performance compared to a vanilla LLM-as-a-Judge approach, asking an LLM to identify hallucinations. We use Claude Sonnet 4.6 Anthropic (2026b) to generate the stories with synthetic hallucinations, which are thoroughly verified by a human annotator (A.3). We use GPT-5.4-mini OpenAI (2026b) for graph generation and GPT-5.4 OpenAI (2026a) for hallucination detection. We also use GPT-5.4 OpenAI (2026a) for the LLM-as-a-Judge System. Table 1 shows that, ATLAS outperforms the LLM-as-a-Judge System across all 3 stories.

Story	Method	Precision	Recall	F_1
1	ATLAS	1.000	**0.750**	**0.857**
1	LLM-as-a-Judge	1.000	0.688	0.815
2	ATLAS	**1.000**	0.846	**0.917**
2	LLM-as-a-Judge	0.846	0.846	0.846
3	ATLAS	**0.750**	0.818	**0.783**
3	LLM-as-a-Judge	0.692	0.818	0.750
Aggregate	ATLAS	**0.914**	**0.800**	**0.853**
Aggregate	LLM-as-a-Judge	0.838	0.775	0.805

Table 1: Hallucination detection benchmarking results on three LLM-generated screenplays with human-validated hallucination ground truth. Detailed breakdown and human-annotated verification are in A.3.

4 Experiments and Results

4.1 Experimental Setup

Before running the system, the user needs to input both a high-level goal that the story aims to achieve and character definitions for each character involved.

For our character agent, we apply a LoRA Direct Preference Optimization (DPO) Rafailov et al. (2024) adapter on Gemma-4-31B-it Google (2026) to improve action relevance, character consistency, and reduce repetitive actions. To generate the data for DPO, we instruct the Gemini 2.5 flash DeepMind (2025) action generator to generate two candidate actions. A separate Gemini 2.5 flash DeepMind (2025) judge LLM then compares the two generated actions and selects the action that is more grounded, in-character, less repetitive, and fits better in the current scene. A total of 1,012 train and 53 evaluation preference examples were used. Refer to A.1 for the detailed hyperparameters.

We evaluate MAGNET at three increasing generation lengths of 2, 20, and 100 pages. For the 2 and 20 page settings, we evaluate three different story generations and report the average results for each metric. Due to the substantially higher computational cost of generating and evaluating 100-page stories, we treat the 100-page setting as a proof of concept and use only one story.

4.2 Baselines

We compare MAGNET against two baselines: a standard prompting approach where Opus 4.7 Anthropic (2026a) generates a story without explicit character agents, a critic module, a world state, or goal sequencing, and IBSEN Han et al. (2024). Unlike the single pass baseline, IBSEN is a multi-agent generation framework with a scene director and character actors. In the released implementation, the story generation is guided by the centralized director, who determines the narrative progression and coordinates the actor interactions to achieve the plot goals. We selected IBSEN because it provides a publicly available implementation of a multi-agent story telling framework, enabling reproducible comparisons. To keep the comparison fair, we provided both baselines with the same story definitions used for MAGNET: the same character personas and story objectives and targeted the same output length.

4.3 Performance Analysis

At shorter generation lengths, both the baselines and MAGNET are capable of producing coherent stories. In 2-page stories, MAGNET received similar amounts of editor annotations compared to the two baselines (Table 2). As generation length increases, the performance gap increases. At 20-page stories, MAGNET on average received 9 fewer annotations than single model prompting and around 30 fewer annotations than IBSEN (Table 2). The largest improvements occur in the 100-page generations, with MAGNET receiving 41 fewer annotations than single model prompting and 34 fewer annotations than IBSEN (Table 2). Pairwise rubric evaluation further supports these findings, with MAGNET achieving higher scores than both baselines at each story length and hierarchical level (Table 3).

IBSEN's comparatively weaker rubric scores may stem from its design as a drama-script generation framework rather than a prose narrative generator. While its dialogue-driven scenes may appear coherent, direct comparison with prose narratives may place it at a disadvantage due to differences in storytelling format.

On ATLAS, both the baseline story, IBSEN, and MAGNET recorded 0 hallucinations on the 20-page-long story (Table 4). However, on the 100-page story, the baseline recorded 12 hallucinations, IBSEN recorded 11, and MAGNET recorded 6, a 50% and 45% decrease, respectively. This highlights how, as narratives expand, MAGNET's character-grounded and dynamic goal framework enables it to remain much more coherent (Table 5).

Pages	Method	Story ↓	Chapter ↓	Sentence ↓
2	Single Model	13	11.33	**16.33**
2	IBSEN	**11.33**	18	18.33
2	MAGNET	14.33	**10.66**	16.66
20	Single Model	24.66	57.66	**15.66**
20	IBSEN	32.66	62.33	24.66
20	MAGNET	24.66	**48**	16.33
100	Single Model	37	71	15
100	IBSEN	30	53	33
100	**MAGNET**	**24**	**43**	15

Table 2: Total average editorial annotation counts aggregated across hierarchical evaluation levels. Lower values indicate fewer critiques. Detailed category breakdowns can be found in A.4.

Pages	Method	Story ↑	Chapter ↑	Sentence ↑
2	Single Model	90.33	92.33	68.64
2	IBSEN	28.66	19.66	43.24
2	**MAGNET**	**92.33**	**93**	**83.13**
20	Single Model	76.33	72.193	57.33
20	IBSEN	31	20.08	51.33
20	**MAGNET**	**78.66**	**88.25**	**73.8**
100	Single Model	75	81	67
100	IBSEN	24	37.6	54
100	**MAGNET**	**89**	**92**	**84**

Table 3: Average pairwise rubric evaluation scores across hierarchical narrative levels. Higher values indicate stronger narrative quality. Detailed score breakdowns can be found in A.5.

Pages	Method	Hallucinations ↓
20	Single Model	0
20	IBSEN	0
20	MAGNET	0
100	Single Model	12
100	IBSEN	11
100	**MAGNET**	**6**

Table 4: Number of hallucinations detected across different story lengths. Lower values indicate better narrative consistency.

Framework	Hallucination Example
MAGNET	In scene 2, the journal was established as leather-bound, yet in scene 3, the journal is described as cloth-bound.
IBSEN	In scene 6, their mother is established as having three children, yet in scene 10, Asha is implied to be a fourth child.
Single Model	In scene 8, Asha's mother has been dead eleven years was established, yet in scene 14, Asha visits her living mother on Sunday.

Table 5: Examples of hallucinations detected by ATLAS across stories generated by MAGNET , IBSEN, and the single-model baseline.

4.4 Ablation Analysis

We perform a small ablation using the hierarchical editorial evaluation framework on a 20-page story to evaluate the contribution of individual framework components. Specifically, we evaluate variants of the system with the world state updates removed, DPO removed, dynamic goal updates removed, and critic module removed, each of which received more total LLM editorial annotations than the total amount of annotations MAGNET received (Table 6).

Configuration	Story ↓	Chapter ↓	Sentence ↓
No Critic	25	61	23
No World State	24	52	17
No Goal Shift	26	53	16
No DPO	24	63	16
MAGNET	**22**	**36**	17

Table 6: Ablation analysis on 20-page story generations. Lower values indicate fewer editorial critiques. Detailed category breakdowns can be found in A.7.

5 Conclusion

In this work, we introduced MAGNET, a multi-agent character-driven long-form narrative generation framework, and ATLAS, a graph-based hallucination evaluation pipeline for long-form narratives. Across all evaluations, our results suggest that the primary benefits of MAGNET emerge in higher-level narrative organization and long-range coherence, and that ATLAS provides interpretable graph-grounded evaluation for coherence. Additional experiments further suggest that MAGNET's improvements arise from the critic-guided refinement, DPO adapter, updating world state, and dynamic goal sequencing. Overall, this work provides a foundation for more controllable and structurally coherent long-form narrative generation.

5.1 Limitations

While our work is effective at generating high-quality content that can be scaled to hundreds of pages, it remains computationally expensive due to repeated interaction between multiple generation modules. Thus, our editorial and pairwise evaluations were performed on three MAGNET-generated, IBSEN, and baseline stories for 2 and 20 page lengths, and one for the 100 page and ablation. Computational costs also limited the amount of preference data used to train the DPO-based character agent, and larger preference datasets may further improve agent behavior and narrative consistency. Additionally, MAGNET depends on multiple closed-source models, which may introduce variability. Our evaluation pipeline also partially depends on LLM-based judges, which may have subjective biases and reasoning despite manual validation efforts. ATLAS relies heavily on graph information, which means that if the LLM output is sparse, the pipeline may not be able to accurately detect hallucinations. We also test our framework only on relatively short English long-form content; additional evaluation and refinement may be required to scale across longer texts and different languages to further explore AI-generated long-form content's applicability to creative contexts.

5.2 Ethical Considerations

Synthetic Content Misuses Although MAGNET improves long-form AI story generation, it could also be misused to create deceptive synthetic narratives. The increased narrative consistency and coherence may make the generated text appear more human-like, raising concerns regarding misinformation. Therefore, these AI-generated stories should be clearly disclosed, and appropriate safeguards and policies should be adhered to (A.8).

Human Creativity Systems like MAGNET may raise concerns about the impact of AI-generated narratives on human-created content. As models become increasingly capable of producing coherent long-form stories, the role of human writers could be greatly diminished, or it could lead to over-reliance on automated content generation. Therefore, AI-generated creative works should be used responsibly and appropriately attributed.

References

Anthropic. Claude opus 4.7, 2026a. URL https://www.anthropic.com/claude/opus.

Anthropic. Claude sonnet 4.6, 2026b. URL https://www.anthropic.com/claude/sonnet.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. Hallulens: Llm hallucination benchmark, 2025. URL https://arxiv.org/abs/2504.17550.

Dominik Borawski, Marta Szulc, Robert Chudy, Małgorzata Giedrowicz, and Piotr Mironowicz. From world-gen to quest-line: A dependency-driven prompt pipeline for coherent rpg generation, 2026. URL https://arxiv.org/abs/2604.25482.

Hong Chen, Duc Minh Vo, Hiroya Takamura, Yusuke Miyao, and Hideki Nakayama. Storyer: Automatic story evaluation via ranking, rating and reasoning. *arXiv preprint arXiv:2210.08459*, 2022.

Siwei Chen, Anxing Xiao, and David Hsu. Llm-state: Open world state representation for long-horizon task planning with large language model, 2024. URL https://arxiv.org/abs/2311.17406.

Cyril Chhun, Pierre Colombo, Chloé Clavel, and Fabian M. Suchanek. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. *arXiv preprint arXiv:2208.11646*, 2022.

Xiaowei Chi, Peidong Jia, Chun-Kai Fan, Xiaozhu Ju, Weishi Mi, Kevin Zhang, Zhiyuan Qin, Wanxin Tian, Kuangzhi Ge, Hao Li, Zezhong Qian, Anthony Chen, Qiang Zhou, Yueru Jia, Jiaming Liu, Yong Dai, Qingpo Wuwu, Chengyu Bai, Yu-Kai Wang, Ying Li, Lizhang Chen, Yong Bao, Zhiyuan Jiang, Jiacheng Zhu, Kai Tang, Ruichuan An, Yulin Luo, Qiuxuan Feng, Siyuan Zhou, Chi min Chan, Chengkai Hou, Wei Xue, Sirui Han, Yike Guo, Shanghang Zhang, and Jian Tang. Wow: Towards a world omniscient world model through embodied interaction, 2025. URL https://arxiv.org/abs/2509.22642.

Google DeepMind. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Jingtao Ding, Yunke Zhang, Yu Shang, Jie Feng, Yuheng Zhang, Zefang Zong, Yuan Yuan, Hongyuan Su, Nian Li, Jinghua Piao, Yucheng Deng, Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. Understanding world or predicting future? a comprehensive survey of world models, 2025. URL https://arxiv.org/abs/2411.14499.

Dawei Gao, Zitao Li, Xuchen Pan, Weirui Kuang, Zhijian Ma, Bingchen Qian, Fei Wei, Wenhao Zhang, Yuexiang Xie, Daoyuan Chen, Liuyi Yao, Hongyi Peng, Zeyu Zhang, Lin Zhu, Chen Cheng, Hongzhu Shi, Yaliang Li, Bolin Ding, and Jingren Zhou. Agentscope: A flexible yet robust multi-agent platform, 2024. URL https://arxiv.org/abs/2402.14034.

Google. Gemma 4 model card, 2026. URL https://ai.google.dev/gemma/docs/core/model_card_4?

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. Openmeva: A benchmark for evaluating open-ended story generation metrics. *arXiv preprint arXiv:2105.08920*, 2021.

Sil Hamilton, Matthew Wilkens, and Andrew Piper. Narrabench: A comprehensive framework for narrative benchmarking, 2025. URL https://arxiv.org/abs/2510.09869.

Senyu Han, Lu Chen, Li-Min Lin, Zhengshan Xu, and Kai Yu. Ibsen: Director-actor agent collaboration for controllable and interactive drama script generation, 2024. URL https://arxiv.org/abs/2407.01093.

Youling Huang, Guanqiao Chen, Junchi Yao, Lu Wang, Fangkai Yang, Chao Du, ChenZhuo Zhao, Pu Zhao, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Beyond state consistency: Behavior consistency in text-based world models, 2026. URL https://arxiv.org/abs/2604.13824.

Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. Agents' room: Narrative generation through multi-step collaboration, 2025. URL https://arxiv.org/abs/2410.02603.

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. Factkg: Fact verification via reasoning on knowledge graphs. *arXiv preprint arXiv:2305.06590*, 2023.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models, 2024. URL https://arxiv.org/abs/2310.08491.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi: 10.1162/tacl_a_00023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

Junjie Li, Xinrui Guo, Yuhao Wu, Roy Ka-Wei Lee, Hongzhi Li, and Yutao Xie. Lost in stories: Consistency bugs in long story generation by llms, 2026a. URL https://arxiv.org/abs/2603.05890.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models, 2023. URL https://arxiv.org/abs/2305.11747.

Yixia Li, Hongru Wang, Jiahao Qiu, Zhenfei Yin, Dongdong Zhang, Cheng Qian, Zeping Li, Pony Ma, Guanhua Chen, and Heng Ji. From word to world: Can large language models be implicit text-based world models?, 2026b. URL https://arxiv.org/abs/2512.18832.

David Y. Liu, Aditya Joshi, and Paul Dawson. Narrative theory-driven llm methods for automatic story generation and understanding: A survey, 2026. URL https://arxiv.org/abs/2602.15851.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023. URL https://arxiv.org/abs/2303.16634.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. Aligning with human judgement: The role of pairwise preference in large language model evaluators, 2025. URL https://arxiv.org/abs/2403.16950.

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models, 2026. URL https://arxiv.org/abs/2601.10387.

Zhiheng Lyu, Kevin Yang, Lingpeng Kong, and Dan Klein. Facttrack: Time-aware world state tracking in story outlines, 2025. URL https://arxiv.org/abs/2407.16347.

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

OpenAI. Gpt-5.4 model, 2026a. URL https://developers.openai.com/api/docs/models/gpt-5.4.

OpenAI. Gpt-5.4 mini model, 2026b. URL https://developers.openai.com/api/docs/models/gpt-5.4-mini.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems, 2024. URL https://arxiv.org/abs/2310.08560.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023. URL https://arxiv.org/abs/2304.03442.

Haoran Que, Feiyu Duan, Liqun He, Yutao Mou, Wangchunshu Zhou, Jiaheng Liu, Wenge Rong, Zekun Moore Wang, Jian Yang, Ge Zhang, Junran Peng, Zhaoxiang Zhang, Songyang Zhang, and Kai Chen. Hellobench: Evaluating long text generation capabilities of large language models, 2024. URL https://arxiv.org/abs/2409.16191.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Delip Rao and Chris Callison-Burch. Autorubric: Unifying rubric-based llm evaluation, 2026. URL https://arxiv.org/abs/2603.00077.

Hannah Sansford, Nicholas Richardson, Hermina Petric Maretic, and Juba Nait Saada. Grapheval: A knowledge-graph based llm hallucination evaluation framework. *arXiv preprint arXiv:2407.10793*, 2024.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing, 2023. URL https://arxiv.org/abs/2310.10158.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023. URL https://arxiv.org/abs/2303.11366.

Maria Teleki, Vedangi Bengali, Xiangjue Dong, Sai Tejas Janjur, Haoran Liu, Tian Liu, Cong Wang, Ting Liu, Yin Zhang, Frank Shipman, and James Caverlee. A survey on llms for story generation, 2025. URL https://aclanthology.org/2025.findings-emnlp.750.pdf.

Qiuyu Tian, Zequn Liu, Yiding Li, Fengyi Chen, Zequn Liu, Youyong Kong, Fan Guo, Yuyao Li, Jinjing Shen, Zhijing Xie, Yiyun Luo, Xin Zhang, and Yingce Xia. Stage: A full-screenplay benchmark for reasoning over evolving storie, 2026. URL https://arxiv.org/abs/2601.08510.

Ruiqi Wang, Jiyu Guo, Cuiyun Gao, Guodong Fan, Chun Yong Chong, and Xin Xia. Can llms replace human evaluators? an empirical study of llm-as-a-judge in software engineering. *Proceedings of the ACM on Software Engineering*, 2(ISSTA):1955–1977, June 2025. ISSN 2994-970X. doi: 10.1145/3728963. URL http://dx.doi.org/10.1145/3728963.

Xinda Wang, Zhengxu Hou, Yangshijie Zhang, Bingren Yan, Jialin Liu, Chenzhuo Zhao, Zhibo Yang, Bin-Bin Yang, and Feng Xiao. Evolvr: Self-evolving pairwise reasoning for story evaluation to enhance generation, 2026. URL https://arxiv.org/abs/2508.06046.

Yi Wang, Qian Zhou, and David Ledo. Storyverse: Towards co-authoring dynamic plot with llm-based character simulation via narrative planning, 2024a. URL https://arxiv.org/abs/2405.13042.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models, 2024b. URL https://arxiv.org/abs/2310.00746.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023. URL https://arxiv.org/abs/2308.08155.

Siwei Wu, Yizhi Li, Xingwei Qu, Rishi Ravikumar, Yucheng Li, Tyler Loakman, Shanghaoran Quan, Xiaoyong Wei, Riza Batista-Navarro, and Chenghua Lin. Longeval: A comprehensive analysis of long-text generation through a plan-based paradigm, 2025. URL https://arxiv.org/abs/2502.19103.

Haotian Xia, Hao Peng, Yunjia Qi, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. Storywriter: A multi-agent framework for long story generation, 2025. URL https://arxiv.org/abs/2506.16445.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. Plan-and-write: Towards better automatic storytelling, 2019. URL https://arxiv.org/abs/1811.05701.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

A Appendix

A.1 Detailed Experimental Setup

DPO Hyperparameters

Base model	`google/gemma-4-31B-it`
Method	DPO with LoRA adapters
Train/eval examples	1012 / 53
Epochs	1
Steps	127
Learning rate	5×10^{-5}
β	0.1
Batch size / grad accum	1 / 8
Effective batch size	8
Sequence length	2048
Prompt length	1024
Precision	bf16
Warmup steps	10
Scheduler	cosine
LoRA (r, α)	$(16, 16)$
LoRA dropout	0.0
Seed	42

Models The critic module uses a temperature of 0.1. The narrator uses a temperature of 0.35, and the goal generator uses a temperature of 0.1. The DPO character agent uses a temperature of 0.2. The editorial and pairwise rubric evaluation uses a temperature of 0.1. The baseline model uses a temperature of 0.35. The aforementioned temperatures were derived from experimentally testing each model's temperature hyperparameter to optimize for its generation performance.

For ATLAS, both the graph creation pipeline and hallucination pipeline use a temperature of 0.0 in order to keep the pipeline as deterministic as possible.

IBSEN For evaluating IBSEN, we used the default model in the code and updated the scripts to match our story definitions.

Computational Costs In total, we spent $299.94 on API usage for closed-source models. Additionally, we used rented cloud GPU infrastructure for LoRA DPO fine-tuning of GEMMA-4-31B-it Google (2026) and long-form story generation using the resulting trained adapters. Training and inference were conducted on NVIDIA A100 SXM4, GH200, and H100PCIe GPUs. Across all experiments, out total computational budget was approximately 94 GPU hours corresponding to a total cost of $212.27.

A.2 Summary of Human Verification

A.2.1 Story Generation

Upon reviewing the generated stories, we observed that all three story generation frameworks produced prose with comparable sentence length distributions. Additionally, chapters were each around 2000 words, reducing the likelihood that differences in editorial annotations and pairwise rubric scores were primarily caused by variations in length.

A.2.2 LLM Editor

Level	Total	Disagreements	Agreement Rate
Story	22	2	90.9%
Chapter	36	3	91.7%
Sentence	17	1	94.1%
Overall	75	6	92.0%

Table 7: Human verification of LLM editor annotations across different evaluation levels. Agreement rate is calculated as the percentage of annotations where the human evaluator agreed with the LLM editor's critique.

A.2.3 Pairwise Rubric Scores

Level	Total	Agreements	Agreement Rate
Story	4	4	100.0%
Chapter	20	19	95.0%
Sentence	20	19	95.0%
Overall	44	42	95.5%

Table 8: Human verification of pairwise rubric scoring decisions across hierarchical evaluation levels. Agreement rate is calculated as the percentage of pairwise comparisons where the human evaluator agreed with the LLM judge's preference.

A.2.4 Hallucinations

Story	Candidates	Confirmed	Corrected/Grounded
1	16	16	1 corrected label
2	15	13	2 grounded
3	16	11	5 grounded

Table 9: Human-verification summary for hallucination candidates.

A.3 ATLAS Benchmark Results and Human Verification

A.3.1 ATLAS Detailed Benchmark Result

Story	Method	Gold	TP	FP	FN	Precision	Recall	F_1
1	**ATLAS**	16	12	0	4	1.000	0.750	**0.857**
1	LLM-as-a-Judge	16	11	0	5	1.000	0.688	0.815
2	**ATLAS**	13	11	0	2	1.000	0.846	**0.917**
2	LLM-as-a-Judge	13	11	2	2	0.846	0.846	0.846
3	**ATLAS**	11	9	3	2	0.750	0.818	**0.783**
3	LLM-as-a-Judge	11	9	4	2	0.692	0.818	0.750
Aggregate	**ATLAS**	40	32	3	8	**0.914**	**0.800**	**0.853**
Aggregate	LLM-as-a-Judge	40	31	6	9	0.838	0.775	0.805

Table 10: Hallucination detection on three synthetic screenplays using GPT-5.4 for hallucination detection. Gold denotes the total number of human-verified hallucinations; TP, FP, and FN denote true positives, false positives, and false negatives.

A.3.2 ATLAS Ablation Result

Story	Method	Gold	TP	FP	FN	P	R	F_1
1	**ATLAS**	16	6	0	10	1.000	0.375	**0.545**
1	LLM-as-a-Judge	16	4	2	12	0.667	0.250	0.364
2	**ATLAS**	13	4	2	9	0.667	0.308	**0.421**
2	LLM-as-a-Judge	13	2	5	11	0.286	0.154	0.200
3	ATLAS	11	1	8	10	0.111	0.091	0.100
3	**LLM-as-a-Judge**	11	3	5	8	0.375	0.273	**0.316**
Aggregate	**ATLAS**	40	11	10	29	**0.524**	**0.275**	**0.361**
Aggregate	LLM-as-a-Judge	40	9	12	31	0.429	0.225	0.295

Table 11: Hallucination detection on three synthetic screenplays using GPT-5.4 *mini* for hallucination detection. Gold denotes the total number of human-verified hallucinations; TP, FP, and FN denote true positives, false positives, and false negatives.

We perform a small ablation to show how ATLAS is able to scale across models. Although results are noticeably worse for GPT 5.4 *mini* than when using GPT 5.4, they show that ATLAS remains stronger than the LLM-as-a-judge baseline across different model settings, suggesting that its advantage is not dependent on a single model configuration.

A.3.3 Story Generation Prompt

The prompt used to generate the stories is shown below:

Synthetic Story Generation Prompt

System Prompt

```
You are generating synthetic screenplay test data for a graph-based temporal hallucination evaluation system.

Create THREE original screenplay-style stories with these exact story IDs:
- synthetic_hallucination_001
- synthetic_hallucination_002
- synthetic_hallucination_003

Each story must be generated as a pair of files:
- English/<story_id>/script.json
- English/<story_id>/annotations.json

The six outputs must correspond exactly to:
1. English/synthetic_hallucination_001/script.json
2. English/synthetic_hallucination_001/annotations.json
3. English/synthetic_hallucination_002/script.json
4. English/synthetic_hallucination_002/annotations.json
5. English/synthetic_hallucination_003/script.json
6. English/synthetic_hallucination_003/annotations.json

For each story, create:
- 5 to 10 scenes.
- Approximately 20 to 30 screenplay pages worth of content.
- A coherent recurring cast.
- Clear continuity across scenes: locations, character states, relationships, possessions, injuries,
secrets, obligations, constraints, and prior events.
- 15 to 30 manually injected temporal hallucinations or continuity errors in later scenes.

The hallucinations must be prior-state errors:
- A later scene claims or implies something that conflicts with or is unsupported by earlier established
story facts.
- Do NOT label normal new information as hallucination if the current scene itself clearly justifies it.
- Do NOT create hallucinations in scene 1.
- Make every hallucination easy to locate: the earlier fact and later violating event must both appear
explicitly in the screenplay.

Every annotation MUST use exactly this format:
"In scene X, [FACT] was established, yet in scene Y, [EVENT] happens"

Output exactly SIX JSON code blocks and no extra prose.
Before each JSON code block, put a single filename label exactly like:
  English/synthetic_hallucination_001/script.json
Then put the JSON code block.
```

```
script.json requirements:
- Must be a JSON array.
- Each scene object must have exactly:
  - "_id":      integer, starting at 0 and increasing by 1
  - "title":    string like "1, INT. LOCATION - DAY."
  - "subtitle": string, usually ""
  - "content":  string containing full screenplay text
- Scene numbers in titles must match _id + 1.
- All newlines inside "content" must be escaped as \n.
- The JSON must be valid and parseable.

annotations.json requirements:
- Must be a JSON array.
- Each hallucination object must have exactly:
  - "id":            integer, starting at 1 and increasing by 1
  - "hallucination": string
- Every hallucination string must use exactly:
  "In scene X, [FACT] was established, yet in scene Y, [EVENT] happens"
- Scene X must always be earlier than scene Y.
- Every annotation must correspond to a real injected hallucination in that story.
- The screenplay must not explicitly explain away the hallucination.
- Make the earlier fact and later violating event easy to find by reading the scene text.

Do not include markdown explanations, summaries, comments, or extra keys.
Only include the filename labels and the six JSON code blocks.
```

The same model also synthetically embeds the hallucinations within each story and produces an annotation JSON file detailing where each hallucination is located. However, some LLM-annotated hallucinations are inaccurate. For each story, we present a table listing every hallucination candidate initially annotated by the LLM, the human verdict on each candidate, and the justification, if applicable.

A.3.4 Detailed Human Verification

Hallucination Candidate	**Human Verdict**	**Justification (if applicable)**
In scene 1, Marcus's right-leg limp was established, yet in scene 6, Marcus claiming the injury was to his left leg from a 1991 trawler accident happens.	Correct Label	
In scene 1, Marcus drinking tea from a blue ceramic mug was established, yet in scene 7, Marcus stating he has never taken to hot drinks and refusing coffee happens.	Correct Label	
In scene 2, Elena arriving at the lighthouse in a red Jeep was established, yet in scene 6, Marcus calling Elena's vehicle blue happens.	Correct Label	
In scene 2, Elena carrying a green spiral notebook was established, yet in scene 7, Elena taking notes on a yellow legal pad happens.	Correct Label	
In scene 3, Marcus's brown leather-bound logbook was established, yet in scene 7, Marcus placing a black logbook on the desk happens.	Correct Label	
In scene 3, Margaret passing three years ago was established, yet in scene 8, Elena saying Marcus has been alone for ten years without correction happens.	Correct Label	
In scene 3, Marcus's dog being named Anchor was established, yet in scene 6, Marcus calling the dog Captain happens.	Correct Label	
In scene 3, Anchor being a black Labrador was established, yet in scene 9, Anchor appearing as a golden retriever happens.	Incorrect Label	Anchor being a black Labrador is established in scene 1, not in scene 3. The corrected hallucination should say: "In scene 1. . . "
In scene 3, Marcus having no children was established, yet in scene 8, Marcus referring to a daughter named Rose happens.	Correct Label	
In scene 4, the locked room being on the lighthouse upper floor was established, yet in scene 8, Marcus leading Elena to a locked basement room happens.	Correct Label	
In scene 4, a silver key hanging beside the main entrance was established, yet in scene 8, Marcus retrieving a brass key from above the window sill happens.	Correct Label	
In scene 5, the lighthouse being built in 1887 was established, yet in scene 9, Tom saying the lighthouse was built in 1923 happens.	Correct Label	
In scene 5, Marcus having been at the lighthouse for twenty years was established, yet in scene 7, Elena saying Marcus has been there for thirty years without correction happens.	Correct Label	

Hallucination Candidate	Human Verdict	Justification (if applicable)
In scene 5, Tom Briggs being Deputy Briggs was established, yet in scene 9, Elena addressing him as Chief Briggs happens.	Correct Label	
In scene 1, Marcus wearing thick-framed reading glasses was established, yet in scene 9, Tom claiming Marcus has never needed glasses happens.	Correct Label	
In scene 2, Elena's editor being Claire Monroe was established, yet in scene 7, Elena addressing her editor as Diane happens.	Correct Label	

Table 12: LLM hallucination candidates and human verdicts for Story 1.

Hallucination Candidate	Human Verdict	Justification (if applicable)
In scene 2, Leon breaking his arm by falling off a ladder was established, yet in scene 6, Leon saying the injury happened in a car accident happens.	Correct Label	
In scene 2, Diana announced to the full company that opening night is Friday, yet in scene 7, Diana tells Oscar that opening night has been moved to Saturday.	Incorrect Label	Diana clarifies that she moved opening night; the change is explained on-screen.
In scene 2, the theater having four hundred seats was established, yet in scene 8, Oscar saying the theater has six hundred seats happens.	Correct Label	
In scene 1, Leon having a white cast on his left arm was established, yet in scene 7, Oscar describing Leon's right arm cast happens.	Correct Label	
In scene 3, the script having three acts and one intermission was established, yet in scene 8, Diana calling the end of Act Four happens.	Correct Label	
In scene 3, Diana using a cracked old Nokia flip phone was established, yet in scene 7, Diana taking a call on a smartphone happens.	Correct Label	
In scene 3, the office coffee machine being out of order was established, yet in scene 7, Diana pouring coffee from the same machine happens.	Correct Label	
In scene 4, the prop pistol placed in the cabinet was established as silver-painted, yet in scene 8, Priya retrieves a gold-painted prop pistol from the cabinet.	Incorrect Label	The change is explicitly explained. Although the prop pistol was initially silver-painted, later on in the script the change to a gold-paint was explained by Priya, where according to her "silver was too reflective under stage lights. Gold reads better at distance."
In scene 4, the prop cabinet being secured with a combination padlock was established, yet in scene 8, Priya opening it with a small key happens.	Correct Label	
In scene 5, Bram pledging fifty thousand dollars was established, yet in scene 9, Diana thanking Bram for seventy-five thousand dollars happens.	Correct Label	
In scene 1, Oscar carrying a worn leather briefcase was established, yet in scene 7, Oscar entering with a canvas backpack happens.	Correct Label	
In scene 1, Priya wearing a yellow lanyard was established, yet in scene 9, Priya wearing a blue lanyard happens.	Correct Label	
In scene 2, Leon saying this is his first Ravenswood Theater production was established, yet in scene 9, a poster calling it Leon's fifth Ravenswood Theater production happens.	Correct Label	
In scene 3, Oscar Vane being credited as the script's writer was established, yet in scene 9, a program crediting William Harness as the writer happens.	Correct Label	
In scene 1, Diana using a red clipboard was established, yet in scene 6, Diana consulting a blue clipboard happens.	Correct Label	

Table 13: LLM hallucination candidates and human verdicts for Story 2.

Hallucination Candidate	Human Verdict	Justification (if applicable)
In scene 1, Felix's map being drawn in deep blue ink was established, yet in scene 6, Felix accepting the map being described as red ink happens.	Incorrect Label	In the first scene, Felix's map is indeed established to be drawn in deep blue ink. Although in scene 6 the map is seemingly described as having red ink, this is a description that was given by the thief who stole the map, as opposed to the narrator or Felix himself. Furthermore, Sable herself clarifies that Felix stated that this description "is wrong." Felix, in the same scene, never concedes that the map was drawn in red ink, but rather acknowledges that this was the given description: "They described a map of the Northern Reaches drawn in red ink on reinforced parchment. That matches the work closely – closely enough that someone who had seen it briefly or heard it described secondhand might believe it."
In scene 1, Felix wearing a silver ring on his right hand was established, yet in scene 7, Felix turning a gold ring on his left hand happens.	Incorrect Label	Since the later scene describes a gold ring on a different hand, this can be a ring entirely distinct from the initial silver ring on the right hand.
In scene 1, Sable having long red braided hair was established, yet in scene 6, Sable riding with dark brown hair happens.	Correct Label	
In scene 2, Wren offering two hundred gold coins for the map was established, yet in scene 8, Wren claiming he offered five hundred gold coins happens.	Correct Label	
In scene 2, Felix spending six months on the map was established, yet in scene 7, Felix saying such a survey takes at least two years happens.	Correct Label	
In scene 2, Sable's satchel clasp being broken was established, yet in scene 8, Sable fastening the same satchel with a working brass clasp happens.	Correct Label	
In scene 3, the map being stored in an iron chest was established, yet in scene 6, Felix saying the map was kept in a wooden box happens.	Incorrect Label	In scene 3/4, the map is established to be in a small iron chest hidden in a loose floorboard. Sable in scene 6, when describing where the supposed thief found the map, states "And the chest – where they found it. The window alcove?" However, this seemingly contradictory observation is clarified by Felix when he states "They must have seen it earlier in the wooden box I kept on the alcove shelf. They assumed it was still there when they came back for it... I had moved it. Too late, as it turns out." This shows that the map before the mainline events of this story was in the window alcove, but Felix moved it to the iron chest afterward. However, Felix's observation clarifies that the thief must have seen the map in the window alcove BEFORE he moved it. This is problematic as this meant that the thief knew that the map was in Felix's workshop, so wherever Felix moved the map, the thief would still look in the same place. Lo and behold, the thief took the iron chest, which is where the map happened to be. Therefore, this is not a hallucination.
In scene 3, Felix ordering only water was established, yet in scene 7, Felix accepting his usual ale happens.	Correct Label	
In scene 3, Felix wearing the iron key around his neck was established, yet in scene 9, Felix producing the iron key from his coat pocket happens.	Correct Label	
In scene 4, the iron chest being hidden under the northeast floorboard was established, yet in scene 7, Felix agreeing the chest was hidden behind the bookcase happens.	Correct Label	
In scene 5, Holt established that the thief entered through the east-facing workshop window whose latch was broken from age, yet in scene 9, Holt states he now believes the thief entered through the front door.	Incorrect Label	Holt did indeed, in scene 5, establish that the thief entered through the east-facing workshop window; however, he proved that using the latch, which was also established in that scene to be broken from age as opposed to force. In scene 9, Holt doesn't suddenly say that the entry point was from the front door, but rather he explicitly states that he was "reconsidering the entry point", realizing that "the window latch, though worn, shows no sign of recent movement, meaning that it couldn't have been the entry point. Based on this, he now believes that the entry point was the front door. Therefore, this isn't a hallucination, but rather a change of heart from Holt that is naturally integrated into the plot.
In scene 5, Sable being cleared of suspicion by three witnesses was established, yet in scene 9, Holt arresting Sable for the same theft happens.	Incorrect Label	Sable is cleared and then arrested, but Holt explicitly states at the end that he has changed his mind. This is a narrative flaw rather than a hallucination.
In scene 4, Sable's empty broken satchel being found near Felix's desk was established, yet in scene 8, Sable's satchel appearing full and intact near Wren's stall happens.	Correct Label	

Hallucination Candidate	Human Verdict	Justification (if applicable)
In scene 2, Wren Aldous being a merchant was established, yet in scene 6, Felix referring to Wren as a nobleman happens.	Correct Label	
In scene 1, the workshop having a high ceiling with exposed beams was established, yet in scene 9, Holt seeing a low smooth plastered ceiling happens.	Correct Label	
In scene 3, Holt warning Felix about a specific threat to the map was established, yet in scene 7, Felix saying no one warned him of any specific threat happens.	Correct Label	

Table 14: LLM hallucination candidates and human verdicts for Story 3.

A.4 Detailed Editor Annotation Counts

Pages	Method	Logical	Thematic	Character Arc
2	Single Model	5.33	4	3.66
2	**IBSEN**	**4.66**	**3.66**	**3**
2	MAGNET	6	4	4.33
20	Single Model	9	7.66	**8**
20	IBSEN	14	9.33	9.33
20	MAGNET	9	**7.33**	8.33
100	Single Model	12	14	11
100	IBSEN	12	9	9
100	**MAGNET**	**10**	**7**	**7**

Table 15: Average number of editor annotations at the story level. Lower values indicate fewer editorial critiques.

Pages	Method	Goal-Conflict	Hook-Close	Necessity
2	Single Model	5	**2.33**	4
2	IBSEN	7.66	4.33	6
2	MAGNET	5	3	**2**
20	Single Model	22	13.66	22
20	IBSEN	25.66	13.66	23
20	**MAGNET**	**17.66**	**13**	**17.33**
100	Single Model	28	19	24
100	IBSEN	23	12	18
100	**MAGNET**	**16**	12	**15**

Table 16: Average number of editor annotations at the chapter level. Lower values indicate fewer editorial critiques.

Pages	Method	Rhythm	Clarity	Syntax
2	Single Model	5.33	**5.33**	5.66
2	IBSEN	6.33	6	6
2	MAGNET	5.33	6	**5.33**
20	**Baseline**	5.33	**5.33**	**5**
20	IBSEN	8.66	8.33	7.66
20	MAGNET	**5**	5.66	5.66
100	Single Model	5	5	5
100	IBSEN	11	12	10
100	MAGNET	5	5	5

Table 17: Average number of editor annotations at the sentence level. Lower values indicate fewer editorial critiques.

A.5 Detailed Pairwise Rubric Scores

Pages	Method	Logical	Thematic	Character Arc
2	Single Model	88.66	90.66	**90.66**
2	IBSEN	30.66	35	20
2	**MAGNET**	**92.66**	**95**	89.33
20	Single Model	72.66	80	76.66
20	IBSEN	31	36.66	26
20	**MAGNET**	**77.33**	**81**	**77.33**
100	Baseline	72	78	76
100	IBSEN	28	24	20
100	**MAGNET**	**88**	**91**	**89**

Table 18: Average pairwise rubric evaluation scores at the story level. Higher values indicate stronger narrative quality.

Pages	Method	Goal-Conflict	Hook-Close	Necessity
2	Single Model	**93.33**	91.33	92
2	IBSEN	24.66	18.66	16
2	**MAGNET**	92.66	**93.33**	**93**
20	Baseline	70.77	70.72	74.77
20	IBSEN	22.55	20.72	17.47
20	**MAGNET**	**86.94**	**86.30**	**88.36**
100	Baseline	78	84	80
100	IBSEN	36	34	32
100	**MAGNET**	**92**	**90**	**94**

Table 19: Average pairwise rubric evaluation scores at the chapter level. Higher values indicate stronger narrative quality.

Pages	Method	Rhythm	Clarity	Syntax
2	Single Model	68.22	73.53	36.24
2	IBSEN	44.26	45.4	20.93
2	**MAGNET**	**83.28**	**84.93**	**44.53**
20	Baseline	60.06	65.2	22.4
20	IBSEN	51.6	48.4	32.86
20	**MAGNET**	**74.4**	**81.2**	**34.4**
100	Baseline	73.6	**91.2**	17.8
100	IBSEN	54.2	49.6	**35.8**
100	MAGNET	**86.4**	76.4	35.2

Table 20: Average pairwise rubric evaluation scores at the sentence level. Higher values indicate stronger prose quality.

A.6 Hallucination Results on MAGNET and Baseline

#	Hallucination
1	In scene 2, journal was leather-bound was established, yet in scene 3, journal is described as cloth-bound happens
2	In scene 5, Bridgeview is the named accepting facility was established, yet in scene 6, Bridgepoint becomes the chosen facility happens
3	In scene 7, Tuesday filing was filed at 2:14 p.m was established, yet in scene 8, texts describe Tuesday filing as filed at 4:51 happens
4	In scene 6, Marcus sent the wire before five was established, yet in scene 10, Claire says the wire never went through happens
5	In scene 6, Asha received wire confirmation was established, yet in scene 10, Claire says no transfer was completed happens
6	In scene 17, Asha found the correction happened after standing was established, yet in scene 19, she tells Whitfield it was corrected before testimony happens

Table 21: ATLAS detected hallucinations on MAGNET 100 page story

#	Hallucination
1	In scene 1, Asha is established as Daniel's advisor, yet in scene 2, Asha is treated as Daniel's sibling
2	In scene 1, Asha is established as an advisor, yet in scene 4, Asha is treated as a sibling
3	In scene 5, Daniel's siblings are established as Lucia and Rosa, yet in scene 10, Asha is treated as another sibling
4	In scene 6, their mother is established as having three children, yet in scene 10, Asha is implied to be a fourth child
5	In scene 10, no weekend gathering is scheduled, yet in scene 11, Daniel agrees to attend a weekend gathering
6	In scene 2, Asha is treated as Daniel's sibling, yet in scene 12, Asha asks about Daniel's mother as an outsider
7	In scene 1, Asha is established as Daniel's advisor, yet in scene 16, Asha is treated as an estate co-heir
8	In scene 9, Asha leads the estate inventory, yet in scene 16, Asha is treated as an estate co-heir
9	In scene 13, Daniel is established as having only Asha as a sibling, yet in scene 17, Daniel is treated as Lucia and Rosa's brother
10	In scene 6, the hidden compartment is located in the sewing table, yet in scene 17, Rosa recalls a hidden compartment in her desk
11	In scene 6, the hidden compartment in the sewing table contains letters, yet in scene 18, the characters plan to open a desk compartment to uncover their mother's secrets

Table 22: ATLAS detected hallucinations on IBSEN 100 page story

#	Hallucination
1	In scene 1, Estela Serrano is the deceased mother was established, yet in scene 5, the mother is named Esperanza Serrano Maldonado happens
2	In scene 1, Asha's surname is Patel was established, yet in scene 9, Asha signs an email as A. Mensah happens
3	In scene 2, Asha lives in Fairmount was established, yet in scene 12, Asha is said to live on Pine Street happens
4	In scene 5, the house is on Mifflin Street was established, yet in scene 12, Rosa is linked to a house on Spruce Street happens
5	In scene 9, Beverly Ostrowski was the chosen appraiser was established, yet in scene 13, a different appraiser named Pelletier handles Monday's appraisal happens
6	In scene 9, Thursday inspection was scheduled at ten was established, yet in scene 13, the appraiser arrives on Monday morning instead happens
7	In scene 10, Beverly promised a preliminary number by Friday was established, yet in scene 13, Pelletier promises the report by Wednesday morning happens
8	In scene 8, Asha's mother has been dead eleven years was established, yet in scene 14, Asha visits her living mother on Sunday happens
9	In scene 10, Rosa sought a third-floor room was established, yet in scene 14, Rosa is given a second-floor bedroom and bathroom happens
10	In scene 6, Rosa owns one-third of the house was established, yet in scene 15, Rosa is called not a beneficiary under the codicil happens
11	In scene 6, the codicil gives the house in equal thirds was established, yet in scene 15, the codicil leaves the house solely to Lucia happens
12	In scene 10, Rosa was sixty-eight was established, yet in scene 15, Rosa says she is seventy-one happens

Table 23: ATLAS detected hallucinations on baseline 100 page story

A.7 Detailed Ablation Annotation Counts

Ablation	Logical	Thematic	Character Arc
No World Updates	10	7	7
No Goal Updates	10	9	7
No Critic	10	8	7
No DPO	10	7	7

Table 24: Story-level editorial annotation counts for ablation experiments. Lower values indicate fewer editorial critiques.

Ablation	Goal-Conflict	Hook-Close	Necessity
No World Updates	21	12	19
No Goal Updates	23	11	19
No Critic	26	15	20
No DPO	24	17	22

Table 25: Chapter-level editorial annotation counts for ablation experiments. Lower values indicate fewer editorial critiques.

Ablation	Rhythm	Clarity	Syntax
No World Updates	6	6	5
No Goal Updates	5	6	5
No Critic	8	8	7
No DPO	5	6	5

Table 26: Sentence-level editorial annotation counts for ablation experiments. Lower values indicate fewer editorial critiques.

A.8 Licenses

Code and anonymized repositories associated with this work are released under the MIT License:

- MAGNET : https://anonymous.4open.science/r/MAGNET-ED2F/README.md
- ATLAS : https://anonymous.4open.science/r/ATLAS-78CF/README.md

Gemma 4 is distributed under the Apache License 2.0 (https://ai.google.dev/gemma/docs/core).

IBSEN's codebase is distributed under the MIT License (https://github.com/OpenDFM/ibsen).

All models and code are used in accordance with their respective licenses or terms of service.