From Emergence to Planning: A Triangle Framework for Scalable, Controllable Interactive Storytelling

Lasantha Senanayake
Narrative Intelligence Lab, University of Kentucky
Lexington, Kentucky 40506 USA
lasantha.senanayake@uky.edu

Abstract

Interactive story systems today sit at three extremes. Emergent multi-agent simulations give each character local intelligence but no global view, often losing plot structure. Reactive systems makes fast, state-based decisions. They form plans using hand-authored rules without searching for action sequences, so these systems can respond quickly but can wander if long-term rules are not explicitly authored. Centralized narrative planners reason globally to craft coherent, goal-directed plots, yet are computationally expensive. In my doctoral work I treat these not as isolated choices but as the three corners of a *triangle spectrum* of narrative generation. I propose hybrid, *landmark-guided* approaches that can scale to larger domains. I am also exploring how large language models (LLMs) can be embedded within these hybrid approaches themselves. This paper outlines research questions, methodology, progress to date, evaluation plan, and requested feedback.

Introduction

Interactive narrative environments, such as games and training simulations, often need to maintain a coherent storyline while adapting to user choices. Narrative planning is the use of automated planning to construct, communicate, and understand stories, a form of information to which human cognition and enaction is predisposed (Rivera et al. 2024). I organize this landscape around three points of a triangle, as seen in Figure 1: reactive systems (no search), emergent systems (local agents), and centralized narrative planners (global search and reasoning).

Reactive systems make fast, state-based decisions from hand-authored rules, allowing quick responses but risking shallow or wandering stories without long-term guidance (e.g. Façade (Mateas and Stern 2005)). Because plans are assembled from rules without searching over action sequences, these systems are highly responsive and generally the lowest cost.

Emergent multi-agent simulations give each character local intelligence but no global view, yielding high character believability (e.g. Dwarf Fortress (Adams 2019)) but often losing coherent plot structure. They are typically more expensive than purely reactive approaches and do not guarantee story organization.

Centralized narrative planners reason globally about story logic and character intentions, producing coherent, goal-directed plots, but they are computationally expensive. Systems like Sabre (Ware and Siler 2021) model the intentions of the author as well as the intentions and beliefs of each character, ensuring every action is grounded in goals and beliefs. Narrative planners reason about the logical structure of events using preconditions and effects (Young 1999), exploring many possible story paths to ensure quality and structure. However, forward search in these domains is computationally expensive since planning is P-SPACE hard (Helmert 2006) and does not scale well to longer or more complex stories.

No single paradigm is universally optimal for all storytelling contexts, consistent with the No Free Lunch theorem (Wolpert and Macready 1997). In this framing, reactive systems are low cost and highly responsive. Emergent systems are medium cost with high believability but weaker structure. Classical planners sit inside the triangle (medium cost, high structure, medium believability) and full narrative planners occupy the high-cost, high-structure, high-believability corner. My doctoral research investigates hybrid methods situated along the edges and interior of this triangle such as landmark-guided approaches and tries to fit Large Language Models (LLMs) into each of those corners to improve scalability, coherence, and believability.

Background

Planning is popular for generating interactive stories because it provides a formal, generative framework that reasons about causality and event ordering (Young 1999). Early systems such as Tale-Spin (Meehan 1977), Universe (Lebowitz 1985), and Façade (Mateas and Stern 2005) employed symbolic representations of characters, locations, objects, and conditions, but lacked the full search capabilities of modern narrative planners.

Classical planners were later used to create playable narrative adaptations of *Friends* (Cavazza, Charles, and Mead 2002), *Madame Bovary* (Pizzi and Cavazza 2007), and *The Merchant of Venice* (Porteous, Cavazza, and Charles 2010), and narrative planning has also been applied to training simulations (Fisher, Siler, and Ware 2022). Subsequent nar-

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

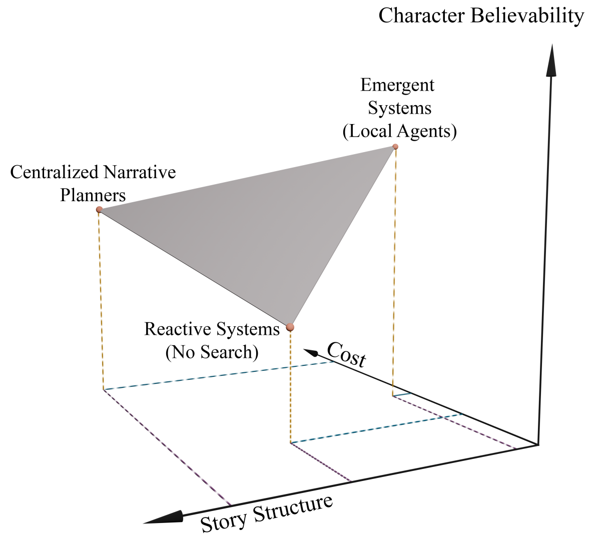
Character Believability
Emergent Systems (Local Agents)
Centralized Narrative Planners
Reactive Systems (No Search)
Cost
Story Structure

Figure 1: The triangle spectrum of interactive narrative generation. Each corner represents a paradigm, emergent multi-agent simulation, reactive decision-making, and centralized narrative planning

rative planners extend classical planning by modeling additional story properties. IPOCL (Riedl and Young 2010) represents character intentions. IMPRACTical (Teutenberg and Porteous 2013) and Glaive (Ware and Young 2014) incorporate conflict and failed plans. HeadSpace (Sanghrajka, Young, and Thorne 2022), Ostari (Eger and Martens 2017), and Sabre (Ware and Siler 2021) allow characters to act under mistaken beliefs. A survey by Young et al. (2013) summarizes these systems. Because the number of possible action sequences grows quickly, much work has focused on efficient search using classical heuristics (Bonet and Geffner 2001) or narrative-specific methods (Teutenberg and Porteous 2013; Ware and Young 2014). Additional structural pruning strategies, such as causal necessity (Ware, Senanayake, and Farrell 2023) and salience-based costs (Farrell, Ware, and Baker 2020; Ware and Farrell 2022), further reduce search effort.

Recent research explores integrating LLMs with planning. Neural story planning techniques keep an LLM goal oriented during generation (Ye et al. 2023). Other work uses an LLM as a cost function to prioritize promising actions (Senanayake and Ware 2025). Studies on the planning abilities of LLMs in symbolic domains (Valmeekam et al. 2023) can transfer to narrative problems. *Drama Llama* introduces an LLM-powered storylets framework that combines the structural benefits of storylet based systems with the generative capabilities of large language models, supporting author responsiveness and coherent, believable character interactions in interactive stories (Sun et al. 2025). Together, these advances motivate hybrid approaches that combine the structural guarantees of symbolic planning with the flexibility of LLM-based reasoning, situated along the triangle spectrum defined above.

Research Questions and Proposed Approach

To provide global scaffolding without over committing to a full trajectory, I use *planning landmarks*, propositions that must become true at some point along every valid solution, typically organized as a partial order that constrains when they should be achieved (Hoffmann, Porteous, and Sebastia 2011). Ordered landmarks are widely used to decompose hard planning problems into smaller subproblems while preserving causal coherence (Hoffmann, Porteous, and Sebastia 2011). In narrative planning, this gives a way to encode high-level plot points that support causal progression and intentional character behavior while leaving flexibility and allowing the specific sequence of low-level actions that achieves each point (Riedl and Young 2010) to be chosen at run time. The research questions I propose to explore are:

1. How do narrative quality (story structure and character believability), controllability, and computational cost vary across pure paradigms (emergent, reactive, centralized) as world complexity scales?
2. Do landmark-guided hybrids achieve higher quality than emergent or reactive baselines while scaling better than centralized narrative planning?
3. How robust are high-level landmark plans to deviations introduced by agents at runtime (e.g., failed actions, LLM hallucinations)?

Hybrid Systems. The following two hybrid systems illustrate my approach. Additional variants will be explored as the work progresses. Neither the LLM nor the planner are trained or tuned to a single story world; both are kept domain agnostic so they can be dropped into new domains with only a change of action schema and initial state.

1. **State trajectory constrained LLM Simulation:** A centralized narrative planner computes a sequence of causal landmarks over an abstracted domain. Each agent is an agent with *sensors* (observations), *internal state* (belief, intention, memory, goals). The agent's actions are controlled by an LLM conditioned on the currently active landmark depending on the agent's observations and the internal state.
2. **Landmark-Guided Classical Planner:** The abstracted plan supplies a sequence of sub-goals. A centralized classical planner expands the next sub-goal into low-level actions, executes them, monitors effects, and either advances to the next sub-goal. I am currently implementing this component as a low-level GPU (CUDA/C++) parallelized planner. After the basic classical version is stable, I will incrementally add a lightweight layer of narrative features (e.g., simplified intention model) so it sits between pure classical search and the full belief/intention model of Sabre. This intermediate planner can then serve as a hybrid planner that respects high-level narrative structure while remaining generic and reusable across story domains.

Progress to Date

This research builds on top of several existing assets and early implementations:

- **Narrative Planner:** The Sabre planner supports author goals, character intentions, and mistaken beliefs (Ware and Siler 2021). Its code base will be extended to export abstracted domain models for landmark computation.
- **Scalable Narrative Domain:** A parameterized story domain allows controlled variations of:
  - number of characters,
  - available actions,
  - items and locations,
  - author goals.

  These parameters can be changed to produce different story domains for scaling studies.
- **Baselines:**
  - *Pure narrative planning* using Sabre with classical and narrative-specific heuristics (e.g., Glaive (Ware and Young 2014)).
  - *Emergent simulation* prototype in which agents act from predefined goals with only local knowledge.
  - *Reactive simulation* in which agents act towards their goals without searching for plans.
- **Instrumentation:** Logging utilities record wall-clock time, memory usage, node expansions and plan length enabling consistent comparison across methods and scales.
- **Landmark Extraction:** An abstract domain representation system (with reduced action and predicate set).
- **Hybrid Infrastructure:** Messaging interfaces are designed so that LLM agents can reason and find action sequences to achieve the current landmark and a classical planner can request the next sub-goal when a landmark is achieved.

These components establish the experimental foundation needed to implement and evaluate the proposed hybrid systems.

Evaluation Plan

I propose to evaluate the system along three dimensions: quality, scalability, and controllability/stability. For each configuration (method and scale parameters), I will run multiple trials with different random seeds and record raw counts, measuring metrics from internal state of the planner or simulation. I also proposed human evaluations of the qualities.

Quality I will evaluate both character believability and story structure using a combination of human subjects evaluation and automatic metrics. For the human study, participants judge how plausible and intentional the characters' actions appear and how coherent the overall plot feels. For the automatic measures, each run logs which author and character goals are satisfied (and how many times) and tracks predefined set of "main plot points" (e.g., discovery events, key relationship changes, critical resource transfers). I record how many of these plot points occur and at what plan depth. Runs with higher goal satisfaction and broader/earlier plot-point coverage are taken to reflect stronger structure and, alongside human ratings, provide quantitative evidence relevant to believability.

Scalability Scalability will be measured by wall-clock time, number of states visited, memory usage, and the maximum feasible scale before timeout or failure. Scale is controlled by increasing the number of characters, actions, items, and locations. For each scale step I will record total search time, memory usage, number of node expansions (for planning components), and whether a complete story reaching at least one author goal was produced. The "maximum feasible scale" is the largest configuration in which the method still produces a valid story within a fixed time limit. This lets me plot how quickly the pure approaches fail versus the hybrids.

Controllability/Stability Finally, I will look at how predictable each method is under the same initial conditions. For hybrids with landmark guidance, I will compute landmark adherence, the fraction of planned landmarks that are achieved in the order predicted by the abstract plan. Deviations (skipped landmarks, reordered landmarks) are logged and give me a direct measure of how robust the method is to failed actions and/or LLM variation. I will also measure variance in final goal counts across seeds. A method that produces wildly different outcomes is less controllable. One that keeps goal completion without much deviation is more stable.

Requested Feedback

I would appreciate feedback and support on the following points:

1. **Triangle Framing.** Is the idea of treating emergent systems, reactive decision-making, and centralized narrative planning as the three corners of a spectrum clear and novel enough to guide a dissertation? If not, how might I sharpen or adjust this framing?
2. **Metrics, Landmark Methodology, and Human Evaluation.** Are the current system metrics (author/character goal completion, main plot point achievement, landmark adherence, variance across seeds, and scalability) together with the human-subjects measures of coherence, intention, and reader preference sufficient to compare hybrids against baselines? I am especially interested in advice on additional metrics that could strengthen the argument about "quality".
3. **Handling LLM Unpredictability.** I welcome suggestions on practical strategies for mitigating LLM drift or hallucination when agents attempt to follow a landmark sequence.

Any comments on other risks I may have overlooked or references to similar evaluation practices in adjacent work would also be very helpful.

References

Adams, T. 2019. Emergent narrative in dwarf fortress. In *Procedural storytelling in game design*, 149–158. AK Peters/CRC Press.

Bonet, B.; and Geffner, H. 2001. Planning as heuristic search. *Artificial Intelligence*, 129(1): 5–33.

Cavazza, M.; Charles, F.; and Mead, S. J. 2002. Character-based interactive storytelling. *IEEE Intelligent Systems special issue on AI in Interactive Entertainment*, 17(4): 17–24.

Eger, M.; and Martens, C. 2017. Practical specification of belief manipulation in games. In *Proceedings of the 13th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 30–36.

Farrell, R.; Ware, S. G.; and Baker, L. J. 2020. Manipulating narrative salience in interactive stories using Indexter's Pairwise Event Salience Hypothesis. *IEEE Transactions on Games*, 12(1): 74–85.

Fisher, M.; Siler, C.; and Ware, S. G. 2022. Intelligent de-escalation training via emotion-inspired narrative planning. In *Proceedings of the 13th Intelligent Narrative Technologies workshop at the 18th AAAI international conference on Artificial Intelligence and Interactive Digital Entertainment*.

Helmert, M. 2006. New complexity results for classical planning benchmarks. In *Proceedings of the 16th International Conference on Automated Planning and Scheduling*, 52–62.

Hoffmann, J.; Porteous, J.; and Sebastia, L. 2011. Ordered Landmarks in Planning. *The journal of artificial intelligence research*.

Lebowitz, M. 1985. Story-telling as planning and learning. *Poetics*, 14(6): 483–502.

Mateas, M.; and Stern, A. 2005. Structuring content in the Façade interactive drama architecture. In *Proceedings of the 1st AAAI international conference on Artificial Intelligence and Interactive Digital Entertainment*, 93–98.

Meehan, J. R. 1977. TALE-SPIN, an interactive program that writes stories. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, 91–98.

Pizzi, D.; and Cavazza, M. 2007. Affective storytelling based on characters' feelings. In *Proceedings of the AAAI Fall Symposium on Intelligent Narrative Technologies*, 111–118.

Porteous, J.; Cavazza, M.; and Charles, F. 2010. Applying planning to interactive storytelling: Narrative control using state constraints. *ACM Transactions on Intelligent Systems and Technology*, 1(2): 1–21.

Riedl, M. O.; and Young, R. M. 2010. Narrative planning: balancing plot and character. *Journal of Artificial Intelligence Research*, 39(1): 217–268.

Rivera, R. E. C.; Jhala, A.; Porteous, J.; and Young, R. M. 2024. The Story So Far on Narrative Planning. *Proceedings of the International Conference on Automated Planning and Scheduling*, 34(1): 489–499.

Sanghrajka, R.; Young, R. M.; and Thorne, B. R. 2022. HeadSpace: Incorporating Action Failure and Character Beliefs into Narrative Planning. In *Artificial Intelligence and Interactive Digital Entertainment Conference*.

Senanayake, L.; and Ware, S. G. 2025. Language Models as Narrative Planning Heuristics. In *Proceedings of the 20th International Conference on the Foundations of Digital Games*, 1–9.

Sun, Y.; Wang, P. J.; Chung, J. J. Y.; Roemmele, M.; Kim, T.; and Kreminski, M. 2025. Drama Llama: An LLM-Powered Storylets Framework for Authorable Responsiveness in Interactive Narrative. arXiv:2501.09099.

Teutenberg, J.; and Porteous, J. 2013. Efficient intent-based narrative generation using multiple planning agents. In *Proceedings of the 2013 international conference on Autonomous Agents and Multiagent Systems*, 603–610.

Valmeekam, K.; Marquez, M.; Sreedharan, S.; and Kambhampati, S. 2023. On the Planning Abilities of Large Language Models - A Critical Investigation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ware, S. G.; and Farrell, R. 2022. Salience as a narrative planning step cost function. In *Proceedings of the IEEE Conference on Games*, 433–440.

Ware, S. G.; Senanayake, L.; and Farrell, R. 2023. Causal Necessity as a Narrative Planning Step Cost Function. In *Proceedings of the 19th AAAI international conference on Artificial Intelligence and Interactive Digital Entertainment*, 155–164.

Ware, S. G.; and Siler, C. 2021. Sabre: A Narrative Planner Supporting Intention and Deep Theory of Mind. In *Proceedings of the 17th AAAI International Conference on Artificial Intelligence and Interactive Digital Entertainment*, 99–106.

Ware, S. G.; and Young, R. M. 2014. Glaive: a state-space narrative planner supporting intentionality and conflict. In *Proceedings of the 10th AAAI international conference on Artificial Intelligence and Interactive Digital Entertainment*, 80–86.

Wolpert, D.; and Macready, W. 1997. Macready, W.G.: No Free Lunch Theorems for Optimization. IEEE Transactions on Evolutionary Computation 1(1), 67-82. *Evolutionary Computation, IEEE Transactions on*, 1: 67 – 82.

Ye, A.; Cui, C. Z.; Shi, T.; and Riedl, M. 2023. Neural Story Planning. In *The AAAI-23 Workshop on Creative AI Across Modalities*.

Young, R. M. 1999. Notes on the use of plan structures in the creation of interactive plot. In *Proceedings of the AAAI Fall Symposium on Narrative Intelligence*, 164–167.

Young, R. M.; Ware, S. G.; Cassell, B. A.; and Robertson, J. 2013. Plans and planning in narrative generation: a review of plan-based approaches to the generation of story, discourse and interactivity in narratives. *Sprache und Datenverarbeitung, Special Issue on Formal and Computational Models of Narrative*, 37(1-2): 41–64.