Orchestrating Emergent Storytelling with Embodied Multi-Agent Systems

Parag K. Mital
The Garden in the Machine, Inc.
Los Angeles, CA
parag@emergentic.ai

Seth Rosetter
The Garden in the Machine, Inc.
Los Angeles, CA
seth@emergentic.ai

Arturo Castro Prieto
The Garden in the Machine, Inc.
Los Angeles, CA
arturo@emergentic.ai

Jacobo Heredia Zurita
The Garden in the Machine, Inc.
Los Angeles, CA
jacobo@emergentic.ai

Breanna Browning
The Garden in the Machine, Inc.
Los Angeles, CA
breanna@emergentic.ai

Abstract

We present a novel approach to emergent storytelling through multi-agent systems powered by large language models (LLMs), advancing beyond current approaches to game AI and interactive storytelling which rely on heavily scripted dialogue systems and moving closer towards genuinely emergent narrative ecosystems. Through two artworks / video games—*Conflicts* and *The Game of Whispers*—we demonstrate how LLM-driven agents with persistent memory, behavioral models, and coordination capabilities generate coherent narratives from simulated social dynamics. Our architecture introduces: (1) a hierarchical memory system integrating working memory, episodic buffers, and consolidated narrative storage; (2) a conversation graph that tracks topic centroids, engagement, and unresolved questions; (3) a hybrid orchestrator that directs autonomy by fusing LLM reasoning with the conversation graph; and (4) their integration within embodied agents with a streaming multimodal action-perception loop that enables spatial awareness and environmental responsiveness. Experiments reveal emergent behaviors including strategic deception, coalition formation, the spread of disinformation, and meta-narrative awareness. Our contributions include several architectural patterns for producing stable emergent narrative systems.

1 Introduction

Computational narrative has evolved from ELIZA's pattern matching [Weizenbaum, 1966] through TALE-SPIN's planning-based generation [Meehan, 1977] to recent neural approaches that are fundamentally reimagining storytelling. While games like Dwarf Fortress and The Sims have demonstrated emergent narrative through simulation, they rely on human players to sift through and curate the interesting stories from the dull ones [Ryan, 2018]. Recent multi-agent frameworks like CAMEL [Li et al., 2023], MetaGPT [Hong et al., 2024], or AutoGen [Wu et al., 2023] demonstrate unprecedented collaborative capabilities. Parallel efforts in interactive narrative research have explored balancing authored content and player agency, from Façade's pioneering drama manage-

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Creative AI Track.

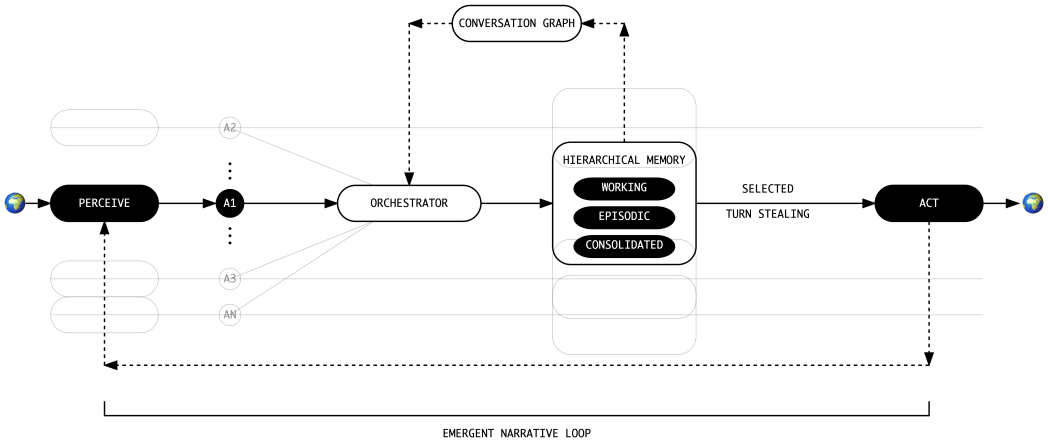

Figure 1: The cognitive architecture of our multi agent network comprises an action / perception feedback loop intercepted by several main components: a perception module, orchestrator, hierarchical memory, action module, and a conversation graph. Agents are indicated by the circular nodes labeled "A1", "A2", ..., "AN".

ment [Mateas and Stern, 2003] and IDtension's narrative planning [Szilas, 2003] to contemporary approaches using neural language models for story generation [Mirowski et al., 2023] and emergent narrative evaluation [Kreminski et al., 2024]. However, sustained narrative generation with multi-agent architectures that balances both a storyteller's input and emergent complexity where the system itself curates meaningful narratives from agent interactions remains an open and growing technical challenge.

Park et al.'s [Park et al., 2023] seminal work with generative agents achieved believable social behaviors through architectural innovations in memory streams, reflection, and planning with 25 agents spontaneously organizing a Valentine's Day party. While their system excels at emergent social dynamics, it lacks mechanisms for authorial guidance or narrative shaping. Maintaining coherence and narrative control across thousands of interactions with limited context windows requires explicit engineering for long-horizon consistency [Wang et al., 2024]. We address both challenges through a hybrid orchestration agent guided by a conversation graph informed scheduler that anticipates narrative flow while balancing pre-designed plot structure with emergent complexity, giving creators tools to shape narrative without sacrificing emergent complexity.

Conflicts and *The Game of Whispers* serve as two showcases of our unique and comprehensive approach to emergent storytelling showcasing the possibilities of multi-agent narrative storytelling. In *Conflicts*, two embodied agents engage in an endless debate while reconstructing intimate relationship dynamics from personal data and generating uncanny recognitions alongside novel interpretations. *The Game of Whispers* extends this to 13 embodied agents modeled after Mughal court figures whose political machinations emerge through social dynamics and power structures. We demonstrate how these systems can produce surprising, emotionally resonant narratives that reflect the complexity of human relationships and deepen our understanding of complex social dynamics.

2 System Architecture

The main system architecture is comprised of an action / perception feedback loop intercepted by several main components shown in Figure 1 and described in more detail in the following sections. We have also developed a web platform[1] for enabling authors to simulate their own stories using our multi-agent architecture, allowing them to create their own agents, define their backstories, and enable them to connect the agents to any external environment via an API in which they can interact. This platform allows for rapid iteration and testing of different narrative scenarios, enabling authors to explore the possibilities of emergent storytelling in a controlled environment.

[1] `https://emergentic.ai`

2.1 Perceive

The perception module encapsulates agent state, environment actors, locations, objects, and optional image captures. Each agent maintains a differential perception mechanism that compares previous and current payloads to detect environmental changes, preventing redundant LLM requests when the environment remains static. Each agent is blocked by the orchestrator and once selected, is given limited information as structured text which can include image blobs, text data, customizable agent schema data, and additional outputs from the hierarchical memory system described in the next sections.

2.2 Conversation Graph

The conversation graph is a directed graph where nodes represent agents and edges represent interactions. Each edge is weighted by the frequency of interactions and emotional valence, allowing us to track engagement and narrative flow. The graph also maintains topic centroids for each conversation thread, enabling agents to reference past discussions and maintain coherence over long narratives. This graph-based approach allows us to model conversational dynamics, identify unresolved questions, and guide the orchestrator in selecting the next agent to act. The output of the conversation graph is a set of probabilities for each agent to speak next along with a set of unresolved questions that the agents can use to guide their actions. This allows for a more structured and coherent narrative flow, while still allowing for emergent behaviors to arise from the interactions between agents.

2.3 Orchestrator

We introduce a hybrid orchestrator composed of a scheduler orchestrator and a runtime orchestrator with configurable modes (LLM, round-robin, random). The scheduler orchestrator fuses LLM reasoning with the conversation graph's next-speaker probabilities in order to make a choice as to which agent, if any, should go next. It maintains a limited context window and implements anti-monopolization through the conversation graph's model. The scheduler also generates context-aware prompts that include agent locations, available actions/tools for each agent, and graph-based insights such as unanswered questions and silent agents. The runtime orchestrator enforces turn-level mutual exclusion with a timestamp-based turn stealing mechanism that allows dramatic interruptions while preventing simultaneous utterances. Player actions override the schedule through explicit turn stealing, maintaining responsiveness in interactive scenarios.

2.4 Hierarchical Memory

Memory coherence across extended narratives requires moving beyond flat context windows due to limited attention Maharana et al. [2024]. In order to address this, we create a hierarchical memory composed of a three-tier storage system inspired by CoALA Sumers et al. [2024]: working memory with immediate context; episodic buffer with recent interactions indexed by temporal-semantic similarity; and consolidated memory with compressed narratives maintaining causal chains. We further track narrative continuity with the conversation graph's topic centroids, engagement, and coherence scores per edge, and compress long histories into topic-aware summaries. This prevents the context collapse observed in unconstrained systems while enabling callbacks to prior events.

2.5 Act

Once an agent is free to execute by the runtime orchestrator, they can act in the world either by streaming their text output, selecting a tool call (e.g. a specified action or tool call), or in the case of an embodied agent, through their visualization inside of a game engine. Actions are streamed from the LLM-backend including viseme and procedural body animation generation from the text and speech data, while the client streams back in text and image data used for the multimodal chat completion, completing an action-perception feedback loop for a single agent in the multi-agent system. Embodied agents also possess a spatial and visual image awareness that can further affect conversation dynamics through speaking probability, environmental triggers that can catalyze narrative events, and shared memory based on shared conversational graphs due to proximity leading to eavesdropping or direct conversation cues.

Figure 2: Stills from the artwork *Conflicts* where two former partners modeled as agents interact in an endless debate over their past relationship.

3 Implementations

We explore two implementations of our multi-agent architecture, each demonstrating a balance between authorial control and emergent complexity in narrative storytelling. In the first, *Conflicts*, we reconstruct past relationship dynamics through agents that engage in an endless debate and focus on the conversational dynamics of agent to agent narrative emergence. In the second, *The Game of Whispers*, we simulate Mughal court politics with historical figures, allowing emergent political narratives to unfold through social dynamics and power structures. This narrative opens the door to emergent storytelling where agents are both embodied and spatially aware, allowing for emergent actions based on their environment and the actions of other agents.

3.1 *Conflicts*: Reconstructing Past Relationship Narratives

Conflicts[2] is an artwork and video game born from asking: what happens when the messy, intimate details of past relationships are fed into an algorithmic crucible of large language models (LLMs) and reflected back onto us? Can a simulation of past relationship conflicts illuminate patterns of failed relational dynamics, acting as a mirror of oneself, perhaps even teaching us to empathize with one anothers perspective? Or does it merely create an uncanny, unsettling echo, potentially even furthering the effects of the trauma we sought to escape by breaking up with one another?

The work stages an encounter between two agents engaged in an endless debate of past conflicts intended to personify Mital and his former romantic partners (see Figure 2). The stage is a city scene inspired by downtown Los Angeles, where the two agents can interact. The agent's personalities, conversational styles, and even their perspectives on each other are derived from a so-called "relationship exit survey", questionnaires initially tested on Mital's friends before being sent to the actual ex-partners involved.

The form contained a wide range of optional and open-ended questions stemming from the introductory "describe yourself" to the more pointed and personal "describe the main sources of conflict you had with Parag." There were also questions surrounding what data would be allowed such as "would you be willing to use previous direct messages with Parag in the training of the LLM?", questions covering attachment styles, conflict patterns, and communication preferences. The system resulted in startling emergent behaviors with agents identifying recurring relationship patterns, resurrecting specific arguments with uncanny accuracy, and even developing meta-awareness of their situation with one agent challenging: "Why didn't you just go to therapy like a normal person?"

3.2 *The Game of Whispers*: Mughal Palace Historical Simulation

The Game of Whispers[3] is an artwork and video game, commissioned by the LACMA Art + Technology Lab, that reimagines the artistry of Mughal miniature painting through the lens of artificial intelligence and agents in an emergent storytelling exercise. Set within a reimagined version of Delhis historic Red Fort (see Figure 3), the piece parallels the political intrigue of Shah Jahans court with the role of AI in shaping modern disinformation. By weaving together historical narratives and

[2] `https://pkmital.com/home/works/conflicts/`

[3] `https://pkmital.com/home/works/the-game-of-whispers/`

Figure 3: Stills from the artwork *The Game of Whispers* where 13 agents modeled after historical figures from the time of Shah Jahan reenact a succession narrative through an emergent multi-agent orchestration. In the left, three agents are shown near one of the buildings in the map, while the right shows an overhead view of the entire map which was inspired by the Red Fort.

generative storytelling, the work explores how rumors, manipulation, and shifting power dynamics, which were once wielded by emperors and courtiers, now echo in an increasingly digital world. The artwork highlights the ways technology distorts truth through the amplification of misinformation, raising urgent questions about the mechanics of influence in both past and present societies.

Thirteen historical figures navigate Mughal succession politics, producing narratives set during the time period of Shah Jahan and his succession. The historical record is that Dara Shikoh, the eldest prince son of Shah Jahan succeeds the throne, and represents ideals of pluralism and mysticism, while his younger more ambitious brother Aurangezeb represents an orthodox rule governed by military rule and kills his brother and imprisons his father. Much of this will be known to the language models, though we additionally include authored elements of character backstories, locations, possible actions, and key plot events spanning 5 marjor arcs including intelligence gathering, conspiring to meet at a location, moving conspirators to the location, ambush, and spread of rumors to cover up the crime.

We observe a number of key emergent phenomena include strategic deception cascades leading to successful disinformation campaigns; assassination timing after approximately 15 minutes into a simulation where agents orchestrate a murder and its coverup based on alliance structures; poetic emergence where after fratricide, the unprompted reflection integrates the visual input of the nearby river with the ongoing narrative to produce the startling line: "The river runs red with the price of ambition." Lastly, by having spatial awareness and the freedom to move with direct feedback on their changing environment, the agents exhibited emergent phenomena where they did not want their plans overheard by other nearby agents:

> **Begum Roshanara** (to Roshan): Is everything prepared for the dinner with Dara? The wine must be poisoned without fail. We cannot afford any mistakes.
> *Hakim Khan appears.*
> **Roshan** (to Begum Roshanara): Shall we adjourn to the Khas Mahal? It provides a more secure atmosphere for discussing our plans without interruption.
> **Begum Roshanara** (to Roshan): The Khas Mahal is wise indeed. We must ensure our conversation there is secure. Once we arrive, make sure the wine is ready and the guests unbothered. We must execute everything perfectly tonight.

In another example, we see the aftermath of the assassination of Dara Shikoh, where the agents must plan their next steps and cover up the crime. The conversation illustrates how they coordinate their actions and manage the narrative to avoid suspicion:

> *Hakim Khan poisons Dara Shikoh.*
> **Roshan** *(internal)*: To act surprised and feign sorrow in front of others while assessing the situation.
> **Hakim Khan** (to Roshan): It is done. Dara will no longer be a threat. We must ensure that

this remains hidden and plan our next steps carefully.
Roshan (to Hakim Khan): This is a devastating loss for the court. We must ensure that those loyal to Dara do not suspect foul play. How can we maintain control without drawing attention?
Hakim Khan (to Roshan): We must declare this a sudden illness. The court will expect a cause, and we must control the narrative before whispers spread.
Roshan (to Hakim Khan): We must circulate that Prince Dara had been unwell for days, possibly even weeks, before his sudden demise. It will serve to distract any suspicion from us.

4 Discussion

Our work demonstrates multi-agent narrative systems that balance both authorship and emergent complexity and can produce compelling narrative works. We make three important contributions critical to maintaining this balance: (1) constraining emergence with precise and limited context through a simplified hierarchical memory system; (2) directing autonomy through hybrid orchestration layers enabling mutual exclusion while allowing for turn-stealing; and (3) maintaining narrative coherence through a conversation graph that tracks topic centroids, engagement scores, and unresolved questions; and (4) ensuring tight coherence of embodied agents with their environment through a feedback loop of action and perception. These components work together to create a system that can generate coherent narratives while still allowing for the complexity and unpredictability of emergent storytelling.

However, significant challenges remain. As agents become more equipped with the ability to act autonomously in complex environments, it will become increasingly important to maintain authorial control and orchestration so as to not let these system diverge beyond aligned intents. Yet, narrative coherence is still hard to measure. Worlds, agents, and stories and their interactions all change depending on the use case, making comparison across different architectures difficult. This will require further efforts such as [Kreminski et al., 2024], which seek to measure player experience in AI-based games. Future efforts should seek to develop standardized sandbox environments and coherence metrics for evaluating emergent narrative quality, while enabling measurement of: (1) causal consistency measuring how well agents follow their backstory and established motivations; (2) coherences over long term interactions or when their behavior starts to break down; (3) emotional resonance measuring how well the narrative evokes emotional responses; and (4) surprise and unpredictability measuring how well the narrative surprises both players and its creators over repeated runs. Further, by expanding testing and evaluation to other domains such as education, therapeutics, or conflict resolution, these systems could serve as foundational tools for research and innovation or even enterprise applications.

5 Conclusion

Multi-agent LLM systems enable genuinely emergent storytelling through careful architectural design. Our agent orchestrator, hierarchical memory architecture, conversation graph, and action/perception feedback loop work in concert to maintain an embodied agent's narrative coherence with the author while preserving their own emergent complexity. *Conflicts* and *The Game of Whispers* demonstrate these systems are capable of generating emotionally resonant stories that surprise even their creators. This paradigm shift from authoring content to designing possibility spaces is an exciting new frontier in computational creativity, interactive media, and world building that will have profound implications for the future of storytelling, not just in games, but likely in all forms of media where complex social interactions matter.

Acknowledgements

The authors would like to thank Joel Ferree and the LACMA Art + Technology Lab for their support in the development of *The Game of Whispers*. We also thank the many individuals who provided feedback on early versions of this work, including our colleagues at The Garden in the Machine, Inc. and the Serendipity Arts Festival in Goa, India, where the first version of *The Game of Whispers* was presented.

References

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. Metagpt: Meta Programming for a Multi-Agent Collaborative Framework. *12th International Conference on Learning Representations, ICLR 2024*, 2024.

Max Kreminski, Ben Samuel, Edward Melcer, and Noah Wardrip-Fruin. Evaluating AI-Based Games through Retellings. In *Proceedings of the Fifteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE-19)*, pages 45–51, 2024.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for mind exploration of large language model society. *Advances in Neural Information Processing Systems*, 36(NeurIPS), 2023. ISSN 10495258.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870. Association for Computational Linguistics, 2024. URL `https://aclanthology.org/2024.acl-long.747/`.

Michael Mateas and Andrew Stern. Façade: An experiment in building a fully-realized interactive drama. In *Game Developers Conference*, 2003.

James R. Meehan. Tale-spin, an interactive program that writes stories. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'77, page 9198, San Francisco, CA, USA, 1977. Morgan Kaufmann Publishers Inc.

Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581225. URL `https://doi.org/10.1145/3544548.3581225`.

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. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3586183.3606763.

James Ryan. *Curating Simulated Storyworlds*. PhD thesis, University of California, Santa Cruz, 2018.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive Architectures for Language Agents. *Transactions on Machine Learning Research*, 2024:1–32, 2024. ISSN 28358856.

Nicolas Szilas. Idtension: A narrative engine for interactive drama. In *Proceedings of the Technologies for Interactive Digital Storytelling and Entertainment (TIDSE) Conference*, pages 187–203. Fraunhofer IRB, 2003. URL `https://api.semanticscholar.org/CorpusID:15972197`.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Jim Fan, and Anima Anandkumar. Voyager: An Open-Ended Embodied Agent with Large Language Models. *Transactions on Machine Learning Research*, 2024:1–42, 2024. ISSN 28358856.

Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

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. pages 1–43, 2023. URL `http://arxiv.org/abs/2308.08155`.