# Coordinated Goal Evolution Restores Structural Quality in Deterministic Narrative Simulation Chains

Jack Chaudier Gaffney
jackchaudier@gmail.com

## Abstract

Multi-agent narrative simulations that carry persistent world state across episodes face a structural degradation problem: as agents' fixed goals become misaligned with the evolved story world, arc quality declines. We present NarrativeField, a deterministic simulation-extraction pipeline for multi-protagonist stories, and identify *information depletion*—the exhaustion of secrets, unresolved tensions, and belief asymmetries—as the root cause. A controlled perturbation experiment provides evidence that, within our pipeline, belief-state content causally affects extracted arc structure, with structurally central propositions producing order-of-magnitude larger quality effects than peripheral controls. Coordinated goal evolution across all six agents closes 89% of the quality gap between degraded and fresh simulations, revealing a collective threshold effect: single-agent evolution *degrades* quality, while coordinated evolution restores it. All experiments run deterministically at zero LLM cost with full reproducibility from fixed random seeds.

## 1 Introduction

Emergent narrative systems promise stories that arise from simulated character interactions rather than pre-authored plot structures [3, 13]. Unlike plan-based approaches that search over action sequences toward a desired ending [12, 17], simulation-first systems let narrative structure emerge from agent decision-making under social pressure. This approach has produced compelling results in single-episode settings [8, 11], but a fundamental challenge arises when simulations must produce *sequential* stories sharing persistent world state.

We identify and formalize a *chain degradation* problem: when world canon—accumulated beliefs, relationship states, and resolved secrets—carries forward between episodes, agents with fixed goal vectors find themselves in an information-saturated environment where the dramatic fuel that powered earlier stories has been consumed. Secrets are known, alliances are settled, and the decision engine's utility landscape flattens. The result is measurable: across four independent chain orderings of five-story sequences, mean arc quality scores show an overall downward trend by position, from 0.696 at position 0 to 0.637 at position 4 (Figure 2).

This paper makes four contributions:

1. We demonstrate *chain degradation* as a systematic, ordering-robust phenomenon in multi-episode narrative simulation, reducing the plausibility that the observed decline is a seed-order artifact.
2. We introduce a *belief-state perturbation* experiment providing evidence that, within our deterministic pipeline, narrative structure is causally sensitive to belief propositions, with structurally central claims producing order-of-magnitude larger quality effects than peripheral controls.
3. We present *coordinated goal evolution* as a mechanism for restoring narrative quality in information-depleted environments, closing 89% of the degradation gap.
4. We show that goal evolution exhibits a *collective threshold*: single-agent evolution degrades quality ($-0.027$), while coordinated six-agent evolution improves it ($+0.047$), supporting the hypothesis that narrative quality in our system is strongly shaped by the agent interaction topology.

All experiments are fully deterministic and reproducible from fixed random seeds.[1] The simulation, metrics, and arc extraction pipeline runs without LLM calls, at $0.00 compute cost per experimental condition via skip-narration mode.

[1] Chain seeds: 42, 51, 7, 13, 29 (four orderings). Evolution experiment: seed 7. Configuration files and experiment scripts available upon request.

## 2 Related Work

**Multi-agent narrative simulation.** Character-driven narrative has a long history from TALE-SPIN [9] through FAÇADE [8] and THESPIAN [14]. Recent work on generative agents demonstrated that LLM-backed characters produce emergent social behavior in open-ended sandbox environments [11]. Our system differs in using deterministic utility-based agents with explicit goal vectors rather than LLM inference for decision-making, enabling reproducible experiments and controlled causal analysis within the pipeline.

**Narrative planning.** Plan-based approaches search action spaces for sequences satisfying narrative goals [12, 17]. C2PO introduced causal commonsense ordering constraints [2]; neural planners use learned models to propose plot outlines [16]. These systems optimize for coherent plot structure but typically lack agent autonomy. Goal-Oriented Action Planning (GOAP) [10] bridges planning and agent autonomy but focuses on single-agent tactical behavior. Our approach inverts the pipeline: simulate first, extract structure post-hoc.

**Story sifting and extraction.** Story sifting [5] retrieves narrative patterns from simulation logs. WINNOW [6] provides a domain-specific language for incremental sifting queries. Our Rashomon extraction generalizes sifting by extracting complete per-protagonist arc structures with grammar validation and composite scoring, rather than pattern-matching individual story moments.

**Long-form narrative coherence.** Maintaining coherence across extended narratives is recognized as a key challenge [1]. Multi-agent collaborative generation with specialized roles has shown promise for document-length text [4]. BDI architectures for interactive narrative [15] provide a related formalism for persistent agent state. Our work specifically addresses the under-explored problem of *sequential* story coherence when world state persists across episodes.

## 3 System Description

NARRATIVEFIELD is a four-stage pipeline (Figure 1): simulation, metrics computation, arc extraction, and optional prose rendering. Steps 1–3 are fully deterministic; step 4 uses an LLM for prose generation but is not required for evaluation. All results in this paper use only the deterministic stages.

### 3.1 Simulation Engine

The simulation operates on discrete ticks with an event queue for simultaneous action resolution. Six agents—Thorne, Elena, Marcus, Lydia, Diana, and Victor—interact across five locations (dining table, kitchen, balcony, foyer, bathroom) with privacy, adjacency, and overhearing rules.

**Agent state.** Each agent maintains a `GoalVector` with seven utility dimensions: *safety*, *status*, *closeness* (a per-target dictionary), *secrecy*, *truth_seeking*, *autonomy*, and *loyalty*. Relationships are modeled as per-target triples of *trust*, *affection*, and *obligation* on $[-1, 1]$. Each agent holds categorical beliefs over propositions: `unknown`, `suspects`, `believes_true`, or `believes_false`.

**Decision engine.** At each tick, the engine scores candidate actions via:

$$u(a) = u_{\text{base}}(a) + b_{\text{flaw}}(a) + m_{\text{pace}}(a) + m_{\text{rel}}(a) - p_{\text{rec}}(a) + \epsilon \quad (1)$$

where $u_{\text{base}}$ evaluates goal-weighted action effects across all seven dimensions, $b_{\text{flaw}}$ applies character flaw biases, $m_{\text{pace}}$ enforces pacing physics (dramatic budget, stress, composure, hysteresis), $m_{\text{rel}}$ modifies for relationship state, $p_{\text{rec}}$ penalizes recently taken actions, and $\epsilon$ is drawn from a seeded pseudorandom generator. The agent selects $a^* = \arg\max_a u(a)$; the seeded noise term $\epsilon$ serves as a deterministic tiebreaker, ensuring identical action sequences given identical seeds. Additionally, SOCIAL_MOVE actions receive a fixed $-0.10$ penalty to discourage unproductive movement, and INTERNAL actions are capped at $u \leq 0.45$ to prevent inaction from dominating dramatic choices.

**Event types.** The simulation produces ten event types: CHAT, OBSERVE, SOCIAL_MOVE, REVEAL, CONFLICT, INTERNAL, PHYSICAL, CONFIDE, LIE, and CATASTROPHE. Each run generates approximately 200 events over 30 simulation ticks.

**Wounds.** A *wound* is a persistent conflict dyad: a pair of agents who engage in repeated CONFLICT events at the same location during a simulation. The *wound topology* is the set of active wounds across a run; we use it to track how the social conflict graph restructures under different experimental conditions (Section 6).

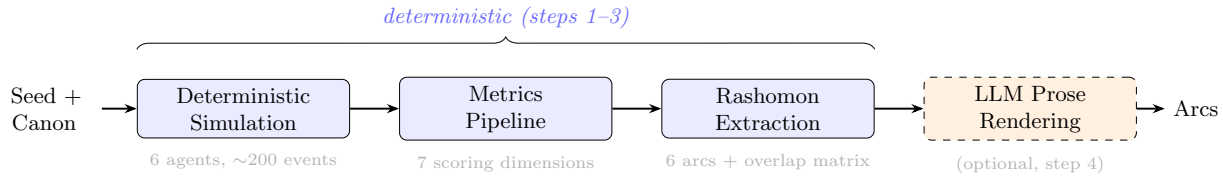


Figure 1: System pipeline. Steps 1–3 are fully deterministic and produce all metrics reported in this paper. Step 4 (LLM prose rendering) is optional and was disabled for all experiments.

**World canon.** A `WorldCanon` object accumulates persistent state between sequential stories: per-location tension residue, agent belief states with confidence scores, and committed texture facts. The engine supports configurable exponential decay between episodes (tension at rate $\alpha_t$ and belief confidence at rate $\alpha_b$, with beliefs reverting to `unknown` below confidence 0.1). However, all chain experiments reported in this paper use no-decay settings ($\alpha_t = \alpha_b = 1.0$); a separate decay experiment found zero structural effect (Section 4.2).

### 3.2 Metrics Pipeline

Arc quality is measured by a composite score $Q$ aggregating seven normalized dimensions ($M_i \in [0, 1]$):

$$\begin{aligned} Q = {} & 0.20\, M_{\text{var}} + 0.15\, M_{\text{peak}} + 0.15\, M_{\text{shape}} \\ & + 0.15\, M_{\text{sig}} + 0.15\, M_{\text{theme}} \\ & + 0.10\, M_{\text{irony}} + 0.10\, M_{\text{prot}} \end{aligned} \tag{2}$$

where $M_{\text{var}}$ is tension variance (normalized by a fixed constant of 0.013), $M_{\text{peak}}$ is peak tension magnitude, $M_{\text{shape}}$ rewards rising-then-falling tension profiles, $M_{\text{sig}}$ measures the impact of events at turning points, $M_{\text{theme}}$ evaluates thematic thread coherence across events, $M_{\text{irony}}$ rewards dramatic irony (events where audience knowledge exceeds participant knowledge), and $M_{\text{prot}}$ measures how central the focal agent is to their own arc. All $M_i$ are normalized to $[0, 1]$ using fixed constants defined in the scoring implementation; $M_{\text{var}}$ uses a normalization constant of 0.013 derived from population-level tension statistics. Tension-related dimensions ($M_{\text{var}} + M_{\text{peak}} + M_{\text{shape}}$) collectively account for 50% of the composite; the remaining 50% comes from structural, thematic, and dramatic irony dimensions. Weights were fixed across all experimental conditions; no per-condition tuning was performed. All metrics are computed deterministically from event logs.

### 3.3 Rashomon Extraction

Following the principle that a single sequence of events contains multiple valid narratives depending on whose perspective is centered [7], we extract one arc per agent per simulation. Each arc selects up to 20 events and validates against a beat grammar with the following constraints:

- At least one SETUP, one CONSEQUENCE, and exactly one TURNING_POINT beat.
- At least one development beat (COMPLICATION or ESCALATION; these are interchangeable within the development phase).
- Monotonic phase progression: setup → development → crisis → resolution (no phase reversals).
- Beat count between 4 and 20 (inclusive).
- Protagonist appears in ≥60% of arc events.
- Causal continuity: each event shares a causal link or participant overlap with its predecessor.
- Temporal span ≥10 sim-minutes and ≥15% of total simulation time.

A `RashomonSet` contains all six per-protagonist arcs, their composite scores, and a Jaccard overlap matrix quantifying narrative divergence. Population-level validation across 50 seeds yields 294/300 valid arcs (98%) with 44/50 seeds producing six valid arcs (88%).

## 4 The Chain Degradation Problem

### 4.1 Observation

We run five-story chains ($A \to B \to C \to D \to E$) where each story's simulation loads the accumulated canon from all prior stories. To control for seed-specific effects, we run four independent orderings of the same five seeds (42, 51, 7, 13, 29) and average scores by chain position. Table 1 and Figure 2 show the results.

Three of four orderings show negative first-to-last slope; one ordering (seeds 42, 51, 7, 13, 29) exhibits a U-shaped recovery at position 4 (0.697 → 0.672 →

Table 1: Mean arc score by chain position, averaged across four independent seed orderings. Overall downward trend with local variation at position 3.

| Position | Mean Score | Δ from Pos. 0 |
|---|---|---|
| 0 (first story) | 0.696 | — |
| 1 | 0.679 | −0.017 |
| 2 | 0.644 | −0.052 |
| 3 | 0.650 | −0.046 |
| 4 (fifth story) | 0.637 | −0.059 |

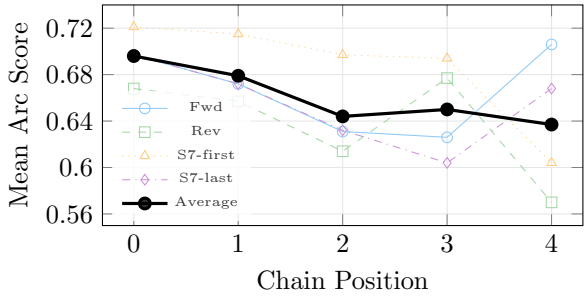


Figure 2: Mean arc score by chain position. Bold black line: position-averaged means across four seed orderings. Thin lines: individual orderings with distinct dash patterns for grayscale readability. Three of four orderings show negative first-to-last slope; one (Forward, solid) exhibits a U-shaped recovery at position 4 that does not replicate under averaging.

$0.631 \to 0.626 \to 0.706$) that did not replicate under seed-order controls, substantially reducing the likelihood that the observed degradation is only a seed-order artifact. The position-averaged means show a clear overall decline of $-0.059$ from position 0 to position 4 despite a local rebound at position 3.

For context, across the 50-seed Rashomon sweep at depth 0, per-seed mean arc scores have standard deviation $\sigma = 0.037$ (294 valid arcs, mean 0.674). The full-evolution lift of $+0.047$ reported in Section 6 corresponds to approximately $1.3\sigma$ of this baseline population distribution, providing a rough effect-size calibration despite the single-seed limitation of the evolution experiment.

## 4.2 Failed Interventions

Three prior interventions targeting different canon fields all produced zero measurable structural effect on arc quality, converging on the diagnosis that the decision engine consumes a narrower state interface than the full canon schema suggests.

Table 2: Belief states after Story B (seed 7). Four of seven claims are universally `believes_true` (BT), leaving only one structurally mixed claim. U = unknown, S = suspects, BF = believes_false.

| Claim | Di | El | Ly | Ma | Th | Vi |
|---|---|---|---|---|---|---|
| affair | BT | BT | BT | BT | BT | BT |
| diana_debt | BT | BT | BT | BT | BT | BT |
| lydia_knows | BT | BT | BT | BT | BT | BT |
| victor_inv | BT | BT | BT | BT | BT | BT |
| embezzle | U | U | S | BT | U | BF |
| guild_press | U | U | U | U | U | U |
| thorne_health | U | U | U | U | U | U |

**Texture accumulation.** Over a five-story chain, 69 narrator-invented texture facts (wine labels, physical gestures, room descriptions) accumulated in the canon. These feed into subsequent narration prompts but have zero structural effect: skip-narration runs produce identical metrics with or without texture. Texture and simulation structure are orthogonal channels.

**Belief confidence decay.** A float-valued confidence score was added to each belief, decayed by $0.85\times$ between episodes. This produced zero structural effect because the simulation's decision engine reads categorical `claim_states` (string enums), not confidence floats. Decaying a field that no decision path reads changes nothing.

**Tension residue decay.** Per-location tension residue was decayed by $0.60\times$ between episodes. Zero structural effect: within-run tension decay ($0.97\times$ per tick) dominates any starting residue, making inter-episode tension memory irrelevant to action selection.

These null results motivated the perturbation experiment (Section 5), which directly probes which state fields the decision engine is actually sensitive to.

## 4.3 Diagnosis

The degradation mechanism is *information depletion*. Table 2 shows the concrete belief state after two episodes: four of seven claims are universally `believes_true`—every agent knows every major secret. Only one claim (`embezzle`) retains meaningful disagreement. When agents carry forward this saturated knowledge, the space of dramatically productive actions shrinks: there is nothing to reveal,

no secret worth protecting, and alliances are settled. Fixed goals in this changed information landscape produce *goal–world misalignment*: agents optimize for objectives (e.g., Elena maintaining affair secrecy) that are no longer achievable or meaningful given what is universally known.

# 5 Experiment 1: Belief-State Perturbation

## 5.1 Method

To test whether agent belief states causally affect narrative structure within our deterministic pipeline, we design a controlled perturbation experiment. Starting from the post-Story-B world state (seed 7), we introduce two single-belief mutations and measure their downstream impact on Story C.

**Primary mutation.** The selection script ranks mutable beliefs by three criteria: (1) whether the claim involves a secret (thematic centrality to the scenario's primary conflicts), (2) belief diversity across agents (number of distinct belief states, favoring claims with active disagreement), and (3) agent importance (protagonist weighting). The intended primary target was Thorne's belief about `secret_affair_01`—the most thematically potent claim—but after two chain steps this belief is already `believes_true` (Table 2), making it immutable. The fallback selected Thorne × `secret_embezzle_01` as the highest-ranked remaining candidate: the only secret with active belief disagreement (three `unknown`, one `suspects`, one `believes_true`, one `believes_false`). Thorne's belief was mutated from `unknown` to `suspects`. That the most potent candidate was already spent is itself evidence of information depletion.

**Control mutation.** Diana's belief about `claim_guild_pressure`, a peripheral non-secret proposition with universal `unknown` status, mutated from `unknown` to `suspects`. Control selection avoids central tokens and prefers claims with low belief diversity.

**Instrumentation.** A `TracedBeliefs` wrapper intercepts all belief-state reads during Story C simulation, recording agent, claim, value, and access method for every read operation.

Table 3: Perturbation experiment results. Primary mutation to a structurally central belief produces ∼16× the quality impact of the peripheral control (exact ratio 15.8 from unrounded values), with complete event-level divergence.

| Metric | Primary | Control |
|---|---|---|
| Mean score Δ | −0.040 | −0.0025 |
| First divergence tick | 0 | None |
| Event sequences identical | No | Yes |
| Wound topology changes | +2, −1 | None |
| Valid arcs | 6/6 | 6/6 |

## 5.2 Results

Table 3 summarizes the results. The primary mutation produces a mean score delta of −0.040 (baseline 0.668 → 0.628), while the control mutation produces −0.0025 (0.668 → 0.666). Event traces diverge at tick 0 for the primary mutation—the very first decision cycle reads the mutated belief and produces a different action—while the control mutation generates an *identical* event sequence to baseline.

**Belief trace analysis.** The instrumented run recorded 13,895 total belief reads across 42 unique agent-claim fields. The most-read field was Diana × `secret_affair_01` with 913 reads, confirming that even fully-resolved propositions continue to dominate the decision engine's information consumption. The `secret_embezzle_01` claim appeared in the top-25 most-read fields for three agents (Elena: 321 reads, Diana: 319, Marcus: 279), consistent with its structural centrality as the sole asymmetric secret.

**Wound topology shift.** The primary mutation caused two new conflict patterns to emerge (Diana–Thorne and Marcus–Thorne at the dining table) and one to disappear (Diana–Victor), restructuring the social conflict graph. The control mutation left the wound topology unchanged.

## 5.3 Analysis

These results provide evidence for two claims within our deterministic pipeline. First, narrative structure is *causally sensitive* to belief-state content: changing a single categorical belief from `unknown` to `suspects` produces measurable structural consequences. Second, this sensitivity is *specific to structurally central propositions*: peripheral beliefs can be mutated

without observable effect, providing a natural control. The $\sim 16\times$ effect size ratio ($-0.040$ vs. $-0.0025$) and the all-or-nothing divergence pattern (tick 0 vs. identical) suggest a sharp, threshold-like sensitivity pattern in this scenario, where central beliefs act as bifurcation points in the simulation's trajectory space.

The perturbation experiment confirmed that beliefs are causally active within our pipeline, but showed that individual belief mutations cannot rescue an information-depleted system—the problem is systemic. This motivates testing whether the root cause is not individual belief states but the *goal structures* that interpret them.

# 6 Experiment 2: Goal Evolution

## 6.1 Method

Given the evidence that beliefs causally affect structure within our pipeline (Section 5), we hypothesize that agents whose goals have not adapted to the post-canon information landscape are the proximate cause of chain degradation. We test this by evolving agent goal vectors to reflect the narrative consequences of prior episodes.

**Experimental conditions.** Starting from the same post-Story-B canon (seed 7), we run five conditions:

1. **Depth-0 (fresh)**: New canon, no accumulated beliefs. Upper bound on quality.
2. **Baseline (depth 2)**: Loaded canon from Stories A→B. No goal changes.
3. **Thorne-only**: Only Thorne's goals and relationships evolved.
4. **Targeted (T+E)**: Thorne and Elena evolved.
5. **Full evolution**: All six agents' goals, relationships, and commitments evolved to reflect post-canon state.

Goal mutations modify `GoalVector` dimensions, per-target `closeness` values, relationship triples, and active commitments. Table 4 shows the most dramatically significant changes. Thorne's evolution reflects the consequences of discovering the affair in prior stories: `truth_seeking` rises sharply, closeness and trust toward Elena and Marcus invert, and loyalty drops. Elena's evolution reflects a shift from concealment to self-preservation: `secrecy` drops by half, her commitment `maintain_affair_secrecy` is removed, and she pivots social bonds from Marcus toward Diana and Lydia.

Table 4: Key goal mutations for Thorne and Elena in the full evolution condition. Values are representative fields; 15+ fields were modified per agent.

| Agent | Field | Before | After |
|---|---|---|---|
| Thorne | truth_seeking | 0.60 | 0.95 |
| Thorne | closeness[elena] | 0.70 | −0.35 |
| Thorne | closeness[marcus] | 0.60 | −0.55 |
| Thorne | rel.elena.trust | 0.80 | −0.25 |
| Thorne | loyalty | 0.80 | 0.55 |
| Elena | secrecy | 0.90 | 0.45 |
| Elena | commitment | removed | |
| Elena | closeness[marcus] | 0.80 | 0.20 |
| Elena | safety | 0.70 | 0.85 |
| Elena | closeness[diana] | 0.60 | 0.80 |

Table 5: Arc quality across conditions. Full evolution closes 89% of the gap between baseline and fresh simulations. Means are over valid arcs only.

| Condition | Mean | Δ Base | Δ Fresh |
|---|---|---|---|
| Depth-0 (fresh) | 0.721 | +0.053 | — |
| Full evolution | 0.715 | +0.047 | −0.006 |
| Targeted (T+E) | 0.683 | +0.015 | −0.038 |
| Baseline (depth 2) | 0.668 | — | −0.053 |
| Thorne-only† | 0.641 | −0.027 | −0.080 |

† 5/6 valid arcs (Marcus arc invalid).

**Determinism verification.** The baseline was run twice with identical seeds; the repeat produced identical event sequences with metric deltas below $10^{-12}$, confirming deterministic reproducibility.

## 6.2 Results

Figure 3 and Table 5 present aggregate scores. Full evolution achieves a mean of 0.715, closing 89% of the 0.053-point gap between baseline (0.668) and fresh (0.721). The relationship between agent count and quality is non-monotonic: evolving one agent degrades quality below baseline ($-0.027$), two agents partially recover ($+0.015$), and six agents approach the fresh ceiling ($+0.047$). This collective threshold pattern—partial evolution is counterproductive, while comprehensive evolution restores quality—is the central empirical finding of this paper.

Table 6 shows per-agent breakdowns. Table 7 reveals the mechanism: full evolution nearly doubles CONFLICT events ($7 \rightarrow 13$) and more than triples CATASTROPHE events ($2 \rightarrow 7$), while reducing low-tension CHAT ($24 \rightarrow 13$) and routine PHYSICAL

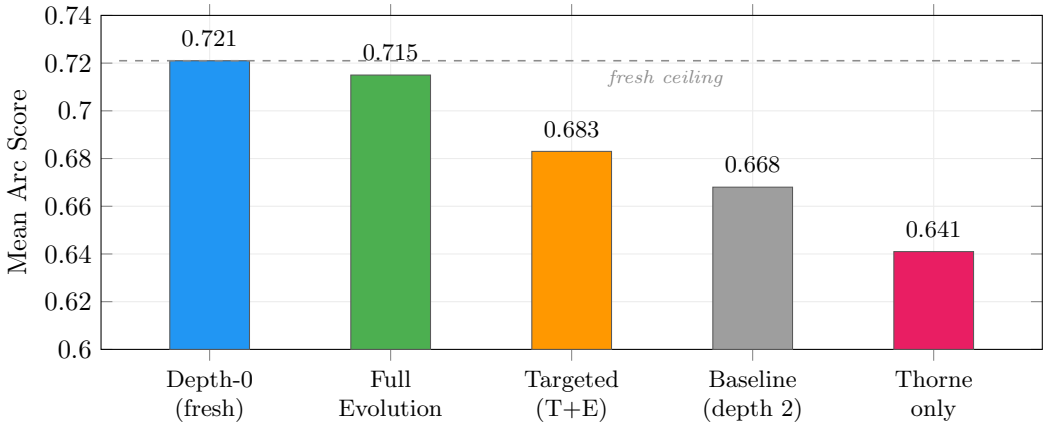


Figure 3: Mean arc quality scores across five experimental conditions. Dashed line marks the depth-0 (fresh) ceiling. Full evolution closes 89% of the gap between baseline and fresh. Thorne-only evolution *degrades* quality below baseline, revealing a collective threshold effect.

Table 6: Per-agent arc scores across conditions. Elena shows the largest improvement under full evolution (+0.258 vs. baseline); Marcus shows the largest decline (−0.067). "—" indicates an invalid arc (excluded from condition means).

| **Condition** | **Thorne** | **Elena** | **Marcus** | **Lydia** | **Diana** | **Victor** |
|---|---|---|---|---|---|---|
| Depth-0 (fresh) | 0.727 | 0.718 | 0.735 | 0.718 | 0.713 | 0.716 |
| Full evolution | 0.718 | 0.726 | 0.656 | 0.728 | 0.743 | 0.720 |
| Targeted (T+E) | 0.676 | 0.728 | 0.665 | 0.679 | 0.678 | 0.672 |
| Baseline (depth 2) | 0.709 | 0.468 | 0.723 | 0.719 | 0.706 | 0.685 |
| Thorne-only | 0.663 | 0.614 | — | 0.665 | 0.626 | 0.639 |

Table 7: Event type distribution (reported types; T-only = Thorne-only evolution). Full evolution shifts the genre profile: conflict and catastrophe increase while chat and physical decrease.

| **Type** | **D-0** | **Base** | **T+E** | **Full** | **T-only** |
|---|---|---|---|---|---|
| chat | 21 | 24 | 21 | 13 | 20 |
| conflict | 7 | 7 | 9 | 13 | 9 |
| reveal | 6 | 2 | 4 | 0 | 3 |
| lie | 6 | 6 | 5 | 5 | 4 |
| confide | 10 | 4 | 1 | 4 | 0 |
| catastrophe | 3 | 2 | 7 | 7 | 4 |
| physical | 16 | 20 | 13 | 11 | 11 |
| observe | 42 | 43 | 76 | 59 | 66 |

(20 → 11). The baseline's suppressed REVEAL count (2 vs. 6 in fresh) reflects the information depletion problem directly: with most secrets already known, there is nothing left to reveal.

## 6.3 Key Findings

**Finding 1: Elena as canary.** Elena's baseline arc score (0.468) is a dramatic outlier—the only agent below 0.60. Her original goal structure (secrecy 0.90, commitment to maintain secrecy) produces incoherent behavior when the affair is universally known (Table 2). Full evolution restores her to 0.726 (+0.258), the largest single-agent improvement, by replacing secrecy-oriented goals with post-revelation motivations.

**Finding 2: Collective threshold.** Evolving Thorne alone *degrades* mean quality by −0.027 and invalidates Marcus's arc entirely (5/6 valid). Evolving Thorne and Elena together recovers only +0.015. Full six-agent evolution achieves +0.047. This non-monotonic pattern suggests a coordination requirement: partial evolution is counterproductive, while comprehensive evolution achieves near-complete recovery. Locating the precise threshold requires 3-, 4-, and 5-agent conditions (future work). The finding supports the hypothesis that, in our system, nar-

rative quality is strongly shaped by the agent interaction topology: Thorne's evolved hostility toward Marcus produces good narrative structure only when Marcus, Lydia, Diana, and Victor have adapted their goals to respond coherently.

**Finding 3: Wound topology restructuring.** Table 8 shows the full wound topology across conditions. Full evolution causes three wound patterns to disappear (Thorne–Victor, Lydia–Marcus, Lydia–Thorne) and two to appear (Diana–Thorne, Marcus–Thorne), producing a net reduction from 6 to 5 active wounds. The persistent wound Marcus–Victor (population frequency 0.94) survives all conditions, suggesting it reflects a structural property of the scenario rather than a belief-contingent pattern.

**Finding 4: Overlap redistribution.** Baseline arcs show high pairwise overlap (top Jaccard: 0.481, Lydia–Thorne), indicating convergence toward similar events. Full evolution reduces maximum overlap to 0.379, producing more narratively distinct per-protagonist arcs—the different agents' stories diverge more under coordinated evolution.

**Finding 5: Immediate divergence.** All evolution conditions diverge from baseline at tick 0—the first decision cycle. Because goal vectors are direct inputs to $u_{\text{base}}$ (Eq. 1), any goal change immediately alters action selection. This contrasts with the perturbation experiment's control mutation, which produced zero divergence.

# 7 Discussion

## 7.1 Information Equilibrium

Chain degradation can be understood through an information-theoretic lens. A fresh simulation begins with high *dramatic potential*: secrets are unknown, relationships are ambiguous, and the space of information-revealing actions is large. As episodes accumulate and secrets resolve, the system approaches an information equilibrium. The canon state in Table 2 makes this concrete: after just two episodes, 24 of 42 agent-claim pairs (57%) are `believes_true`, and four of five secret claims are universally known. The decision engine still reads these beliefs every tick (13,895 reads in our traced run), but they no longer create utility gradients that differentiate between actions.

Goal evolution operates as a *gradient injection* mechanism: by changing what agents value, it creates new utility gradients without requiring new secrets or propositions. Thorne shifting from loyalty-oriented to truth-seeking goals makes previously low-utility actions (confrontation, investigation) newly attractive, restoring the conditions for dramatic arc formation.

## 7.2 Collective Threshold

The non-monotonic relationship between evolved agent count and quality recovery suggests a collective threshold phenomenon in our system. One evolved agent in a system of five unevolved agents produces *worse* outcomes than no evolution, because the evolved agent's new behavior is met with response patterns calibrated for the old interaction dynamics. Two evolved agents achieve partial recovery. Six achieve near-complete recovery. The pattern suggests that a substantial fraction of the interaction graph must be updated for coherent collective behavior to emerge; pinpointing the precise threshold between 2 and 6 agents requires additional intermediate conditions (3, 4, and 5 evolved agents) left for future work. Whether this coordination requirement generalizes beyond our six-agent scenario is an open question.

## 7.3 Implications for Serial Narrative

These results have practical implications for systems generating season-length narrative content. Rather than accumulating secrets and resolutions that gradually deplete dramatic potential, such systems could implement periodic goal evolution passes—analogous to character development between seasons of a television series—where agent motivations are updated to reflect the narrative consequences of prior episodes. The 89% quality recovery demonstrated here suggests that goal evolution alone, without introducing new characters, locations, or secrets, may be sufficient to sustain narrative quality across longer episode sequences than fixed-goal baselines can support.

## 7.4 On Hand-Tuned Evolution

A natural objection is that hand-tuned goal evolution simply encodes good character development by authorial fiat. However, the critical finding is not that particular goal values produce quality recovery, but that the *topology* of evolution matters: identical

Table 8: Wound topology across conditions. ✓ indicates wound pattern present; "—" indicates absent. Population frequency (Pop. Freq.) is the fraction of 50 baseline seeds exhibiting each pattern. Full evolution restructures the conflict graph, losing 3 wounds and gaining 2.

| Wound Pattern | Pop. Freq. | Depth-0 | Baseline | Targeted | Full | Thorne-only |
|---|---|---|---|---|---|---|
| Marcus–Victor | 0.94 | ✓ | ✓ | ✓ | ✓ | ✓ |
| Diana–Thorne | 0.90 | ✓ | — | ✓ | ✓ | ✓ |
| Thorne–Victor | 0.90 | ✓ | ✓ | ✓ | — | ✓ |
| Elena–Thorne | 0.86 | ✓ | ✓ | ✓ | ✓ | — |
| Lydia–Marcus | 0.78 | ✓ | ✓ | ✓ | — | ✓ |
| Lydia–Thorne | 0.66 | — | ✓ | ✓ | — | ✓ |
| Marcus–Thorne | 0.56 | ✓ | — | ✓ | ✓ | ✓ |
| Diana–Victor | 0.50 | — | ✓ | — | ✓ | — |
| **Total** | | **6** | **6** | **7** | **5** | **6** |

evolution style applied to one agent degrades quality and invalidates a co-protagonist's arc, while the same style applied across all agents achieves near-complete recovery. This coordination requirement is not an artifact of the specific mutation values—it reflects a structural property of how multi-agent utility landscapes interact. Automated goal evolution (Section 8) would test whether machine-generated mutations exhibit the same coordination requirement.

### 7.5 Limitations

1. **Single scenario.** All experiments use the Dinner Party scenario with six agents and five locations. Generalization to larger populations, different social dynamics, or non-social scenarios remains untested.
2. **Hand-tuned mutations.** Goal evolution profiles were manually designed to reflect plausible character development. Automated goal evolution is future work.
3. **Single seed per condition.** The Rashomon sweep validates baseline quality across 50 seeds ($\sigma = 0.037$), but the goal evolution experiment uses seed 7 only. Seed-specific interactions between initial conditions and goal mutations cannot be ruled out.
4. **Composite scoring.** Arc quality is a weighted composite of seven dimensions (Eq. 2). Different weightings could change condition rankings, though tension-related and non-tension dimensions are balanced at 50/50.
5. **Depth-2 chain.** We test at canon depth 2. Deeper chains may require more aggressive evolution or additional mechanisms.
6. **No human evaluation.** All quality judgments are automated metrics; perceptual quality of narrated output has not been assessed.

## 8 Future Work

**Stance machines.** Hand-tuned evolution profiles could be replaced by pre-defined goal archetypes—*stance profiles*—with deterministic transitions (e.g., Elena.THE_CONCEALER → Elena.THE_SURVIVOR). Selection between stances at episode boundaries could use constrained classifiers conditioned on prior-episode metrics, providing structured goal evolution without per-field tuning.

**Structural gradient ascent.** At $0.00 per deterministic run, brute-force sensitivity sweeps over the goal vector space are computationally feasible. Systematic evaluation of goal mutations by their downstream arc quality impact could identify optimal evolution configurations without hand-tuning.

**Affordance injection.** Goal evolution restores quality by creating new utility gradients over *existing* information. Complementarily, introducing new information objects—evidence, leverage, commitments—at episode boundaries would provide fresh dramatic fuel alongside evolved goals.

**Alternative scenarios.** The Dinner Party relies on information asymmetry (secrets, social alliances) as its primary dramatic fuel. Whether chain degradation occurs in exploration or quest scenarios—where dramatic potential derives from spatial discovery rather than social revelation—remains an open question that would test generality of the information depletion mechanism.

**Scaling experiments.** Testing with larger agent populations, deeper chains (5–10 episodes), and multiple scenario types would establish the generality

of the degradation and collective-threshold phenomena.

**Human evaluation.** A perceptual quality study comparing narrated output from baseline versus evolved conditions would validate whether the automated arc score improvements correspond to human judgments of narrative quality.

# 9 Conclusion

We have demonstrated that multi-episode narrative simulations with persistent world state face a systematic chain degradation problem rooted in goal–world misalignment. Through controlled perturbation experiments, we provided evidence that, within our deterministic pipeline, belief-state content causally affects narrative structure, with structurally central propositions producing order-of-magnitude larger effects than peripheral controls. Coordinated goal evolution across all agents closes 89% of the quality gap, approaching fresh-simulation quality without discarding accumulated canon. The collective nature of this recovery—single-agent evolution degrades quality while coordinated evolution restores it—suggests that narrative quality in multi-agent simulation is strongly shaped by the interaction topology rather than by individual character properties alone.

These findings suggest that the longstanding challenge of maintaining narrative quality in extended multi-agent simulations may be addressable through principled goal adaptation mechanisms, opening a path toward simulation-driven serial narrative systems that sustain dramatic quality across longer episode sequences than fixed-goal baselines can support.

# References


[1] Nader Akoury, Shufan Wang, Josh Whittaker, Doug Burdick, and Mohit Iyyer. A framework for document-level text generation via planning. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, 2023.

[2] Prithviraj Ammanabrolu, Wesley Cheung, William Tu, and Mark O. Riedl. Automated storytelling via causal, commonsense plot ordering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[3] Marc Cavazza, Fred Charles, and Steven J. Mead. Character-based interactive storytelling. *IEEE Intelligent Systems*, 17(4):17–24, 2002.

[4] Fantine Huot Crippa et al. Agents' room: Narrative generation through multi-step collaboration. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[5] Jacob Garbe, Max Kreminski, Ben Samuel, Noah Wardrip-Fruin, and Michael Mateas. Story sifting. In *Proceedings of the International Conference on the Foundations of Digital Games (FDG)*, 2019.

[6] Max Kreminski, Ben Samuel, Edward Melcer, and Noah Wardrip-Fruin. Winnow: A domain-specific language for incremental story sifting. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*, 2020.

[7] Akira Kurosawa. Rashomon, 1950. Film. Daiei Film.

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

[9] James R. Meehan. TALE-SPIN, an interactive program that writes stories. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 91–98, 1977.

[10] Jeff Orkin. Three states and a plan: The A.I. of F.E.A.R. *Game Developers Conference*, 2006.

[11] 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 *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, 2023.

[12] Mark O. Riedl and R. Michael Young. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39:217–268, 2010.

[13] James Owen Ryan, Michael Mateas, and Noah Wardrip-Fruin. Open design challenges for interactive emergent narrative. In *Proceedings of the International Conference on Interactive Digital Storytelling (ICIDS)*, pages 14–26. Springer, 2015.

[14] Mei Si, Stacy C. Marsella, and David V. Pynadath. Thespian: Modeling socially normative behavior in a decision-theoretic framework. In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA)*, pages 369–382, 2005.

[15] Tim Wadsley and James Owen Ryan. A BDI agent architecture for character-driven interactive narrative. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*, 2013.

[16] Anbang Ye, Christopher Nair, Prithviraj Ammanabrolu, and Mark O. Riedl. Neural story planning. In *Findings of the Association for Computational Linguistics (ACL)*, 2022.

[17] R. Michael Young, Stephen G. Ware, Brad A. Cassell, and Justus Robertson. 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*, 37:41–64, 2013.