Beyond Text Intelligence: Evaluating Narrative Intelligence in LLM Story Systems with the Dramatica Platform, Narrova, Subtxt, and NCP

Whitepaper

Dramatica Narrative Platform

Subtxt · Narrova · Narrative Context Protocol (NCP)

Abstract

Large Language Models (LLMs) make fluent, localized story generation trivial—but not meaningful narrative. A recent EMNLP survey on LLMs for story generation catalogs a growing ecosystem of systems that scaffold LLMs with prompts, templates, and multi-agent orchestration, yet still report persistent failures in global coherence, emotional depth, and thematic clarity. These systems operate primarily at the level of text intelligence: they optimize the probability and style of sequences of words.

This whitepaper introduces a complementary paradigm: narrative intelligence—the ability to model, evaluate, and generate stories as structured systems of inequity, perspective, and resolution. We present the Dramatica narrative platform—comprising Dramatica theory, the multi-agent Narrova narrative intelligence layer, the deep exploratory Subtxt workspace, and the open-source Narrative Context Protocol (NCP)—as a concrete implementation of narrative intelligence at scale.

We focus on three pillars:

1. Evaluation and story metrics grounded in formal narrative structure rather than surface text alone.
2. Narrative intelligence vs text intelligence, and why modern story systems require both—but must be led by narrative.
3. NCP as an open standard that other systems can adopt, while acknowledging that its core model and semantics ultimately trace back to Dramatica's narrative framework.

We argue that the EMNLP survey's own findings—on weak evaluation practices, lack of benchmarks, and structural limitations of LLM-generated stories—implicitly point toward exactly what Dramatica, Narrova, Subtxt, and NCP already implement. All roads eventually lead to a Dramatica-style story model; the only question is when each practitioner chooses to step onto that path.

1. Introduction: A Landscape Dominated by Text Intelligence

The EMNLP 2025 survey "A Survey on LLMs for Story Generation" organizes recent work into a taxonomy of independent story generation (LLM as primary author) and author-assistance (LLM as support), with further breakdowns into constrained, prompt-based, outline-based, multi-agent, and adaptive systems. It compares these systems along four dimensions:

• Methodology and architecture (prompt templates, multi-agent pipelines, planning modules). • Datasets (WritingPrompts, Story Commonsense, domain-specific scraped corpora). • Evaluation methods (user studies, automatic metrics, LLM-as-a-judge). • LLM usage patterns (API models vs open-source models, single vs multi-agent setups).

Across this diversity, several themes recur:

• Systems rely heavily on templated prompts, human-in-the-loop editing, and ad hoc constraints. • Evaluation practices are fragmented, small-scale, and rarely tied to any formal model of narrative quality. • Even the most sophisticated pipelines acknowledge that current LLMs struggle with story arcs, turning points, character development, and emotional dynamics.

In other words, the surveyed ecosystem is rich in text intelligence—models that are extremely good at generating locally fluent and stylistically plausible text—but almost entirely lacks a formal conception of narrative intelligence.

The Dramatica narrative platform, Narrova, Subtxt, and NCP were built precisely to fill this gap.

2. Limitations of Current LLM Story Systems

2.1 Taxonomy as interaction, not narrative structure

The survey divides systems by interaction pattern and authorship: • Independent vs author-assistance. • Single-agent vs multi-agent. • Constrained vs prompt-based vs outline-based. • Adaptive vs static.

This is a useful engineering classification, but it is structurally agnostic. It says nothing about:

• What constitutes a complete narrative argument. • How perspectives (Main Character, Influence Character, Objective Story, Relationship Story) interplay. • How conflicts are organized and resolved over time.

Every system surveyed uses LLMs as powerful text engines and then tries to patch story-level issues with more scaffolding.

2.2 Datasets: text without storyforms

The survey notes that: • Many systems use no formal dataset at all, instead relying on live user input or synthetic prompts. • Others use generic corpora (WritingPrompts, Story Commonsense) or domain-specific scraped data (e.g., postpartum-maternal narratives).

Critically, none of these datasets encode narrative structure explicitly—no storyforms, no throughlines, no model of the inequity being explored. At best, a few systems incorporate:

• Emotion/action tags. • Outlines, premises, or “abstract acts” for interactive games.

These are still surface- or plot-level representations, not a formal narrative model.

2.3 Evaluation: metrics without a theory of “good”

Evaluation in the surveyed systems is fragmented:

• User studies (often with $N \leq 35$), focusing on engagement, usability, or perceived creativity. • Automatic metrics like BLEU, ROUGE, and BERTScore for sentence-level tasks. • LLM-as-a-judge setups that ask a model to rate or compare stories.

The survey itself highlights a notable gap:

• No standardized benchmark for story generation. • No widely adopted story-specific metrics that capture arcs, turning points, or thematic development.

In other words, the community lacks an agreed definition of what it means for a story to be structurally good.

2.4 Structural and emotional limitations of LLM-generated narratives

The survey cites recent findings that:

• GPT-4 and Claude fail to match human narratives on story arc development, turning points, and affective measures; outputs are structurally uniform and emotionally shallow, especially for darker plots. • LLMs struggle to capture subtext and narrator reliability in literary summaries.

No amount of prompt engineering or multi-agent orchestration has yet produced human-level narrative structure. This is precisely the problem Dramatica was originally designed to solve.

3. Narrative Intelligence vs Text Intelligence

3.1 Text intelligence

Text intelligence is a model's ability to: • Predict and generate plausible sequences of tokens. • Match style, tone, and local coherence. • Satisfy lexical or topical constraints (e.g., include vocabulary words, mimic a genre).

LLMs excel at text intelligence. But narrative quality is not reducible to local token probabilities.

3.2 Narrative intelligence

Narrative intelligence is the capacity to:

• Represent inequities—the tensions and imbalances that drive a story. • Organize perspectives on those inequities (e.g., Main Character vs Objective Story) and maintain their interactions over time. • Determine and maintain story arcs, dynamics, and structural completeness. • Evaluate whether a given story instance realizes a coherent narrative argument.

The Dramatica theory of story provides such a model. Dramatica treats narrative as the processing and resolution of inequity, encoded in an interconnected Storyform—a specific configuration of narrative dynamics, storypoints, and storybeats.

Narrative intelligence requires:

• A symbolic model of story (the Storyform). • A way to expose this model to tools and agents (NCP). • Operational systems that can generate, inspect, and revise stories in terms of that model (Subtxt, Narrova).

3.3 Why narrative intelligence must lead

Text intelligence answers: "What token should come next?" Narrative intelligence answers: "What needs to happen next, given the current inequity and the narrative argument we're making?"

LLMs answer the first question exceptionally well. Without a narrative model, they can only approximate the second question by pattern matching over training data. This is why we see:

• Strong local coherence but weak global structure. • Surface-level emotion words but shallow emotional journeys. • Stylistic creativity but fuzzy or contradictory thematic messages.

By contrast, a Dramatica-based system knows before any text is generated:

• Which throughlines are in play. • What the structural progression must be. • What problem/solution dynamics are being argued.

Text intelligence should be a servant to narrative intelligence—not the other way around.

4. The Dramatica Narrative Platform

The Dramatica narrative platform unifies three major components:

1. Dramatica theory as the underlying narrative model.
2. Subtxt as the deep exploration and authoring workspace.
3. Narrova as the multi-agent narrative intelligence layer.

These are connected via the Narrative Context Protocol (NCP), an open standard for encoding Storyforms as portable, machine-readable data.

4.1 Dramatica and Storyform

At its core, the platform builds on Dramatica's Storyform:

• Dynamics: high-level choices about narrative resolution (change vs steadfast, success vs failure, etc.). • Storypoints: nested quads of conflict (domains, concerns, issues, problems, solutions, etc.). • Storybeats: the temporal unfolding of these conflicts (signposts and journeys).

A valid Storyform encodes a complete, coherent narrative argument. This gives the platform:

• An objective sense of structural completeness and consistency. • A shared representation for all downstream tools (Subtxt, Narrova, external engines via NCP).

4.2 Subtxt: deep exploration and structural authoring

Subtxt is the platform's deep exploratory and authoring environment. It allows writers to:

• Define core meaning, theme, and structure through a Dramatica-powered Storyforming engine. • Express that intent in a workspace that translates Storyforms into Storybeats and scene-level guidance. • Use AI-powered assistance that is always constrained by the Storyform, ensuring that generated text serves the intended narrative argument.

Subtxt is where narrative intelligence and text intelligence first meet: authors operate at the level of meaning and structure; LLMs help articulate that meaning in prose.

4.3 Narrova: multi-agent narrative intelligence

Narrova is a multi-agent system built on top of Dramatica. Each agent specializes in a distinct phase of the storytelling process, aligned with Dramatica's four stages of communication: Storyforming, Story Encoding, Storyweaving, and Story Reception.

Key properties:

• A narrative expertise layer that interprets user questions and story state through an in-depth understanding of narrative theory, rather than generic pattern matching. • Automatic routing: authors do not have to select agents; Narrova infers where they are in the workflow and routes requests to the appropriate specialized agent. • Tight integration with Storyforms (via NCP) so that Narrova can reason about and manipulate narrative structure, not just text.

In the taxonomy of the EMNLP survey, Narrova subsumes multiple roles: an author-assistance system with static and adaptive capabilities, and a role-based multi-agent architecture, but with roles defined by narrative function rather than ad hoc prompting.

4.4 NCP: an open standard grounded in Dramatica

The Narrative Context Protocol (NCP) is an open-source, community-driven standard for encoding Storyforms and their narrative context in a portable, machine-readable way.

NCP:

• Builds explicitly on Dramatica's narrative model and its Storyform structure. • Represents narrative structure (inequity, perspectives, storypoints, storybeats) as a first-class object separate from any

particular telling. • Enables multi-agent ecosystems—writing tools, game engines, visual generators, analysis systems—to share a single narrative intent and remain coherent.

Because NCP is open, any storytelling system—academic or commercial—can adopt it. But because NCP's semantics are Dramatica's semantics, any serious attempt to standardize narrative structure ultimately converges on the same conceptual ground. In that sense, all roads eventually lead back to Dramatica; NCP simply makes that destination accessible and interoperable.

5. Evaluation and Story Metrics in a Dramatica/NCP World

The EMNLP survey calls explicitly for better benchmarks and story-specific metrics. Dramatica plus NCP provides an immediate path to such metrics.

5.1 Gaps in current evaluation

Today's systems evaluate LLM story generation using:

• Small user studies, often domain- and interface-specific. • Surface metrics (BLEU/ROUGE/BERTScore) that correlate poorly with narrative quality. • LLM-as-a-judge scores that are scalable but opaque and themselves limited in narrative understanding.

The survey's own discussion acknowledges that:

• There is no comprehensive benchmark of story capabilities. • Current metrics do not capture arc development, turning points, or affective trajectories.

This is exactly what a formal Storyform model is designed to express and evaluate.

5.2 Structural metrics derived from Storyforms

With Dramatica and NCP, we can define structural metrics that are:

• Interpretable: grounded in explicit narrative theory. • Portable: encoded in NCP and usable across tools. • Model-agnostic: applicable to any LLM or generation pipeline.

Examples include:

1. Storyform Validity Score – Does the encoded story correspond to a valid Storyform (i.e., a legal configuration of dynamics, storypoints, and storybeats)? – Metrics: proportion of required structural positions filled; number of conflicting assignments; degree of deviation from a consistent quad structure.
2. Throughline Completeness and Balance – Are all four canonical perspectives (MC, IC, OS, RS) represented and developed? – Metrics: coverage of required signposts for each throughline; relative word count or beat allocation; detection of missing or prematurely resolved throughlines.
3. Inequity and Resolution Coherence – Does the story's resolution actually address the inequity encoded in the Storyform? – Metrics: alignment between final beats and the specified problem/solution pair; semantic similarity between realized and intended resolutions.
4. Thematic Consistency – Are scene-level conflicts and decisions consistent with the encoded Issues and Counterpoints? – Metrics: classifier-based or embedding-based alignment between scene content and the thematic quadrants defined in the Storyform.

5 Beat-to-Beat Structural Alignment – Do specific Storybeats (e.g., Signpost 2 of the Objective Story) actually express the intended type of conflict? – Metrics: mapping generated scenes back into Storypoint space; measuring divergence from expected conflict categories or emotional profiles.

These metrics do not replace human judgment, but they provide formal, repeatable diagnostics grounded in a tested narrative model—something missing in current work.

5.3 Hybrid metrics: integrating narrative and text intelligence

Structural metrics can be paired with:

• Text-level metrics (fluency, grammar, style). • LLM-as-a-judge scores (engagement, surprise) like those used in ASE-style frameworks.

A Dramatica/NCP-based evaluation pipeline might:

1 Take a generated story (from any system).

2 Use Narrova/Subtxt to infer or refine a compatible Storyform.

3 Compute structural metrics (validity, completeness, thematic consistency).

4 Compute text-level and LLM-judge metrics.

5 Combine them into a multi-dimensional evaluation profile.

This yields a richer picture: a story might be textually polished but structurally incoherent, or structurally sound but stylistically weak. Interventions can then be targeted accordingly: revise the Storyform vs revise the prose.

5.4 Benchmarking with NCP

Because NCP encodes Storyforms as portable data, it can underpin story benchmarks that finally meet the survey's call for standardized evaluation:

• Gold-standard NCP instances for canonical stories (e.g., films, novels, games) with verified Storyforms. • Generation benchmarks: – Given an NCP Storyform, generate a telling and evaluate structural fidelity plus text quality. – Given a telling, infer an NCP Storyform and evaluate reconstruction accuracy. • Interactive benchmarks: – Measure how well a system maintains Storyform coherence under player/reader interventions.

These benchmarks are independent of any particular LLM, tool, or UX—a direct answer to the survey's call for reproducible, model-agnostic evaluation.

6. Comparative Analysis: Dramatica vs Surveyed LLM Systems

6.1 Mapping the survey's taxonomy onto the Dramatica platform

Using the survey's categories, we can position the Dramatica platform as follows:

• Independent Story Generation – Systems like DOME, MoPS, SWAG, and COLLABSTORY attempt long-form or premise-based story generation with scaffolding (outlines, action lists, multi-agents). – A Dramatica/NCP-based generator would treat the Storyform as a program: once defined, it determines

the possible narratives; LLMs then render those possibilities in text.

• Author Assistance – Tools like STORYVERSE, CHARACTERMEET, MATHEMYTHS, and MEAT support specific aspects of the process (adaptive plots, character exploration, educational stories). – Narrova plus Subtxt provide a full-stack authoring environment, from premise and structure to beat-level expression, all anchored in a single Storyform.

Functionally, you can reproduce almost any interaction pattern in the survey using the Dramatica platform—but with narrative semantics baked into the core, not bolted on afterward.

6.2 Narrative coherence and emotional depth

Surveyed systems tackle coherence and emotion with:

• Emotion/action labels. • Knowledge-graph-based temporal consistency checks. • Feedback loops over action labels or critic personas.

These are important but limited. They operate on short windows (sentence or paragraph level) and shallow notions of emotion (category tags without full psychological journeys).

Dramatica's Storyform, as exposed via NCP and surfaced in Subtxt/Narrova, models coherence at the level of entire story arcs, not just local transitions. Emotional depth is understood as the product of how perspectives process an inequity over time, not just the presence of emotion words.

When paired with LLMs, this lets the Dramatica platform:

• Detect when a draft violates its own Storyform. • Suggest structurally appropriate revisions. • Evaluate emotional trajectories relative to intended dynamics (e.g., change vs steadfast Main Character).

6.3 Multi-agent systems: arbitrary roles vs narrative roles

The survey's multi-agent systems use roles such as:

• "Story writer" vs "action discriminator" (SWAG). • "Critic personas" for creativity (CRITICS). • Turn-taking agents in COLLABSTORY.

These roles are engineering roles, not narrative roles.

Narrova defines agents by narrative function: Storyforming, Story Encoding, Storyweaving, Story Reception and other specialized helpers. And because every agent has access to a shared Storyform via NCP, they collaborate on the same narrative intent and structural ground truth.

This is a fundamentally different kind of multi-agent system: not many LLMs improvising together, but many narrative specialists coordinating around the same story brain—Dramatica.

7. Ecosystem and Interoperability: Why All Roads Lead to Dramatica

7.1 NCP as cross-agent backbone

NCP is deliberately:

• Open-source and community-driven, inviting adoption by tools across film, games, literature, and AI storytelling. • Model-agnostic, able to drive or constrain any LLM or multi-agent system.

As more systems seek:

• A common narrative format, • Reliable cross-tool story coherence, • Guardrails for generative workflows,

they converge on the need for something like NCP. And NCP's semantics are explicitly those of Dramatica's Storyform.

So whether researchers start from multimodal storytelling, interactive game design, educational story tools, or author-assistance systems, any move toward a structural narrative standard leads inevitably toward the same conceptual space that Dramatica has explored and validated for decades. NCP simply turns that space into a shared protocol.

7.2 Integration across media and agents

In a mature Dramatica/NCP ecosystem:

• Subtxt is where authors sculpt and refine Storyforms and Storybeats. • Narrova orchestrates multi-agent narrative work—analysis, drafting, revision—on top of those Storyforms. • External tools (game engines, visual storytelling systems, music scoring AIs, educational platforms) consume NCP to ensure their outputs remain aligned with the same underlying narrative intent.

This answers several of the survey's "future work" directions at once:

• Inference-time constraints → NCP-specified guardrails during decoding. • Multimodal coherence → cross-modal agents all reading from the same Storyform. • Benchmarking and metrics → NCP-based evaluation pipelines and shared Storyform-driven datasets.

7.3 Adoption paths: when practitioners choose narrative intelligence

Practitioners can approach this in stages:

1. Use Narrova as a smarter, narrative-aware co-author on top of existing workflows.
2. Adopt Subtxt to formalize story intent into Storyforms and Storybeats.
3. Gradually encode those Storyforms into NCP to make them portable across tools and agents.

But no matter where they start—prompt engineering, multi-agent orchestration, multimodal pipelines—once they seek stable structure, reusable evaluation, and cross-tool coherence, they arrive at the same conclusion: you need a formal narrative model. And the most fully-articulated, tooling-ready model available today is Dramatica, operationalized through Subtxt, Narrova, and NCP.

8. Conclusion

The EMNLP survey on LLMs for story generation documents an impressive variety of systems and techniques, but also exposes a common limitation: virtually all current work treats LLMs as engines of text intelligence and then layers prompts, templates, and agents on top to approximate narrative structure.

By contrast, the Dramatica narrative platform—anchored in Storyform theory, implemented through Subtxt, operationalized in Narrova, and opened to the world via NCP—starts from narrative intelligence:

• Stories are not just sequences of words; they are structured explorations of inequity. • Quality is not just fluency; it is structural coherence, thematic clarity, and emotional completeness. • Evaluation is not

just automated metrics or LLM preferences; it is diagnostics against an explicit narrative model.

NCP offers the open-standard bridge for the broader community. Any system can adopt it. But in doing so, it implicitly accepts Dramatica's underlying narrative semantics—because that is what NCP encodes.

In that sense, the field is already moving in Dramatica's direction: calling for structure, metrics, and standards that Dramatica, Subtxt, Narrova, and NCP already embody. All that remains is for researchers and builders to decide when they are ready to step beyond text intelligence and embrace narrative intelligence as the foundation of serious story AI.