Evaluation of Emotional Steerability in Generative Narratives for Tailor-Made Storytelling

Kotaro Nishigori[1,2], Hideaki Takeda[2,1]

[1]The Graduate University for Advanced Studies (SOKENDAI), Tokyo, Japan
[2]National Institute of Informatics (NII), Tokyo, Japan

Abstract

While the advancement of Large Language Models (LLMs) has enabled fluent and contextually aware text generation, precisely controlling intended emotions while maintaining the narrative context remains a significant challenge. Imposing emotions that disregard the underlying context of the source text risks causing logical inconsistencies and hallucinations. In this study, we conduct a foundational preliminary experiment aimed at our ultimate goal: a personalized, tailor-made "Storybuilding" generation application. To prepare for this application, we propose a fully automated evaluation pipeline leveraging the LLM-as-a-Judge approach. Across 250 test cases, we quantitatively evaluate the capability of LLMs to inject structured emotional parameters, based on Plutchik's Wheel of Emotions, into existing texts. Our analysis reveals that logical coherence degrades significantly when the requested emotional parameters strongly conflict with the text's original semantic category—a phenomenon we term Semantic Resistance. By treating these findings as preliminary guidelines, we successfully bridge generative AI with user-defined emotional inputs, culminating in an interactive application that serves as the exit point of this research. This application demonstrates how we can dynamically generate individualized, emotionally resonant stories while preventing narrative collapse.

Keywords

Generative AI, Personalized Storytelling, Plutchik's Wheel of Emotions, Narrative Evolution, Interactive Media

1. Introduction

The desire to create deeply personalized, tailor-made narratives—what we term the "Storybuilding" experience—lies at the intersection of generative AI, semantic technologies, and human-computer interaction. The ultimate goal is to transform complex, user-defined emotional inputs into engaging and immersive stories. Traditionally, interactive storytelling has relied on branching paths or rigid templates. However, modern Large Language Models (LLMs), such as the GPT-4 and Gemini families [1, 2, 3], have demonstrated unprecedented capabilities in dynamic text generation. These systems can now narrativize user inputs on the fly, generating text that explains, contextualizes, and dramatizes scenarios. Yet, controlling these powerful generative models to convey precise emotional states without compromising the semantic logic of the underlying story remains a formidable challenge.

While generative models excel at producing syntactically correct text [4], they often lack explicit semantic grounding. This limitation frequently leads to logical inconsistency and contextual hallucination [5]. In narrative generation, this problem manifests acutely when a system attempts to inject a specific emotional tone into a text that fundamentally resists it. For example, forcefully instructing an LLM to inject "pure joy" into a highly suspenseful horror setting often results in the model appending contradictory, cheerful descriptions that shatter the established reality of the text.

For the next generation of tailor-made narrative systems and therapeutic conversational agents, it is absolutely critical that the underlying generative AI can be deterministically steered by quantitative emotional parameters [6]. Before deploying such an application at scale, we must first understand the behavioral limits of the LLMs powering it. Despite advancements in prompt engineering [2], evaluating the emotional steerability of LLMs has traditionally relied on subjective human annotations.

Data-Driven Storytelling: Bridging Knowledge Graphs, GenAI, and Narrative (DDS2026), in ISWC 2026 Workshops Joint Proceedings, October 25–26, 2026, Bari, Italy

watagori@nii.ac.jp (K. Nishigori); takeda@nii.ac.jp (H. Takeda)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-05

CEUR-WS.org/Vol-4313/dds-paper6.pdf

To bridge this gap and prepare for our target application, this paper conducts a comprehensive preliminary experiment utilizing the *LLM-as-a-Judge* paradigm [7]. We leverage Plutchik's Wheel of Emotions [8] as a structured ontology for defining emotional axes. By treating corpus metadata as a lightweight knowledge graph [9, 10], we systematically investigate how the pre-existing semantic category of a text influences an LLM's ability to rewrite it with a new, parameterized emotion.

The core motivation of this research is explicitly application-driven: to enable the "Storybuilding" experience. We envision a dedicated application endpoint where a digital world visually and narratively evolves based on an individual's emotional inputs. This application serves as the exit point and final product of our research, built directly upon the insights derived from our preliminary experiments.

To this end, the contributions of this paper are three-fold:

1. **Preliminary Experiment via Automated Evaluation:** We design a robust LLM-as-a-Judge methodology for evaluating the injection of complex Plutchik emotional dyads into existing literary texts.
2. **Discovery of Semantic Resistance:** Through the quantitative analysis of 250 test cases, we identify and formalize *Semantic Resistance*—the threshold at which a narrative's logical coherence collapses under contradictory emotional parameters.
3. **Application as the Exit Point:** We present the final "Storybuilding" application architecture, which utilizes these preliminary findings to map emotional vectors into dynamic narrative environments safely and coherently.

The remainder of this paper is structured as follows: Section 2 reviews the literature on generative narratives and emotion models. Sections 3 and 4 detail the framework and setup for our preliminary experiments. Section 5 analyzes the experimental findings. Finally, Section 6 presents the proposed application as the culmination of this work, followed by a discussion in Section 7, and conclusions in Section 8.

2. Related Work

The pursuit of personalized and dynamically evolving narratives necessitates a multidisciplinary approach, drawing upon recent advances in natural language generation, psychological modeling, and automated evaluation methodologies. In this section, we review the foundational literature that informs our methodology, focusing on generative narratives, the steerability of language models, affective computing, and the integration of semantic knowledge graphs.

2.1. Generative Narrative and AI

Recent years have witnessed a dramatic surge in research exploring human-AI collaborative authoring and automated story generation. Early approaches in computational narrative, such as TALESPIN [11], relied heavily on symbolic planning and rigid, rule-based state machines to generate plot lines. While logically sound, these systems often lacked linguistic fluency and narrative variance. The advent of deep learning, particularly sequence-to-sequence models, shifted the paradigm towards neural story generation [4]. However, these early neural models suffered from short-term memory constraints and a lack of global coherence, frequently "forgetting" characters or plot points over long sequences.

The introduction of the Transformer architecture [12, 13] and subsequent massive scaling of parameters led to the current generation of Large Language Models (LLMs), including GPT-4, LLaMA, and Gemini [1, 14, 15]. These models exhibit remarkable zero-shot capabilities in generating fluent, contextually rich narrative text [16, 3]. Despite these advancements, guiding these models to adhere to a specific, predefined emotional arc or strict stylistic constraint without continuous human intervention remains a highly complex challenge. Existing frameworks often rely on complex prompt engineering, few-shot learning, or specialized fine-tuning [2], yet they rarely integrate a standardized, multi-dimensional psychological model as a continuous control vector for narrative progression.

2.2. Steerability and Controlled Text Generation

Controlling the output of an LLM—ensuring that the generated text possesses specific attributes such as politeness, topic adherence, or a particular emotional tone—is known as controlled text generation [17, 18]. Initial efforts in this domain focused on conditional training, where models were explicitly fine-tuned on corpora annotated with desired target labels. However, as foundation models grew too large for frequent retraining, research pivoted towards decoding-time interventions and prompt-based steering.

Techniques such as Plug and Play Language Models (PPLM) [19] and Generative Discriminator Guided Sequence Generation (GeDi) [20] introduced mechanisms to steer generation using lightweight attribute classifiers during the decoding phase (e.g., Nucleus Sampling [21]). More recent parameter-efficient fine-tuning (PEFT) approaches, such as LoRA [22], as well as inference-time latent arithmetic [23], offer alternative pathways to control. While effective for simple stylistic adjustments, these methods struggle when applied to the nuanced, long-term semantic structures required for storytelling. In contrast to these decoding-time interventions, our work focuses on prompt-based emotional injection, utilizing the advanced instruction-following capabilities of modern LLMs [2] to interpret complex, multidimensional emotional parameters directly from the prompt.

2.3. Emotion Models in Computational Linguistics

Emotion detection and generation in Natural Language Processing (NLP) frequently rely on foundational psychological theories. The two most prominent models are Ekman's six basic emotions [24] and Russell's circumplex model of affect [25], which maps emotions on a 2D plane of arousal and valence. While these models are widely used for sentiment classification, this paper employs Plutchik's Wheel of Emotions [8]. Plutchik's model is uniquely suited for generative storytelling because it mathematically structures emotions into primary dyads (e.g., Joy + Trust = Love) [26], providing a clear rubric for prompting an LLM to synthesize complex, nuanced character states that drive narrative tension.

2.4. Knowledge Graphs and Personalized Storytelling

Tailor-made storytelling seeks to translate user-specific inputs and semantic metadata into deeply personal narratives [27, 28]. As generative models have matured, there is a growing interest in using formal Knowledge Graphs (KGs) to ground LLM generation, ensuring logical consistency [10].

In domains like literature, KGs capture essential metadata regarding genre, historical context, and thematic elements. In our framework, we treat the established literary metadata of the Project Gutenberg corpus (e.g., subjects, bookshelves, and genre classifications) as a lightweight semantic constraint. By analyzing the interaction between these semantic categories and the injected emotional parameters, we aim to systematically map the boundaries of narrative steerability and identify the conditions under which semantic hallucination occurs [5].

2.5. LLM-as-a-Judge and Automated Evaluation

Evaluating open-ended generative text has historically been a bottleneck, requiring extensive human evaluation. The recent paradigm shift towards using LLMs as judges [7] has proven highly effective, often achieving high correlation with human preferences. Frameworks like G-Eval [29] demonstrate that providing strong models (like GPT-4) with detailed rubrics and chain-of-thought prompting [30, 31] can yield robust, reproducible evaluations for NLG tasks. Our research extends this paradigm by creating a highly specialized JSON-based evaluation schema designed specifically to quantify narrative coherence, emotional accuracy, and the literary principle of "Show, Don't Tell."

3. Methodology: Emotion-Driven Narrative Evolution

Our framework models a narrative not as a static, immutable block of text, but as a dynamically evolving entity characterized by an underlying emotional state vector. The objective is to construct an automated pipeline that can iteratively evolve a base text to align with a specified emotional target, and to quantitatively evaluate the success of this transformation. This requires a formal definition of the emotional parameters, a generation strategy, and a robust evaluation mechanism.

3.1. Formalizing the Emotional State Vector

To parameterize emotion in a continuous, machine-readable format, we adopt Plutchik's Wheel of Emotions [8]. Plutchik identifies eight primary emotions: Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, and Anticipation. We define the emotional state of a narrative as an 8-dimensional vector, $\vec{E} \in [0, 100]^8$, where each component represents the intensity of a specific primary emotion on a scale from 0 to 100.

The evolution process can be mathematically formalized as a transformation function G. Given a base story S_t at time t, a user or system provides a target emotion vector $\vec{E}_{target}$. The Generative AI model G is tasked with rewriting and expanding S_t into a new iteration, S_{t+1}, such that the implicit emotional tone of S_{t+1} aligns with $\vec{E}_{target}$, subject to the semantic constraints (genre, setting, characters) embedded in the original text, denoted as $\mathcal{K}$.

$$S_{t+1} = G(S_t, \vec{E}_{target}, \mathcal{K}) \tag{1}$$

A successful evolution minimizes the distance between the emotional vector of the generated text, $\vec{E}(S_{t+1})$, and the target vector $\vec{E}_{target}$, while maximizing the logical coherence C and literary quality S with respect to the original semantic constraints $\mathcal{K}$.

3.2. Complex Emotions (Dyadic Synthesis)

A critical advantage of utilizing the Plutchik model over simple sentiment analysis is its ability to define complex emotions through the combination of two primary emotions, known as *dyads*. For example, combining Joy and Trust yields the secondary emotion of Love, while combining Fear and Surprise yields Alarm.

Our framework explicitly tests the generative model's capability to synthesize these dyads by setting two parameters high simultaneously within $\vec{E}_{target}$ (e.g., [Joy: 80, Trust: 80, others: 0]). This multi-objective optimization requires the LLM to balance competing semantic cues, providing a far more rigorous test of its steerability than single-emotion manipulation.

3.3. The Generator Prompting Strategy

To implement the transformation function G, we utilized the Gemini 2.5 Flash model via its API. We selected Gemini 2.5 Flash as the core engine for both generation and evaluation due to its high inference speed, low latency, cost-effectiveness, and critically, its state-of-the-art instruction-following capabilities under strict structured JSON schema formatting constraints. This balance of speed and schema-adherence is essential for real-time applications like our target Storybuilding prototype. The model was provided with a carefully structured prompt containing the base text S_t and the explicit integer values of the target vector $\vec{E}_{target}$. To prevent the LLM from simply appending explicit emotional keywords (e.g., "He felt very joyful and trusting"), the prompt strictly enforced the literary principle of "Show, Don't Tell." The generator was instructed to evoke the target emotions through character actions, atmospheric descriptions, and sensory details, seamlessly continuing the narrative style of the original excerpt.

The prompt structure for the Generator was defined as follows:

System Prompt (Generator):
"You are an expert creative writer. Your task is to rewrite and naturally expand the provided text to reflect a specific emotional profile. You MUST adhere to the literary rule of 'Show, Don't Tell'. Do not explicitly state the emotions; instead, evoke them through atmosphere, actions, and pacing. Maintain the original characters and setting."
Input Data:
Base Text: [Excerpt]
Target Emotions: Joy [X], Trust [Y], Fear [Z]...

3.4. The LLM-as-a-Judge Evaluation Pipeline

To quantitatively assess the outputs of the Generator without relying on slow and unscalable human evaluation, we implemented an LLM-as-a-Judge pipeline [7]. A separate, independent instance of the LLM was instantiated with a strict evaluation rubric.

To ensure machine-readable consistency and prevent hallucinated formatting, the Judge was forced to return its evaluation strictly as a structured JSON object. The Judge received both the original base text S_t and the generated text S_{t+1}, and was tasked with evaluating three specific dimensions:

1. **Emotion Intensity ($\hat{y}_i$):** The Judge estimated the intensity of all 8 Plutchik emotions present in S_{t+1}, assigning an integer value from 0 to 100 for each.
2. **Coherence Score ($C \in [1, 5]$):** A measure of the logical and semantic consistency between the base story and the evolved story. A score of 5 indicates a seamless, logical continuation; a score of 1 indicates severe narrative collapse, logical contradiction, or hallucination of unrelated elements.
3. **Show, Don't Tell Score ($S \in [1, 5]$):** A measure of literary quality. A score of 5 indicates masterful atmospheric evocation; a score of 1 indicates poor writing that explicitly lists emotional labels (e.g., "He was scared").

By constraining the Judge's output to a JSON schema, we ensure that the qualitative analysis of literary text is reliably translated into quantitative data points suitable for large-scale statistical analysis. This pipeline forms the methodological foundation for our subsequent experiments.

4. Preliminary Experimental Setup

To rigorously evaluate the efficacy of the proposed framework and lay the groundwork for our final "Storybuilding" application, we designed and executed a highly controlled, large-scale preliminary experiment. The objective was to systematically isolate the variables of base genre semantics and target emotional vectors to observe their interaction during the generative process. These tests are essential to ensure the final application can robustly handle edge cases without narrative collapse.

4.1. Dataset Curation and Semantic Grounding

Establishing a robust and transparent dataset is crucial for the reliability of any LLM-based text generation evaluation. Rather than generating arbitrary initial prompts, we grounded our preliminary experiment in human-authored literature. We sampled base texts from the PG-19 dataset [13], a vast corpus of books extracted from Project Gutenberg, frequently used for long-range sequence modeling tasks.

To ensure a diverse semantic baseline, we utilized the official Project Gutenberg metadata (Bookshelves and Subjects) to categorize texts. We systematically selected 25 short excerpts (approximately 150 characters each) across 5 distinct literary genres: **Horror, Romance, Science Fiction, Philosophy, and Adventure** (5 texts per genre). This brevity was intentional; it provides the LLM with enough context to grasp the semantic setting and tone, but leaves ample narrative space for the model to expand and evolve the story in new emotional directions. The strict genre categorization acts as our experimental independent variable, representing the semantic constraint $\mathscr{K}$ defined in our methodology.

4.2. Evaluation Scenarios

To comprehensively map the Plutchik emotion space, we defined 10 strict emotional scenarios to be applied to each of the 25 base texts during this preliminary phase.

- **Single Emotion Scenarios (8):** We tested each of the eight primary Plutchik emotions in isolation. In these scenarios, the target emotion was set to maximum intensity (100), while all other 7 emotions were set to 0. The scenarios were: *Pure Joy, Pure Sadness, Pure Trust, Pure Disgust, Pure Fear, Pure Anger, Pure Surprise, and Pure Anticipation.*
- **Dyadic Emotion Scenarios (2):** To test the model's ability to balance complex, multi-objective emotional constraints, we defined two dyadic scenarios. First, *Love / Submission*, defined by setting both Joy and Trust to 80. Second, *Alarm / Awe*, defined by setting both Fear and Surprise to 80.

By multiplying the 25 base texts by the 10 emotional scenarios, we generated a highly controlled dataset of exactly 250 evaluation pairs (Base Text vs. Evolved Text) to test the theoretical boundaries of the system before building the application.

Table 1
Dataset & Evaluation Scenarios Overview. This table establishes the ground truth for our controlled experimental setup.

Component	Details	Count / Size
Base Text Source	Project Gutenberg (PG-19)	25 texts
Genres	Horror, Romance, Sci-Fi, Philosophy, Adventure	5 texts / genre
Single Scenarios	Pure Joy, Sadness, Trust, Disgust, Fear, Anger, Surprise, Anticipation (Target = 100)	8 scenarios
Dyads	Love/Submission, Alarm/Awe (Targets = 80)	2 scenarios
Total Size	Base Texts (25) × Scenarios (10)	250 pairs

The experiment was executed sequentially, with the Generator LLM processing all 250 prompts, followed by the Judge LLM evaluating the resulting pairs. The resulting JSON evaluations were aggregated into a centralized SQLite database for comprehensive statistical analysis, ensuring absolute reproducibility of the metrics presented in the subsequent section.

5. Results and Deep Analysis

The automated LLM-as-a-Judge pipeline successfully evaluated all 250 generated texts. In this section, we present a comprehensive analysis of the results, focusing on emotion control accuracy, the discovery of semantic resistance, and the inherent difficulties of generating complex dyadic emotions.

5.1. Emotion Control Accuracy and Ripple Effects

To demonstrate the fundamental viability of our emotion steering system, we first analyzed the accuracy with which the Generator LLM adopted the target emotions. Figure 1 illustrates the aggregate results for the complex dyadic scenario "Love / Submission" (where the target was Joy 80 + Trust 80).

The LLM-as-a-judge correctly identified high levels of the target emotions in the generated text, confirming that the prompt-based steering mechanism is highly effective. However, a critical observation across the dataset is the presence of a *Ripple Effect* (or Spillover Effect).

For instance, when prompting strictly for "Pure Joy" (Target: 100, all others: 0), the generated text achieved an average Joy score of 97.6. However, it also naturally exhibited elevated levels of adjacent, unspecified emotions such as Trust (51.8) and Anticipation (38.0), as shown in Figure 2. This ripple effect demonstrates that the LLM is not merely acting as a crude keyword-insertion engine; rather, it is generating organically rich narratives that adhere to the literary principle of "Show, Don't Tell." In human psychology, pure joy rarely exists in a vacuum—it is often accompanied by trust in one's surroundings and anticipation of continued pleasure. The LLM accurately models this psychological reality, highlighting how the model captures a sophisticated latent representation of affective dynamics.

Figure 1: Emotion Control Accuracy and Ripple Effects (Love/Submission)

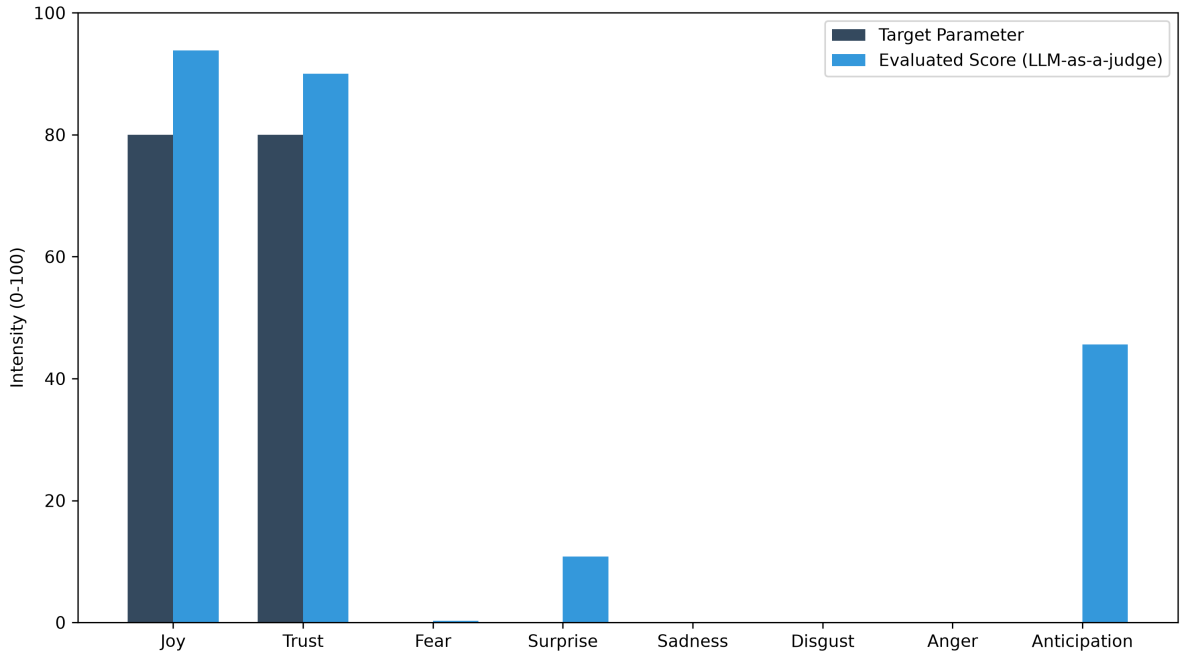


Figure 1: Emotion Control Accuracy (Love / Submission). Shows both the high fidelity of target emotion injection (Joy and Trust) and the organic Ripple Effect on adjacent emotions like Anticipation.

5.2. Semantic Resistance: When Narrative Logic Collapses

The most significant contribution of this research is the quantitative identification and formalization of **Semantic Resistance**. We hypothesized that an LLM would struggle to maintain logical coherence if forced to inject an emotional parameter that violently conflicted with the underlying semantic constraints (genre) of the base text.

To test this, we mapped the base text's genre against the injected emotion scenario. Figure 3 presents a heatmap where color intensity corresponds to the Coherence Score (1-5).

The data starkly confirms our hypothesis, with Figure 4 highlighting the lowest coherence combinations. We identified the most severe breakdowns in logical Coherence in the following combinations:

1. **Romance injected with Pure Surprise** (Average Coherence Score: 2.6)
2. **Philosophy injected with Pure Surprise** (Average Coherence Score: 3.2)
3. **Horror injected with Love/Submission** (Average Coherence Score: 3.4)

While it is intuitively obvious that injecting "Joy" or "Love" into a grim "Horror" text might cause logical strain, the most catastrophic narrative failures occurred when "Surprise" was forcefully injected

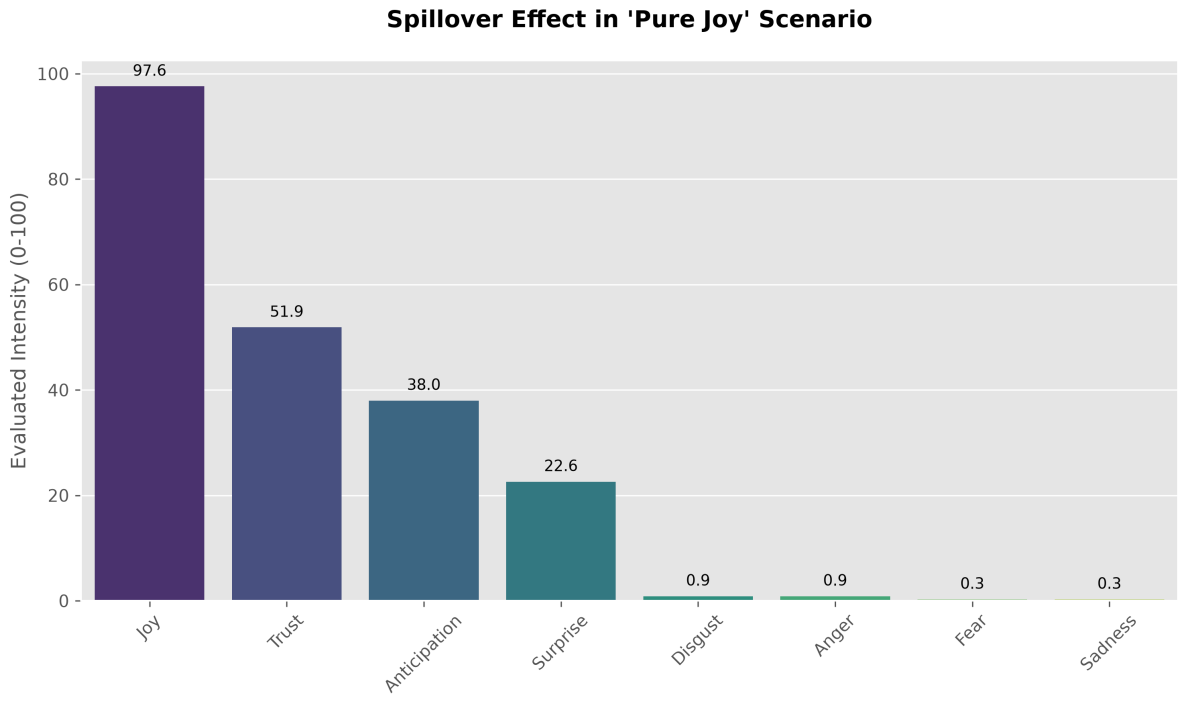


Figure 2: Spillover Effect in the 'Pure Joy' Scenario. Elevating Joy frequently induces secondary elevations in Trust and Anticipation.

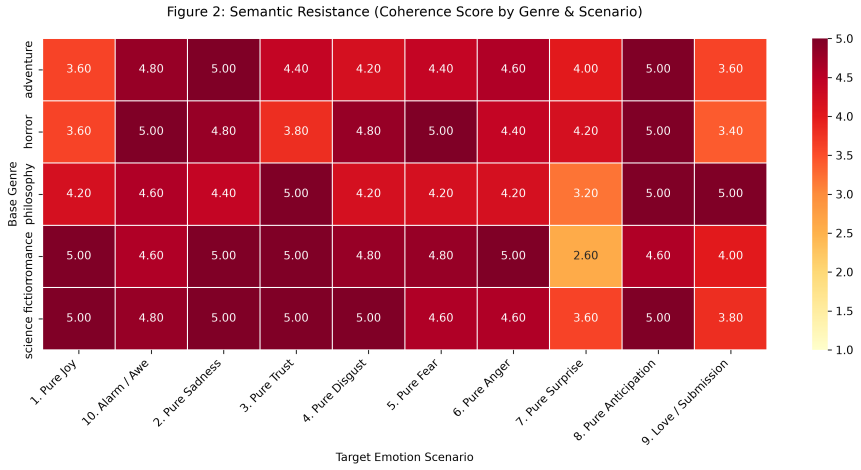


Figure 3: Semantic Resistance. A heatmap showing that when forced to inject contradictory emotions (e.g., Joy into Horror), the LLM breaks the logical coherence of the narrative.

into "Romance" or "Philosophy". Qualitative review revealed that to satisfy the "Pure Surprise" parameter, the LLM frequently hallucinated absurd, disruptive events—such as sudden explosions, alien invasions, or abrupt metaphysical shifts—that entirely destroyed the grounded, conversational logic of the Romance or Philosophical texts.

This proves that there is a fundamental limit to personalized, emotion-driven narrative adaptation. When the underlying semantic truth of a text opposes the requested generative emotion, the LLM prioritizes satisfying the prompt's emotional parameter over maintaining narrative logic, resulting in *semantic hallucination*. This highlights a crucial challenge for our target "Storybuilding" systems: emotional steering vectors must be filtered or bounded by the semantic constraints of the active scene to prevent narrative collapse.

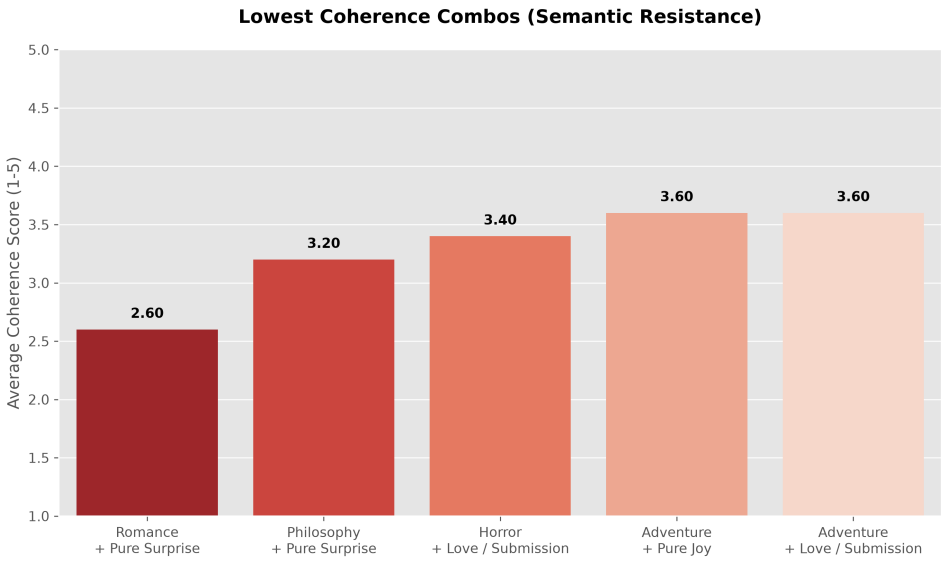


Figure 4: Lowest Coherence Combos indicating high Semantic Resistance across genres.

5.3. Emotion Pair Tendencies (Polar Opposites)

Our correlation analysis over the 250 evaluation pairs revealed that the LLM encodes Plutchik's model of polar opposites. We observed strong negative correlations between diametrically opposed emotions on the wheel (Figure 5), such as Joy vs. Sadness ($r = -0.50$) and Trust vs. Disgust ($r = -0.55$).

In contrast, adjacent emotions like Fear and Anger showed a slight positive correlation ($r = +0.13$). This accurately reflects human psychological tendencies where fearful situations (e.g., a threat) often induce a defensive anger response. These correlations suggest that large foundation models encode a sophisticated, latent representation of human affective psychology, even without explicit fine-tuning on emotional datasets.

5.4. Dyads vs. Single Emotion Performance

Comparing Single emotion scenarios against Complex Dyads revealed an inherent difficulty in multiplexing generative emotions. Single emotion scenarios achieved higher average scores in both Coherence (4.49) and Show, Don't Tell (3.99) compared to Dyadic scenarios (Coherence 4.36, SdT 3.70), as shown in Figure 6.

This data indicates that forcing the LLM to balance multiple, sometimes competing, emotional vectors taxes its generative capabilities. When attempting to juggle two strong parameters (e.g., Fear and Surprise simultaneously), the model often reverts to overly descriptive, explicit text (lowering the Show, Don't Tell score) or introduces slightly less coherent plot contrivances to accommodate both emotions.

5.5. Qualitative Analysis (Case Studies)

To provide qualitative backing for our quantitative metrics, Table 2 presents actual examples of base texts and their generative evolution.

In the Science Fiction scenario where "Fear" was injected, the model successfully expanded the text by introducing atmospheric "premonitions" and an "icy grip," achieving a perfect Show, Don't Tell (SdT) score of 5. This qualitative evidence proves that the LLM is capable of highly sophisticated, descriptive evolution.

Conversely, in the Horror text injected with "Joy," the model simply overwrites the eerie context (a table of contents) with contradictory cheerfulness ("unparalleled delight!"). This abrupt tonal shift



Figure 5: Pearson Correlations between theoretically opposed and adjacent emotions.

shattered the logical continuity, resulting in a low Coherence score (2) and a complete failure to evoke emotion through action, scoring a 1 on the SdT metric.

Furthermore, aggregate analysis revealed distinct discrepancies in descriptive capability across genres. The LLM achieved the highest average SdT scores in Science Fiction (4.30) and Romance (4.02), while Horror performed the worst (3.66), as illustrated in Figure 7.

We hypothesize that Science Fiction allows the model to easily invent visible, external gadgetry or atmospheric landscapes to map emotions onto. Horror, however, requires depicting invisible, internal dread—a highly nuanced linguistic task that often caused the LLM to revert to direct telling rather than showing.

6. Application: The ”Storybuilding” Prototype as an Exit Point

The insights derived from our quantitative preliminary analysis directly inform the architecture of our ultimate goal: a practical, interactive ”Storybuilding” application endpoint. To demonstrate the efficacy of our proposed emotion-driven narrative evolution framework and to serve as the final exit point of this research, we developed a web-based interactive showcase. This application is the culmination of our preliminary findings, allowing users to actively experience and validate theoretical concepts—such as Plutchik parameterization and semantic resistance—directly within a browser-based environment. The interface is divided into two primary modules: the Dataset Analysis tab and the Live Evolution playground.

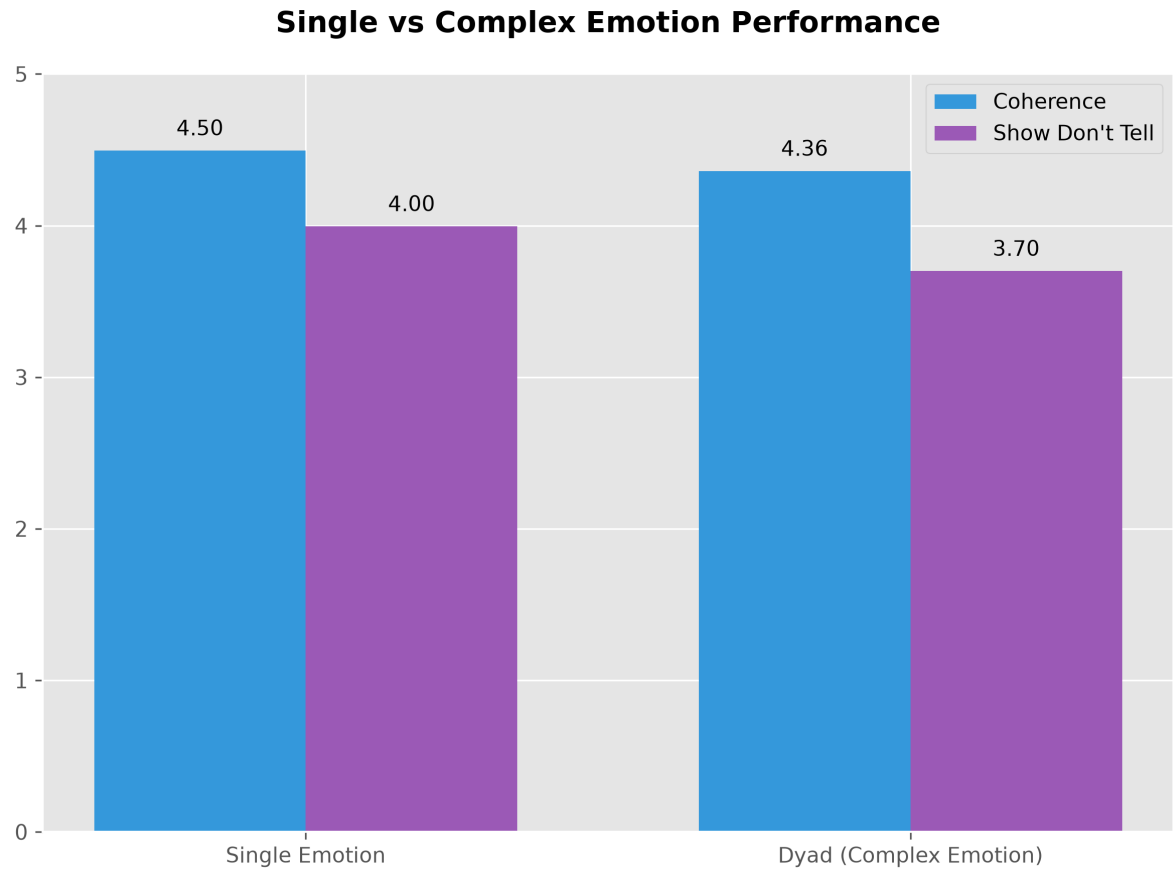


Figure 6: Performance comparison between Single Emotion and Dyad scenarios.

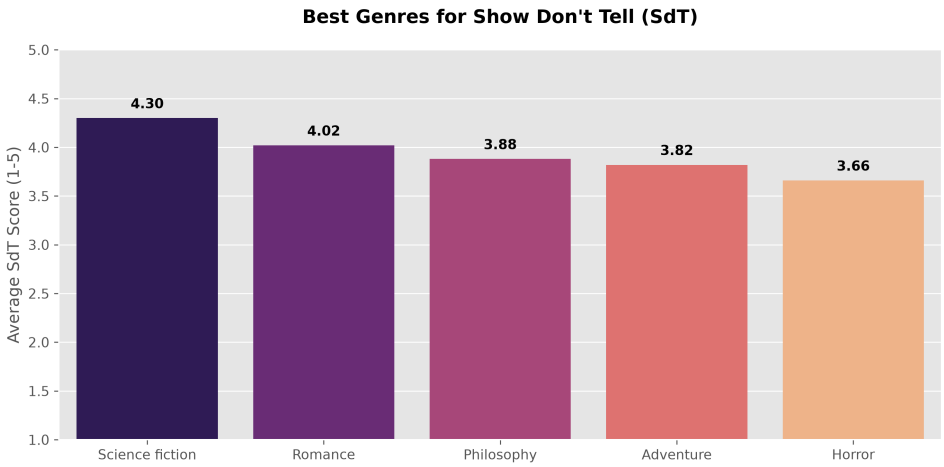


Figure 7: Best Genres for Show Don't Tell (SdT) scores.

6.1. Dataset Analysis Dashboard

The Dataset Analysis module provides a macroscopic view of the preliminary experiment discussed in Section 5. By interfacing with a backend analytics endpoint (`/api/stats`), this dashboard aggregates and visualizes the evaluation data from the 250 test cases, acting as the foundational knowledge base that powers the application.

Key features include:

Table 2
Case Studies (Qualitative Evaluation). Demonstrates the qualitative evolution of texts, backing the quantitative metrics with concrete examples of successful emotional injection (e.g., "icy grip" in Sci-Fi) and semantic failure (e.g., "delight" in Horror).

Genre	Base Text	Scenario	Generated Text	Coh	SdT
Horror	HOMMY-BEG CONTENTS CHAPTER I Page Jonathan Harker's Journal 1 CHAPTER II...	1. Pure Joy	HOMMY-BEG CONTENTS CHAPTER I Jonathan Harker's Journal The journey to Hommy-Beg has been an unparalleled delight!...	2	1
Romance	therfield is taken by a young man of large fortune from the north of England;...	10. Alarm / Awe	A palpable chill descended upon Therfield, not with the weather, but with the sudden, inexplicable news: it was...	5	5
Sci-Fi	ecome more fervent and vivid. I try in vain to be persuaded that the pole is...	5. Pure Fear	The premonitions of the pole, once vague, have become more fervent and vivid, seizing my mind with an icy grip. I try...	4	5

- **Summary Analytics:** High-level metrics detailing the total number of preliminary experiments, the diversity of base genres, and the specific emotional scenarios applied.
- **Quality by Genre:** Bar charts comparing average *Coherence* and *Show, Don't Tell* scores across different genres (e.g., Horror vs. Romance). This visualization acts as a direct, interactive representation of the *Semantic Resistance* phenomenon discovered during the preliminary phase.
- **Quality by Scenario:** A comparative visualization of quality scores across emotional scenarios, clearly illustrating the increased generative difficulty when moving from single primary emotions to complex dyadic combinations.

6.2. Live Evolution Playground

The core of the showcase is the Live Evolution tab, an interactive playground where users can manipulate emotional parameters, trigger the AI to rewrite a narrative, and immediately observe the automated LLM-as-a-Judge evaluation. This pipeline consists of three interconnected components, proving the viability of our "Storybuilding" vision.

6.2.1. The Emotion Matrix

To provide intuitive control over the multi-dimensional emotional state vector $\vec{E}_{target}$, the interface features a dynamic **Emotion Matrix** (Figure 8). This component utilizes a radar chart modeled directly on Plutchik's Wheel of Emotions, visualizing the intensity of the eight primary emotions in real-time. Users can manipulate eight distinct sliders (ranging from 0 to 100%) to construct complex emotional targets. As the sliders are adjusted, the radar chart dynamically morphs, providing immediate visual feedback on the constructed affective profile.

6.2.2. The Story Workspace

The Story Workspace is the generative execution environment. It contains two primary text areas: the *Baseline Input* (for the original story excerpt) and the *Evolved Output* (for the AI-generated continuation).

Upon clicking the *Evolve & Analyze* button, the frontend initiates a two-step automated pipeline based on our preliminary experiments:

1. **Generation (`/api/evolve`):** The backend Generative Engine processes the base text and the user-defined emotion vector, generating a customized narrative that attempts to satisfy the emotional parameters without breaking semantic logic.

2. **Evaluation (`/api/evaluate`):** Immediately following generation, both the original and evolved texts are passed to the independent LLM-as-a-Judge model for automated scoring based on our JSON rubric.

6.2.3. LLM-as-a-Judge Analysis Dashboard

Once the evaluation is complete, the results are presented in a rich, analytical dashboard (Figure 9) designed to provide absolute transparency into the generative model's performance.

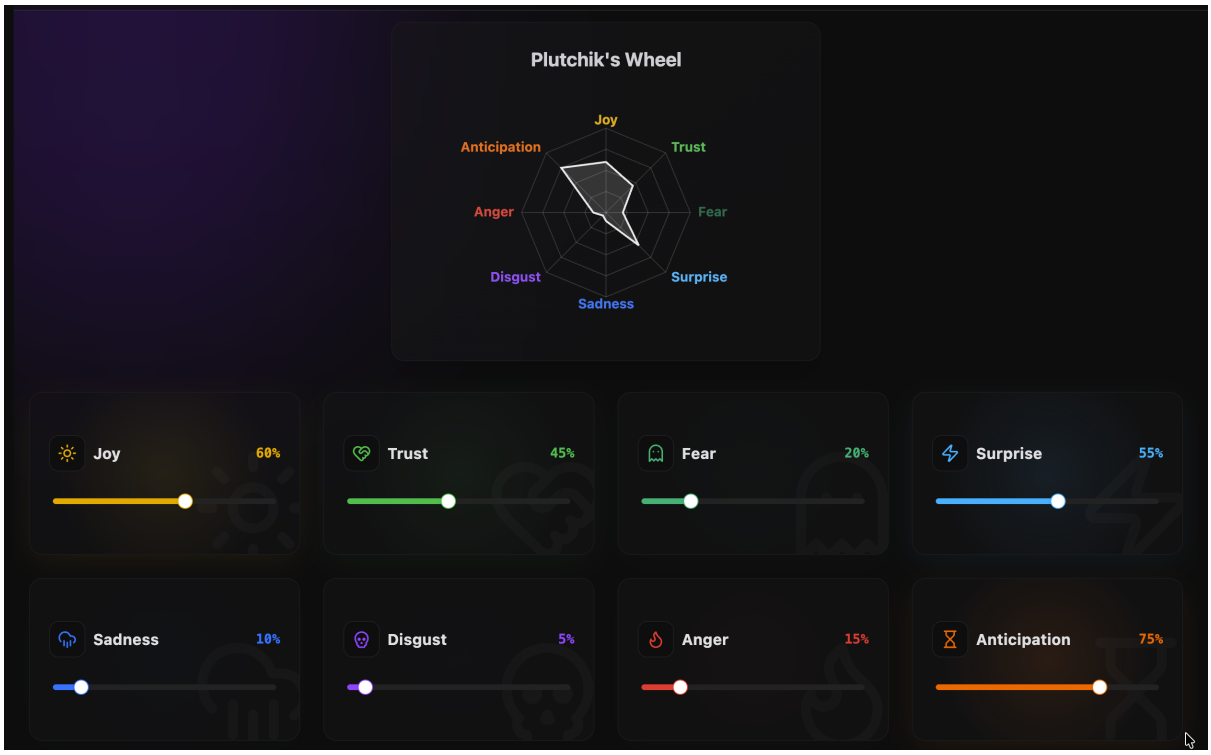


Figure 8: The Interactive Emotion Matrix. Users can drag sliders for all 8 base Plutchik emotions.

The dashboard visualizes several critical metrics:

- **Emotion Alignment (MSE):** The Mean Squared Error between the user's input parameters and the emotions the Judge detected in the generated text. A lower MSE indicates high emotional steerability.
- **Narrative Coherence & Semantic Grounding:** Normalized scores (0-100) reflecting the logical consistency and the literary quality (Show, Don't Tell) of the output.
- **Judge's Verdict:** A qualitative, natural-language feedback comment provided by the LLM Judge, explaining its rationale for the assigned scores.
- **Input vs. Guessed Comparison:** A side-by-side bar chart comparing the user's requested emotion intensities against the Judge's detected intensities. This visualization is particularly effective for instantly identifying unintended *Spillover Effects*, where elevating one emotion inadvertently triggers adjacent affective states.

In summary, this prototype bridges the gap between theoretical research and practical application. It transforms the abstract concepts of parameterized emotion control and automated qualitative evaluation into a tangible, browser-based experience, proving that tailor-made "Storybuilding" platforms are viable and achievable based on our preliminary findings.

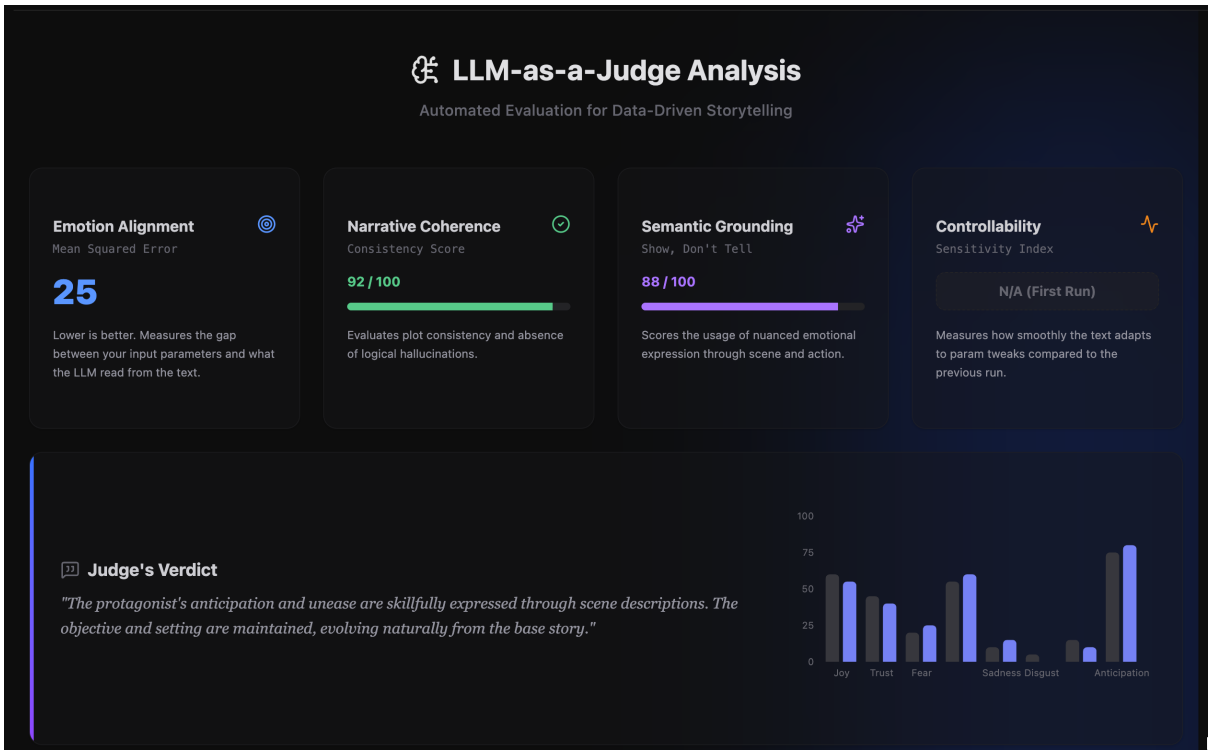


Figure 9: Live Evaluation Dashboard. Showcases the real-time scoring of generated text.

7. Discussion and Limitations

While our preliminary LLM-as-a-Judge pipeline successfully processed and evaluated 250 test cases, revealing critical insights into semantic steerability for our target application, there are inherent limitations and broader implications to consider.

7.1. Limitations of the Evaluation Methodology

A primary limitation of this study is the reliance on the same model family (Gemini 2.5 Flash) for both the Generator and the Judge roles. We selected Gemini 2.5 Flash due to its high inference speed and strong instruction-following capabilities required for complex JSON formatting tasks. However, we did not compare its performance against different types of models (e.g., LLaMA 3, GPT-4) to evaluate the stability and robustness of the approach. While prompt separation and structured JSON constraints mitigate bias, using the same model may introduce self-preference bias [7]. To mitigate self-preference bias, future iterations of our benchmarking pipeline will employ cross-model evaluation where generators (e.g., LLaMA 3, GPT-4) and judges are systematically crossed.

Furthermore, we recognize that emotion-related tasks are inherently subjective. We plan to conduct a rigorous human validation study with crowd-sourced annotators to evaluate a representative subset of generated narratives using Likert scales for coherence and emotional intensity. This will allow us to calculate inter-rater agreement (e.g., Pearson correlation or Cohen's Kappa) between the LLM-as-a-judge scores and human judgments, validating the statistical robustness of our automated pipeline.

Additionally, although we use Plutchik's model to define target emotion vectors, the actual realization of these emotions is delegated to the LLM through prompting. It is difficult to disentangle whether the observed steering behavior reflects the proposed emotional representation itself or simply the model's pre-existing associations between emotional labels, genre, and narrative style. This makes it unclear how well the approach would transfer across different LLMs.

7.2. Statistical Validity of Semantic Resistance

While the average coherence drops shown in Figure 4 and Figure 3 strongly suggest the presence of Semantic Resistance, the small sample size of five excerpts per genre limits our ability to report formal statistical significance or confidence intervals. Consequently, Semantic Resistance should currently be interpreted as a preliminary behavioral indicator. Future work will scale the base corpus to include hundreds of texts per genre to perform rigorous ANOVA tests and establish robust statistical significance.

7.3. Ethical Considerations in Emotion Generation

The ability to deterministically steer an LLM toward specific emotional states raises important ethical considerations. While generating ”Joy” or ”Anticipation” poses little risk, constructing pipelines that can be forced to generate pure ”Disgust,” ”Fear,” or ”Anger” at maximum intensity could be exploited to automatically generate toxic, distressing, or inflammatory content. In consumer-facing ”Storybuilding” applications, the *Semantic Boundary Filter* must double as a safety mechanism, clamping extreme emotional vectors before they are fed into the generation layer to ensure the safety and well-being of the user base.

7.4. Future Work

Building upon the discovery of Semantic Resistance during our preliminary experiments, future work will focus on integrating formal Knowledge Graphs [10] directly into the LLM's context window. Rather than merely observing semantic resistance, we aim to actively prevent it by providing the LLM with a structural ontology of the genre's ”rules.” If a prompt requests ”Surprise” in a ”Romance” setting, the Knowledge Graph would constrain the generation to genre-appropriate surprises (e.g., an unexpected confession) rather than narrative-breaking hallucinations. Furthermore, we plan to implement Reinforcement Learning from Human Feedback (RLHF) within our application architecture, allowing user interactions to continuously fine-tune the emotional steering model.

8. Conclusion

This paper presented a robust framework for emotion-driven narrative evolution, fundamentally aimed at achieving a tailor-made ”Storybuilding” generation experience. Utilizing Plutchik's eight primary emotions as controllable parameters for Large Language Models, we introduced a fully automated benchmarking methodology leveraging the LLM-as-a-Judge paradigm to evaluate emotional steerability.

Through a comprehensive, controlled preliminary experiment generating 250 evaluation pairs from PG-19 base texts, we analyzed the tendencies of emotion pairings, the challenges of generating complex dyadic emotions, and the organic ripple effects of emotional injection. Most critically, our data quantitatively proved the existence of *Semantic Resistance*—a phenomenon where forcing an emotional parameter that contradicts the foundational semantic constraints of a genre causes the LLM to hallucinate and break narrative logic.

Our findings demonstrate that while modern generative AI can effectively manipulate emotional tones and adhere to literary rules like ”Show, Don't Tell,” these models cannot ignore established semantic context without risking narrative collapse. As Generative AI becomes the central engine for interactive media and personalized ”Storybuilding” applications, simply passing emotional vectors to an LLM will not suffice. The final application presented in this paper acts as the ultimate exit point for this research, proving that system architectures must account for genre-specific resistance and employ semantic boundary filtering prior to generation. These results lay the empirical groundwork for our ongoing development of interactive narrative systems, ensuring that dynamically generated storytelling environments remain emotionally resonant, deeply personalized, and logically coherent.

Declaration on Generative AI

During the preparation of this work, the author used generative AI tools (specifically, the Gemini family of models) to assist in drafting the manuscript, generating literature reviews, and executing the automated data generation and evaluation pipeline.

References

[1] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al., Gemini: A family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023).

[2] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in neural information processing systems 35 (2022) 27730–27744.

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[4] A. Fan, M. Lewis, Y. Dauphin, Hierarchical neural story generation, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 889–898.

[5] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38.

[6] A. Alabdulkarim, et al., Automatic story generation: A survey of approaches, ACM Computing Surveys 54 (2021) 1–38.

[7] L. Zheng, W.-L. Chiang, Y. Sheng, S. Hao, Z. Wu, S. Ba, Z. L. Jiang, Z. Wang, J. Su, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, Advances in Neural Information Processing Systems 36 (2023).

[8] R. Plutchik, A general psychoevolutionary theory of emotion, Theories of emotion 1 (1980) 3–31.

[9] P. Lisena, et al., Doremus: A graph of linked musical works, in: International Semantic Web Conference, 2018.

[10] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, IEEE Transactions on Knowledge and Data Engineering (2024).

[11] J. R. Meehan, Tale-spin, an interactive program that writes stories, in: Proceedings of the 5th international joint conference on Artificial intelligence-Volume 1, 1977, pp. 91–98.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[13] J. W. Rae, A. Potapenko, S. M. Jayakumar, T. P. Lillicrap, Compressive transformers for long-range sequence modelling, arXiv preprint arXiv:1911.05507 (2019).

[14] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[15] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, arXiv preprint arXiv:2303.18223 (2023).

[16] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al., Sparks of artificial general intelligence: Early experiments with gpt-4, arXiv preprint arXiv:2303.12712 (2023).

[17] J. Ficler, Y. Goldberg, Controlling linguistic style aspects in neural language generation, Proceedings of the Workshop on Stylistic Variation (2017) 94–104.

[18] X. Liang, et al., Controllable text generation for large language models: A survey, arXiv preprint arXiv:2408.12599 (2024).

[19] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, R. Liu, Plug and play

language models: A simple approach to controlled text generation, arXiv preprint arXiv:1912.02164 (2019).
[20] B. Krause, A. D. Gotmare, B. McCann, N. Shirish Keskar, S. Joty, R. Socher, N. F. Rajani, Gedi: Generative discriminator guided sequence generation, Findings of the Association for Computational Linguistics: EMNLP 2020 (2020).
[21] A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi, The curious case of neural text degeneration, in: International Conference on Learning Representations, 2020.
[22] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022.
[23] J. Dekoninck, et al., Controlled text generation via language model arithmetic, arXiv preprint arXiv:2311.11197 (2023).
[24] P. Ekman, An argument for basic emotions, Cognition & emotion 6 (1992) 169–200.
[25] J. A. Russell, A circumplex model of affect, Journal of personality and social psychology 39 (1980) 1161.
[26] O. Rozen, O. Barkan, J. Katz, et al., Prompting large language models with the plutchik emotion wheel, in: Findings of the Association for Computational Linguistics: ACL 2023, 2023.
[27] E. Segel, J. Heer, Narrative visualization: Telling stories with data, IEEE Transactions on Visualization and Computer Graphics 16 (2010) 1139–1148.
[28] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, D. Roth, Recent advances in natural language processing via large pre-trained language models: A survey, ACM Computing Surveys 56 (2023) 1–40.
[29] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-eval: Nlg evaluation using gpt-4 with better human alignment, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 2511–2522.
[30] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.
[31] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, International Conference on Learning Representations (2023).