A Theoretical Framework for Evaluating Narrative Surprise in Large Language Models

Annaliese Bissell [1]
McGill University

Ella Paulin [1]
McGill University

Andrew Piper
McGill University

Abstract

Narrative surprise is a core element of storytelling for engaging audiences, and yet it remains underexplored in the context of large language models (LLMs) and narrative generation. While surprise arises from events that deviate from expectations while maintaining retrospective coherence, current computational approaches lack comprehensive frameworks to evaluate this phenomenon. This paper presents a novel framework for assessing narrative surprise, drawing on psychological theories of narrative comprehension and surprise intensity. We operationalize six criteria—initiatoriness, immutability violation, predictability, post-dictability, importance, and valence—to measure narrative surprise in story endings. Our study evaluates 120 story endings, generated by both human authors and LLMs, across 30 mystery narratives. Through a ranked-choice voting methodology, we identify significant correlations between reader preferences and four of the six criteria. Results underscore the continuing advantage of human-authored endings in achieving compelling narrative surprise, while also revealing significant progress in LLM-generated narratives.

1 Introduction

Narrative surprise represents a fundamental mechanism through which stories engage and captivate audiences, yet our understanding of how to systematically measure this phenomenon in large language models (LLMs) remains limited. While traditional narratology has long recognized surprise as one of three key components of narrative tension alongside suspense and curiosity (Brewer and Lichtenstein, 1980; Sternberg, 1990; Hoeken and Van Vliet, 2000; Bermejo-Berros et al., 2022), the emergence of LLMs as storytelling agents presents novel challenges in quantifying their ability to generate genuine narrative surprise.

Recent work in computational story generation has focused on two key challenges relevant to this area that have nevertheless remained distinct from one another. Narrative *coherence* is essential for establishing narrative meaning by ensuring continuity among multiple narrative elements such as setting, characters, and events (Guan et al., 2019; Gupta et al., 2019). Narrative *surprise*, on the other hand, depends on the introduction of novel information while also maintaining narrative coherence. As Sternberg (1990) argues, for surprise to be effective, the unexpected turn of events must be *retrospectively coherent*.

From this perspective, recent approaches to evaluating narrative surprise in computational storytelling have important limitations. While researchers have made progress in developing word-level surprise metrics (Huang et al., 2023; Wilmot and Keller, 2020) and tracking narrative turning points through sentiment analysis (Tian et al., 2024; Knight et al., 2024; Elkins, 2022), these methods do not capture the complex temporal relationships that make stories coherent and meaningful. Specifically, they do not address how surprising events must deviate from expectations while remaining logically consistent within the broader narrative framework. This disconnect between the evaluation of local surprise and global coherence represents a significant gap in the field, underscoring the need for a more comprehensive theoretical framework that can assess both the unexpectedness of generated story elements and the success of their narrative integration.

In this paper, we present a novel theoretical framework for evaluating narrative surprise, grounded in psychological research on narrative comprehension and evaluation. Our framework introduces six key metrics that capture different dimensions of cognitive surprise in narrative understanding. To validate this framework, we conduct an analysis of 120 story endings, generated by both

[1] These authors contributed equally to the paper.

Proceedings of the The 7th Workshop on Narrative Understanding, pages 26–35
May 4, 2025 ©2025 Association for Computational Linguistics

human authors and LLMs, focusing on 30 mystery stories sourced from the Reedsy fiction platform. These stories are manually truncated before their pivotal revelations to enable controlled testing of ending generation. Using a ranked-choice voting methodology, we assess the relative quality of different endings and examine how our proposed metrics correlate with reader preferences.

Our analysis reveals that four of our six variables demonstrate significant associations with reader preferences, providing initial validation of our theoretical framework. We compare LLM and human performance using both voting data and our six-metric framework. We conclude by discussing future directions for enhancing narrative surprise evaluation in computational storytelling and share our underlying data.[2]

2 Prior Work

2.1 Theories of Narrative Surprise

Contemporary theoretical frameworks consistently identify cognitive surprise as an emotion triggered by the disparity between expected and actual events or information revelation (Ortony and Partridge, 1987; Brewer and Lichtenstein, 1980; Celle et al., 2017). In the context of narrative comprehension, Structural Affect Theory (SAT) provides a theoretical foundation for understanding surprise generation (Brewer and Lichtenstein, 1980). SAT posits that presentation of a *surprise event* (SE) without the presentation of its corresponding *initiating events* (IE) or causal antecedents can provoke surprise. Thus, in order to provoke *surprise* as defined in SAT, the initiating event (IE) or “expository information" must be withheld, while maintaining readers’ unawareness of this omission. It is this lack of awareness of the omission that distinguishes surprise from curiosity, which arises when readers consciously perceive an information gap (Brewer and Lichtenstein, 1980).

Moreover, Ortony and Partridge (1987) propose that the intensity of the surprise is contingent on the type of expectation subverted. They categorize propositions into two types: *immutable* (fixed within the story’s universe) and *mutable* (which can change without breaking the story’s logic). In a murder mystery, for example, an immutable element is that the victim is dead—this is a real-world condition of the story’s universe. Changing this would break the internal logic of the mystery. A mutable element is how the detective solves the case—whether uncovering a hidden letter, analyzing forensic evidence, or interrogating suspects, the path to the solution can vary without altering the story’s basic premises. (Ortony and Partridge, 1987).

The framework also differentiates between *deducible* outcomes—which the reader would have been able to predict given a combination of evidence presented in the story and general world knowledge—and non-deducible outcomes, which could not have been predicted. The latter are often outcomes that lack a clear antecedent, e.g. a rock flying through a window without warning (Ortony and Partridge, 1987). They posit that a contradiction of an immutable expectation will elicit maximal surprise, while contradiction of a mutable expectation may elicit high but not maximal surprise.

Bae and Young (2013) provide a concrete set of criteria to check whether a story provokes surprise in the reader. Their Prevoyant story plan generation architecture implements a reader-modeling evaluator that assesses story plans across four dimensions: expectation failure, importance, emotional valence, and incongruity resolution. They posit that emotional valence (positive or negative) influences surprise quality, with higher surprise provoked by an outcome with negative valence than that of one with positive valence. They define incongruity resolution as the presentation of events or information that resolves any apparent contradictions in the story.

These works and concepts will function as the foundation of our annotation framework described in Section 3.

2.2 Language Model Narrative Generation

Prior work has identified significant limitations in LLM-generated narratives, particularly regarding narrative coherence and plot development. Tian et al. (2024) demonstrate that while readers appreciate logical and well-motivated plot developments, LLM outputs frequently default to simplistic positive trajectories or miraculous twists and may suffer from a lack of coherence.

Several methods have been proposed to provide coherent and surprising output. Huang et al. (2023) developed the Affective Story Generator (AffGen), which implements two key mechanisms to enhance narrative engagement: favouring less predictable words and using an Affective Reranking system

[2] https://doi.org/10.5683/SP3/TOXZSQ

that prioritizes heightened emotional intensity in generated content.

See et al. (2019) demonstrated that while GPT2-117 outperformed neural story generation systems in awareness of story context and lexical diversity, it produced similarly repetitive narratives. Building on this work, Akoury et al. (2020) explored domain adaptation through fine-tuning GPT-2 on data from the online storytelling platform STORIUM. They found that while the model achieved linguistic fluency, it struggled with maintaining narrative coherence, frequently introducing inconsistent story events or characters.

Although contemporary LLMs are more fluent and coherent, they continue to lack the ability to generate well-paced and diverse narratives. Tian et al. (2024) investigate the narrative generation ability of commercially available LLMs, finding that despite recent advances in LLM capabilities, story arcs in LLM output are more poorly paced than human narratives. Moreover, LLMs' tendencies toward homogeneous, positive plot trajectories lead to less suspenseful output.

Chakrabarty et al. (2024) find that LLM-generated narratives achieve only 10-33% of human-level performance across four dimensions of creativity. LLMs perform badly on tasks related to narrative surprise, containing "turns that are both surprising and appropriate" only between 22% and 34% as often as human narratives (Chakrabarty et al., 2024). Specific narrative surprise-related problems identified by Chakrabarty et al. (2024)'s annotators include illogical events, inconsistent characterization, clichés, unrealistic happy endings, unexpected surreal elements and failure to deliver on potential of a premise. However, when basing their analysis on amateur short stories on Reddit Zhou et al. (2024) show that GPT-4 rivals human ability to produce engaging, provocative and narratively complex short stories, which suggests model performance may vary based on the specific narrative generation task and evaluation context.

3 A Theoretical Framework for Measuring Narrative Surprise

We evaluate six criteria for narrative surprise, drawing from foundational work on story comprehension and narrative affect discussed above (see Table 1 for an overview). Our framework integrates Bae and Young (2013)'s work on computational models for generating surprising narratives and Ortony and Partridge (1987)'s framework for surprise intensity. The framework relies on narratives segmented into two structural components: the 'stem,' encompassing the beginning and middle of the narrative, and the 'ending,' which resolves earlier narrative events, often in the form of a 'big reveal'. Note that we assume the surprising event with unknown causes (SE) is presented in the 'stem,' while its initiating events (IEs), i.e. causes, are presented in the 'ending.'

Category	Description
Initiatory	Ending describes a novel event that temporally precedes and causes the SE.
Immutability Violation	Ending contradicts an immutable fact of the story world.
Predictable	A typical reader could have predicted the ending given the stem.
Post-dictable	Looking backwards at the whole story, the events are explainable, i.e. there are neither loose ends nor contradictions.
Important	Events of the ending meaningfully impact the protagonist.
Valence	Events of the ending are positive for the protagonist.

Table 1: Definitions of Surprise Criteria

The first criterion, *initiatoriness*, which operationalizes Brewer and Lichtenstein (1980)'s surprise generation hypothesis, examines whether initiatory events are presented in the ending which offer a causal explanation for the SE that occurred in the stem. A highly initiatory narrative ending will provide key initiating events that explain how the surprising event(s) of the narrative stem occurred.

The second criterion, *immutability violation*, builds on Ortony and Partridge (1987) theoretical framework concerning proposition violation. This dimension assesses the degree to which narrative events challenge established axioms within the story world's logical framework. Immutability violations occur when narratives contradict fundamental beliefs about the world (such as the absence of flying pigs). Narratives contradicting more flexible beliefs, such as the belief that employers always hire by merit, are less *immutability violating* and easier to accept as plausible.

The third criterion, *predictability*, builds on the observation of Ortony and Partridge (1987) that an expectation-reality discrepancy is required to elicit surprise. Our framework posits outcome predictability to be inversely related to surprise magnitude, while acknowledging that to ensure reader satisfaction, narrative surprises must not be totally impossible to predict. This suggests an optimal zone of moderate predictability.

The fourth dimension, *post-dictability*, is drawn from Bae and Young (2013). It measures the degree to which the narrative maintains internal consistency and fully explains plot events in order to leave readers with the feeling that the story makes sense in retrospect. This aligns with Sternberg (1990)'s argument that surprise necessitates events to be *retrospectively coherent*.

The final two criteria, *valence* and *importance*, are taken directly from the framework of Bae and Young (2013), where negativity and importance are hypothesized to be positively correlated with surprise.

4 Methods

4.1 Dataset

We construct a dataset of 30 mystery short stories drawn from the story prompt website Reedsy, written after October 2023. We choose this date as it post-dates our selected models' training period, ensuring that the LLMs are evaluated on new data. We use mysteries because surprisingness is both inherent to the genre and also highly structured. Each narrative begins with an unexplained event, followed by a systematic revelation of details that lead readers to the ultimate solution, i.e. all necessary information has been revealed.

Mysteries thus provide a controlled pattern for the study of narrative surprise, one that aligns with prior work on story ending generation (Guan et al., 2019). However, whereas prior work on story ending generation has typically focused on very short sequences–Zhou et al. (2024) focus on stories with an average length of 450 words, while Mostafazadeh et al. (2016) look at stories of 6 sentences in length–our stories are considerably longer by comparison posing a more challenging task (Table 2).

To prepare our data for evaluation, we manually divide each story into a "stem" and "ending," truncating the story at the point where the author begins to answer the central question posed at the beginning by the unexplained event (e.g. "who killed the protagonist's brother," "why is food going missing from the kitchen when nobody in the family is touching it," etc.), which we hypothesize to be where the "big reveal" happens. As can be seen in Table 2, stem and ending lengths are not only of considerably different lengths, but the two categories themselves contain considerable variance.

Story Portion	**Mean**	**SD**
Stem	2056	511
Human Ending	339	220
GPT Zero Shot Ending	447	82
GPT Few Shot Ending	577	144
Phi3 Zero Shot Ending	424	186

Table 2: Stem and Ending Lengths

4.2 Story Ending Generation

We then prompt two language models, one large frontier model, gpt-4o-2024-08-06, and one small open-weight model with a large-enough context window to handle our texts, Phi3-mini-128k-instruct. Both models were trained prior to our cut-off date for our stories. In order to generate an ending given a stem, we use two prompting strategies:

1. Zero Shot: "Your task is to write a surprising twist ending for a given incomplete mystery short story. The story does not need to have a moral, and the ending should be about 300 words. Here is the story: ..."

2. Few Shot: the same prompt as above was used, with the addition of "When writing your ending, follow these examples: ..." and 2 example stem/ending pairs.

We found that using a chain of thought approach, where the model was prompted to analyze the characters and plot points and brainstorm possible twist endings before generating a final ending, provided no improvement over the outputs of the zero shot or few shot approaches. We also found that the few shot approach diminished the quality of endings for our Phi3 model. Thus our final dataset consisted of 120 story endings, consisting of endings generated by GPT4 (Zero Shot), GPT4 (Few Shot), and Phi3 (Zero Shot) along with the original human-authored ending.

4.3 Narrative Surprise Annotation

A team of four undergraduate student annotators were assembled, all of whom have prior experience in literary studies and text annotation. They were given a codebook, included in the data repository, with explicit descriptions for each criterion and instructions for rating endings on a 5-point Likert scale. This approach follows Chhun et al. (2022)'s recommendations for using human annotations in automatic story generation evaluation, while the explicit scale descriptions help reduce subjectivity in the labeling process. Students were then asked to identify the ending that they felt was the "most" and "least" surprising.

5 Results

5.1 Inter-Annotator Agreement

To measure inter-annotator agreement on our Likert scale annotations, we use the average deviation index (ADI) as suggested by O'Neill (2017). As can be seen in Table 3, for all criteria the ADI is <1 on a five-point scale suggesting good levels of agreement.

Category	ADI
Predictable	0.715
Post-Dictable	0.639
Immutability Violation	0.598
Initiatory	0.559
Important	0.466
Valence	0.459

Table 3: Average Deviation Index across all surprise criteria.

We analyzed inter-annotator agreement on story-ending preferences using Kendall's W, a non-parametric statistic particularly suited for ranked ordinal data. The analysis revealed moderate consensus among the four raters (W = 0.552, $\chi^2(119)$ = 263, p < 0.001). This coefficient, ranging from 0 to 1, indicates reliable but subjective judgments in evaluating ending quality, with the highly significant p-value confirming non-random agreement.

Given prior research on the variation of the experience of surprise (Juergensen et al., 2014), a medium degree of agreement is expected. To address this, we add random effects to the regression model discussed in Section 5.3 to control for annotator variability when analyzing correlations between surprise criteria ratings and reader preferences.

5.2 Model Preference

To assess the performance of the generated endings, we compare the number of most/least surprising votes each model received across all annotators and endings along with the odds ratio of observed voting behaviour relative to a random baseline of equal votes across all models.

Model	Most	OR	Least	OR
Phi3	4	0.13	87	2.90
GPT4 (Zero)	24	0.80	10	0.33
GPT4 (Few)	34	1.13	8	0.27
Human	58	1.93	15	0.50

Table 4: Counts of reader preferences with accompanying odds ratio of observed votes relative to a random baseline of expected votes for each model.

As can be seen in Table 4, our analysis reveals clear preferences among story endings. Human-authored endings were most preferred, selected at nearly twice the random baseline rate. Combined GPT-4 endings received comparable preference (58 selections total), though few-shot prompting proved more effective than zero-shot generation. In contrast, Phi3-generated endings were rarely preferred, suggesting significant quality differences between large and small language models for this task. Mixed-effects logistic regression confirmed these patterns, showing human-authored endings were 2.94 times more likely to be chosen than GPT-4 endings (p < 0.001), while Phi3 endings were significantly less preferred (OR = 0.11, p < 0.001).

5.3 Correlation with Reader Annotations

As a first step, we analyze the relationship between the distribution of surprise criteria across endings for our different models versus human endings. Using Spearman correlation coefficients, which are appropriate for ordinal Likert scale data, we find correlations of 0.60, 0.49, and 0.03 for GPT4-Zero Shot, GPT4-Few Shot, and Phi3, respectively with human-authored endings.

Fig. 1 illustrates the specific levels of correlation for each criteria and model comparison, indicating some meaningful degree of variance. GPT achieved the highest correlation on the initiatoriness of story endings and the lowest on post-dictability, i.e. the ability to explain ending events given prior story elements.

As a way of further illustrating the degree of correlation between human ratings and our LLMs,

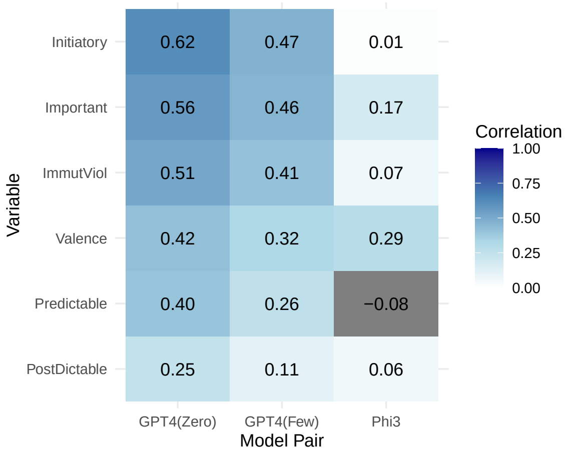

Figure 1: Spearman's correlation coefficient between human endings and each LLM.

Fig. 2 shows the distributions of annotator ratings across all six variables for human endings and GPT4 (Zero Shot), our highest correlated model.

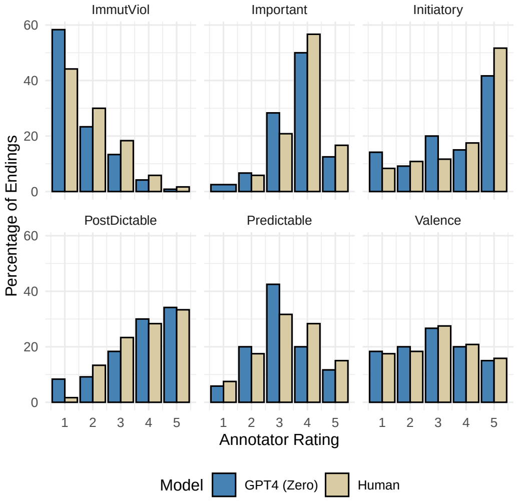

Figure 2: Distributions of annotator ratings for story endings authored by humans and GPT4 (Zero Shot) across all six variables.

We conducted a conditional logistic regression analysis to examine the relationship between our six predictor variables of surprise and the binary outcome of being the most preferred ending or not. We stratified the analysis by Stem to control for potential group-level effects. To assess model fit, we compared our model to a null model using likelihood ratio tests and evaluated the model's discriminative ability using the concordance index (C-index).

The conditional logistic regression model demonstrated strong overall fit (likelihood ratio test: $\chi^2(6) = 59.79$, $p < .001$). The model showed good discriminative ability with a C-index of 0.714 (SE = 0.028), indicating successful distinction between outcomes.

Bootstrap validation (100 resamples) suggested moderate model stability (SE = 9.81) with some potential for overfitting (bias = 14.88). We also compared our model to a random-effects model including annotator effects, but the lower AIC value for our primary model (AIC = 402.61 vs. 594.98) supported its selection as the final model.

As can be seen in Table 5, four predictors showed significant associations with being selected the most surprising ending, with two positively associated (Initiatory and Post-Dictable) and two negatively associated (Predictability and Valence). In Fig. 3, we illustrate the odds ratios and 95% confidence intervals showing how a one-unit increase in each variable affects the likelihood of being selected as winner.

	Dependent variable	
	Most Surprising	Odds-Ratio
ImmutViol	0.053 (0.133)	1.05
Important	0.256 (0.159)	1.30
Initiatory	0.352 (0.117)***	1.42
Post-Dictable	0.283 (0.133)**	1.33
Predictable	-0.496 (0.124)***	0.61
Valence	-0.263 (0.118)**	0.77

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5: Logistic regression results analyzing the relationship between our surprise features and being selected the most surprising ending. We translate coefficients into the increased odds of winning with a one unit increase/decrease of a given variable.

Initiatoriness demonstrates the strongest positive influence, with each unit increase raising the odds of an ending being selected as most surprising by 42%. This effect is most pronounced when comparing extreme cases: endings with maximal initiatoriness were more than four times as likely to be chosen compared to those with minimal initiatoriness. Post-dictability shows a similar positive relationship, with each unit increase raising selection odds by 32%. At the extremes, maximally post-dictable endings were preferred over three times as often as minimally post-dictable ones.

On the other hand, both predictability and valence demonstrate significant negative relationships

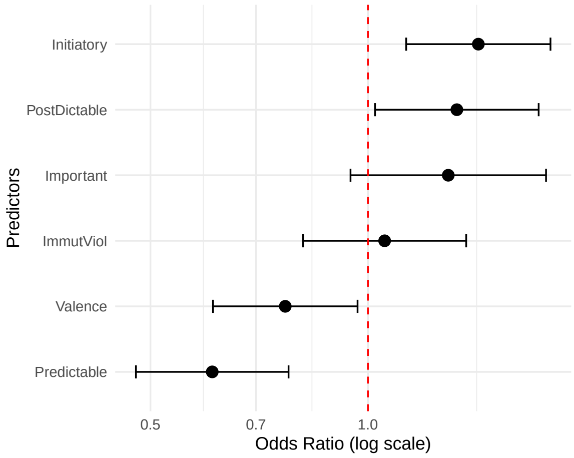

Figure 3: Odds-ratios with confidence intervals of being associated with the most surprising ending for our six surprise variables.

with surprise selection. Each unit increase in predictability reduces an ending's selection odds by 60%, with maximally predictable endings experiencing a sevenfold reduction in selection likelihood compared to minimally predictable ones. Valence shows a more moderate negative effect, with each unit increase (i.e., more positive outcomes) reducing selection odds by 24%. At the extremes, highly positive endings are 2.8 times less likely to be selected as surprising compared to highly negative ones.

6 Discussion

6.1 Understanding Narrative Surprise

Our analysis validates four of the six theoretically proposed criteria as significant predictors of narrative surprise intensity. These include endings with strong causal relationships to the main surprising event of the story (Initiatoriness); strong explanatory power of prior events (Post-Dictability); low predictability of reported events (Predictability); and negative valence (Valence).

In Story #10, for example, whose preferred ending was rated 4.75 (out of 5) for Initiatoriness, the story stem focuses on a protagonist who discovers a crumpled letter addressed to them. The preferred ending (human-authored) reveals that the protagonist had written and discarded the letter years prior. This demonstrates high initiatoriness by revealing a causal event that precedes the story stem's central surprising event.

Endings with high post-dictability are characterized by more complete and coherent resolution of narrative uncertainties from the story stem. Story #10's preferred ending was also rated highly for post-dictability by providing a coherent resolution that explains the letter's origin without contradicting established narrative elements.

Conversely, endings with higher predictability and more positive emotional valence had significantly reduced chances of being selected as the most surprising ending, in keeping with Bae and Young (2013) on the importance of negative valence for surprise intensity. Predictability had the strongest overall effect on reader preference, with each unit increase in predictability reducing an ending's odds of selection by 60%. As an example of this preference, consider story #6, which centres on an interaction between a menacing crime writer and his admiring fan during an alleged 'improv exercise.' When the writer lunges at the fan with a knife, claiming it's for creative inspiration, two possible endings emerge. Annotators consistently preferred the less predictable outcome—where the fan becomes the killer and achieves literary fame—over the more obvious ending where the writer kills his fan.

Interesting, immutability violations and event importance did not show meaningful associations with reader preferences. While most stories did not exhibit immutability violations (see Fig. 2), it is interesting and worth further consideration as to why this feature did not strongly factor into reader preferences. Although Ortony and Partridge (1987) hypothesized that more immutability-violating stories would provoke more surprise than less immutability-violating stories, we provide an initial hypothesis that there are two distinct pathways to narrative surprise: through immutability violations and through unexpected resolutions of mutable variables. We propose that readers can experience intense surprise when mutable variables—those naturally capable of taking different values—resolve to unexpected states. Consider a mystery narrative where evidence strongly implicates character A, but the ending reveals the seemingly innocent character B to be the perpetrator. The resulting surprise may derive not from violating any fundamental story-world constraints (immutable propositions) but from strategically subverting reader expectations about the specific value a mutable variable will resolve to, although future work is needed to evaluate this potential additional pathway by which a story without an immutability violation can produce intense narrative surprise.

6.2 Comparing LLM and Human Endings

When it comes to comparing model-generated and human endings, our analysis reveals significant preference disparities between human-authored endings and those generated by large language models. Human-authored endings were preferred almost three-times more often than even our best-performing model (GPT-4 Few-Shot). At the same time, GPT-4 generated endings were chosen about as often as human-authored endings, suggesting that the generation task is indeed feasible.

As an example of human/LLM differences, Story #30 provides a useful case. This story stem focuses on a British intelligence agent who follows a KGB spy who wears a red scarf. After learning of the double-agent's murder, the protagonist spots a red scarf in his colleague's car. The human ending reveals that the colleague, himself a double agent working for the KGB, killed the KGB spy with the red scarf because she had defected. In contrast, GPT-4's ending introduces unexplained elements—the double-agent is revealed to be alive, and she and the protagonist apparently have known each other the whole time.

This example illustrates a pattern with the GPT-4 endings where new details and backstory are often introduced which are not coherent with the existing story elements, potentially indicating the way the problem of hallucination infects narrative generation. In this ending, GPT-4 also fabricates details to create a more optimistic tone that deviates from the human version, a fact also noted by prior work (Tian et al., 2024).

In addition to these problems of positivity and coherence, GPT-4 endings were also on average more predictable than human-authored endings. For example, in Story #29, a man is trapped in a VR game show seeking funds for his son's medical treatment. When approached by a figure in white attempting to wake him, GPT-4's ending describes a straightforward rescue, while the human-authored ending reveals that the figure was the protagonist's son, producing significantly higher narrative surprise. This is a good example of the challenges of balancing novelty plus coherence that is the hallmark of successful narrative surprise. Too much new information risks damaging coherence (post-dictability), while too little risks being too predictable.

Future work will want to explore further prompt-engineering approaches to assess pathways towards more successful surprising narrative endings. It could also be the case that fine-tuning approaches might also facilitate a deeper understanding of the conditions of surprise. Given the small-scale of our evaluation experiment, further work exploring more diverse stories as well as larger evaluator pools will help solidify our understanding of the concept of narrative surprise.

7 Conclusion

This paper presents a novel theoretical framework for evaluating narrative surprise in stories generated by large language models (LLMs) and human authors. By integrating theoretical insights from narrative comprehension and cognitive surprise, we develop six key metrics to assess narrative surprise. Our analysis of mystery story endings highlights the value of these metrics in understanding reader preferences, with initiatoriness and post-dictability emerging as particularly significant factors in driving narrative surprise.

While our findings underscore the potential of LLMs to produce engaging narrative surprises, they also reveal limitations in their current ability to match the complexity and nuance of human-authored endings. The preference for human-authored stories suggests that LLMs need further advancements in generating unexpected yet coherent twists. In particular, enhancing the ability to generate causal relationships (Initiatoriness) and logically coherent endings (Post-dictability) and avoiding overly positive endings that are highly predictable offer promising avenues for improving the quality of machine-generated narratives.

Future research should go beyond the mystery genre to explore how narrative surprise varies across different storytelling traditions and audience expectations. Incorporating multilingual datasets will also be essential for understanding how cultural and linguistic factors shape perceptions of surprise, coherence, and narrative quality. Additionally, employing more diverse evaluation methodologies, such as real-time audience engagement tracking or large-scale reader surveys will help capture the multifaceted nature of narrative surprise. These efforts will not only refine our understanding of narrative dynamics but also advance the development of computational storytelling systems that are better equipped to create more nuanced and interesting stories.

Limitations

Limitations of our methodology include the inherent subjectivity of surprise assessment, which resulted in moderate inter-annotator agreement. Evaluating surprise, particularly in narrative contexts, is deeply influenced by individual differences in reader expectations, cultural backgrounds, and personal preferences, making it challenging to establish universally consistent criteria. While we employed a codebook and explicit descriptions to standardize the evaluation process, the inherently subjective nature of surprise likely contributed to the variability in ratings. Future work could explore ways to mitigate this limitation, such as integrating physiological measures of surprise (e.g., eye-tracking, galvanic skin response) or employing larger and more demographically diverse annotator pools to capture a broader range of reactions.

Second, our corpus composition—English-language mystery narratives from non-professional authors—may limit generalizability across different languages and literary traditions. Mystery stories, particularly those written in English, tend to follow culturally specific narrative structures and conventions that may not align with storytelling patterns in other languages or regions. Additionally, the use of non-professional authors introduces variability in narrative quality and style, which may not reflect the complexity and craftsmanship of professionally written texts. Expanding future datasets to include stories from diverse linguistic and cultural backgrounds, as well as works authored by professionals, would provide a richer foundation for analyzing narrative surprise and its universality.

Finally, our experimental design, focusing on ending completion, captures only a subset of the complex processes involved in constructing narrative surprise. While our approach allowed for controlled testing, it did not account for the broader aspects of storytelling and their relationship to surprise, such as plot architecture, pacing, or more local moments of surprise. These elements play a critical role in building tension, shaping expectations, and delivering impactful surprises. Future studies could incorporate a more holistic approach by analyzing full narratives, from their inception to resolution, and examining how surprise is cultivated across the entire arc of the story. Additionally, incorporating methods to evaluate narrative planning and the interplay of suspense, curiosity, and surprise could provide a more comprehensive understanding of the storytelling process.

Acknowledgements

We wish to thank the Social Sciences and Humanities Research Council of Canada (435-2022-089) for funding to support this research.

References

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. Storium: A dataset and evaluation platform for machine-in-the-loop story generation. *arXiv preprint arXiv:2010.01717*.

Byung-Chull Bae and R Michael Young. 2013. A computational model of narrative generation for surprise arousal. *IEEE Transactions on Computational Intelligence and AI in Games*, 6(2):131–143.

Jesús Bermejo-Berros, Jaime Lopez-Diez, and Miguel Angel Gil Martínez. 2022. Inducing narrative tension in the viewer through suspense, surprise, and curiosity. *Poetics*, 93:101664.

William F Brewer and Edward H Lichtenstein. 1980. Event schemas, story schemas, and story grammars. *Center for the Study of Reading Technical Report; no. 197*.

Agnès Celle, Anne Jugnet, Laure Lansari, and Emilie L'Hôte. 2017. *Expressing and describing surprise*. John Benjamins Amsterdam.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Cyril Chhun, Pierre Colombo, Chloé Clavel, and Fabian M Suchanek. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. *arXiv preprint arXiv:2208.11646*.

Katherine Elkins. 2022. *The shapes of stories: sentiment analysis for narrative*. Cambridge University Press.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6473–6480.

Prakhar Gupta, Vinayshekhar Bannihatti Kumar, Mukul Bhutani, and Alan W Black. 2019. Writerforcing: Generating more interesting story endings. *arXiv preprint arXiv:1907.08259*.

Hans Hoeken and Mario Van Vliet. 2000. Suspense, curiosity, and surprise: How discourse structure influences the affective and cognitive processing of a story. *Poetics*, 27(4):277–286.

Tenghao Huang, Ehsan Qasemi, Bangzheng Li, He Wang, Faeze Brahman, Muhao Chen, and Snigdha Chaturvedi. 2023. Affective and dynamic beam search for story generation. *arXiv preprint arXiv:2310.15079*.

James Juergensen, Joseph S Weaver, Kevin J Burns, Peter E Knutson, Jennifer L Butler, and Heath A Demaree. 2014. Surprise is predicted by event probability, outcome valence, outcome meaningfulness, and gender. *Motivation and Emotion*, 38:297–304.

Samsun Knight, Matthew D Rocklage, and Yakov Bart. 2024. Narrative reversals and story success. *Science Advances*, 10(34):eadl2013.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Thomas A O'Neill. 2017. An overview of interrater agreement on likert scales for researchers and practitioners. *Frontiers in psychology*, 8:777.

Andrew Ortony and Derek Partridge. 1987. Surprisingness and expectation failure: what's the difference? In *IJCAI*, volume 87, pages 106–108.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. 2019. Do massively pretrained language models make better storytellers? *arXiv preprint arXiv:1909.10705*.

Meir Sternberg. 1990. Telling in time (i): Chronology and narrative theory. *Poetics today*, 11(4):901–948.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? *arXiv preprint arXiv:2407.13248*.

David Wilmot and Frank Keller. 2020. Modelling suspense in short stories as uncertainty reduction over neural representation. *arXiv preprint arXiv:2004.14905*.

Hanyu Zhou, Sreya Muppalla, Zeynep Yildiz, Miryung Kim, Amit Sahai, Nanyun Peng, et al. 2024. Measuring psychological depth in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17162–17196.