Quantifying Narrative Surprise

Devashish Juyal
Department of EECS
University of Michigan
juyal@umich.edu

Yuganshi Agrawal
Department of Statistics
University of Michigan
yuganshi@umich.edu

1 Project Goals

Why do readers stay absorbed in a novel? A key driver is suspense, the anticipation of surprise created by narrative twists. With the rise of large language models (LLMs), we can revisit a classic question in narrative theory: Can plot twists be detected and quantified computationally?

Our goals are threefold: (i) formalize "suspense" and "twists" within an NLP framework, (ii) quantify their magnitude via probabilistic divergence, and (iii) analyze thriller stories to study narrative structure and build twist-based ranking systems. This approach extends recent work on statistical models of narrative surprise (Underwood, 2024; Piper et al., 2023), while positioning twist detection as a sequential anomaly detection problem with implications for computational literary studies, creative AI, and cultural analytics.

From an academic perspective, operationalizing "narrative surprise" contributes to the growing field of computational literary studies, where researchers seek to formalize literary phenomena in measurable terms. Moreover, from an NLP and machine learning perspective, this project addresses the general challenge of expectation modeling and anomaly detection in sequences. Plot twists are essentially semantic anomalies within narrative flow, making them an ideal testbed for evaluating how well language models capture surprise. Ultimately, by uniting literary theory, NLP methodology, and industry relevance, this project demonstrates how computational tools can shed light on practical applications.

2 NLP Task Definition

A plot twist can be broadly defined as a narrative event that subverts audience expecations, creating surprise and altering the perceived trajectory of the story. Formally, the task can be defined as follows:

Input: A narrative text (e.g., a small story), segmented into sentences, paragraphs, or chapters.

Process: The procedure involves four stages. First, we construct a probabilistic or embedding-based model of "expected" narrative progression, using language models to capture baseline coherence. Second, we identify points in the narrative where the observed text exhibits statistically significant divergence from the expected trajectory, which we interpret as candidate twist events. Third, we assign each candidate twist a quantitative score based on the magnitude of deviation, measured for example through surprisal, embedding shifts, or prediction error. Finally, we scale this approach to the corpus level, applying the method to collections of thriller stories, ranking texts by cumulative or average twist intensity, and visualizing suspense trajectories across the narrative.

Output: A ranked list of detected twist events within a novel, together with an aggregated "twist score" that characterizes the overall twist intensity of the narrative.

This formulation situates the project within broader NLP challenges such as narrative modeling, anomaly detection, and information-theoretic measures of surprise. By applying these methods to thriller fiction, we aim to provide a replicable computational framework for analyzing one of the genre's most salient features.

2.1 Formal Task Formulation

Let a story be represented as a sequence of narrative units:

$$X = (x_1, x_2, \ldots, x_n),$$

where each x_i corresponds to a sentence, paragraph, or chapter, depending on the chosen segmentation granularity.

The task of plot twist detection and scoring can be formalized as follows.

We define a narrative model M that estimates

the probability distribution of the next unit given its preceding context:

$$P(x_i \mid x_1, \ldots, x_{i-1}; \theta_M).$$

From this distribution, we compute the surprisal of each unit as defined by Levy (2008):

$$s_i = -\log P(x_i \mid x_1, \ldots, x_{i-1}; \theta_M).$$

Higher s_i values correspond to lower predictability relative to the modeled baseline.

Candidate twist events are defined as units where the observed text exhibits statistically significant deviation from expectation. Formally, a set of twist candidates $\mathcal{T}$ is given by:

$$\mathcal{T} = \{\, x_i \in X \;\mid\; s_i \geq \tau \,\},$$

where τ is a learned or heuristic threshold.

Alternatively, semantic shifts can be used. Let $f(\cdot)$ be an embedding function and k the context window. Then the semantic deviation is:

$$d_i = \left\| f(x_i) - \mathbb{E}[f(x_{i-k:i-1})] \right\|.$$

A unit x_i is considered a candidate twist if either s_i or d_i exceeds its respective threshold.

The total twist intensity is then defined as the aggregate score of individual twist:

$$\text{TotalTwist}(X) = \frac{1}{|\mathcal{T}|} \sum_{x_i \in \mathcal{T}} \text{TwistScore}(x_i).$$

3 Data

For this project, we will be collecting thriller short stories from two online sources: Reedsy Prompts and Short Fiction Break. These platforms feature hundreds of community-submitted stories that are freely accessible.

Each short story contains approximately 200–250 sentences, with an average length of 2,000–3,500 words. Across both sources, we estimate having access to around 100 short stories.
Example excerpt (Reedsy Thriller):

> "The air was still. A floorboard creaked upstairs, and his heart stopped. He wasn't supposed to be home yet."

This kind of narrative content makes the dataset well-suited for *plot-twist detection* and *suspense modeling*, since the stories are rich in unexpected turns and emotional intensity. Table 1 summarizes rough statistics of the dataset.

Source	Stories	Avg. sent./story	Total sent.	Words
Reedsy Thrillers	50	200–250	∼11,000	∼135k
Short Fiction Break	50	200–250	∼11,000	∼135k
Combined	100	200–250	∼22,000	∼270k

Table 1: Estimated dataset statistics for thriller short stories.

4 Related Work

Prior work on suspense, surprise, and revelation in narrative has explored these problems from different perspectives. Wilmot and Keller (2020) model suspense as the reduction of possible paths to desired outcomes, using a hierarchical GPT–RNN framework to encode story-level representations and achieve high correlation with human judgments of suspense. However, their approach was computationally expensive and lacked deeper event-structural modeling. Bissell et al. (2025) focused on narrative surprise by constructing a dataset of mystery short stories from Reedsy and evaluating model-generated endings against human-authored ones. They used human annotations across multiple surprise metrics, but their work was limited to ending completion and subject to annotation subjectivity. Finally, Piper et al. (2023) quantified narrative revelation using Kullback–Leibler divergence across adjacent text windows in a large corpus of contemporary prose, capturing dynamics of novelty over time but overlooking global narrative structures.

Our work differs by targeting plot twist detection directly in thriller short stories. Unlike prior approaches that rely on either local language patterns or subjective evaluations of endings, we aim to model the divergence between narrative expectations and realized events in a way that is computationally tractable and aligned with literary theories of suspense. By focusing on shorter thriller texts, our method avoids the scalability issues of long-form modeling while emphasizing surprise in the central turning points of narratives.

5 Evaluation

The proposed system will be evaluated at both the event level, corresponding to individual twist occurrences, and the corpus level, reflecting overall twist intensity. At the event level, standard classification metrics such as F1-score will assess the system's ability to identify twists relative to a manually annotated standard.

At the corpus level, aggregated twist scores will

be compared to human judgments or secondary sources, such as literary critiques, using correlation metrics (eg: Pearson) to assess overall narrative twist intensity. Genre sensitivity will be evaluated by verifying that thriller stories consistently receive higher scores than non-thriller stories.

To contextualize performance, we will establish two baselines: a random baseline assigning twist events uniformly across narrative units, and a simple heuristic baseline, such as labeling longer sentences or passages with rare words as twists. These baselines provide meaningful reference points, ensuring that improvements demonstrate the system's ability to capture genuine narrative surprise rather than relying on trivial cues. Together, these evaluation strategies offer a rigorous framework for assessing both the accuracy and interpretability of the proposed twist detection.

6 Work Plan

The tentative plan of work is as follows:

1. Data Acquisition and Preprocessing: Collect thriller stories and corpora. Clean and segment text into sentences, paragraphs, or chapters.

2. Problem Formalization and Modeling: Define computational criteria for plot twists as deviations from expected narrative progression. Select NLP models for expectation modeling and develop scoring metrics for twist magnitude.

3. Model Implementation and Training: Train on narrative corpora. Implement twist detection using surprisal, embedding divergence, or anomaly detection.

4. Evaluation and Validation: Measure performance with precision, recall, F1-score, and ranking correlations at event and story levels. Compare against random and heuristic baselines.

5. Analysis and Visualization: Visualize twist scores over narrative time. Analyze performance across genres, story lengths, and narrative complexity.

Multi-person Team Justification

The project requires two team members to cover complementary NLP perspectives. One member will focus on the probabilistic approach, modeling narrative expectations to compute surprisal values while the second member will focus on the embedding-based approach, computing semantic deviations using contextual embeddings this division will allow both members to engage in core NLP tasks while exploring distinct, theoretically grounded methods for twist detection.

References

Annaliese Bissell, Ella Paulin, and Andrew Piper. 2025. A theoretical framework for evaluating narrative surprise in large language models. In *Proceedings of the The 7th Workshop on Narrative Understanding*, pages 26–35, Albuquerque, New Mexico. Association for Computational Linguistics.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Andrew Piper, Hao Xu, and Eric D. Kolaczyk. 2023. Modeling narrative revelation. In *CHR 2023: Computational Humanities Research Conference*, Paris, France. CEUR Workshop Proceedings. CEUR-WS, Vol. 3558.

Ted Underwood. 2024. Can language models predict the next twist in a story? Blog post, "The Stone and the Shell".

David Wilmot and Frank Keller. 2020. Modelling suspense in short stories as uncertainty reduction over neural representation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1763–1788, Online. Association for Computational Linguistics.