

Are Large Language Models Capable of Generating Human-Level Narratives?

Yufei Tian^{1*} Tenghao Huang^{2*} Miri Liu¹ Derek Jiang¹ Alexander Spangher²
Muhao Chen^{2,3} Jonathan May² Nanyun Peng¹

¹University of California, Los Angeles, ²University of Southern California

³University of California, Davis

<https://github.com/PlusLabNLP/Narrative-Discourse>

yufeit@cs.ucla.edu tenghao@usc.edu

Abstract

This paper investigates the capability of LLMs in storytelling, focusing on narrative development and plot progression. We introduce a novel computational framework to analyze narratives through three discourse-level aspects: i) story arcs, ii) turning points, and iii) affective dimensions, including arousal and valence. By leveraging expert and automatic annotations, we uncover significant discrepancies between the LLM- and human-written stories. While human-written stories are suspenseful, arousing, and diverse in narrative structures, LLM stories are homogeneously positive and lack tension. Next, we measure narrative reasoning skills as a precursor to generative capacities, concluding that most LLMs fall short of human abilities in discourse understanding. Finally, we show that explicit integration of aforementioned discourse features can enhance storytelling, as is demonstrated by over 40% improvement in neural storytelling in terms of diversity, suspense, and arousal.

1 Introduction

Storytelling serves as an integral part in shaping our understandings of ourselves, our society and our world (Langer, 1942; Kaniss, 1991). As large language models (LLMs) grow in capabilities (Minaee et al., 2024) and are integrated into our daily communicative routines (Kasneci et al., 2023), assessing the narrative structures of the stories they tell is crucial to understanding the ways they will shape our society.

Humans incorporate discourse structures that span local and global levels to captivate audiences, evoke emotions, convey complex messages, and share unique perspectives (Vonnegut, 1995; Van Dijk, 1980). A recent HCI study has pointed to gaps in machine storytelling ability at the global-level: despite being able to craft *fluent* narratives,

*The two authors contributed equally.

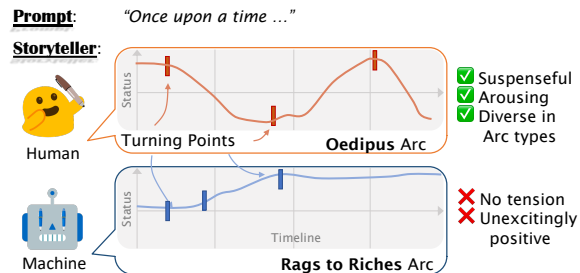


Figure 1: The story arc and turning point positions of human- and LLM- generated narratives. The vertical axis shows the character’s fortune (bad to good), and the horizontal axis represents timeline (beginning to end). Compared with human storytellers, LLMs tend to (1) adopt homogeneously happier, less complex story arcs, (2) introduce plot turning points earlier in the timeline, and (3) have less suspense or fewer setbacks in their storylines. The impact of these differences grow as LLMs gain greater prominence in communicative patterns.

LLMs such as GPT-4 and Claude exhibit plot holes or produce repetitive themes that are less preferred by human critics (Chakrabarty et al., 2024).

However, a computational framework remains to be established for quantitative assessments of narratives at the global or discourse level. We take a step towards the much-desired *analytical framework* and attempt to answer pertinent questions such as: do stories generated by LLMs exhibit the same narrative complexity and diversity as human storytelling? Do LLMs have the capacity to comprehend narrative structures? Concretely, we measure narrative discourse structures at three distinct levels: 1) story arcs (*i.e.*, macro-level narrative development), 2) turning points (*i.e.*, meso-level shifts) and 3) arousal & valence (*i.e.*, micro-level dynamics). We collect a dataset of movie synopses, on which we conduct a wide range of human and automated annotations for each of these levels, with the goals of (1) contrasting LLM and human storytelling, and (2) probing LLM narrative structure comprehension.

First, we explore LLM storytelling abilities. As

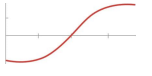
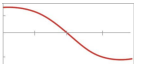
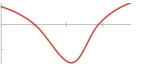
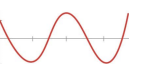
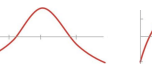


| Rags to Riches | Riches to Rags | Man in a Hole | Double Man in a Hole | Icarus | Cinderella | Oedipus |
|---|---|---|---|--|---|---|
|  |  |  |  |  |  |  |
| Starts low and gradually rises, ending in a high state. | Starts high and gradually falls, ending in a low state. | Starts high, has a dilemma or crisis and finally finds a way out. | Two cycles of fall and rise. | A rise followed by a sharp fall. | A rise, followed by a fall, ending with a significant rise. | A fall, followed by a rise, ending with a significant fall. |

Table 1: **Story Arc Types:** We define and visualize the seven story arc types in our macro-level narrative discourse schema. Story arc types are derived from [Vonnegut \(1995\)](#), and are characterized by transformations of the story’s protagonist(s) across the plot progression.

shown in Figure 1 and Section 4, LLMs such as GPT-4 exhibit notable deficiencies in *narrative pacing*. These models often struggle to adequately develop critical turning points in a story, such as the major setback and climax, diminishing two key qualities for an engaging story: suspense and arousal. Additionally, machines are biased towards certain types of macro-level story-arcs and show a *lack of narrative diversity*, particularly in avoiding negative plot progressions.

Next, we probe LLMs’ *narrative structure comprehension ability* (Section 5) to test the hypothesis that poor comprehension affects LLMs’ narrative generation abilities. To achieve this, we develop two benchmarks, (1) story arc identification and (2) turning point identification in stories, and evaluate Gemini Pro, Claude 3 Opus, Llama3, GPT-3.5 and GPT-4 ([Reid et al., 2024](#); [Anthropic, 2024](#); [AI@Meta, 2024](#); [OpenAI, 2022, 2023](#)). Again, we observe a substantial gap between most LLMs and human abilities, which matches our hypothesis. Interestingly, we find that the different discourse-levels reinforce each other: we can improve turning point identification including story arc information in the input, and vice versa.

Motivated by this finding, we explore whether we can improve machine story-telling by leveraging the aforementioned discourse features, and we find promising results. Incorporating discourse reasoning in prompts can serve as important guidance towards better story generation (Section 6). In two parallel experiments, we demonstrate that integrating awareness of story arcs enhances model diversity (outperforming vanilla prompting by 45%), whereas incorporating turning point information significantly improves narrative suspense and engagement (outperforming vanilla by 40%).

In summary, our contributions are threefold:

1. We unify three levels of discourse in narrative

analysis: story arc, turning point, and affective dimension. Based on this, we present the first quantitative analysis framework to benchmark narrative development, and demonstrate that it can be operationalized by humans on benchmark dataset which we release (§2 and §3).

2. We use this discourse framework to provide a novel comparison of LLM and human generative capacities by examining story-telling (§4) and story-comprehension abilities (§5). We find that LLMs’ abilities fall short of human abilities in both, but especially in story-telling.
3. We demonstrate that a discourse-aware generation process with LLMs (§6) — i.e. incorporating and reasoning about the story arc or turning points—enhances their overall narrative construction, as is reflected in improved suspense, emotion provocation, and narrative diversity. This lays the groundwork for future research to refine models to incorporate complex narrative structures for storytelling and beyond.

2 Background: Discourse in Narratives

We identify three aspects of plot progression in story-telling: story-arcs (macro-level), turning-points (meso-level) and arousal/valence (micro-level), each representing a different level on which storytellers develop their narratives ([Van Dijk, 1980](#)). We describe each of them before describing how we collect data to measure them in stories.

2.1 Three Aspects of Story-Telling

Aspect 1: Story Arcs. A narrative’s story arc charts the transformation of a story’s protagonist(s) across a plot’s progression. [Vonnegut \(1995\)](#) developed a five-part schema to categorize story arcs. Following [Reagan et al. \(2016\)](#); [Wu et al. \(2023\)](#), we adopt an expanded seven-part schema as shown

| Turning Point (TP) | Description |
|---------------------------------|---|
| TP1 - Opportunity | The introductory event that sets the stage for the narrative. |
| TP2 - Change of Plans | A pivotal moment where the main goal of the narrative is defined or altered. |
| TP3 - Point of No Return | The commitment point beyond which the protagonists are invested in goals. |
| TP4 - Major Setback | A critical juncture where the protagonists face significant challenges or failures. |
| TP5 - Climax | The peak of the narrative arc, encompassing the resolution of the central conflict. |

Table 2: **Turning Point (TP) Types:** We describe the 5 TP types in our meso-level narrative discourse schema. A turning point is an event (or plot moment) that significantly influences a plot progression (Papalampidi et al., 2019). These turning points are generally in sequential order in a narrative (*i.e.*, TP1 happens first; TP5 happens last).

in Table 1. This schema captures various positive and negative transitions, such as ‘Rags to Riches’ (*i.e.* a character ascends from adverse conditions to prosperity), or ‘Cinderella’ (*i.e.* a character ascends from adversity, falls and then ascends again). Despite its simplicity, the story-arc classification schema has become a useful tool in writing (Härmä et al., 2021) and computational story-telling research (Reagan et al., 2016; Chu et al., 2017).

Aspect 2: Turning Points. A turning point in a narrative, as conceptualized by Papalampidi et al. (2019), is an event (or more generally a plot-moment) that significantly influences the plot progression. Typically, a turning point represents a protagonist’s transition between rises and falls and serves to demarcate different stages of the plot. Turning points are crucial to narratives for providing a sense of dynamism and maintaining momentum. The types of turning points, identified by Papalampidi et al. (2019), are shown in Table 2. Some, like “Opportunity”, “Change of Plans” and “Point of No Return” are designed to capture exposition, or rising actions of the plot. “Major Setback” further develops the conflict and “Climax” describes the resolution. In general, we consider these last two to be the most important in determining the arc of the story.

Aspect 3: Affective Dimensions. Two affective dimensions: arousal (*i.e.*, the intensity of emotions conveyed in a sentence) and valence (*i.e.*, the positivity or negativity of the emotions expressed) play crucial roles in shaping the emotional impact of narratives (Medhat et al., 2014). This is quantified using the NRC-VAD lexicon (Mohammad, 2018), which provides arousal and valence scores for individual tokens from a 0 to 1 scale. Affective dimensions provide a more nuanced analysis of sentence-level dynamics, capturing subtle shifts in emotional intensity and polarity that may not be fully represented in broader narrative structures, such as story arcs and turning points.

3 Data Collection and Annotation

Films are our culture’s *Gesamtkunstwerk* (or “total work of art”) according to Michelson (1991), and our mass market vehicle for telling narratives (Balio, 2013). Thus, we take films as a basis for exploring the stories our culture tells itself. The narratives we examine are condensed versions of some of the most intricate storylines humans create—those found in movies. While these synopses focus on key plot developments, they should not be considered simple or straightforward. In this section, we will describe first how we built our dataset of film plots, and then we will describe our annotation approach to study the plots’ discourse structures.

3.1 Data Preparation

We crawl the recent English-language films category on Wikipedia¹ to collect the titles, genres, release dates and synopses of these films. To avoid data leakage, we filter out well-known movies using the lengths of Wikipedia pages as an approximate indicator of popularity, resulting in 819 synopses. To further avoid data contamination, we rephrase the titles and initial settings by altering all the unique identifiers such as proper nouns. Finally, we instruct GPT-4 using the rephrased titles, initial settings, and the genres to generate a paired synopsis for each collected film, resulting in 1638 synopses. All human and machine narratives are roughly of the same length.

3.2 Analysis Approaches

Annotating Turning Point and Story Arc We seek to collect human annotations for each synopsis. To do so, we design annotation tasks for input narratives to label each with a story arc, and locate the sentential position where each of the five turning points occurs. We introduce a few key

¹https://en.m.wikipedia.org/wiki/Category:2020s_English-language_films

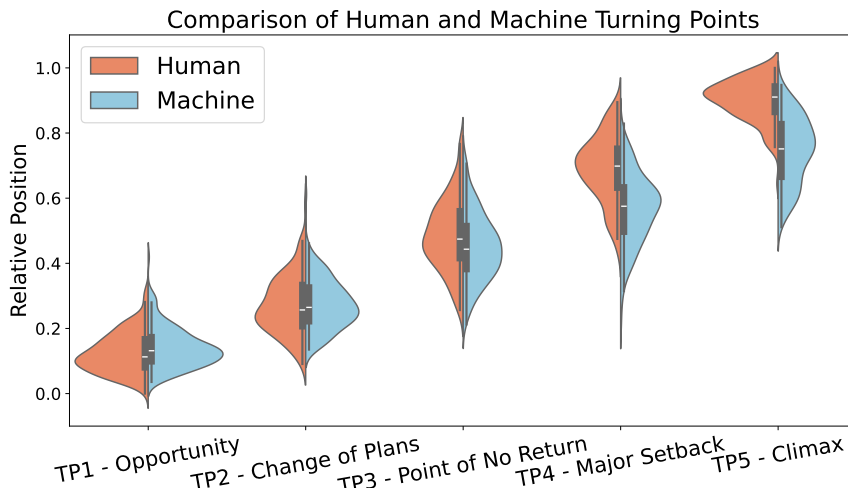


Figure 2: Violin plots showing the positions of five turning points: TP1 - opportunity, TP2 - change of plans, TP3 - point of no return, TP4 - major setback, and TP5 - climax. Relative positions (y-axis) are calculated by $\frac{\text{Index}(\text{TP}_i)}{\text{Total Length}}$. For example, 0.5 means that the turning point occurs exactly in the middle of the whole story. We observe early arrival for TP 4-5 in AI outputs, indicating bad pacing and a lack of intensity.

modifications to the turning point schema of [Papalampidi et al. \(2019\)](#) to handle more complex narratives: 1) flexible positioning for TP3 and TP4, and 2) optional, though discouraged, missing or multiple positions for TP2, TP3, and TP4 when annotators are uncertain.

We recruited 16 annotators who either hold (or are pursuing) a bachelor’s degree in English or have prior experience in story analysis. To ensure the reliability of our annotators, we conducted multiple training sessions to fully onboard our annotators and administered a qualification task, exemplified in Figure 10 in the appendix. We also designed short questions to assist annotators in determining the story arcs more accurately. Two example pairs of human and GPT-4 written narratives, along with corresponding human annotations, are compiled in Appendix B.1. We also detail our annotation guidelines in Appendix C.1.

The narratives we studied have an average of 705.6 words and 37.8 sentences, longer and more complex compared to most other papers quantitatively studying narratives, such as ([Sap et al., 2022](#)) with 300 words and ([Li et al., 2018](#)) with 18.5 sentences. We had a total of 440 narratives annotated, with each narrative annotated by three workers in-depth. The inter-annotator agreements (IAA) for the two tasks are measured at 0.90 (using Spearman’s Correlation) and 0.62 (using Cohen’s Kappa), which indicate a substantial agreement and speaks to the quality of our annotation process. Considering extensive labor for an in-depth human study at scale, our annotators limited their evaluations to stories produced by humans and GPT-

4-0613, one of the most powerful LLMs, which should approximate the upper bound of current LLM capabilities.

Measuring Arousal and Valence We take an agentic analysis of arousal and valence, as in the previous work by [Field et al. \(2019\)](#). To do this, we first instruct GPT-4 to identify the main character of the story. Then for each sentence s_i in a narrative, we ask the same LLM to infer three adjectives, $W_i = \{w_{i1}, w_{i2}, w_{i3}\}$, that describe the protagonist’s emotions as the plot progresses (e.g., amused, relaxed, anxious). We then utilize the NRC VAD lexicon ([Mohammad, 2018](#)) to obtain the arousal and valence scores of w_{ij} ranging from 0 to 1, $j \in \{1, 2, 3\}$. For each sentence, we use the average scores of w_{ij} to represent the arousal and valence of s , obtaining $A(s_i)$ and $V(s_i)$.

In our analysis, a narrative with N sentences is evaluated using the discrete arousal (A) and valence (S) values at the sentence level. These values are plotted on scatter plots with sentence relative position $\frac{i}{N}$ on the x-axis, and $A(s_i)$ or $V(s_i)$ on the y-axis. To facilitate comparison across narratives of varying lengths, we interpolate these plots to generate smooth curves. The mean of these individual curves is then calculated to derive an aggregated curve that represents the arousal or valence across multiple narratives.

4 Human vs. AI Narratives: A Discourse-Level Comparison

Having described our framework for analyzing narratives, our data collection and our measurement

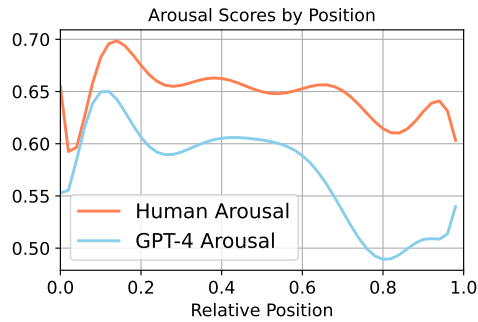


Figure 3: Arousal of human and GPT-4. Human stories consistently exhibit higher levels of suspense (greater arousal). The gap enlarges from the midpoint to the end.

approaches, we now describe insights we derived.

LLMs incorrectly pace their storytelling relative to human writers. Figure 2 shows paired violin plots comparing the sentential position of turning points in human- and AI-generated stories. As shown, while the positioning of TP1 through TP3 is consistent between human and AI narratives, we observe a substantial advancement (*i.e.*, early occurrence) of TP4 and TP5 in AI outputs. This suggests that while LLMs grasp the correct pacing to establish the initial setup (TP1, opportunity) and introduce the main goal (TP2, change of plans), they still *struggle to unfold the narrative’s most crucial junctures adequately* : major setback (TP4) and climax (TP5).

Poor pacing leads to flat narratives without suspense. The pacing we observed in AI narratives, as discussed prior, is unnatural compared to human writers. It often results in less narration being spent on the last two turning points in a story (*i.e.* Major Setback and Climax). Anecdotally, we notice that when these two elements are introduced briefly and then resolved rapidly, the resulting arc feels flatter less exciting, and is more lacking in intensity. To further verify this hypothesis, we draw arousal curves in Figure 3 to visualize the suspense level throughout the whole story. We find that human-written stories consistently exhibit higher levels of suspense (*i.e.*, greater arousal), but the gap begins to enlarge as the plot progresses from the midpoint (0.5 relative position) towards the end. All these observations indicate that AI-generated stories tend to be less arousing and lack suspense, especially after the introductory events are established and the action begins to build.

LLMs are biased towards story arcs with positive endings and lack narrative diversity.

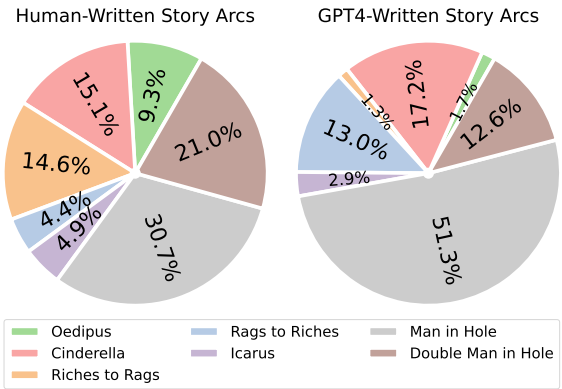


Figure 4: The share of story arcs between human and GPT-4 generated stories show significant differences. GPT-4 is much more likely to generate story arcs with less inflections and happier endings than human stories.

Pie charts in Figure 4 contrast the share of story arcs between human and GPT-4 generated stories. Notably, GPT-4 augments the human bias, by writing Man in Hole, the most popular arc type in human stories, more than half of the time.

Moreover, story arcs that traditionally end negatively, such as Riches to Rags (gradual fall) and Oedipus (fall then rise then fall), which represent 14.6% and 9.3% of human narratives, are almost missing in GPT-4 outputs (1.3% and 1.7%). On the other hand, Rags to Riches (gradual rise), which is scarcely found among human stories (4.4%), now disproportionately accounts for 13.0% of AI-generated stories. Such patterns lead to the conclusion that *LLMs such as GPT-4 exhibit a distinct bias, strongly favoring positive outcomes and avoiding negative plot progressions*. One possible explanation is that the effect of RLHF on an LLM’s language distribution pushes it more towards a positive, helpful generative stance. Figure 5 also shows human-written stories contain more setbacks or negative events (less valence) while GPT-4 narratives are much more positive. Similar to arousal curves, the gap is more pronounced from the midpoint to the ending of the story.

5 Benchmarking Narrative Comprehension

We hypothesize that poor narrative comprehension of LLMs lead to their poor generative outcomes, as much evidence exists for these skills being tied (Collobert and Weston, 2008; Raffel et al., 2020). Therefore, we designed and conducted two benchmark tests to measure narrative reasoning. We start by outlining our tasks (§ 5.1) and the methodologies employed to evaluate performance (§5.2). We

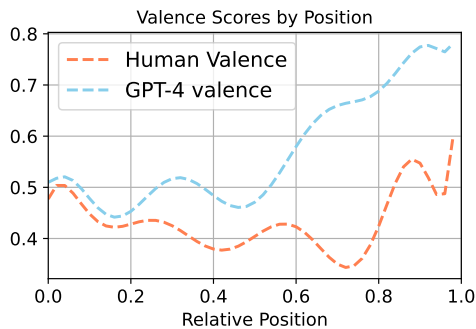


Figure 5: Valence of human and GPT-4. Human-written stories have more setbacks than GPT-4 (lower valence). The gap enlarges from the midpoint to the end.

finally report the benchmarking results over popular foundation LLMs (§ 5.3).

5.1 Benchmark Tasks

Task 1: Story Arc Identification Given a narrative text, our primary task is to classify it into one of several predefined story arcs. This task tests the ability of the model to understand and categorize overarching narrative structures. The effectiveness of the model is measured by its accuracy in matching these arcs against expert annotations.

Task 2: Turning Point Identification Formally, this task can be defined as follows. Given a sequence of n sentences $S = \{s_1, s_2, \dots, s_n\}$ that make up the narrative, the model needs to determine a set of five turning points $T = \{t_1, t_2, t_3, t_4, t_5\}$, where each t_i is a tuple (p, d) . Here, p denotes the position of the sentence within S representing the turning point, and d is a label from the predefined set of turning point types.

Task Variants We formulate two settings for each task: (1) we seek to identify turning points or story arcs given *just* the text of the narrative (2) we give the model additional discourse-level features to aid in each task. Prior research has found that additional discourse information can improve narrative reasoning (Spangher et al., 2021, 2024a): we hypothesize that macro-level story discourse and meso-level information are related. More specifically, for turning point identification, *information about the overarching story arc type is provided*. Conversely, when identifying story arcs, *descriptions of key turning points within the narrative are included*. To assess how well models are able to identify story arc and turning points, we compare the model’s classifications with ground truth annotations provided by human experts.

| Model | TP1 | TP2 | TP3 | TP4 | TP5 | Avg. |
|--------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Human | 59.6 | 40.3 | 37.0 | 45.4 | 50.4 | 46.6 |
| Gemini | 40.5 | 27.3 | 15.5 | 29.4 | 43.8 | 31.3 |
| GPT-4 | 43.9 | 20.1 | 13.8 | 23.3 | 25.4 | 25.3 |
| GPT-3.5 | 28.7 | 19.5 | 8.2 | 14.9 | 23.1 | 18.8 |
| Claude | 46.5 | 24.5 | 16.3 | 30.1 | 35.7 | 30.6 |
| Llama3 | 24.6 | 14.4 | 9.2 | 21.0 | 32.3 | 20.3 |
| <i>with arc as prior</i> | | | | | | |
| Gemini | 38.0 | 26.1 | 12.7 | 26.1 | 40.8 | 28.7 |
| GPT-4 | 38.2 | 27.6 | 13.2 | 30.3 | 27.6 | 27.3 |
| GPT-3.5 | 34.6 | 16.0 | 5.1 | 11.5 | 19.9 | 17.4 |
| Claude | 47.4 | 27.3 | 16.9 | 27.9 | 33.1 | 30.5 |
| Llama3 | 33.5 | 15.5 | 11.0 | 20.1 | 31.0 | 22.2 |

Table 3: The success rates of five language models and humans on the task of turning point identification, presented as percentages (%). The five turning points are TP1 - Opportunity, TP2 - Change of Plans, TP3 - Point of No Return, TP4 - Major Setback, TP5 - Climax. We use boldface to denote the best machine result.

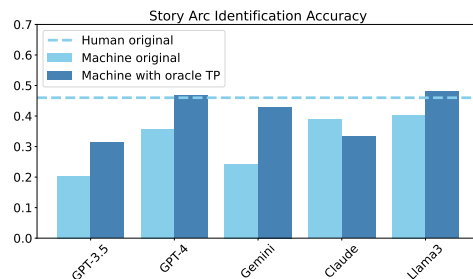


Figure 6: Story arc identification accuracy. Human judgements (blue line) are made without access to turning point information. Language models approach human accuracy *only* when provided with such ground truth information, indicating the *conceptual overlaps* in these discourse structures.

5.2 Models

We collect classifications from multiple state-of-the-art language models, including GPT-3.5, GPT-4,² Gemini 1.0 Pro, Claude3, Llama3-8B (Touvron et al., 2023). For turning point identification, we instruct a model to analyze and tag key turning points in a movie synopsis with explanations. To enhance the model’s counting ability, all narratives are tagged with a sentence index. The exact prompts used are shown in Appendix D.2. We use accuracy as the metric to measure how well the models’ predictions align with expert annotations.

5.3 Benchmark Findings

Larger, closed models identify turning point identification with higher accuracy Table 3 reports the model performance on the turning point identification task without incorporating story arc

²April 29th, 2024 version

| | SUSPENSE | | | EMOTION PROVOKING | | | OVERALL PREFERENCE | | |
|----------------------------|--------------|--------|--------------|-------------------|--------|--------------|--------------------|--------|--------------|
| | Best (↑) | Medium | Worst (↓) | Best (↑) | Medium | Worst (↓) | Best (↑) | Medium | Worst (↓) |
| <i>Outline-Only</i> | 7.9% | 10.1% | 82.0% | 14.6% | 24.7% | 60.7% | 13.5% | 25.8% | 60.7% |
| + <i>Self-generated TP</i> | 48.3% | 42.7% | 9.0% | <u>39.3%</u> | 42.7% | 18.0% | <u>43.8%</u> | 37.1% | 19.1% |
| + <i>Human TP</i> | <u>46.1%</u> | 42.7% | <u>11.2%</u> | 48.3% | 28.1% | <u>23.6%</u> | 44.9% | 32.6% | <u>22.5%</u> |

Table 4: Human evaluated results in suspense, emotion provocation, and overall preference. We compare machine generations with and without the awareness of turning points (TP3, TP4, and TP5).

information as a prior. Gemini and Claude demonstrate the highest performance, with average accuracies over 30%, respectively. GPT-4 also performs reasonably well with an average accuracy of 26%. However, GPT-3.5 and Llama3 lag behind. All models perform below human levels, emphasizing the challenge of accurately identifying story arcs using current LLMs.

Story arc identification also lags human performance Figure 6 shows each model’s performance on identifying story arc types. The original model performances (light blue bars) reveal that accuracy is generally low across all models. For instance, GPT-3.5 achieves an exact agreement score of approximately 0.2 (random guessing being $\frac{1}{7} = 0.14$). GPT-4, Claude, and Llama3 perform better, with exact agreements above 0.35. Similar to turning point, human still achieves higher accuracy than the LLMs without additional knowledge.

Incorporating additional discourse information improves model comprehension We find that finer-grained information will benefit the coarse-grained task more than the reverse. For example, when the ground-truth, macro-level discourse tag (*i.e.*, story arc) is provided to the meso-level task (*i.e.*, turning point identification), the average accuracy of a few LLMs (GPT-4 and Llama3) improves by 2%. However, not all models benefit from such hints. On the other hand, the dark blue bars in Figure 6 demonstrate a significant improved performance in story arc identification when turning point information is given across all models. Both results support our hypothesis that *incorporating discourse-level features can enhance the machine’s narrative reasoning capabilities*.

6 Towards Better Machine Storytelling

Finally, we investigate whether incorporating the discourse aspects into the generation stage enhances machine’s storytelling ability.

Reasoning about TPs improves overall narrative construction, including reduced plot holes

and enhanced suspense and emotion provocation. Motivated by our observations that a major flaw in vanilla LLM story-telling is narrative pacing (in §4) we hypothesize that integrating discourse features can improve pacing and significantly improve narratives.

We test three variations of a planning-first (Yao et al., 2019) approach (*i.e.* generating first the outline and then the narrative). Each variation incorporates different degrees of explicit structure.

- *Outline-Only*: We simply instruct the model to generate an outline, then expand it to a full story.
- + *Self TPs*: We instruct the LLM to generate an outline that marks the 3rd, 4th, and last turning points (*i.e.* “Point of No Return”, “Major Setback”, and “Climax”), with their detailed definitions, and then to write the full-length story.
- + *Human TPs*: We replace the machine-generated major setback and climax with the oracle, human-crafted equivalents (which are typically more compelling and intriguing than their machine-generated counterparts).

We annotate comparatively, ranking three randomly shuffled narratives in terms of suspense, emotion provocation, and overall preference. Table 4 shows win-rates over the above three approaches. Both + *Self TP* and +*Human TP* achieve significant gains over *Outline-Only*, highlighting the efficacy of incorporating TPs in LLM-generated narratives. Interestingly, we find that while +*Human TP* scored highly, especially for emotional engagement (48.3%), it is not significantly preferred over + *Self TP*. Upon further investigation, we realize that the enforcement of external events in +*Human TP* could disrupt the machine’s narrative flow, making the whole plot illogical at times. + *Self TP*, which maintained the natural flow of LLM with its own generations, emerged as the most balanced and least disliked approach. This indicates that future work in the domain of human-machine collaborative writing must be careful to integrate human creativity in beneficial ways.

| Requested Arc | Acc. | Requested Arc | Acc. |
|--------------------|------|---------------|------|
| Cinderella | 33% | Oedipus | 64% |
| Riches to Rags | 33% | Icarus | 67% |
| Double Man in Hole | 54% | Man in Hole | 71% |
| Rags to Riches | 57% | Average | 54% |

Table 5: GPT-4 shows poor accuracy in generating narratives with specified story arc types, although is better for arcs that have one inflection point (e.g. “Man in the Hole”) compared with two (e.g. “Cinderella”).

| Diversity | THEME | SETTING | CONFLICT | CHARACTER | OVERALL |
|---------------------|------------|------------|------------|------------|------------|
| <i>Outline-Only</i> | 5% | 32% | 5% | 23% | 23% |
| <i>Tie</i> | 32% | 36% | 41% | 27% | 9% |
| <i>Arc-Enhanced</i> | 64% | 32% | 55% | 50% | 68% |

Table 6: Win rates of the outline-only stories and story-arc enhanced stories. We focus on four specific aspects of diversity: theme, setting, conflict, and character.

Incorporating explicit directives about story arcs helps improve narrative diversity Motivated by our observations, in figure 4, that LLM generations lack arc-level diversity, we explore whether explicit instruction can induce a more human-like story arcs. We design another variant, *Arc Enhanced*, that explicitly instructs the model to generate story with a specified story arc, specifies the number of major rises and falls and details the initial and ending state of the protagonists.

First, we evaluate how well LLMs are able to follow the requested story-arcs. The results, in Table 5, show that GPT-4 achieves an average success rate below 55%, suggesting that LLMs’ capability to mirror human narrative distributions is limited, even with explicit instructions. Notably, these models struggle with story arcs that depict negative plot progressions (e.g., riches to rags), humble starts (e.g., cinderella), and those with complex narrative dynamics (e.g., double man in hole).

Next, we compare the narrative diversity between sets of *Arc Enhanced* stories and *Outline-Only* stories³. As seen in Table 6, *Arc Enhanced* significantly outperforms *Outline-Only* across most aspects of diversity that we considered: thematic, conflict-type, and character. We conclude that story arc discourse is a significant driver of many aspects of narrative diversity, affirming the basic truth

³We instruct annotators to examine diversity in the following aspects: 1) **Thematic**: the central ideas conveyed; 2) **Setting**: when and where these stories take place; 3) **Conflict type**: including but not limited to character with self, other characters, society, nature, technology, fate; 4) **Character**: including but not limited to personality, background, development, and relationships

of [Vonnegut \(1995\)](#)’s assertion that stories can be broadly categorized into story arc types.

7 Qualitative Study: Understanding Human Preferences

After completing all annotation tasks, we conducted two follow-up interviews to gather qualitative feedback from our annotators. These interviews focused on annotators who worked on the task in § 6 — reading pairs of machine generated narratives (*Outline-Only* vs *Arc-Enhanced*) and examining diversities plus overall preference. They were encouraged to freely provide justifications or comments on any of the readings. After the interviews, we reconstructed their feedback and presented representative comments in Figure 7 and Figure 9. Overall, the human annotators prefer concrete narratives with twists in plot development that are logical and well-motivated. They dislike straightforward, positive plots or those with ‘miracle turns’ that are not adequately justified.

8 Related Work

Discourse-Aware Evaluation. In contrast to conventional story evaluation frameworks, which primarily focus on fluency and coherence, *discourse-aware* evaluation focuses on critiquing the structural and creative quality of machine-generated content ([Harel-Canada et al., 2024](#); [Spangher et al., 2024b](#); [Tian et al., 2024](#)). [Liu et al. \(2024\)](#) introduced a model that assesses stories by embedding conventional narrative structures within the evaluation process. Complementing this, [Chakrabarty et al. \(2024\)](#) explore narrative differences between humans and AI through a qualitative study and ([Li et al., 2024](#)) reveal that RLHF-aligned language models are less diverse than the base LMs.. [Bergus \(2023\)](#) delves into the creative outputs of LLMs, questioning their true creativity versus their capacity to merely replicate observed patterns, which encourages further exploration of the creative limits of these models. Additionally, [Wang et al. \(2023\)](#) introduces the Positional Discourse Coherence metric to quantitatively assess logical narrative progression.

However, prior works have been limited by the vague definitions of creativity and discourse structure. We take inspiration from the literature theories of discourse analysis of drama and fiction ([Labov and Waletzky, 1967](#); [Vonnegut, 1995](#)). For

| |
|---|
| <p>Annotator General Comment</p> <p>(pointing to multiple GPT-4 generated stories) ... The authors are unwilling to put characters in real risk in the sense that characters make a plan and everything goes according to plan. Moreover, oftentimes conflicts are just resolved by "talking it out" or some "community effort" which I dislike because it doesn't allow the characters to grow stronger or improve themselves.</p> |
| <p>Excerpt (Arc Enhanced)</p> <p>... As the fire raged around him, Tom stumbled through the manor, the woman in white appearing amidst the flames, guiding him deeper into the inferno ... In his final moments, Tom realized the woman in white was not guiding him to safety, but delivering him to his fate.</p> <p>Annotator Comment</p> <p>This was a neat little twist which was especially unexpected in a AI story since others tend to be very straightforward and extremely positive.</p> |
| <p>Excerpt (Arc Enhanced)</p> <p>... She turns to Alex, her voice steady with conviction, "Let's go shoot the sunset" a daily ritual that has become their shared passion. Together, they head towards the horizon, their steps light and sure, the camera slung over Alex's shoulder, ready to capture the golden hues painting a new chapter in their lives...</p> <p>Annotator Comment</p> <p>... this was a nice example of the kind of specificity I mean wrt dialogue -- this is a very specific reference to photography which other stories didn't include.</p> |
| <p>Annotator General Comment</p> <p>For the stories [pointing to multiple set ID], I ultimately couldn't choose one that was better because both had very positive endings, which was annoying to me and this was probably my least favorite set for that reason.</p> |

Figure 7: Human annotators' feedback on machine-generated stories when comparing the *Outline-Only* vs *Arc-Enhanced* strategy. They were blind to the prompting strategies and all presented stories were randomly shuffled. We reconstruct their comments and color-code with green for favorable ones and red for unfavorable ones. Continued in Figure 9.

instance, Freytag (1894) proposed a a five-part dramatic structure, now commonly understood as Exposition, Rising Action, Climax, Falling Action, and Dénouement. Li et al. (2018) combined such literature theories and annotated story macrostructures. In non-fiction storytelling like news writing, Choubey et al. (2020); Spangher et al. (2021, 2022) demonstrated that language models can classify similar elements, with artificial news differing from human-generated content. Likewise, we define specific story arcs and turning points in creative stories and examine how stories generated by LMs structurally differ from human ones.

Discourse-Aware Generation with LLMs. Attempts to incorporate discourse features into story

generation include Yao et al. (2019); Han et al. (2022); Yang et al. (2022) that focus on generating coherent, logical, and interesting stories. Huang et al. (2023) incorporates affective dimensions to foster the creation of more captivating stories. Brei et al. (2024) examines the efficacy of "bookends" as a structural enhancement in narrative quality. Further studies have also ventured into embedding elements of suspense to forge more engaging narratives (Zehe et al., 2023; Xie and Riedl, 2024). Different from previous endeavors, our study enhances narrative construction through the systematic incorporation of discourse elements, similar to Spangher et al. (2022) who focus on news structures. Our approach seeks to bridge the gap between human-like storytelling and the capabilities of current AI systems through three levels of discourse elements.

9 Conclusion

This work aims to advance the understanding and generation of narratives through the lens of three discourse elements: story arcs at macro-level, turning points at meso-level, and affective dimensions at micro-level. We contribute an expert-annotated dataset, based on which we conducted quantitative comparison between human and AI in terms of narrative generation and comprehension: LLMs fall short especially in story writing. We find models lack narrative diversity, and struggle at develop crucial turning points, such as major setback and climax, leading to less engaging stories. We also show promising results that discourse-aware generation improves AI's story-telling ability in terms of suspense, emotion engagement, and narrative diversity.

We view our effort as a useful starting point towards a systematic analysis of narrative discourse. We hope the collected dataset and experimental results, along with our proposed perspective, will attract wider academic interest in discourse studies and provide insights into better narrative generation, comprehension, and evaluation.

Limitations

For both discourse-level comparison (§ 3) and better machine storytelling (§ 6) which require in-depth human annotation, we limit our experiment to GPT-4 generated narratives. While we believe the conclusions are applicable to other LLMs such as Claude, Llama, Gemini, etc., their generations are not direct assessed.

Another limitation is that our research primarily focuses on English-based LLMs and resources. Our initial focus on English allows us to establish the discourse-level framework before expanding to others. Future research can look into expanding this scope to include multilingual language models and diverse linguistic resources. This expansion could help to better understand and predict flavors across different cultural and linguistic contexts, potentially uncovering unique insights and flavor combinations that are specific to various cuisines and regional preferences.

Earlier studies, such as [Huang et al. \(2021\)](#), have identified the presence of gender biases within this dataset. Consequently, we want to point out that generating stories using our pipeline may also be at risk of perpetuating and intensifying these biases.

Acknowledgement

The authors would like to thank the PlusLab members at UCLA and the anonymous reviewers for their valuable feedback and helpful discussions. This research is partly supported by National Science Foundation CAREER award #2339766, an Amazon AGI foundation research award, and a Google Research Scholar grant. Tenghao Huang and Jonathan May are supported by the Defense Advanced Research Projects Agency with award HR00112220046.

References

- AI@Meta. 2024. [Llama 3 model card](#).
- AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1.
- Tino Balio. 2013. A major presence in all of the world’s important markets: The globalization of hollywood in the 1990s. In *Contemporary Hollywood Cinema*, pages 58–73. Routledge.
- Nina Begus. 2023. Experimental narratives: A comparison of human crowdsourced storytelling and ai storytelling. *arXiv preprint arXiv:2310.12902*.
- Anneliese Brei, Chao Zhao, and Snigdha Chaturvedi. 2024. [Returning to the start: Generating narratives with related endpoints](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 101–112, Mexico City, Mexico. Association for Computational Linguistics.
- 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.
- Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Eric Chu, Jonathan Dunn, Deb Roy, Geoffrey Sands, and Russell Stevens. 2017. Ai in storytelling: Machines as cocreators. *McKinsey & Company Media & Entertainment*.
- Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167.
- Anjalie Field, Gayatri Bhat, and Yulia Tsvetkov. 2019. Contextual affective analysis: A case study of people portrayals in online# metoo stories. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 158–169.
- Gustav Freytag. 1894. *Die technik des dramas*. S. Hirzel.
- Rujun Han, Hong Chen, Yufei Tian, and Nanyun Peng. 2022. Go back in time: Generating flashbacks in stories with event temporal prompts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1450–1470.

- Fabrice Y Harel-Canada, Hanyu Zhou, Sreya Muppalla, Zeynep Senahan Yildiz, Miryung Kim, Amit Sahai, and Nanyun Peng. 2024. Measuring psychological depth in language models. In *Proceedings of The 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Kimmo Härmä, Sirpa Kärkkäinen, and Eila Jeronen. 2021. The dramatic arc in the development of argumentation skills of upper secondary school students in geography education. *Education Sciences*, 11(11):734.
- Qiyue Hu, Bin Liu, Mads Rosendahl Thomsen, Jianbo Gao, and Kristoffer L Nielbo. 2021. Dynamic evolution of sentiments in never let me go: Insights from multifractal theory and its implications for literary analysis. *Digital Scholarship in the Humanities*, 36(2):322–332.
- Tenghao Huang, Faeze Brahman, Vered Shwartz, and Snigdha Chaturvedi. 2021. [Uncovering implicit gender bias in narratives through commonsense inference](#). *ArXiv*, abs/2109.06437.
- Tenghao Huang, Ehsan Qasemi, Bangzheng Li, He Wang, Faeze Brahman, Muhao Chen, and Snigdha Chaturvedi. 2023. [Affective and dynamic beam search for story generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11792–11806, Singapore. Association for Computational Linguistics.
- Phyllis Kaniss. 1991. *Making local news*. University of Chicago Press.
- Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.
- William Labov and Joshua Waletzky. 1967. Narrative analysis. In *Essays on the Verbal and Visual Arts*. University of Washington Press, Seattle, WA.
- Susanne K Langer. 1942. *Philosophy in a new key: A study in the symbolism of reason, rite, and art*. Harvard University Press.
- Boyang Li, Beth Cardier, Tong Wang, and Florian Metzger. 2018. Annotating high-level structures of short stories and personal anecdotes. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Margaret Li, Weijia Shi, Artidoro Pagnoni, Peter West, and Ari Holtzman. 2024. Predicting vs. acting: A trade-off between world modeling & agent modeling. *arXiv preprint arXiv:2407.02446*.
- Yinhong Liu, Yixuan Su, Ehsan Shareghi, and Nigel Collier. 2024. Unlocking structure measuring: Introducing pdd, an automatic metric for positional discourse coherence. *arXiv preprint arXiv:2402.10175*.
- Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.
- Annette Michelson. 1991. "where is your rupture?": Mass culture and the gesamtkunstwerk. *October*, 56:43–63.
- Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.
- Saif Mohammad. 2018. [Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.
- OpenAI. 2022. [Introducing chatgpt](#).
- OpenAI. 2023. [Gpt-4 technical report](#).
- Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2019. [Movie plot analysis via turning point identification](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1707–1717, Hong Kong, China. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ data science*, 5(1):1–12.
- Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Maarten Sap, Anna Jafarpour, Yejin Choi, Noah A Smith, James W Pennebaker, and Eric Horvitz. 2022. Quantifying the narrative flow of imagined versus autobiographical stories. *Proceedings of the National Academy of Sciences*, 119(45):e2211715119.
- Alexander Spangher, Kung-Hsiang Huang, Hyundong Cho, and Jonathan May. 2024a. Newsdits 2.0: Learning the intentions behind updating news.

- Alexander Spangher, Jonathan May, Sz-Rung Shiang, and Lingjia Deng. 2021. Multitask semi-supervised learning for class-imbalanced discourse classification. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 498–517.
- Alexander Spangher, Yao Ming, Xinyu Hua, and Nanyun Peng. 2022. Sequentially controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6848–6866.
- Alexander Spangher, Nanyun Peng, Sebastian Gehrmann, and Mark Dredze. 2024b. Do llms plan like human writers? comparing journalist coverage of press releases with llms. In *Proceedings of The 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Yufei Tian, Zeyu Pan, and Nanyun Peng. 2024. Detecting machine-generated long-form content with latent-space variables. *Findings of The 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Teun A Van Dijk. 1980. *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition*. Lawrence Erlbaum.
- Kurt Vonnegut. 1995. *Shapes of stories*. Accessed June 2024.
- Danqing Wang, Kevin Yang, Hanlin Zhu, Xiaomeng Yang, Andrew Cohen, Lei Li, and Yuandong Tian. 2023. Learning personalized story evaluation. *arXiv preprint arXiv:2310.03304*.
- Winston Wu, Lu Wang, and Rada Mihalcea. 2023. Word category arcs in literature across languages and genres. In *Proceedings of the The 5th Workshop on Narrative Understanding*, pages 36–47.
- Kaige Xie and Mark Riedl. 2024. *Creating suspenseful stories: Iterative planning with large language models*. In *Conference of the European Chapter of the Association for Computational Linguistics*.
- Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4393–4479.
- Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.
- Albin Zehe, Juliane Schröter, and Andreas Hotho. 2023. *Towards a computational analysis of suspense: Detecting dangerous situations*. *ArXiv*, abs/2305.06818.

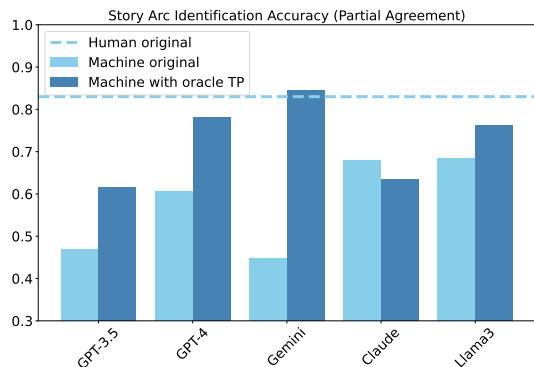


Figure 8: Story arc fuzzy matching results.

A Additional Results

Due to the level of abstraction from detailed stories to arcs, we recognize that even human annotators experience discrepancies when categorizing a story into a specific story arc type (Hu et al., 2021). For example, what one person might dismiss as a moderate obstacle that will *not* affect the overall arc could be perceived by another as more significant, leading to variations in the number of falls assigned to the chosen story arc. To address this, we identified story arcs that by nature blur together or are easily confused, which we call “hard pairs” as illustrated in Table 7. We then employed fuzzy matching to assess the models’ capability in identifying story arcs, by granting credit to a model if its predicted label falls within the hard pairs, even if it is not an exact match with the ground truth. Figure 7 presents the results of the fuzzy matching on story arc identification. Under this setting, all models perform significantly below human levels, highlighting the inherent challenges in accurately identifying story arcs using LLMs. It is worth noting that the performance gap is **larger** than story arc exact match results, as shown in Figure 6.

We also apply a similar fuzzy matching approach to the task of turning point identification. Instead of crediting the model exactly selecting the groundtruth sentence for a turning point, we now credit a model for choosing a sentence that is within ± 3 positions of the groundtruth sentence. We report turning points fuzzy matching results in Figure 8. We observe that the overall trend is not significantly different compared to the results from exact matches.

B Examples of Generated Narratives and Annotations

B.1 Examples of standard annotation

We exemplify two pairs of human and machine written narratives along with their marked annota-

| |
|--|
| <p>Excerpt (Arc Enhanced) <i>...Mike immediately suggested calling 911, but James hesitated, voicing concerns about potential racial implications and the mistrust between law enforcement and African-American communities...</i></p> <p>Annotator Comment <i>I noted that in the set of [Outline-Only] stories, the LLM definitely really leaned away from suggesting fear of racial bias as a motive for the characters. However the fact that the Set of [Arc-Enhanced] stories did include this motivation made the characters feel more real and authentic, even if the characters acted the same in each of the stories. Basically I cared less about character diversity in motivations here because it felt realer.</i></p> |
| <p>Excerpt (Arc Enhanced) <i>... And so, amidst the debris of a life once dreamed, Ben found his true voice, illustrating that sometimes, the deepest falls lead to the most poignant tales. [End]</i></p> <p>Annotator Comment <i>This is the final few lines of the "miraculous turn" I was talking about and dislike. I think in this story basically everyone suffers some kind of life-altering tragedy and the main character is put in jail but the AI justifies it at the end with "the deepest falls lead to the most poignant tales" which I really disliked and this swayed me to [Outline-Only] instead.</i></p> |
| <p>Annotator Comment <i>In stories [ID], ..., what struck me most about the two sets and what led me to choose [Arc-Enhanced] was that there were very few obstacles for the main character in [Outline-Only] -- whereas in [Arc-Enhanced] I believe [character name] relapses once or twice in some of the stories, in [Outline-Only] stories it's generally a straight line upwards. This was disappointing to me, ...</i></p> |

Figure 9: Figure 7 Continued. Prior to the interviews and after the annotation tasks were completed, annotators were already informed that the presented stories were generated by AI, but the specific methods and models were hidden from them.

tions in Table 9 to Table 12. In both cases, human-written narratives have more suspenseful and arousing events; the major setbacks and climax arrive earlier in machine generated narratives.

B.2 Detailed feedback from interviewed annotators

We are lucky to have conducted interviews and collected detailed feedback from two annotators who worked for the last task— reading pairs of machine generated narratives (*Outline-Only* vs *Arc-Enhanced*) and examining diversities plus overall preference (§ 6). Before the interview, annotators were told that all the presented stories were generated by AI, but with different methods which were hidden from them. They were asked to freely provide any justification

| Story Arc | Hard Label Pairs |
|--------------------|---|
| Man in Hole | Double Man in Hole, Cinderella |
| Double Man in Hole | Man in Hole, Cinderella |
| Cinderella | Rags to Riches, Man in Hole, Double Man in Hole |
| Rags to Riches | Cinderella |
| Riches to Rags | Oedipus |
| Oedipus | Riches to Rags |

Table 7: Hard label pairs for story arcs.

| Model | TP1 | TP2 | TP3 | TP4 | TP5 |
|--------------------------|-------------|-------------|-------------|-------------|-------------|
| Human | 88.2 | 72.3 | 68.9 | 77.3 | 82.4 |
| GEMINI | 89.2 | 65.5 | 60.8 | 65.5 | 83.0 |
| GPT4 | 82.5 | 54.5 | 50.8 | 58.2 | 71.4 |
| GPT35 | 78.5 | 57.4 | 40.5 | 51.8 | 74.4 |
| CLAUDE | 87.2 | 64.8 | 51.0 | 60.7 | 79.6 |
| LLAMA3 | 73.8 | 46.7 | 43.6 | 59.5 | 79.5 |
| <i>with arc as prior</i> | | | | | |
| GEMINI | 85.2 | 69.0 | 53.5 | 62.7 | 82.4 |
| GPT4 | 78.9 | 78.9 | 57.9 | 60.5 | 76.3 |
| GPT35 | 78.2 | 52.6 | 32.7 | 44.9 | 76.3 |
| CLAUDE | 87.7 | 68.8 | 55.8 | 62.3 | 77.3 |
| LLAMA3 | - | - | - | - | - |

Table 8: The fuzzy matching success rates of five language models and humans on the task of turning point identification, presented as percentages (%).

or comment on any of the readings. We reconstruct their comments and report representative ones in Figure 7 and Figure 9. Overall, the human annotators prefer concrete narratives with twists in plot development that are logical and well-motivated. They dislike straightforward, positive plots or those with ‘miracle turns’ but are not adequately justified.

C Experimental Details

C.1 Human Annotation Interfaces

Recall that in § 3 we design annotation tasks for input narratives to 1) label each with a story arc, and 2) locate the sentential position where each of the five turning points occurs. An example of the task interface can be found in Figure 10.

Figure 11 to Figure 14 list the detailed annotation guideline and examples of story arc categorization. Figure 15 to Figure 18 list the detailed annotation guideline and examples of turning point identification.

D Prompt details

D.1 Prompts Used in Data Preparation

We consider the introductory part (first 1-3 sentences in the human-written narrative) as the initial setting. We asked GPT-4 to rephrase the setting by replacing all proper nouns (names, places, anything unique), and then change the phrasing slightly to return the new setting. Based on the newly rephrased initial setting and the original title, we asked the model to generate a similar but not identical title.

D.2 Prompts Used in Benchmarking

We show the prompt of story arc identification task in Figure 19, and prompt of turning point identification task in Figure 20.

Table 9: Example 1 of human written narratives and the annotated story arc, turning points.

| Source: Human Title: The Dark and the Wicked Genre: Horror Annotated Story Arc: Riches to Rags | |
|---|--|
| 1 | Siblings Louise and Michael return to their family farm in Texas when their father's chronic illness seems to be reaching its last stages. |
| 2 | Their mother seems disturbed at their arrival, and expresses a desire for the children to leave. |
| 3 (tp1) | That night, she hangs herself in the barn after (apparently involuntarily) cutting off her own fingers in the kitchen. |
| 4 | As time goes on, Louise and Michael start to understand what happened to their mother. |
| 5 | Their father's nurse confides in them that she heard their mother whispering to their father, but it seemed as if she was speaking not to him, but some other presence. |
| 6 (tp2) | Michael finds their mother's diary, which describes her fears of an unnamed and possibly demonic presence preying on her husband. |
| 7 | At their mother's burial, Louise and Michael meet Father Thorne, a priest who claims to have known their mother. |
| 8 | Later that night, Father Thorne appears at the farm, beckoning them from outside, before vanishing before their eyes. |
| 9 | Meanwhile, Charlie, a ranch hand who lives on a nearby plot of land in his RV, witnesses a vision of what appears to be Louise, speaking indistinctly and cutting herself repeatedly with a kitchen knife. |
| 10 | The entity drives a distraught Charlie to shoot himself in the head with his shotgun. |
| 11 | Louise is subsequently unable to reach Charlie by phone, unaware that he is dead. |
| 12 | Louise calls the phone number that Father Thorne gave her to ask why he visited the farm the night prior. |
| 13 | The man who answers claims to have never met her, and says that he lives in Chicago and has never been to Texas. |
| 14 (tp3) | Worried for their father's safety, the siblings summon a doctor for a house call and request that he be moved to a hospital. |
| 15 | The doctor determines that their father's health is grave, and that he is on his deathbed. |
| 16 | He tells the siblings he cannot relocate him to a hospital, as moving him could result in him dying en route. |
| 17 | On the farm, Louise and Michael find that their large herd of goats have all been brutally killed. |
| 18 | The two start a bonfire to dispose of the numerous animal carcasses. |
| 19 | That night, Michael is approached in the barn by an apparition of his nude mother, who disappears as she approaches him. |
| 20 | Later, while Louise lies in bed beside her father, she has a nightmare in which the entity attempts to possess her, but she manages to resist it, before witnessing her father levitating against the ceiling. |
| 21 | In the morning, Charlie's granddaughter arrives at the farm and informs Louise that he killed himself two days prior. |
| 22 | The girl's forlorn demeanor soon turns malevolent, and Louise realizes it is in fact the entity taking the shape of Charlie's granddaughter. |
| 23 | She too disappears before Louise's eyes. |
| 24 | The nurse arrives moments later to care for Louise and Michael's father. |
| 25 (tp4) | Meanwhile, Louise finds that Michael has fled the farm to return to his wife and daughters, leaving her behind. |
| 26 | Michael calls Louise from his cell phone, and tells her she too should leave. |
| 27 | Moments later, the nurse, possessed by the entity, begins stabbing herself with a pair of knitting needles, attacking Louise in the process, before stabbing herself in the eyes, killing herself. |
| 28 (tp5) | Michael arrives at his home and finds the dead bodies of his daughters and wife in the kitchen, in what appears to have been a murder-suicide. |
| 29 | A distraught Michael cuts his own throat, and moments after, observes that the bodies of his wife and daughters have disappeared. |
| 30 | He realizes the entity has tricked him when his wife and daughters enter the house moments later, and find him bleeding to death. |
| 31 | Back at the farm, Louise regains consciousness at nightfall and finds that her father is dying. |
| 32 | Moments after he dies, the demon in turn attacks and claims her. |

Table 10: Example 1 of machine written narratives and the annotated story arc, turning points.

| Source: Machine Title: The Dark and the Wicked (slightly rephrased to avoid memorization) Genre: Horror Annotated Story Arc: Man in Hole | |
|---|---|
| 1 | The lives of siblings Amy and John Mackenzie, leading peaceful lives amid the ups and downs of New York City, put a vibrant appeal to the audience. |
| 2 | As the story unfolds, Amy, a well-renowned interior designer, graces her office with an artistic flair. |
| 3 | At the same time, John, engrossed in complex coding, generates buzzing innovative software for a top-tier tech firm. |
| 4 (tp1) | All's well in their city life until an unanticipated phone call from a family lawyer pierces their harmonious routine like a sharp edged blade. |
| 5 | The news delivered shakes their peaceful existence, revealing their estranged father's terminal pancreatic cancer diagnosis. |
| 6 | In response to the heart-wrenching news, they decide to leave their hustle-bustle life centered in New York and retreat to their tranquil origins of Monterey, California. |
| 7 | The transition from the city to the contemplative countryside period calls them to assist their ailing father. |
| 8 | Greeted by a frail figure that once used to be their lively father, they begin to comprehend the severity of the situation. |
| 9 | The ancestral ranch, once verdant with lush vineyards, had transformed into a deserted expanse with signs of virtually no upkeep. |
| 10 (tp2) | Amid all the despair, the siblings start noticing a series of weird incidents that were out of sync with the placid ranch. |
| 11 | Favorite books getting relocated from their accustomed spots, the sudden rush of cold gust, and cellar doors creaking open during the silent night hours become regular events. |
| 12 | The duo, fueled by heightened curiosity, takes up the challenge to unfold the source of these unnatural occurrences. |
| 13 | Their quest leads them to an overshadowed wooden cabin hidden under years of vegetation, shrouded in dust and sprinkled with ancestral memories. |
| 14 (tp3) | The cabin unveils a century-old diary hidden beneath a rotting carpet, narrating tales of their ancestors intertwined with hints of supernatural events. |
| 15 | The diary begins painting centuries-old tales of unsettling shadows and persistent cold gusts that have made the ranch their abode- striking a chilling accord with the incidents happening on the ranch currently. |
| 16 | A cryptic message on a diary page sends shivers down their spine and reveals a terrifying secret about their bloodline curse. |
| 17 (tp4) | Days turn into nights and the paranormal activities begin escalating, correlating with their father's alarming health deterioration. |
| 18 | The haw-haw of their father rapidly morphs into weak choking gasps for breath. |
| 19 | Seeking guidance from the compassionate local Sacred Heart Church priest, the siblings begin decoding the family curse etched within the old diary. |
| 20 | Consumed by fear and a worryingly less amount of time, they connect the dots that point towards their ancestors' grave mistake- the theft of land from Native Americans. |
| 21 | As they make amends for their ancestors' misdeeds, they notice a significant reduction in the paranormal activities around the ranch. |
| 22 (tp5) | More importantly, their father's gasps for air and trembling substantially reduce, and his breaths start stabilizing. |
| 23 | Transformations take over the ranch. |
| 24 | Nights become serene, and paranormal events decline, allowing peace to bask onto the property. |
| 25 | The siblings are warmly embraced as brave warriors by the locals for enduring and persevering through the haunting incidents and lifting the ancestral curse from their family. |
| 26 | The ranch starts buzzing with life as locals lend their hands in restoring the property. |
| 27 | Laughs echo through the vast expanse, replacing the eerie events that had earlier dominated the place. |
| 28 | The bustling New York City life that was earlier their home becomes eventually replaced with vivid memories of open landscapes and the haunting experience at the ranch. |
| 29 | The twilight, once a symbol of dread and fear, now stands as a testament of their bravery and acquired wisdom. |
| 30 | The movie concludes with a panoramic view of the ranch, brimming with life against the beautiful sunset. |
| 31 | The Mackenzies' ranch, once residing under the shadow of a curse, had finally seen a new dawn. |

Table 11: Example 2 of human written narratives and the annotated story arc, turning points.

| Source: Human Title: Fair Play Genre: Thriller Annotated Story Arc: Riches to Rags | |
|---|---|
| 1 | Emily Meyers and Luke Edmunds, analysts at the cutthroat Manhattan hedge fund One Crest Capital, are in a secret passionate relationship unbeknownst to their coworkers. |
| 2 | Luke proposes to Emily while at his brother's wedding, and she happily accepts. |
| 3 | The next day, one of the company's portfolio managers is fired. |
| 4 | Emily tells Luke she overheard her colleagues mentioning Luke being considered as a replacement and they celebrate that night. |
| 5 (tp1) | However, at a late-night meeting with Campbell, the firm's CEO, Emily learns she will be receiving the promotion. |
| 6 | Emily reluctantly breaks the news to Luke, but he expresses his support. |
| 7 (tp2) | As Emily settles into her new job, Luke's resentment over not being promoted becomes increasingly apparent, leading to tensions in his relationship with Emily. |
| 8 | Luke becomes consumed with the work of a self-help guru coaching people on how to assert themselves in the workplace. |
| 9 | When Emily questions his choice to spend \$3,000 on the course, Luke suggests she could benefit from becoming more assertive, to which she becomes defensive. |
| 10 | Luke rebuffs Emily's attempts to initiate sex and goes to bed. |
| 11 | While out for drinks with Campbell and Paul, a senior executive at the fund, Emily learns Campbell is seeking to get rid of Luke, considering him ineffectual. |
| 12 | Emily attempts to advocate more for Luke in the workplace, but it backfires when Luke makes a poor trading call that loses the company \$25 million, leading to Campbell insulting her. |
| 13 | Luke attempts to rectify himself by feeding Emily insider information confirming the alleged collapse of a company whose stock the fund can short. |
| 14 | Concerned about the trade being illegal, Emily recommends Campbell to short another company, which proves successful. |
| 15 | When the short sale is closed, Emily receives a \$575,000 commission check. |
| 16 (tp3) | Emily considers celebrating her success with Luke, who is in her office after hours to discuss strategies for future trades but opts to go to a strip club with her male co-workers. |
| 17 | She comes home intoxicated while Luke, after seeing the check, has no interest in having sex with her. |
| 18 | When another portfolio manager is fired the next day, Luke wants Emily to recommend him for the role, but she hints Campbell is not interested in promoting him. |
| 19 | Luke goes to Campbell's office and makes an elaborate speech pledging his loyalty to him, only to learn Campbell has already hired a new portfolio manager. |
| 20 | That night, Emily learns her mother had planned a surprise engagement party for them that Friday. |
| 21 | A drunken Luke accuses Emily of stealing his job, but Emily reveals Campbell wanted to fire him, leading Luke to storm out. |
| 22 (tp4) | The next day, while Emily, Campbell, and Paul pitch to overseas investors, Luke barges into the conference room intoxicated and causes a scene, berating Campbell for denying him a promotion and revealing his relationship with Emily, which has violated company policy since her promotion. |
| 23 | An infuriated Emily is unable to reach Luke over the phone, only to find him at the engagement party. |
| 24 | The two argue in front of their families, and Emily smashes a bottle on Luke's head when he suggests she had traded sexual favors for the promotion. |
| 25 | Emily retreats to a bathroom where Luke finds her and the two argue before having sex. |
| 26 | During sex, Luke forces Emily forward twice, causing her face to slam against the bathroom counter. |
| 27 | Emily tells Luke to stop, but he does not. |
| 28 | The next morning, to protect her job, Emily tells Campbell she was being stalked by Luke and they were never in a relationship. |
| 29 | Emily returns home to find Luke there, having packed up his belongings and planning to move in with his brother. |
| 30 (tp5) | Infuriated by his nonchalant attitude and demanding an apology for raping her, Emily threatens Luke with a knife. |
| 31 | She attacks Luke with the knife until he apologizes and breaks down crying. |
| 32 | Luke begs for her forgiveness and Emily orders him to leave before dropping the knife and smiling. |

Table 12: Example 2 of machine written narratives and the annotated story arc, turning points.

| Source: Machine Title: Fair Play (slightly rephrased to avoid memorization) Genre: Thriller Annotated Story Arc: Man in Hole | |
|---|---|
| 1 | The story unfolds with a picturesque view of Wall Street’s towering skyscrapers, where our protagonists, Madison Carter and Noah Mitchell, are immersed in complex financial reports. |
| 2 | The scene transitions to display Madison and Noah at the heart of Two Peak Enterprises. |
| 3 | Their relationship, albeit business-oriented, is a vital component of the firm. |
| 4 | Despite the demands of their high-pressure jobs, the two manage to cultivate a budding romance. |
| 5 (tp1) | Suddenly, a transaction of unprecedented proportions starts to shake the foundations of Two Peaks. |
| 6 | This billion-dollar deal begins to dominate the conversation, raising stakes across the financial landscape. |
| 7 | Madison and Noah’s covert relationship comes under scrutiny. |
| 8 | There’s a simultaneous tension between their fear of exposure and the stress of the impending billion-dollar deal. |
| 9 (tp2) | In their analysis of the gargantuan deal, Madison and Noah identify a discrepancy that raises alarm bells. |
| 10 | They launch a secret investigation, delving furiously into the labyrinth of deception that infects the higher echelons of Two Peaks. |
| 11 | Their once-familiar office morphs into unwelcoming territory as unknown adversaries send chilling threats in response to their probing. |
| 12 | Their interpersonal bonds are tested further as they wrestle with a moral quandary surrounding their loyalty to their employer and their duty as conscientious citizens. |
| 13 (tp3) | A new roadblock appears; an unexpected internal audit is announced at Two Peaks. |
| 14 | Madison and Noah escalate their investigation. |
| 15 (tp4) | Their actions lead to an earth-shattering revelation: the very person they respected and emulated, their boss, is the puppeteer behind the corruption. |
| 16 | Upon the discovery of their boss’s puppeteering, retaliation is swift and vicious. |
| 17 | A gripping chase ensues with Madison and Noah racing against time to secure irrefutable evidence of the crimes. |
| 18 | They manage to secure the valuable evidence necessary to expose their nefarious boss. |
| 19 | They take a moral stand, choosing their ethical responsibility over their coveted careers. |
| 20 (tp5) | Armed with damning evidence, Madison and Noah confront their colleagues at Two Peaks’ annual gala, unveiling their CEO’s unscrupulous activities to an astounded audience. |
| 21 | Amid the chaos, a physical altercation erupts. |
| 22 | Madison and Noah courageously battle the onslaught to protect their evidence. |
| 23 | After the ordeal, they hand over their evidence to the authorities. |
| 24 | The downfall of Two Peak Enterprises sends shockwaves across Wall Street. |
| 25 | Madison and Noah, hailed as righteous heroes, decide to distance themselves from the aggressive world of finance. |
| 26 | The narrative closes with the couple embarking on a new life in a bucolic setting. |
| 27 | A note of suspense strikes as hints point at an omnipresent surveillance. |
| 28 | The screen pans to a computer monitor, with Two Peaks’ now-defunct website displayed. |
| 29 | The narrative ends leaving a lasting sense of suspense. |

| Instructions | | Examples | |
|---|------------------|-------------|--|
| For each synopsis, please fill in the squares highlighted yellow according to their labels | PID | line | synopsis |
| | Title | Jaws | 1 In the New England beach town of Amity Island, a young woman goes for a late-night ocean swim during a beach party. |
| Turning points: Enter the corresponding line number | Turning Point 1 | 2 | 2 An unseen force attacks and pulls her underwater. |
| | Turning Point 2 | 6 | 3 Her remains are found washed up on the beach the next morning. |
| | Turning Point 3 | 19 | 4 After the medical examiner concludes it was a shark attack, newly hired police chief Martin |
| | Turning Point 4 | 31 | 5 The coroner, apparently under pressure, now concurs with the mayor's theory that it was a |
| | Turning Point 5 | 37 | 6 Brody reluctantly accepts their conclusion until a young boy, Alex Kintner, is killed |
| Humble Start, Significant gain: Select TRUE or FALSE Rises & Falls: Enter a number for each | Humble start? | FALSE | 7 A bounty is placed on the shark, causing an amateur shark-hunting frenzy. |
| | Significant gain | FALSE | 8 Quint, an eccentric and roughened local shark hunter, offers his services for \$10,000. |
| Arc Categorization: Select an option from the dropdown. If there are two appropriate categories, select the best category in Arc Category and the second category in Arc Category 2 | # of rises | 1 | 9 Consulting oceanographer Matt Hooper examines the girl's remains, confirming that an abominably large shark killed her |
| | # of falls | 1 | 10 When local fishermen catch a tiger shark, the mayor declares the beaches safe. |
| Detailed Story Arc Instructions | Arc Category | Man in... | 11 Mrs. Kintner confronts Brody and blames him for her son's death. |
| | Arc Category 2 | | 12 A skeptical Hooper dissects the tiger shark and, finding no human remains inside its stomach, determines a larger shark killed the victims. |
| Detailed Turning Point Instructions | | | 13 While searching the night waters in Hooper's boat, Hooper and Brody find the half-sunken vessel of Ben Gardner, a local fisherman. |
| Annotation Guideline for Story Arcs | | | 14 Underwater, Hooper removes a sizable shark tooth from the boat's hull, but accidentally drops it after discovering Gardner's severed head. |
| Annotation Guideline - Turning Point | | | 15 Vaughn dismisses Brody and Hooper's assertions that a huge great white shark caused the deaths, and refuses to close the beaches, allowing only increased safety precautions. |
| | | | 16 On the Fourth of July weekend, tourists pack the beaches. |
| | | | 17 The shark enters a nearby lagoon, killing a boater. |
| | | | 18 Brody then convinces a guilt-ridden Vaughn to hire Quint. |
| | | | 19 Despite tension between Quint and Hooper, they and Brody head to sea on Quint's boat to hunt the shark. |
| | | | 20 As Brody lays down a chum line, the shark suddenly appears behind the boat. |
| | | | 21 Quint, estimating it is 25 feet (7.6 m) long and weighs 3 tonnes (3.0 long tons; 3.3 short tons), harpoons it with a line attached to a flotation barrel, but the shark pulls the barrel underwater and disappears. |
| | | | 22 At nightfall, Quint and Hooper drunkenly exchange stories about their assorted body scars. |
| | | | 23 One of Quint's is a removed tattoo, and he reveals that he survived the attack on the USS Indianapolis. |
| | | | 24 The shark returns, ramming the boat's hull and disabling the power. |

Figure 10: Human Annotation Interface for Turning Point and Story Arc.

Annotation Guideline for Story Arc Categorization

Objective

Your task is to read each story and determine which of the six story arcs it best aligns with. Use the descriptions below as a guide to categorize each story.

Task

Review the provided story corpus and accurately determine which one (or, in more complex situations, two) of the story arcs most aptly fits the narrative. Refer to the provided examples for clarification. Indicate top one (or top two when unsure) types of story arc that fits the story.

Background

The concept of the story arc, originally proposed in Kurt Vonnegut's rejected master's thesis—which he deemed his most significant contribution—charts the trajectory of a narrative along two axes: 'Beginning-End' and 'Ill Fortune-Great Fortune'. In assessing the fortunes of the protagonist, it is recommended to base your analysis on their tangible circumstances, such as wealth versus poverty, safe versus in danger, or authority versus powerlessness.

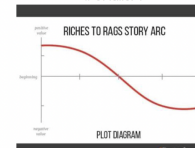
1. Rags to Riches: Look for stories where the main character starts off in a poor or unfavorable situation and ends up in a much better or prosperous condition. E.g. J.K. Rowling

Reference for short questions:
 Humble start: Yes
 Significant gain at the end: Yes
 # of rises: 1
 # of falls: 0



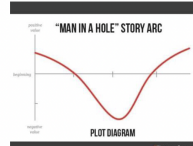
2. Riches to Rags: Identify stories where the protagonist begins in a position of wealth or high status and ends in a significantly lower or impoverished state. E.g. Mike Tyson

Reference for short questions:
 Humble start: No / Neutral
 Significant gain at the end: No
 # of rises: 0
 # of falls: 1



3. Man in a Hole: Focus on stories where the character finds themselves in a dilemma or crisis and must find a way out, ending slightly better off than at the beginning.

Reference for short questions:
 Humble start: No / Neutral
 Significant gain at the end: No
 # of rises: 1
 # of falls: 1



4. Icarus (Rise then Fall): Recognize stories that showcase a character's ascent to success or happiness, followed by a drastic downfall or tragedy.

Reference for short questions:
 Humble start: Yes/ Neutral
 Significant gain at the end: No
 # of rises: 1
 # of falls: 1

Figure 11: Detailed Annotation Guideline for Story Arc Categorization, Page 1-2.

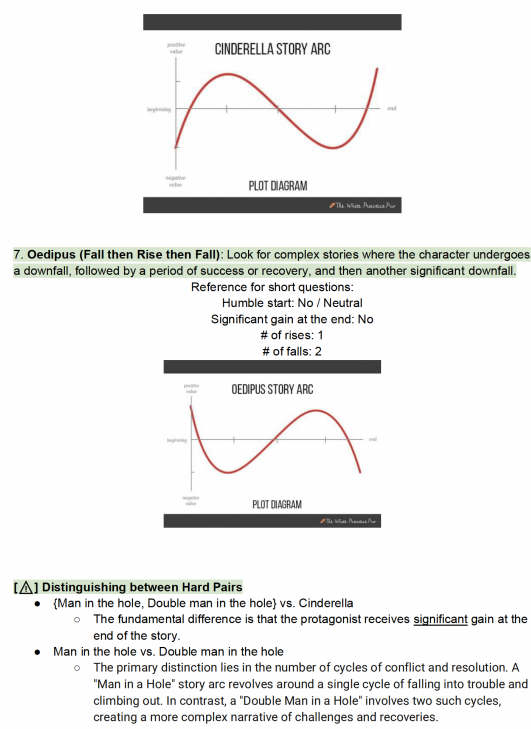
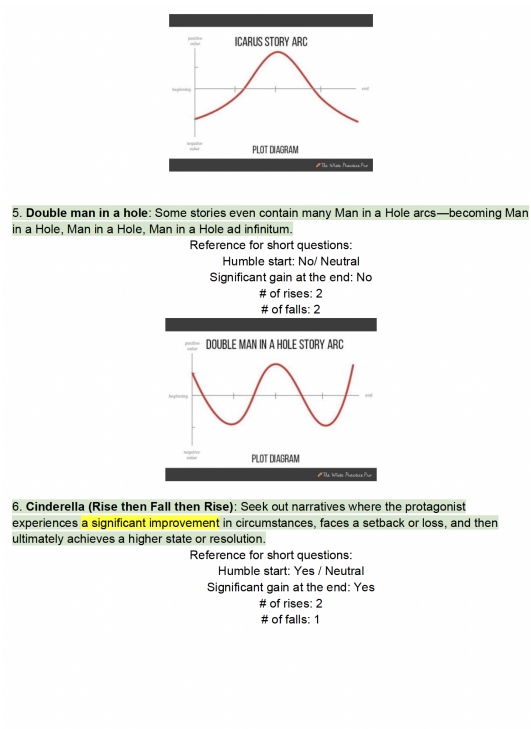


Figure 12: Detailed Annotation Guideline for Story Arc Categorization, Page 3-4

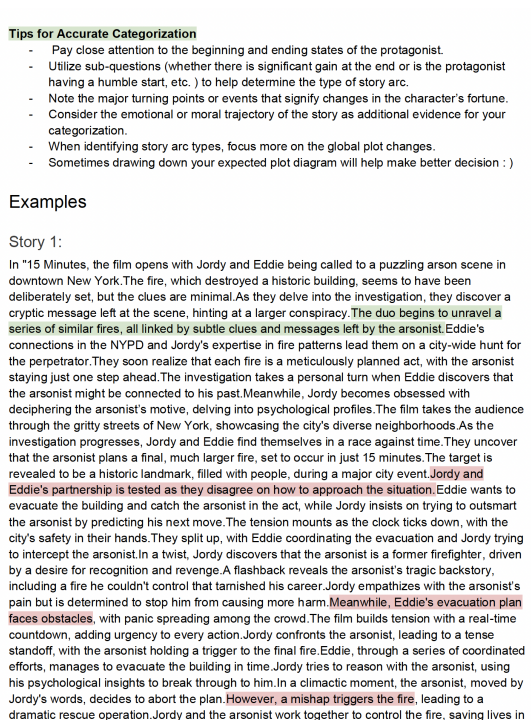
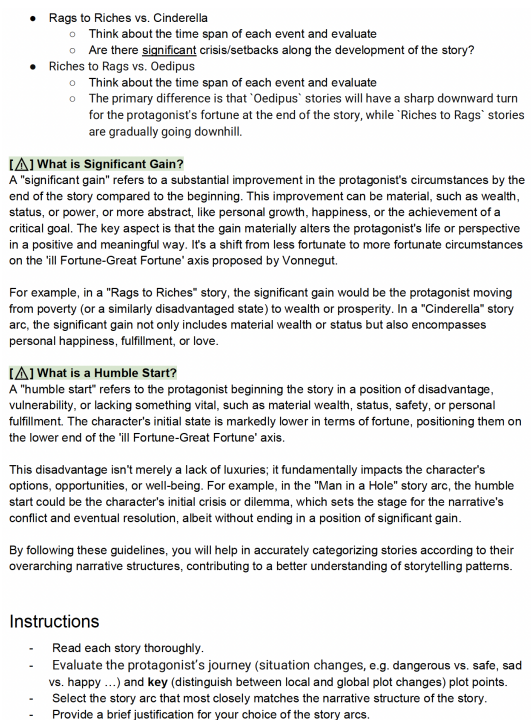
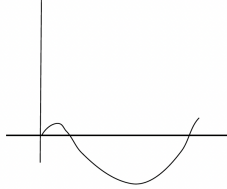


Figure 13: Detailed Annotation Guideline for Story Arc Categorization, Page 5-6.

the process. The film concludes with the arsonist being taken into custody, showing remorse for his actions. **Jordy and Eddie are hailed as heroes,** with their partnership stronger than ever. The aftermath shows the city recovering, with a focus on the importance of community and resilience. Jordy visits the arsonist in prison, promising to help him find redemption. Eddie reflects on the case, realizing the complexities of human nature and justice. **The film ends with Jordy and Eddie receiving commendations for their bravery and dedication.**

Reasoning:

This story starts with the protagonists being summoned to a crime scene, where the general tone of the story is nervous and intense. The investigation went well for a while until they discovered the arsonist's next attack. After going through a series of really dangerous events, they arrested the arsonist. At the end of the story, they are rewarded for their bravery. The overall story arc looks like this:



So the relative ranking of story arcs that fit this story best would be: **Man in a hole > Cinderella** (Provide two choices when very close)

Story 2:

"I Am Legend" begins in October 2009 with Dr. Alice Krippin, played by Emma Thompson, a **medical researcher, announcing her breakthrough in cancer treatment.** She has developed a genetically-engineered measles virus, which has shown remarkable results in curing cancer. The film focuses on a brilliant epidemiologist, Dr. Michael Harris, who becomes involved in the project. Initially, the treatment is hailed as a medical miracle, with cancer patients showing significant improvement. Dr. Harris, however, notices unusual patterns in the recovery of patients. The patients, though cured of cancer, begin to exhibit strange behavioral changes and enhanced physical abilities. Dr. Harris starts investigating these anomalies, suspecting a link to the Krippin Virus. **His research reveals that the virus, while curing cancer, has unintended side effects, altering human DNA.** As the situation escalates, the patients, now referred to as "the Enhanced," start to form a distinct group in society. Dr. Harris faces opposition from the pharmaceutical company backing Dr. Krippin's research. The company is more interested in profits than the potential risks of the treatment. Dr. Harris teams up with another researcher, Dr. Sarah Johnson, to understand the full implications of the virus. They discover that the virus has a

compound effect, growing stronger with each generation. The Enhanced individuals begin to experience a collective consciousness, connecting them in ways previously thought impossible. Dr. Harris uncovers that the Enhanced are developing advanced cognitive abilities, surpassing normal human limits. The public becomes divided on how to view the Enhanced, with some seeing them as a threat, and others as the next step in human evolution. Dr. Harris and Dr. Johnson face ethical dilemmas over the implications of their discovery. Governments around the world start to take an interest in the Enhanced for military purposes. The Enhanced, feeling ostracized and threatened, start to withdraw from society. Dr. Harris advocates for the rights and humane treatment of the Enhanced. He becomes a liaison between the Enhanced and the rest of society, striving for peaceful coexistence. **Tensions rise as incidents involving Enhanced individuals lead to public fear and outcry.** The pharmaceutical company tries to manipulate the situation to their advantage. Dr. Harris discovers a plan to exploit the Enhanced for biological warfare. He and Dr. Johnson work to expose the company's unethical practices. The Enhanced, led by a charismatic figure, begin to organize and demand recognition. Dr. Harris finds himself caught between the government, the pharmaceutical company, and the Enhanced. He realizes that the fate of humanity and the Enhanced is intertwined. The film explores the moral and ethical challenges posed by scientific advancements. Dr. Harris uncovers a potential way to reverse the effects of the virus. However, the reversal has risks, potentially leading to the death of Enhanced individuals. A global debate ensues about whether to reverse the virus's effects. The Enhanced community is split, with some wanting to return to normalcy, and others embracing their new abilities. Dr. Harris faces personal threats as he becomes more involved in the controversy. He develops a close bond with some of the Enhanced, understanding their perspective. The movie builds to a climax with a showdown between the pharmaceutical company, the government, and the Enhanced. Dr. Harris plays a crucial role in negotiating a resolution. **The film ends with a compromise, allowing the Enhanced to live with their abilities, with strict regulations and monitoring.** Dr. Harris continues his research, dedicated to ensuring the safety and rights of both humans and the Enhanced. "I Am Legend" is praised for its thought-provoking narrative and exploration of complex themes.

Reasoning:

The story has a positive beginning where the cure of cancer is developed and patients get cured. However, the accident happens when side effects of the drug is discovered. The story ends with the protagonist solves the crisis.

So the relative ranking of story arcs that fit this story best would be: **Man in a hole > Rags to Riches** (Provide two choices when very close)

Story 3:

Thor and Loki are space traveling with all the Asgardians. Thanos and his lieutenants—Ebony Maw, Cull Obsidian, Proxima Midnight, and Corvus Glaive—intercept the spaceship carrying the survivors of Asgard's destruction. [a] After subduing Thor, Thanos extracts the Space Stone from the Tesseract, overpowers the Hulk, and kills Loki. Thanos also kills Heimdall after he sends

Figure 14: Detailed Annotation Guideline for Story Arc Categorization, Page 7-8.

Introduction

In this first task, you will be presented with a movie synopsis. Your task is to identify the sentences in each synopsis which correspond to the **five turning points** of a movie plot.

Please read carefully through each synopsis and evaluate which sentence most closely matches each turning point. The turning points are defined as follows:

- **Opportunity (Turning Point 1):** the event that introduces the central narrative after the initial setup. This turning point follows the introduction of the setting and main characters and represents the beginning of the story's plot.
 - "An unseen force attacks and pulls her underwater." - post setup this is the first major event
 - "The vigilante Batman, district attorney Harvey Dent, and police lieutenant Jim Gordon ally to eliminate Gotham's organized crime." - represents the first step taken to combat the Joker
- **Change of Plans (Turning Point 2):** this is the moment where the primary objective or goal of the story is defined. After the Change of Plans, the action begins to increase.
 - "Brody reluctantly accepts their conclusion until a young boy, Alex Kintner, is killed at a crowded beach." - a death forces the humans to take the shark threat seriously
 - "Although Gordon saves the mayor, the Joker threatens that his attacks will continue until Batman reveals his identity." - the threat of more attacks forces those with power in Gotham to take the Joker's threats more seriously
- **Point of No Return (Turning Point 3):** this is the event that encourages or forces the protagonists to commit fully to their goal.
 - "Despite tension between Quint and Hooper, they and Brody head to sea on Quint's boat to hunt the shark." - the humans commit to hunting the shark
 - "Batman races to save Rachel while Gordon and the other officers go after Dent, but they discover the Joker gave their positions in reverse." - Batman makes a choice to go after Rachel, which shapes the rest of the story
- **Major Setback (Turning Point 4):** this is the point in the story where the protagonist(s) encounter their greatest obstacle, a setback that significantly complicates their journey.
 - "To Brody's relief, Quint heads toward shore to draw the shark into shallower waters, but the overtaxed engine fails." - humans are trapped with the shark

- "As panic grips the city, the Joker reveals two evacuation ferries, one carrying civilians and the other prisoners, are rigged to explode at midnight unless one group sacrifices the other." - Joker makes his biggest threat yet
- **Climax (Turning Point 5):** this event is the pinnacle of the story—the final and most significant event that resolves the main narrative and represents the story's peak dramatic tension.
 - "Trapped on the sinking vessel, Brody shoves a scuba tank into the shark's mouth and, climbing onto the crow's nest, shoots the tank with a rifle." - the humans finally triumph over the shark
 - "Batman subdues the Joker but refuses to kill him." - Batman will not be corrupted by the Joker

For each turning point, you should choose one and only one sentence which corresponds to that turning point. If some event spans over multiple sentences and you feel that the turning point could be any one of those sentences, choose the **earliest** sentence that is part of the event.

In general, the turning points should be in sequential order (e.g., the sentence you identify for the Change of Plans should come before the Point of No Return). However, if you feel that the Major Setback (4) comes before the Point of No Return (3), it is acceptable to interchange **these two turning points only**.

Figure 15: Detailed Annotation Guideline for Turning Point Identification, Page 1-2.

Examples

Example 1: Jaws

1. In the New England beach town of Amity Island, a young woman goes for a late-night ocean swim during a beach party.
2. **An unseen force attacks and pulls her underwater.**
 - a. Turning Point 1 "Opportunity"
 - b. We have been introduced to the setting (Amity Island) and this is the "inciting incident" which sets the rest of the plot in motion. If the young woman were not attacked by the "unseen force," there would be no movie.
3. Her remains are found washed up on the beach the next morning.
4. After the medical examiner concludes it was a shark attack, newly hired police chief Martin Brody closes the beaches; Mayor Larry Vaughn persuades him to reconsider, fearing the town's summer economy will suffer.
5. The coroner, apparently under pressure, now concurs with the mayor's theory that it was a boating accident.
6. **Brody reluctantly accepts their conclusion until a young boy, Alex Kintner, is killed at a crowded beach.**
 - a. Turning Point 2 "Change of Plans"
 - b. The main goal of eliminating the shark, which we now know for certain to be dangerous, is defined, and the stakes are raised through a child's death. The action increases as a result.
7. A bounty is placed on the shark, causing an amateur shark-hunting frenzy.
8. Quint, an eccentric and roughened local shark hunter, offers his services for \$10,000.
9. Consulting oceanographer Matt Hooper examines the girl's remains, confirming that an abnormally large shark killed her.
10. When local fishermen catch a tiger shark, the mayor declares the beaches safe.
11. Mrs. Kintner confronts Brody and blames him for her son's death.
12. A skeptical Hooper dissects the tiger shark and, finding no human remains inside its stomach, determines a larger shark killed the victims.
13. While searching the night waters in Hooper's boat, Hooper and Brody find the half-sunken vessel of Ben Gardner, a local fisherman.
14. Underwater, Hooper removes a sizable shark tooth from the boat's hull, but accidentally drops it after discovering Gardner's severed head.
15. Vaughn dismisses Brody and Hooper's assertions that a huge great white shark caused the deaths, and refuses to close the beaches, allowing only increased safety precautions.
16. On the Fourth of July weekend, tourists pack the beaches.
17. The shark enters a nearby lagoon, killing a boater.
18. Brody then convinces a gull-ridden Vaughn to hire Quint.
19. **Despite tension between Quint and Hooper, they and Brody head to sea on Quint's boat to hunt the shark.**

- a. Turning Point 3 "Point of No Return"
- b. Despite knowing the risk (the shark has already killed multiple people), the three fully commit to their goal of hunting the shark.
20. As Brody lays down a chum line, the shark suddenly appears behind the boat.
21. Quint, estimating it is 25 feet (7.6 m) long and weighs 3 tonnes (3.0 long tons; 3.3 short tons), harpoons it with a line attached to a flotation barrel, but the shark pulls the barrel underwater and disappears.
22. At nightfall, Quint and Hooper drunkenly exchange stories about their assorted body scars.
23. One of Quint's is a removed tattoo, and he reveals that he survived the attack on the USS Indianapolis.
24. The shark returns, ramming the boat's hull and disabling the power.
25. The men work through the night, repairing the engine.
26. In the morning, Brody attempts to call the Coast Guard, but Quint, obsessed with killing the shark without outside assistance, smashes the radio.
27. After a long chase, Quint harpoons the shark with another barrel.
28. The line is tied to the stern cleats, but the shark drags the boat backward, swamping the deck and flooding the engine compartment.
29. As Quint is about to sever the line to save the boat's transom, the cleats break off.
30. The barrels stay attached to the shark.
31. **To Brody's relief, Quint heads toward shore to draw the shark into shallower waters, but the overtaxed engine fails.**
 - a. Turning Point 4 "Major Setback"
 - b. The trio are now trapped with the shark. They have to figure out how to confront the shark without a working engine, a major setback.
32. As the boat takes on water, the trio attempts a riskier approach.
33. Hooper suits up and enters a shark-proof cage, intending to lethally inject the shark with strychnine via a hypodermic spear.
34. The shark viciously attacks the cage, causing Hooper to drop the spear.
35. While the shark destroys the cage, Hooper escapes to the seabed.
36. The shark leaps onto the sinking boat's stern, subsequently devouring Quint.
37. **Trapped on the sinking vessel, Brody shoves a scuba tank into the shark's mouth and, climbing onto the crew's nest, shoots the tank with a rifle.**
 - a. Turning Point 5 "Climax"
 - b. This is the final confrontation with the shark and the resolution of the hunt, so this is the climax.
38. The resulting explosion kills the shark.
39. Hooper resurfaces and he and Brody paddle back to Amity Island, clinging to the remaining barrels.

| Movie Name | Tp1 | Tp2 | Tp3 | Tp4 | Tp5 |
|------------|-----|-----|-----|-----|-----|
| Jaws | 2 | 6 | 19 | 31 | 37 |

Figure 16: Detailed Annotation Guideline for Turning Point Identification, Page 3-4.

Example 2: The Dark Knight

1. A gang of masked criminals robs a mafia-owned bank in Gotham City, betraying and killing each other until the sole survivor, the Joker, reveals himself as the mastermind and escapes with the money.
2. **The vigilante Batman, district attorney Harvey Dent, and police lieutenant Jim Gordon ally to eliminate Gotham's organized crime.**
 - a. Turning Point 1 "Opportunity"
 - b. Dent, Batman, and Gordon are our main protagonists against the Joker. Their decision to ally is the inciting incident, setting the rest of the story in motion.
3. Batman's true identity, the billionaire Bruce Wayne, publicly supports Dent as Gotham's legitimate protector, as Wayne believes Dent's success will allow Batman to retire, allowing him to romantically pursue his childhood friend Rachel Dawes, despite her relationship with Dent.
4. Gotham's mafia bosses gather to discuss protecting their organizations from the Joker, the police, and Batman.
5. The Joker interrupts the meeting and offers to kill the Batman for half of the fortune their accountant, Lau, concealed before fleeing to Hong Kong to avoid extradition.
6. With the help of Wayne Enterprises CEO Lucius Fox, Batman finds Lau in Hong Kong and returns him to the custody of Gotham police.
7. His testimony enables Dent to apprehend the crime families.
8. The bosses accept the Joker's offer, and he kills high-profile targets involved in the trial, including the judge and police commissioner.
9. **Although Gordon saves the mayor, the Joker threatens that his attacks will continue until Batman reveals his identity.**
 - a. Turning Point 2 "Change of Plans"
 - b. From this sentence, we know the main goal of the story: to determine if Batman can be corrupted/changed by the Joker. Since the protagonists want to stop the attacks on Gotham's citizens, this sentence's events raise the stakes of the story.
10. He targets Dent at a fundraising dinner and throws Rachel out of a window, but Batman rescues her.
11. Wayne struggles to understand the Joker's motives, but his butler Alfred Pennyworth says "some men just want to watch the world burn."
12. Dent claims he is the Batman to lure out the Joker, who attacks the police convoy transporting him.
13. Batman and Gordon apprehend the Joker, and Gordon is promoted to commissioner.
14. At the police station, Batman interrogates the Joker, who says he finds Batman entertaining and has no intention of killing him.
15. Having deduced Batman's feelings for Rachel, the Joker reveals she and Dent are being held separately in buildings rigged to explode.
16. **Batman races to save Rachel while Gordon and the other officers go after Dent, but they discover the Joker gave their positions in reverse.**
 - a. Turning Point 3 "Point of No Return"

- b. Batman is forced to make a choice, thus committing to his goal, but the false information the Joker provides gives his choice even greater weight. This is a point of no return since this decision will follow Batman and the protagonists for the rest of the movie.
17. The bombs detonate, killing Rachel and severely burning Dent's face on one side.
18. The Joker escapes custody, extracts the fortune's location from Lau, and burns all of it, killing Lau in the process.
19. Wayne Enterprises accountant Coleman Reese deduces Batman's identity and attempts to expose it, but the Joker threatens to blow up a hospital unless Reese is killed.
20. While the police evacuate hospitals and Gordon struggles to keep Reese alive, the Joker meets with a disillusioned Dent, persuading him to take the law into his own hands and avenge Rachel.
21. Dent defers his decision-making to his half-scarred, two-headed coin, killing the corrupt officers and the mafia involved in Rachel's death.
22. **As panic grips the city, the Joker reveals two evacuation ferries, one carrying civilians and the other prisoners, are rigged to explode at midnight unless one group sacrifices the other.**
 - a. Turning Point 4 "Major Setback"
 - b. Things fall apart here -- we assume, along with the Joker, that at least one of the ferries will explode. For the protagonists of the story, who aim to protect all of Gotham's citizens, this represents a major setback.
23. To the Joker's disbelief, the passengers refuse to kill one another.
24. **Batman subdues the Joker but refuses to kill him.**
 - a. Turning Point 5 "Climax"
 - b. This is the resolution of the main question of the story, which is whether Batman will be corrupted by the Joker.
25. Before the police arrest the Joker, he says although Batman proved incorruptible, his plan to corrupt Dent has succeeded.
26. Dent takes Gordon's family hostage, blaming his negligence for Rachel's death.
27. He flips his coin to decide their fates, but Batman tackles him to save Gordon's son, and Dent falls to his death.
28. Believing Dent is the hero the city needs and the truth of his corruption will harm Gotham, Batman takes the blame for his death and actions and persuades Gordon to confess the truth.
29. Pennyworth burns an undelivered letter to Wayne from Rachel, who said she chose Dent, and Fox destroys the invasive surveillance network that helped Batman find the Joker.
30. The city mourns Dent as a hero, and the police launch a manhunt for the Batman.

| Movie Name | Tp1 | Tp2 | Tp3 | Tp4 | Tp5 |
|-----------------|-----|-----|-----|-----|-----|
| The Dark Knight | 2 | 9 | 16 | 22 | 24 |

Figure 17: Detailed Annotation Guideline for Turning Point Identification, Page 5-6.

Example 3: Candy Cane Lane

1. The craftsman, Bob Fletcher, from the light-knit town of Lakewood, can be seen meticulously working in his small, wooden house that serves as his humble workshop.
2. His scientifically weathered hands mold thin, wooden twigs into intricately shaped snowflakes, elves, reindeers, and Christmas trees.
3. **At the same time, the Lakewood Community Center exudes a palpable excitement with the announcement of the much-awaited holiday decor contest.**
 - a. Turning Point 1 "Opportunity"
 - b. The initial setup is our introduction to Bob Fletcher, and this sentence introducing the holiday decor contest shows us what the actual story will be about.
4. John Marley, a resident of Lakewood and a five-time champion of the holiday decor contest, registers his name as a participant.
5. **Spurred on by this, Bob decides to enroll himself in the decor contest.**
 - a. Turning Point 2 "Change of Plans"
 - b. This is Bob's initial commitment to his goal, which will define the rest of the story: winning the holiday decor contest.
6. Walking back home, Bob is interrupted by John's grand display of decoration.
7. His mansion stands as a stark contrast to Bob's humble, handcrafted pieces.
8. Bob sticks to his rigorous crafting schedule, creating unique Christmas ornaments before and after his day job at the local hardware store.
9. To bring in a variety of unique craft materials for his decor, Bob ventures out to Potter's Craft Shop, managed by Iris Potter.
10. Understanding Bob's mission, she empathizes and pledges to stand by him in his quest against commercial holiday deco.
11. John is seen placing orders for extravagant, mass-produced decorations.
12. **Bob and Iris, unfazed, hold a community crafting event in Bob's workshop.**
 - a. Turning Point 3 "Point of No Return"
 - b. Although John's order may be intimidating to Bob, he doubles down and recommit to his goal, which is to win the holiday decor contest. Thus, this is a Point of No Return.
13. **However, John's unchecked ostentation starts eating into Bob's determination, instilling a creeping sense of stress and doubt in him.**
 - a. Turning Point 4 "Major Setback"
 - b. This is Bob's moment of real doubt, and the closest Bob's holiday decor entry comes to falling apart (his lowest point). Therefore it is the major setback in this story.
14. Iris reminds him that the holiday season's true spirit revolves around spreading joy, love, and fostering unity.
15. Her wise words reinvigorate Bob.
16. Bob comes up with a bold new plan.
17. He resolves to outshine John's extravagant decorations with the warmth of community spirit and unity.
18. He painstakingly crafts individual invitations to the entire community.

19. His once quiet and small home quickly turns into a bustling hub of artistic expression, filled with enthusiastic residents of Lakewood.
20. Bob's house comes alive with shared experiences and cooperation.
21. Despite the outpouring of community involvement and appreciation, Bob finds himself wavering on the edge of uncertainty about the impending decor competition.
22. The day of the contest arrives, finding both John's mansion and Bob's small house bathed in the festive glow.
23. John's mansion gleams with professionally created props and extensive neon light installations, while Bob's hands are filled with ornaments reflecting individual stories.
24. **The Judges announce Bob as the winner, praising his ingenious idea of symbolizing community spirit and unity through his innovative holiday decor initiatives.**
 - a. Turning Point 5 "Climax"
 - b. Bob's community-driven decor triumphing over John's impersonal, extravagant decor resolves the main question of the story. It is the culmination of everyone's efforts in the earlier parts of the movie.
25. Bob feels immensely grateful to the community members for their support in his endeavor.
26. John, taken aback by this unforeseen turn of events, recognizes his mistake and congratulates Bob on his victory.
27. The narrative ends with Bob and John lighting painted twigs together - a symbol of community over rivalry.
28. They are joined by other community members, who usher in the holiday season by singing traditional carols.
29. The film ends emphasizing the triumph of community spirit over materialism.
30. The closing shot of Bob's beautifully decorated cottage, echoing with laughter, camaraderie, and unity presents a conclusion to the narrative.

| Movie Name | Tp1 | Tp2 | Tp3 | Tp4 | Tp5 |
|-----------------|-----|-----|-----|-----|-----|
| Candy Cane Lane | 3 | 5 | 12 | 13 | 24 |

Figure 18: Detailed Annotation Guideline for Turning Point Identification, Page 7-8.

Story Arc Identification Prompt

Review the provided story corpus and accurately determine which one (or, in more complex situations, two) of the story arcs most aptly fits the narrative. Refer to the provided example for clarification.

{Story Arc taxonomy explanations}

- Evaluate the protagonist's journey (situation changes, e.g. dangerous vs. safe, sad vs. happy ...) and key (distinguish between local and global plot changes) plot points.
- Select the story arc that most closely matches the narrative structure of the story.
- Provide a brief justification for your choice of the story arcs.

{Demonstrations}

Input Story Instance: {S}

Output: {Story Arc Type}

Figure 19: Prompt for Story Arc Identification Task.

Turning Point Identification Prompt

The movie title is *{title}* and the synopsis is: *{synopsis}*.

Please identify ONE (and only one) sentence in the summary as one of the five turning points.

{Turning points taxonomy explanations}

- You should have all and exactly five turning points in this order.
- Additionally, the sentences you pull need to be in this order; you should not choose a sentence for Climax which precedes the sentence you chose for Change of Plan, for example, because Climax comes after Change of Plan.
- The format of the output should be a JSON object with the turning points as keys and the sentence numbers as values.
- Explain why each is an appropriate selection.

{Demonstrations}

Input Story Instance: {S}

Output: {Turning Points}

Figure 20: Prompt for Turning Point Identification Task.