Are LLMs Good at Finding Turning Points in a Movie Narrative?

Abstract

Understanding narrative structure is central to film studies and computational literary analysis. A key element of narrative structure is identifying turning points—critical events (such as opportunities, changes of plans, points of no return, major setbacks, and climaxes) that define the progression of a movie's plot and segment it into coherent thematic units. In this paper, we evaluate whether large language models (LLMs) can reliably detect these turning points. We employ a variety of approaches (Training and Non-Training) to improve the performance of the LLMs for turning point identification on both, plot synopsis and full-length screenplays. Our findings reveal two major challenges: (1) turning point identification is inherently difficult due to its abstract, subjective, and context-dependent nature; and (2) applying these methods directly on full-length screenplays overwhelms current LLM reasoning abilities because of long context limitations. Our code and instruction tuning dataset is open-sourced.

1 Introduction

Narrative structure has long been central to both classical narratology and modern screenwriting theory. Scholars such as Freytag and Syd Field have demonstrated that pivotal plot events (e.g., the inciting incident, point of no return, major setbacks, and the climax) drive the narrative forward and segment the story into distinct thematic units. These turning points capture essential shifts in character motivation and plot direction, making them valuable markers for summarization, segmentation, and retrieval tasks in computational analysis. However, identifying these turning points remains a challenging task due to their latent nature, the need for deep narrative reasoning, and the complexity of processing long-form screenplays. Turning points are not explicitly marked in the text; rather, they must be inferred from subtle narrative cues such as shifts in tone, character goals, or plot direction.

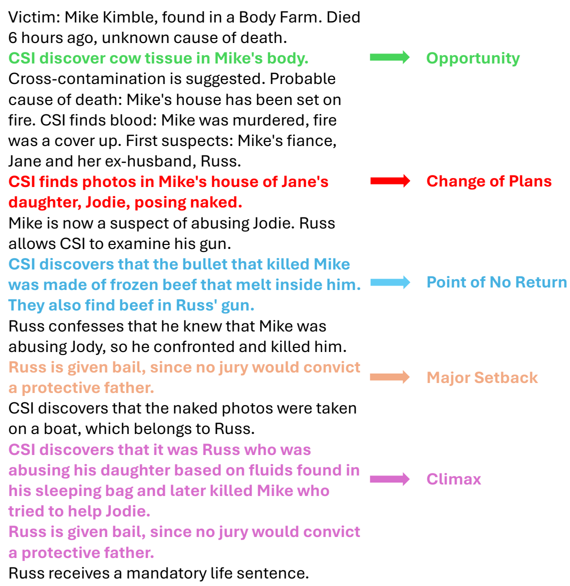

Figure 1: Shows Turning Points over the plot synopsis of a crime thriller.

Unlike concrete extraction tasks (e.g., named entity recognition), turning point detection requires a model to grasp structural elements such as suspense or dramatic reversals. Recent studies (e.g., (Tian et al., 2024), (Hauge, 2017)) have underscored that while large language models (LLMs) generate fluent narratives, their ability to recognize these nuanced transitions remains limited. Evaluations in zero-shot and few-shot settings indicate that most models do not surpass simple distribution-based or theory-informed baselines, highlighting a gap between language generation and true narrative understanding.

Additionally, turning point identification becomes even more complex when applied to full-length screenplays. Screenplays often span over a hundred pages, with key turning points dispersed throughout extensive context. Most transformer-based LLMs struggle with such long sequences due to the inherent limitations of the attention mechanism and fixed input window sizes. This constraint is particularly evident in narrative discourse tasks, where models perform reasonably well on shorter plot synopses but lose effectiveness when processing full-length screenplays (Papalampidi et al., 2019). Recent computational work, such as that of Papalampidi has attempted to operationalize turning point theory using the TRIPOD dataset, which contains screenplays and plot synopses annotated with turning points (Papalampidi et al., 2019). Their end-to-end neural model, despite leveraging state-of-the-art sentence representations and heuristic positional models, demonstrated that turning point detection remains a difficult task, particularly when applied to raw screenplay text rather than plot synopses. These challenges underscore the complexity of computational narrative understanding and the need for improved methods capable of reasoning over long-form narratives.

2 Dataset

To evaluate the ability of state-of-the-art Large Language Models (LLMs) to identify turning points in narratives, we utilize the TRIPOD dataset(Papalampidi et al., 2019). TRIPOD is a structured dataset designed to model and analyze narrative progression in feature-length films. It provides scene-level annotations of turning points (TPs)—key moments in a story that signal transitions between major narrative stages.

TRIPOD consists of 99 movie screenplays, annotated to align with a widely recognized six-stage narrative framework. This framework defines five turning points as shown in Figure 7, which segment a story into thematic acts: (1) Opportunity, introducing a significant event that disrupts the initial status quo; (2) Change of Plans, where the protagonist's main objective is defined; (3) Point of No Return, committing the protagonist to their goal; (4) Major Setback, a crisis that challenges the protagonist's progress; and (5) Climax, where the final confrontation and resolution occur.

The dataset was originally developed for training models to detect turning points in screenplays using distant supervision from plot synopses. Each scene in TRIPOD is assigned a probability distribution over turning points, allowing models to infer narrative structure without explicit human annotation at the screenplay level.

In our study, we leverage TRIPOD as a benchmark to assess whether current LLMs can recognize and predict turning points in short-form and long-form narratives. Given that LLMs have demonstrated strong performance in various text classification and summarization tasks, we aim to determine if they also exhibit an emergent understanding of narrative progression.

Following the same methodology as the paper (Papalampidi et al., 2019), we evaluate LLMs on two tasks:

- **Turning Point Identification in Plot Synopses:** Since plot synopses provide a condensed version of the full narrative, they serve as a natural first step in evaluating whether LLMs can detect turning points from high-level summaries.

- **Turning Point Identification in Full Screenplays:** Unlike synopses, screenplays contain scene-by-scene details, dialogue, and incremental plot development. Identifying turning points in this setting presents a greater challenge, requiring reasoning in long context inputs.

3 Method

3.1 Turning Point Identification in Plot Synopsis

For plot synopsis, the task is to find indices for 5 sentences that act as Turning Points in the movies. To assess LLM's capabilities, we employ a range of approaches:

- **Zero-shot:** Models are asked to output the 5 sentence numbers as the output directly, without any explanation.

- **Few-shot:** Given a few shot examples from the training set, we ask LLMs to provide the index numbers for the 5 sentences.

- **Supervised Finetuning with the Training set:** We create a supervised instruction tuning dataset consisting of 128 samples, in two methods, first, the LLMs are forced to output just the sentence numbers, and the second, using a bigger LLM, we create a reasoning based instruction tuning dataset, in which the model outputs both the reasoning, as well the sentence number for each turning point, then we finetune an LLM on this dataset.

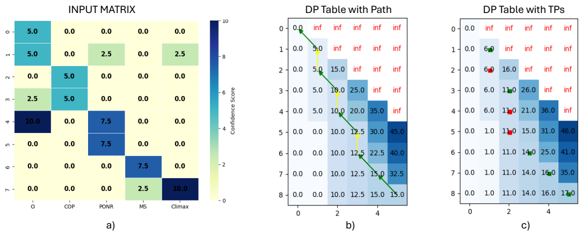

Figure 2: Example for QA + DP approach; a) Represents the input matrix; b) shows the generated DP table; c) Shows the selected TPs, i.e., the green dots are selected as TPs.

3.2 Turning Point Identification for Full Screenplay

The TRIPOD dataset contains 15 movies in the test set, each accompanied by gold standard annotations for the screenplays. Since no dedicated training set is available for the full screenplay, we split the test set into two subsets, with a 4:1 ratio, designating 80% as the test set and the remaining 20% as the development set.

Screenplays present a challenge due to their extensive context length. To address this, we utilize a LLM(LLama-3.1-8B) to generate summaries for each scene. This step effectively reduces the input length from an average of 65k tokens to 15k tokens. Building on the scene-level summaries, we further generate a summary for the entire screenplay to serve as the global context.

Once the screenplay summaries are prepared, we apply **two distinct approaches** to identify Turning Points within the screenplays, without the use of plot synopsis.

3.2.1 Question Answering + Dynamic Programming Approach

In this method, we generate a set of four questions for each potential Turning Point, based on the screenplay theory. The context provided to the LLM (LLM) consists of the following elements:

The LLM receives a context that includes descriptions of the five Turning Points, a summary of the entire screenplay, a summary of the current scene, summaries of the scenes immediately preceding and following the current scene, and a total of twenty questions (four for each Turning Point).

After obtaining answers to each of the twenty questions for each scene, we convert the responses into confidence scores. Each "Yes" answer is assigned 2.5 points, leading to a score out of 10 for each Turning Point.

Further, we employ in-context learning in two ways:

5-Shot, where we provide examples of all five Turning Points from a single movie based on the development set, and **1-Shot**, where we provide an example of only one scene (specifically, the climax) from a single movie based on the development set.

To identify the final set of predicted Turning Points, we apply a dynamic programming (DP) algorithm that enforces two key constraints: all five Turning Points must be present, and they must occur in a fixed sequential order. Once the confidence scores for each scene are obtained, the DP algorithm processes them to determine the final predicted Turning Points (TPs). The input is the confidence score matrix, and the output is the five selected TPs. The algorithm follows a forward pass to construct a DP table based on the rule:

$$dp[i][j] = \min(dp[i-1][j], dp[i-1][j-1] + \text{cost}(i, j))$$

where $\text{cost}(i, j) = 10 - \text{score}(i, j)$. Here, i represents the current scene, and j represents the index of the Turning Point (from 1 to 5). The DP table tracks the minimum cost of selecting TPs up to scene i, ensuring that all five Turning Points are selected in sequence. The term $dp[i][j]$ stores the minimum cumulative cost of selecting the j-th Turning Point up to scene i, while $dp[i-1][j-1]+\text{cost}(i, j)$ considers the cost of choosing scene i as the j-th Turning Point. Once the DP table is constructed, a backward pass (backtracking) is performed to identify the Turning Points. The condition

$$\text{if } dp[i][j] \neq dp[i-1][j] \quad \Rightarrow \quad \text{select } (i, j)$$

indicates that scene i is selected as the j-th Turning Point, whereas if this condition is not satisfied, we move up to the previous scene and continue backtracking. The DP algorithm ultimately identifies the five Turning Points that minimize the total cost while maintaining the correct sequential order. Fig 2 illustrates an example of input confidence scores, the DP algorithm, and the predicted TPs.

3.2.2 Chain of Thought Multi-Turn Conversation

In this method, we maintain the scene-level summaries in context at all times. The large language model (LLM) is tasked with both reasoning and retrieval. The total context provided to the model consists of concatenated summaries of all scenes,

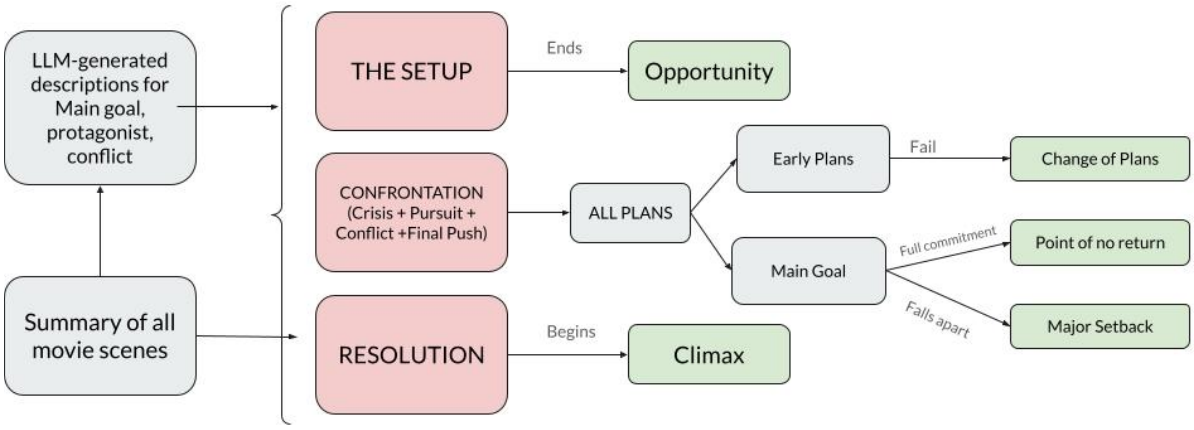

Figure 3: Pipeline for the Chain of Thought Multi-Turn Conversation Method.

along with descriptions for each Turning Point and additional relevant data. The process is carried out in the following order:

1. Describe the main goal, the protagonist, and their desire.
2. Identify the moment when the introduction of the main character concludes, marking the end of the Setup (this is the Opportunity Turning Point).
3. Describe the pivotal moment or climax of the movie.
4. While concealing the climax, outline the main plan that the protagonist is following.
5. Based on the main plan, determine the moment when everything seems to collapse as the protagonist pursues this plan.
6. Describe the alternative plans the protagonist attempts and how they fail, which aids in identifying the remaining Turning Points.

After the LLM has outlined these moments, it is prompted to retrieve the corresponding scene number for each moment described.

Fig **??** shows the flowchart for the complete process.

4 Evaluation Metrics

Tables 1, 2 compare our approaches against baselines set in the paper(Papalampidi et al., 2019), based on three metrics, the total agreement, partial agreement, and the mean distance.

- **Total Agreement (TA):** measures the percentage of predicted TPs that exactly match the ground truth, indicating the accuracy of the predictions.

$$TA = \frac{1}{T \cdot L} \sum_{i=1}^{T \cdot L} \frac{|S_i \cap G_i|}{|S_i \cup G_i|} \quad (1)$$

- **Partial Agreement (PA):** assesses the proportion of TPs with at least one scene in common between predictions and ground truth, highlighting the extent of agreement even when exact matches are absent.

$$PA = \frac{1}{T \cdot L} \sum_{i=1}^{T \cdot L} [S_i \cap G_i \neq \emptyset] \quad (2)$$

- **Mean Distance (D):** quantifies the average distance between the predicted and ground truth scene indices, normalized by the screenplay length, reflecting the overall alignment

Table 1: Comparison among baselines from (Papalampidi et al., 2019) and SOTA LLMs for Turning Point Identification on Plot Synopsis.

	TA	D
Random	2.00	37.79 (25.33)
Theory baseline	22.00	7.47 (6.75)
Theory baseline (Ours)	16.00	8.31
Distribution baseline	28.00	7.28 (6.23)
Distribution baseline (Ours)	17.33	7.89
TAM	34.67	6.80 (5.19)
+ TP views	38.57	7.47 (7.48)
+ entities	41.33	7.30 (7.21)
Zero Shot		
Llama-3.2 (3B) (AI, 2024a)	11.0	17.8
Llama-3.1 (8B)(Dubey et al., 2024)	16.0	11.9
Phi-4 (14B) (Microsoft, 2024)	25.3	9.02
Qwen 2.5 (14B) (Yang et al., 2024)	16.0	11.7
Mistral Small (24B) (Team, 2025)	21.3	9.3
Llama 3.3 (4 bit) (70B) (AI, 2024b)	20.0	8.9
Few Shot		
1 Shot Llama-3.1-8B	22.3	7.9
5 Shot Llama-3.1-8B	17.33	10.05
Supervised Finetuning		
Llama-3.1-8B	16.00	11.47
Llama-3.1-8B Reasoning	24.00	9.60
Human agreement	64.00	4.30 (3.43)

between predictions and truth.

$$d[S_i, G_i] = \frac{1}{M} \min_{(s \in S_i, g \in G_i)} |s - g| \quad (3)$$

5 Results and Analysis

In this section, we present the results of our experiments on turning point identification in movie narratives. The findings are analyzed based on the evaluation metrics defined earlier: Total Agreement (TA), Partial Agreement (PA), and Mean Distance (D). We focus on the performance of LLM-based methodologies and their ability to generalize across narrative structures.

5.1 Theory and Distribution baselines

The theory and distribution baselines (Row 3,5 in Table 1) are included for comparative analysis. Since the official evaluation code is not publicly available, we attempt to reproduce the baselines using our implementation and report the scores we obtain. The theory baseline represents a manually designed heuristic method that aims to capture turning points based on predefined rules, while the distribution baseline leverages statistical distributions of training set turning points. We get different results in our implementation, and we report our numbers for a fair evaluation.

Table 2: Comparison among baselines from (Papalampidi et al., 2019) and our zero-shot methods for Turning Point Identification on Screenplays.

	TA	PA	D
Theory baseline	8.66	10.67	10.45
Distribution baseline	6.67	9.33	10.84
tf*idf similarity	0.74	1.33	53.07
tf*idf + distribution	4.44	6.67	13.33
CAM	11.11	16.00	10.23
+ entities	*14.18*	*17.33*	12.77
TAM	10.63	13.33	**8.94**
+ entities	10.63	13.33	10.15
QA+DP 5 Shot	*14.87*	*18.46*	*20.94*
QA + Threshold 5 Shot	*3.9*	*38.5*	*22.6*
DP Only 5 Shot	*1.7*	*1.7*	*45.9*
QA+DP 1 Shot	*15.8*	*20.3*	*15.8*
QA+Threshold 1 Shot	*4.8*	*48.3*	*22.6*
DP Only 1 Shot	*2.0*	*4.0*	*36.*
CoT Multi-Turn	***20.6***	***28.3***	***13.1***
Human agreement	35.48	56.67	1.48

5.2 Performance on Plot Synopsis-Based Turning Point Identification

Table 1 summarizes the performance of various methods in identifying turning points within plot synopses.

Human agreement scores remain the highest, with TA reaching 64.00 and a mean distance of 4.30, establishing a robust upper bound for computational methods. Among the automated approaches, LLM-based methods show varied performance.

Zero-shot Llama-3.1 (8B)(Dubey et al., 2024) achieves 16.00 TA with a mean distance of 11.9, indicating a limited ability to detect turning points directly. Few-shot models show improvement, with 1-shot Llama-3.1 (8B) achieving 22.3 TA and a mean distance of 7.9. However, adding more examples does not provide further improvements; the 5-shot model achieves a slightly lower TA of 17.33 with a mean distance of 10.05. This suggests that extended context length negatively impacts performance, likely due to difficulties in retaining relevant details over multiple examples.

Supervised fine-tuning with reasoning-based instruction tuning leads to improvements, with TA reaching 24.00 and a mean distance of 9.60. However, the fine-tuned models still fall significantly short of human agreement, highlighting the limitations of the dataset. The dataset is inherently small, subjective, and sparse, making it difficult for models to generalize effectively. This suggests that while fine-tuning can enhance structured knowledge, it struggles to overcome the subjective nature of turning point identification.

5.3 Performance on Full-Length Screenplay-Based Turning Point Identification

Table 2 presents the results for full screenplay turning point detection. Performance drops significantly due to the complexity of processing long narratives and identifying context-dependent transitions. Compared to synopsis-based detection, screenplay-based detection requires handling longer documents with intricate relationships between scenes, making the task more challenging.

QA + DP (5-shot and 1-shot) achieve 14.87 and 15.8 TA, respectively, with 1-shot slightly out- performing 5-shot. This can be attributed to the extended context length for 5-shot (10k tokens) compared to 1-shot (3k tokens), which might lead to reduced reasoning abilities for the LLMs.

For the DP-only method, the model is not explicitly asked questions but is instead prompted to provide confidence scores for potential turning points. This method yields a TA of 2.0 and a PA of 4.0, with a high mean distance of 36.0, suggesting that LLMs are not good at predicting confidence scores, and our QA approach significantly helps in reasoning across scenes.

In contrast, the QA-only method allows the model to generate as many turning point predictions as possible, applying a threshold above 7. This leads to reduced TA scores of 3.9 (for 5-shot) and 4.8 (for 1-shot) compared to the DP approach, with PA reaching very high scores of 38.5 and 48.3, respectively. The higher PA scores are due to the models predicting a significant number of scenes as TPs, which increases the PA but reduces the TA agreement significantly. This shows that our DP algorithm helps in accurately choosing the best-fit scenes for TP prediction.

The Chain-of-Thought Multi-Turn approach achieves the highest scores, with a TA of 20.6, PA of 28.3, and a mean distance of 13.1. This method significantly outperforms others, suggesting that allowing the model to reason through multiple turns and summarize scenes improves narrative understanding. However, the gap between computational methods and human agreement (35.48 TA, 56.67 PA) remains substantial, underscoring the difficulty of turning point identification in long-form narratives.

Overall, current models struggle with both synopsis-based and screenplay-based turning point identification. One-shot learning provides moderate improvements in short-form narratives, but additional examples do not significantly help due to the extended context length affecting coherence.

6 Conclusion

Our findings demonstrate that while LLMs exhibit some emergent capability for narrative structure understanding, they struggle with turning point identification, particularly in long-form screenplays. Zero-shot and few-shot approaches offer marginal improvements, but increasing context length does not necessarily enhance results. Supervised fine-tuning provides some benefits but is constrained by the sparsity and subjectivity of the available datasets. The best-performing method (Chain-of-Thought Multi-Turn) still falls short of human agreement levels, indicating the need for further improvements in long-context processing, structured learning, and multi-modal integration. Future models should incorporate memory-efficient architectures or retrieval-augmented generation (RAG) for better long-form reasoning. Explicit modeling of narrative progression, similar to story grammar-based approaches, could enhance LLM's understanding of plot dynamics. Leveraging screenplay metadata, such as scene descriptions, dialogue sentiment, and visual cues, may improve turning point detection accuracy. This study comprehensively assesses LLMs' ability to analyze narrative structure and identify key turning points. While existing methods show promise, significant advancements are needed to achieve human-level performance in computational narrative understanding.

7 Prompts

This section contains all the prompts used in our experiments.

QA + DP Turning Point Based Questions

Opportunity

1.Does this scene introduce a new element or character that changes the direction of the story?
2.Is this scene significant in setting up the main conflict or goal of the narrative?
3.Does this scene occur after the initial setup and begin driving the story towards the new situation?
4.Is this scene crucial in moving the plot from the exposition phase to a new narrative phase?

Change of Plans

5.Does this scene clarify or redefine the main character's goal or motivation?
6.Is there a change in the character's plan or approach to solving their main problem in this scene?
7.Is there an unexpected twist that forces adaptation?
8.Is this scene pivotal in escalating the narrative towards higher stakes or new challenges?

Point of No Return

9.Does this scene involve a major commitment or irreversible decision by the character?
10.Is there an evident shift in the character's behavior or attitude as a result of this scene?
11.Does the scene show the character fully committing to a course of action despite risks or consequences?
9.Is there a clear transition where the character's previous options or paths are no longer viable?

Major Setback

13.Does this scene represent a significant failure or obstacle for the main character?
14.Is there a dramatic downturn or crisis that affects the character's goals or plans?
15.Does this scene illustrate a temporary or permanent collapse of the character's progress?
16.Is this scene a key moment where the narrative takes a turn towards increased difficulty or tension?

Climax

17.Does this scene represent the highest point of tension or conflict in the story?
18.Is this scene where the main conflict reaches its peak and begins to resolve?
19.Does this scene provide a resolution to the central narrative or thematic elements of the story?
20.Is this scene the final, decisive moment that concludes the main plot arc?

Figure 4: Questions corresponding to each turning point for the QA+DP task.

Plot Synopsis TP Identification Prompt

There are six stages (acts) in a film, namely the setup, the new situation, progress, complications and higher stakes, the final push, and the aftermath, separated by five turning points (TPs). TPs are narrative moments from which the plot goes in a different direction, that is from one act to another. The five turning points are described as:

1.Opportunity: An introductory event that occurs after presenting the setting and background of the main characters, driving the narrative from setup to new situation.

2.Change of Plans: An event where the main goal of the story is defined, leading to increased action, driving the narrative from the introduction of new situation to its progress.

3.Point of No Return: An event that pushes the main character(s) to fully commit to their goal, driving the narrative from the progress to its complications and high-stake difficulties.

4.Major Setback: An event where everything falls apart (temporarily or permanently), progressing the story from complications to the final push.

5.Climax: The final event of the main story, the moment of resolution, and the 'biggest spoiler', progressing the story from the final push to the aftermath of the main plot.

You will be provided with the plot synopsis of a movie, and your task is to find all the 5 turning points.

###**Example Output**###: Turning points: {1,3,15,20,31} where each number is the sentence number from the plot synopsis.

DO NOT provide reasoning for the answer. ONLY give the final sentence numbers as output.

Figure 5: Prompt to obtain Turning Points on the Plot Synopsis, in multiple zero-shot and few-shot settings.

QA + DP Prompt

There are six stages (acts) in a film, namely the setup, the new situation, progress, complications and higher stakes, the final push, and the aftermath, separated by five turning points (TPs). TPs are narrative moments from which the plot goes in a different direction, that is from one act to another. The five turning points are described as:

1.Opportunity: An introductory event that occurs after presenting the setting and background of the main characters, driving the narrative from setup to new situation.
2.Change of Plans: An event where the main goal of the story is defined, leading to increased action, driving the narrative from the introduction of new situation to its progress.
3.Point of No Return: An event that pushes the main character(s) to fully commit to their goal, driving the narrative from the progress to its complications and high-stake difficulties.
4.Major Setback: An event where everything falls apart (temporarily or permanently), progressing the story from complications to the final push.
5.Climax: The final event of the main story, the moment of resolution, and the 'biggest spoiler', progressing the story from the final push to the aftermath of the main plot.

You will be provided with the following inputs:

1.Movie Summary: A brief summary of the entire movie, providing context for the overall plot and character arcs.
2.Current Scene Summary: A summary of the scene currently being analyzed.
3.Scene Before Summary: A summary of the scenes immediately preceding the current scene.
4.Scene After Summary: A summary of the scenes immediately following the current scene.

Based on the movie summary, the summaries for the scenes before and after the present scenes, you will be asked questions related to identifying turning points in the plot synopsis.
Answer the questions with a Yes/No. Do not provide reasoning for the answer."

Figure 6: Prompt for DP + QA method for Turning Point Identification. Given all of the shown context, LLMs are asked a sequence of questions, and LLMs have to answer with a Yes or No.

Movie Summarization Prompt

Read the given screenplay and generate a detailed summary. The summary should include the main plot points, key events, character developments, and any significant turning points in the story.
Ensure to cover the **beginning, middle, and end** of the screenplay, highlighting how the story progresses and resolves. Mention important **dialogues, settings, and any subplots** that contribute to the overall narrative.
Write the summary as a **detailed synopsis** of the movie.
Screenplay: *{Screenplay}*

Scene Summarization Prompt

Your task is to summarize the given scene and convert it into a concise summary. The summary should capture the **main plot points** and **key events** of the scene. Just output the summary, **without any reasoning or your own thoughts on the story.**
Scene: {Scene}

Figure 7: Summarization prompts to get the global context for QA+DP method, and reducing the scene size.

References

Meta AI. 2024a. Llama-3.2. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

Meta AI. 2024b. Llama 3.3 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md.

Abhimanyu Dubey et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Michael Hauge. 2017. *Storytelling Made Easy: Persuade and Transform Your Audiences, Buyers, and Clients—Simply, Quickly, and Profitably*. Indie Books International, Oceanside, CA.

Microsoft. 2024. Phi-4 technical report. https://www.microsoft.com/en-us/research/uploads/prod/2024/12/P4TechReport.pdf.

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.

Mistral AI Team. 2025. Mistral small 3. https://mistral.ai/news/mistral-small-3.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681. Association for Computational Linguistics.

An Yang et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.