LitVISTA: A Benchmark for Narrative Orchestration in Literary Text

Mingzhe Lu[1,2], **Yiwen Wang**[3], **Yanbing Liu**[1,2*], **Qi You**[1,2], **Chong Liu**[3], **Ruize Qin**[4], **Haoyu Dong**[1,2], **Wenyu Zhang**[3], **Jiarui Zhang**[1,2], **Yue Hu**[1,2], **Yunpeng Li**[1,2*]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]University of Science and Technology of China [4]University of Melbourne
{liuyanbing, liyunpeng}@iie.ac.cn

https://huggingface.co/datasets/VivldArc/VISTA

Abstract

Computational narrative analysis aims to capture rhythm, tension, and emotional dynamics in literary texts. Existing large language models can generate long stories but overly focus on causal coherence, neglecting the complex story arcs and orchestration inherent in human narratives. This suggests a structural misalignment between model- and human-generated narratives. We therefore position narrative analysis as a diagnostic proxy for generation and propose VISTA Space, a high-dimensional framework for narrative orchestration that unifies human and model perspectives while jointly characterizing narrative function and structure in a common space. We further introduce LitVISTA, a structurally annotated benchmark grounded in literary texts, which operationalizes VISTA Space for systematic evaluation of models' narrative orchestration capabilities. Under an oracle setting with gold event anchors, we evaluate frontier LLMs including GPT, Claude, Grok, and Gemini. Results reveal systematic deficiencies, as current models struggle to jointly capture narrative function and structure and fail to form an integrated global view of literary narrative orchestration. End-to-end analysis further shows that failures are dominated by anchor identification and localization errors. Even advanced thinking modes yield mixed and often limited gains for literary narrative understanding.

1 Introduction

Computational narrative analysis lies at the intersection of natural language processing and literary studies, aiming to represent the complex phenomena of storytelling in structured, analyzable forms (Mani, 2012; Lakoff and Narayanan, 2010; Bal, 1986). While human meaning-making is articulated through language, in literary narratives, this articulation goes beyond simple action sequences (Bruner, 1991; Herman, 2009). Authors deliberately orchestrate events to externalize perceptions, intentions, and mental states, creating a specific rhythm of experience (Zunshine, 2006; Genette et al., 1980). Accordingly, narrative events are not functional equivalents; they are organized to serve distinct structural roles (Barthes and Duisit, 1975; Chatman, 1979). Capturing these differences is central to modeling the pacing and tension (Brewer and Lichtenstein, 1982) that distinguish compelling literature from mere coherence.

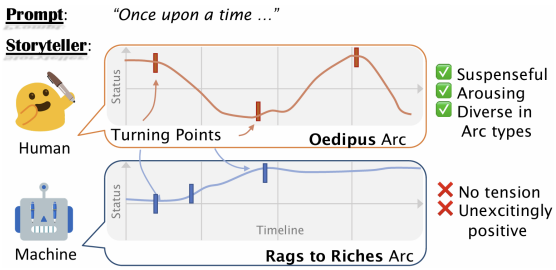

Figure 1: Comparison of story arcs between human and LLM storytellers. This image, reproduced from (Tian et al., 2024), shows that LLM-generated stories often have simpler arcs and earlier turning points, whereas human-authored narratives are more complex.

Existing approaches primarily focus on extending story length while preserving logical consistency (Yi et al., 2025; Park et al., 2024; Xia et al., 2025), but such expansion in scale does not yield a commensurate improvement in the actual reading experience. Recent empirical studies (Tian et al., 2024; Wang et al., 2025) reveal systematic differences between human and model narratives at the level of global story shape. As shown in Figure 1, the vertical axis tracks the protagonist's fortune from bad to good, and the horizontal axis tracks narrative progression from beginning to end. Human-authored stories are distributed across more diverse arc types, whereas model-generated narratives cluster around more uniformly positive and less inflected trajectories, with turning points tending to occur earlier in the story. This pattern sug-

*Corresponding authors

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22371–22396

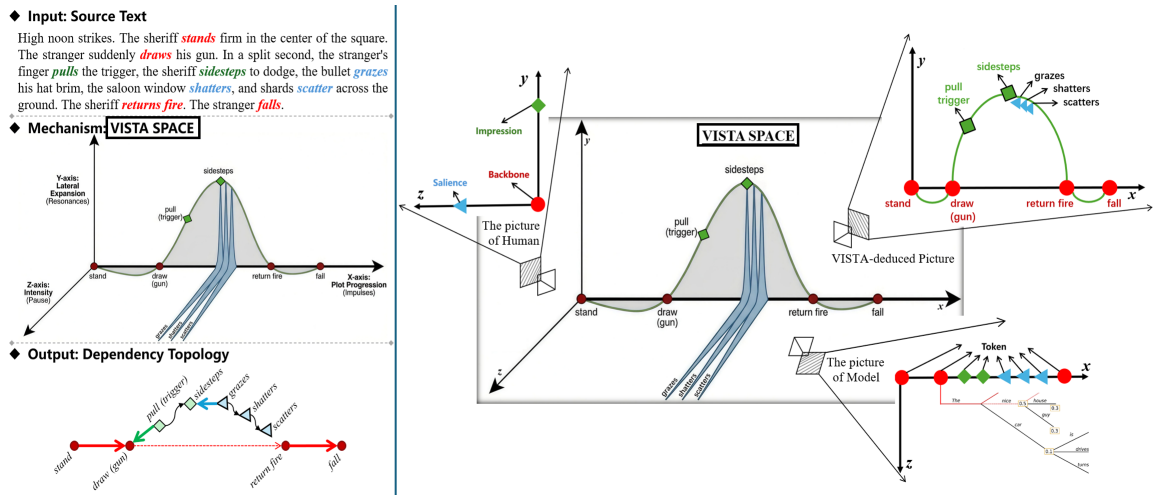

Figure 2: **VISTA Space and its projections.** The center illustrates VISTA Space, a higher-dimensional representation of narrative orchestration. The surrounding panels show three projections: the human picture of narrative experience (left), the LLM picture based on token-level representations (bottom-right), and the VISTA-induced picture (top-right), which situates human and model representations within a unified structural perspective.

gests that current models still underdevelop major setbacks and climactic progression, yielding flatter pacing and weaker suspense than human-authored narratives.

Observations of human reading experience suggest that, after reading, readers do not retain the full textual surface of a story, but instead compress it into a mental picture that preserves the narrative backbone, overall atmosphere, and moments of heightened intensity (van Theo Dijk and Kintsch, 1983). This resonates with *Wittgenstein's picture theory of meaning (Prop. 2.1, 4.01)* (Wittgenstein, 2014), according to which understanding consists in forming internal pictures of facts. Computational models likewise develop internal pictures of stories during understanding and generation through the accumulation of probabilistic signals over text. Although both humans and models form such representational surfaces, the principles governing how these pictures are formed differ, giving rise to a structural misalignment between human narrative experiences and model representations.

To bridge this gap, we introduce VISTA (Viewpoint-Integrated Structural Topology for Analysis) Space, a higher-dimensional framework that situates human and model story pictures in a unified space. Within this space, narrative structure becomes an observable object, and event organization is accessed through a dedicated structural plane. This plane captures how narrative dynamics arise from event arrangement, enabling pacing and tension to be visualized, modeled, and measured, while revealing their effects across human and model representations. These representations must be grounded in concrete, annotatable narrative data (Pustejovsky and Stubbs, 2012) to be empirically accessible.

Importantly, VISTA makes explicit the structural organization that connects narrative production and narrative reception: events are orchestrated to shape pacing, tension, and arc, and these dynamics are in turn reconstructed into a coherent narrative picture. In this work, we position VISTA primarily as a framework for narrative analysis, focusing on recovering narrative topology from a reader-oriented perspective. By rendering narrative orchestration measurable, LitVISTA provides a diagnostic benchmark that reveals where models fail to capture human narrative structure, thereby offering interpretable signals for improving story generation.

Building on this framework, we introduce LitVISTA, a structurally annotated benchmark that makes narrative orchestration explicit in literary texts. LitVISTA represents stories as structured topologies rather than flat sequences, encoding narrative event functions and global dependency relations. Figure 2 illustrates how a literary passage is mapped into VISTA Space, yielding a VISTA-induced dependency topology. To this end, LitVISTA treats Verbs^{+} as minimal narrative anchors, covering canonical verbs and event-denoting nominals, and annotates their roles in propagating narrative structure in a signal-like manner, as man-

ifested in forward progression, lateral expansion, and intensity accumulation. As a result, LitVISTA enables systematic evaluation of models' ability to recover narrative orchestration across events within VISTA Space.

The contributions of this paper can be summarized as follows:

- We propose VISTA Space, a higher-dimensional representational framework that conceptualizes literary narrative understanding as the orchestration of events across structural dimensions, providing a unified view of human and model narrative representations.
- We introduce LitVISTA, a structurally annotated benchmark grounded in literary texts, which operationalizes VISTA Space for empirical evaluation by mapping narratives into structured event topologies.
- Through extensive analysis and evaluation on LitVISTA, we examine the narrative understanding capabilities of existing models, revealing systematic gaps in their ability to orchestrate narrative dynamics.

2 VISTA SPACE

2.1 Narrative Proxy

Human meaning-making is inherently abstract, yet it is expressed through language. In narrative discourse, meaning does not arise from isolated expressions, but from structured configurations that unfold across events. Text therefore serves as the primary medium through which abstract narrative structure is externalized and made observable (Genette et al., 1980). A key step in modeling narrative organization is thus to identify concrete textual anchors that can reliably proxy such structure (Chambers and Jurafsky, 2008).

These anchors must be minimal and well-defined, while remaining representative of underlying narrative dynamics. Verbs naturally fulfill this role as primary carriers of action and change, providing a compact interface between textual form and narrative progression (Davidson, 2001).

To support narrative analysis, we extend the notion of verbs beyond strictly grammatical definitions by including event-denoting nominals such as *marriage* and *departure*. These nominal forms retain the argument structure and event semantics of their verbal counterparts (Pustejovsky et al., 2003), allowing them to participate in event representations in a manner similar to verbs.

Terminological Distinction. Throughout this work, we use the term *Verb*$^+$ to denote a broader class of event anchors. We explicitly distinguish narrative events as abstract units of meaning from Verbs$^+$ as their concrete textual anchors used for computational modeling.

2.2 Narrative Configuration

Narrative meaning is not fully captured by the sum of discrete Verbs$^+$; rather, it arises from their configuration across the text (Polkinghorne, 2010). While a list of Verbs$^+$ can describe what happened, it does not account for how information is structured and presented over time, including the ordering, emphasis, and contextual dependencies among events. The essence of narrative, therefore, lies not in the isolated presence of Verbs$^+$, but in their contribution to the structural architecture.

Within the narrative architecture, different Verbs$^+$ assume distinct structural functions. In practice, the same Verbs$^+$ describing the same situation at the same textual position may be assigned different structural roles within different narrative orchestrations (Chatman, 1979), with concrete illustrations provided in Appendix A. These dynamic role assignments go beyond causality. They allow narrative organization to vary independently of action, giving rise to global properties such as pacing, tension, and rhythm (Sternberg, 1992).

2.3 Narrative Computation

To implement this structural architecture, we introduce VISTA Space as a computational topology. A key distinction is made between discrete chronological progression and continuous lateral expansion.

Two variables are introduced to represent these dimensions: a discrete *Narrative Progress Index* (τ) that indexes story stages, and a continuous *Marginal Increment* (δ) that measures descriptive expansion without advancing the stage.

Definition 1 (Metric Domains). *The narrative coordinate space is formally constrained by the following domains:*

$$\tau \in \mathbb{N}, \quad \delta \in (0, 1) \subset \mathbb{R}. \tag{1}$$

Narrative discourse reconfigures underlying events, distinct from a flat chronology. To capture this structure, we define the *orchestration topology*

through a functional mapping that determines how an anchor operates on the narrative state.

Definition 2 (Anchor Topology). *Let E_τ denote the narrative state at progress index τ. The transition logic $\mathcal{F}(v)$ defines the operation of an anchor on this state:*

$$\mathcal{F}(v) = \begin{cases} E_\tau \to E_{\tau+1}, \\ E_\tau \to E_{\tau+\delta}, \\ E_\tau \to E_\tau. \end{cases} \quad (2)$$

This transition logic establishes a three-dimensional narrative space constructed by three primary functional roles, with a residual category for syntactic elements:

Impulses ($\mathcal{V}_I$): Anchors where $\mathcal{F}(v) : E_\tau \to E_{\tau+1}$. These form the narrative backbone (the X-axis), advancing the plot to a new stage.

Resonances ($\mathcal{V}_R$): Anchors where $\mathcal{F}(v) : E_\tau \to E_{\tau+\delta}$. These form the enveloping texture (the Y-axis), expanding descriptively without advancing the stage.

Pauses ($\mathcal{V}_P$): Anchors where $\mathcal{F}(v) : E_\tau \to E_\tau$. These generate vertical intensity (the Z-axis), inducing temporal suspension to maximize the expressive density of the current moment.

Non-Events ($\mathcal{V}_\emptyset$): Syntactic elements that do not contribute to the topology.

Definition 3 (Narrative Dependency). *The narrative topology is a directed graph $G = (\mathcal{V}, \mathcal{E})$. The set of valid edges $\mathcal{E}$ is the union of two hierarchical layers:*

$$\mathcal{E} \subseteq \underbrace{(\mathcal{V}_R \times \mathcal{V}_I)}_{\text{Primary Layer}} \cup \underbrace{(\mathcal{V}_P \times (\mathcal{V}_I \cup \mathcal{V}_R))}_{\text{Recursive Layer}}. \quad (3)$$

This formation dictates that Resonances must attach directly to the Backbone ($\mathcal{V}_I$), whereas Pauses may attach recursively to existing structures ($v_P \to v_R \to v_I$). For benchmark operationalization, local non-backbone material may be serialized through intermediate heads for convenience, while preserving ultimate dependence on the same governing backbone state.

Definition 4 (VISTA Space). *The VISTA Space is a three-dimensional narrative orchestration space, with its projection planes representing human, model, and computational perspectives.*

As shown in Figure 2, we map $\mathcal{V}_I$, $\mathcal{V}_R$, and $\mathcal{V}_P$ into this 3D coordinate system. The X-axis represents the narrative backbone, driven by $\mathcal{V}_I$ and quantified by the index τ. The Y-axis characterizes $\mathcal{V}_R$, which emerges around $\mathcal{V}_I$ and is quantified by $N\delta$, where N denotes the number of $\mathcal{V}_P$ elements along the Z-axis that correspond to the current $\mathcal{V}_R$. The Z-axis is dedicated to $\mathcal{V}_P$, functioning as a unit impulse with amplitude 1, signifying the discrete presence of a pause.

While it might seem intuitive to merge the Z-axis with the Y-axis, as both capture aspects of narrative progression, it is important to note that the VISTA Space is derived from the orthogonal projections of human and model representations. As shown in the left panel of Figure 2, these projections are distinct in the human narrative picture. Consequently, modeling the Z-axis is indispensable for capturing this distinct behavioral feature.

3 LitVISTA

In this section, we formally introduce **LitVISTA**, a structurally annotated benchmark for evaluating and diagnosing models' narrative orchestration capabilities in literary texts.

3.1 Dataset Construction

To ensure rigorous corpus quality, we constructed LitVISTA based on the *LitBank* (Bamman et al., 2020) corpus.

We adopt LitBank because it provides a curated literary corpus and an established event-centric annotation layer that closely matches our Verb$^+$ notion, covering both verbal and event-denoting nominal anchors. This event layer can be treated as a fixed upstream component in realistic pipelines, allowing LitVISTA to focus on higher-level narrative structure.

The dataset consists of complete narrative chapters, enabling unconstrained long-range topological structure with interleaved $\mathcal{V}_I$, $\mathcal{V}_R$, and recursive $\mathcal{V}_P$ attachments to assess holistic event integration capabilities. Given this design focus, LitVISTA is intentionally positioned as a high-precision evaluation benchmark for narrative orchestration rather than a large-scale training corpus.

3.2 Annotation Procedure

3.2.1 Annotator Background

The broader development effort involved seven contributors. In the early stage, the team was led by one PhD researcher with interdisciplinary expertise in computational narratology and AI, who was responsible for protocol design, pilot exploration, and guideline refinement. The large-scale annotation was then carried out by six annotators with

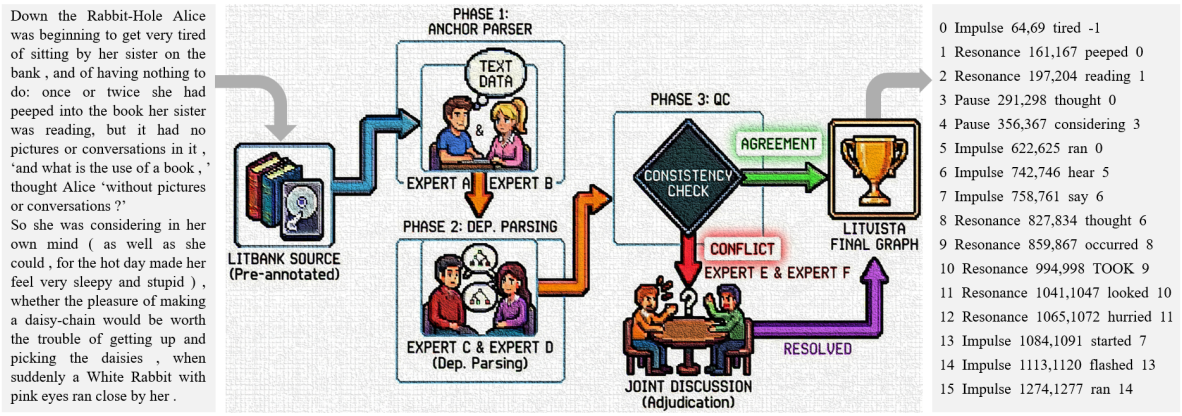

Figure 3: The process begins with LitBank text data. Experts A and B independently annotate Verb$^+$ roles in Phase 1. In Phase 2, dependency parsing is conducted by Experts C and D independently. Phase 3 resolves any conflicts through adjudication, producing the final LitVISTA graph.

NLP backgrounds, including three Master's students and three PhD candidates. This division of labor reflects a practical trade-off between narratological grounding and the efficient execution of a technically constrained annotation protocol.

The annotation was conducted in three phases, as illustrated in Figure 3. Phases 1 and 2 were each completed by a pair consisting of one PhD annotator and one Master's annotator, whereas Phase 3 was carried out by a different pair, again consisting of one PhD annotator and one Master's annotator. All annotators participated on a voluntary basis and received no monetary compensation.

3.2.2 Annotation Details

In Phase 1, we annotated using the LitBank event triggers as the candidate inventory, which was retained almost entirely. Only a very small number of non-eventive or otherwise non-informative items were removed. The main task in this phase was therefore to assign each retained trigger to its topological category ($\mathcal{V}_I$, $\mathcal{V}_R$, or $\mathcal{V}_P$).

In Phase 2, annotators labeled dependencies for these retained triggers. $\mathcal{V}_I$ nodes form the narrative backbone as a single chain, with the first $\mathcal{V}_I$ assigned index -1 to mark the start. $\mathcal{V}_R$ is anchored to the governing $\mathcal{V}_I$ it elaborates, usually locally but sometimes across longer spans, as in flashback or inserted narration. Consecutive $\mathcal{V}_P$ spans are interpreted as co-dependent realizations of the same governing structural state rather than as progressive $\mathcal{V}_P \rightarrow \mathcal{V}_P$ chains. More detailed annotation guidelines are provided in Appendix B.

3.2.3 Annotation Summary

Under this protocol, inter-annotator consistency was 0.49 in Phase 1 and 0.76 in Phase 2. We attribute this gap to the openness of literary interpretation: topological role assignment admits multiple plausible readings, whereas dependency relations are usually more explicit. Disagreements were further resolved in Phase 3 to produce the final consensus-derived gold standard, and no inter-annotator score is reported for this stage because it was adjudicative rather than independent. Based on these finalized annotations, we next summarize the resulting dataset statistics.

Table 1: Statistics of the LitVISTA Dataset. Length is measured in tokens.

Metric	Train	Val	Test
Avg. Length	10.2k	9.9k	10.7k
Avg. # $\lvert V_I \rvert$	13.04	18.20	11.00
Avg. # $\lvert V_R \rvert$	59.90	78.40	49.10
Avg. # $\lvert V_P \rvert$	3.84	3.50	3.90
Avg. Cross Dep.	75.67	100.10	63.90

Using the final adjudicated annotations, we partitioned the dataset into training, validation, and test sets with an 8:1:1 ratio. Table 1 summarizes the average text length, the distribution of Verb$^+$ subtypes ($\mathcal{V}_I$, $\mathcal{V}_R$, $\mathcal{V}_P$), and the number of cross dependencies in each split. The predominance of $\mathcal{V}_R$ reflects the descriptive emphasis commonly observed in literary narrative discourse, while the frequent cross dependencies further indicate the structural complexity of the annotated narratives.

Table 2: **Oracle Evaluation Results on LitVISTA.** We employ a **heatmap visualization** where color intensity corresponds to performance: **Darker** indicates higher scores, and lighter indicates lower scores. Models are sorted by the harmonic mean.

	Oracle Eval						*Overall*
	Anchor Parsing			**Dep. Parsing**			**Harmonic**
	P	R	F1	P	R	F1	**Mean↑**
GPT-5.1	0.4066	0.3393	0.3033	0.0746	0.0464	0.0460	0.0799
GPT-5	0.4823	0.4862	0.4348	0.1006	0.1121	0.0745	0.1272
Doubao-seed-1.6-thinking	0.2914	0.2956	0.2890	0.2066	0.1772	0.1456	0.1936
Claude-opus-4.5-thinking	0.2674	0.2913	0.2646	0.2012	0.1577	0.1641	0.2026
GPT-5.2-pro	0.4543	0.5179	0.4540	0.2090	0.2220	0.1699	0.2473
DeepSeek-v3.2-thinking	0.3123	0.3440	0.3140	0.2564	0.2799	0.2219	0.2600
ChatGLM-4.7	0.3708	0.3225	0.3362	0.2890	0.2314	0.2182	0.2646
Gemini-2.5-pro-thinking	0.3161	0.3819	0.3083	0.2992	0.3285	0.2631	0.2839
Grok-4	0.3297	0.2619	0.2669	0.4185	0.3057	0.3365	0.2977
GPT-5-thinking	0.2327	0.2174	0.1995	0.6771	0.6412	0.6478	0.3051
Claude-sonnet-4.5	0.2377	0.2655	0.2254	0.4981	0.5262	0.4728	0.3053
Qwen3-235B-a22	0.2946	0.3528	0.2701	0.3670	0.4225	0.3538	0.3063
Gemini-2.5-pro	0.3360	0.4178	0.3346	0.3162	0.3562	0.2911	0.3113
Grok-4.1-thinking	0.3930	0.4609	0.4086	0.2798	0.3252	0.2669	0.3229
Doubao-seed-1.6	0.2863	0.2780	0.2815	0.5105	0.4869	0.4618	0.3498
GPT-5.1-thinking	0.2662	0.2458	0.2410	0.8135	0.6441	0.6799	0.3559
Gemini-3-pro-preview-thinking	0.3619	0.3879	0.3285	0.4209	0.4674	0.4061	0.3632
Claude-opus-4.5	0.3058	0.3368	0.2947	0.5147	0.5627	0.5083	0.3731
GPT-4o	0.3169	0.2548	0.2519	0.7807	0.7383	0.7333	0.3750
GPT-5.2	0.4171	0.4776	0.3983	0.4010	0.4085	0.3585	0.3774
Claude-sonnet-4.5-thinking	0.3322	0.3935	0.3309	0.4720	0.5160	0.4575	0.3840
Gemini-3-pro-preview	0.3817	0.4171	0.3495	0.4928	0.5175	0.4736	0.4022
DeepSeek-v3.2	0.3089	0.3403	0.3098	0.5975	0.6222	0.5783	0.4035
Claude-opus-4	0.3868	0.4284	0.3779	0.4603	0.4923	0.4414	0.4072
Claude-sonnet-4	0.2893	0.2987	0.2838	**0.8142**	**0.8115**	**0.7968**	0.4185
Claude-opus-4-thinking	0.3984	0.4426	0.3984	0.5157	0.5197	0.4708	0.4316
Claude-sonnet-4-thinking	**0.4947**	**0.5236**	**0.4914**	0.6104	0.5981	0.5624	**0.5245**

3.3 LitVISTA Task

We define the LitVISTA task as a narrative structure reconstruction problem, and evaluate it under an *oracle event-level setting* that requires reconstructing nodes and edges in a single pass. This one-stage formulation mandates the model to capture a global narrative coherence, moving beyond iterative local refinements that often suffer from error propagation.

3.3.1 Oracle Evaluation

We adopt an oracle event-level setting to isolate models' ability to perform high-level narrative orchestration, under the assumption that candidate event anchors (Verb$^+$) are provided.

Formally, in this oracle setting, the model is provided with the raw text $\mathcal{T}$ along with a set of candidate nodes $\mathcal{V}_{\text{cand}}$ (corresponding to Verb$^+$ tokens).

The model must simultaneously determine the topological roles for these candidates and resolve their dependencies. This joint optimization is described by the following equations:

$$\begin{cases} r^* = \arg \max_{r \in \{\mathcal{V}_I, \mathcal{V}_R, \mathcal{V}_P\}} P(r \mid v, \mathcal{T}), \\ u^* = \arg \max_{u \in \mathcal{V}_{\text{cand}} \setminus \{v\}} P(v \to u \mid v, r^*, \mathcal{T}). \end{cases} \tag{4}$$

where r^* represents the predicted topological role, and u^* represents the predicted parent anchor from the candidates (excluding v itself). This formulation ensures that node classification and dependency resolution are interdependent, reconstructing directed edges that enforce the recursive structure of the narrative.

3.3.2 Eval Metrics

Given the discrete node and edge definitions in LitVISTA, we report standard Precision (P), Recall (R), and F1 for both topological role prediction and dependency reconstruction. These metrics provide an operational measure of how well a model recovers the annotated narrative graph at the levels of both node classification and edge prediction. Higher scores indicate closer agreement with the gold-standard topology, although they should be interpreted as benchmark measures rather than exhaustive measures of literary understanding.

4 Experiments

4.1 Evaluation Setup

We evaluate models' narrative orchestration capabilities on LitVISTA, which renders the VISTA Space computable.

Our main experiments use the oracle event-level setting, where models are evaluated on the provided test set with gold-standard event anchors (trigger + character index). This setting isolates high-level narrative reasoning from upstream extraction errors and enables fair comparison across models. In practical applications, LitVISTA can also be used in a modular pipeline, where a separate event extractor is first applied and its outputs are then passed to LitVISTA for structural prediction; in that case, the resulting scores should be interpreted as reflecting both upstream extraction quality and downstream narrative reasoning.

We consider widely adopted model families, including GPT, Gemini, Grok, and Claude, and compare reasoning-enabled variants with their non-reasoning counterparts. Detailed experimental configurations, including hyperparameter settings and prompt designs, are provided in Appendix D.

We also examine an end-to-end setting in Appendix E, where models are given only raw text. As shown there, current frontier LLMs fail primarily at anchor identification and span localization, which causes cascading failures before role assignment and dependency prediction can be meaningfully assessed. We therefore do not treat end-to-end scores as the primary measure of narrative orchestration ability, but rather as a diagnostic signal of the interaction between upstream extraction and downstream reasoning.

4.2 Result Analysis

We report the performance of all baselines in Table 2, following the oracle evaluation protocol defined in Section 3.3.1. To intuitively reveal the underlying trade-offs and behavioral shifts hidden within these numerical comparisons, we further visualize the performance distribution in Figure 4.

4.2.1 Distribution of Performance

The heatmap visualization in Table 2 provides a clear overview of the overall performance landscape, revealing a pronounced asymmetry between Anchor Parsing and Dependency Parsing across models. Specifically, high performance in one subtask is frequently accompanied by substantially weaker performance in the other, and models that simultaneously achieve strong results on both dimensions are notably scarce. This pattern is most evident in the absence of consistently dark regions across both blocks within the same model row. The same trend is corroborated by the scatter plot in Figure 4, where the upper-right quadrant corresponding to strong performance on both tasks remains largely unpopulated.

4.2.2 Impact of Thinking

The connecting lines in Figure 4 show that enabling thinking induces systematic shifts rather than uniform improvements. In some cases, thinking substantially enhances structural modeling. For example, GPT-5.1-thinking exhibits a large performance gain relative to its base counterpart, while simultaneously reducing Anchor accuracy, indicating a redistribution of modeling capacity rather than a consistent improvement.

However, this behavior does not generalize across models. As shown in Table 2, thinking variants of DeepSeek-v3.2, Claude-opus-4.5, and Gemini-2.5-pro display an overall downward or unstable performance trend when compared with their non-thinking counterparts. Despite isolated improvements in specific configurations, enabling thinking often coincides with broad performance degradation across parsing tasks, suggesting that the induced reasoning process may constrain rather than enrich the model's representational flexibility.

Taken together, these results indicate that thinking primarily reshapes how models allocate capacity, rather than consistently improving narrative understanding. When narrative modeling is dominated by narrow causal reasoning, gains in localized structure may come at the expense of global

event organization. This trade-off is especially limiting for literary narratives, where meaning arises from pacing, tension, figurative relations, and nonlinear structure beyond simple causality.

4.2.3 Family-Specific Patterns

While the above analysis already suggests (i) a scarcity of models that are simultaneously strong on both Anchor and Dependency parsing and (ii) non-uniform shifts induced by enabling thinking, these shifts are not arbitrary. Instead, the explicitly labeled models in Figure 4 exhibit family-specific regularities: within the same model family, the thinking-enabled variants tend to move in a more consistent direction, whereas different families display markedly different trajectories.

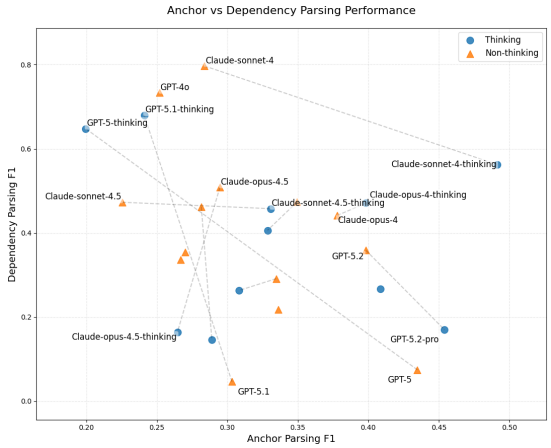

Figure 4: Oracle evaluation results. The scatter plot shows Anchor F1 (x-axis) versus Dependency F1 (y-axis) for each model.

Concretely, Claude variants largely follow a coherent trend in how thinking reshapes the balance between anchor identification and relational reasoning, while GPT variants exhibit a distinct and often contrasting trend. This divergence indicates that "thinking" acts less like a universal improvement knob and more like an amplifier of pre-existing inductive biases encoded by the underlying model family. The connecting lines for the (GPT-5, *-thinking) and (Claude-opus-4.5, *-thinking) pairs appear nearly orthogonal in Figure 4, a pattern that further underscores this conclusion.

5 Further Analysis

In this section, we delve into the unique narrative topologies in LitVISTA to explain why models struggle to comprehend them.

5.1 Long-Range Narrative Dependencies

As shown in Figure 5, narrative dependency frequency varies with the absolute textual distance between dependent Verb$^+$ nodes. If such dependencies primarily followed textual proximity, the distribution would concentrate in short-distance intervals.

The observed data, however, exhibit a marked deviation. Although short-range dependencies are common, a substantial proportion, particularly involving Impulse and Pause nodes, spans hundreds or even thousands of characters. Crucially, for several dependency types, long-range associations persist without attenuation.

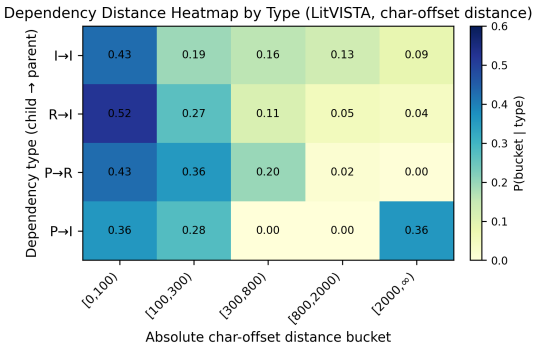

Figure 5: Frequency of narrative dependencies by absolute character offset distance. The X-axis represents distance buckets, and the Y-axis shows different dependency types.

These findings in dependency patterns suggest that textual proximity is a weak predictor in LitVISTA. *Narrative relations frequently link events that are distant in the linear sequence, because the narrative flow disrupts the timeline or plants foreshadowing*, reflecting higher-level discourse organization. This structural mismatch accounts for the difficulty of understanding, as span-local or next-token-biased models are ill-equipped to capture such non-local topology.

5.2 Lexical Grounding of Narrative Roles

Finally, we examine whether narrative roles exhibit lexical regularities by projecting each sufficiently frequent Anchor word's empirical distribution over Impulse, Resonance, and Pause into a two-dimensional role-preference space.

Figure 6 reveals a structured lexical landscape. Action-oriented verbs such as *cast*, *met*, and *reached* cluster in regions strongly biased toward Impulse, while perception and discourse-related verbs (e.g., *looked*, *said*) occupy Resonance-

dominated regions. A smaller set of words aligns with Pause, often corresponding to evaluative or state-descriptive expressions.

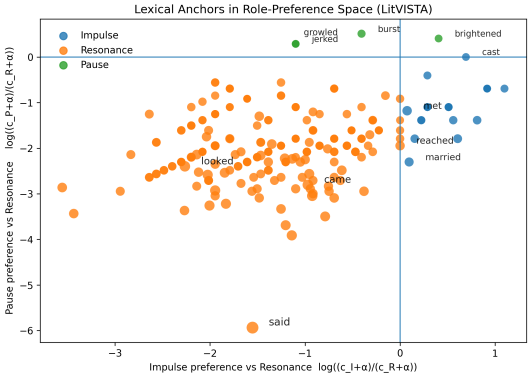

Figure 6: Lexical anchors in role-preference space. The X-axis represents Impulse–Resonance preference, and the Y-axis represents Pause–Resonance preference. Each point corresponds to a lexical item.

Importantly, these clusters emerge without any lexical supervision. The fact that coherent semantic groupings arise purely from narrative role statistics indicates that LitVISTA captures stable associations between lexical items and narrative function. This further supports the claim that the VISTA Space reflects meaningful narrative structure rather than arbitrary annotation artifacts.

6 Related Work

Recent work in computational narrative analysis and computational literary studies has shifted from local semantics toward discourse- and structure-level analysis of narrative phenomena, emphasizing plot organization and narrative dynamics in literary texts (Piper, 2023). This shift is reinforced by methodological surveys that identify narrative structure as a central object of contemporary computational literary research (Hatzel et al., 2023). Related efforts have introduced discourse- and clause-level resources to support large-scale structural analysis of narrative texts (Troiano and Vossen, 2024).

Event-centric representations remain a common foundation for narrative modeling, with recent work examining how event sequences can be organized into coherent storylines or structured graphs (Vijayaraghavan and Roy, 2023). Other studies investigate narrative consistency by modeling global structural constraints over event sequences rather than isolated relations (Zhu et al., 2023).

In parallel, the rise of frontier large language models has motivated evaluations of narrative understanding on long-form inputs, particularly focusing on long-context and multi-step reasoning (Sprague et al., 2024). Additional work analyzes narrative coherence in generated stories, revealing systematic structural failures despite surface fluency (Zhu et al., 2023). More recently, evaluations have probed subtext and implicit meaning comprehension in literary narratives (Subbiah et al., 2024). At a broader level, structured benchmarks (Wu et al., 2025) have also been proposed for evaluating narrative generation and writing quality.

7 Conclusion

This paper introduces VISTA Space, a representational framework that unifies human and model perspectives on narrative structure, and LitVISTA, a high-precision benchmark for evaluating narrative orchestration in literary texts. Under oracle evaluation, current frontier language models exhibit persistent difficulty in jointly recovering narrative function and dependency structure, indicating that literary narrative understanding remains fragmented rather than globally integrated. End-to-end analysis further shows that current failures are dominated by upstream anchor identification and localization errors, suggesting that narrative orchestration and event extraction remain only partially coupled in existing systems. We therefore view LitVISTA not as a benchmark for generation itself, but as a diagnostic benchmark for understanding where narrative generation systems still fall short. We hope LitVISTA will support future work on structure-aware narrative analysis, stronger event-grounded modeling, and ultimately more human-like long-form story generation.

8 Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. U21B2009).

9 Limitations

While LitVISTA serves as a rigorous benchmark for narrative orchestration, we acknowledge several limitations in our current work:

Reliance on Oracle Settings: Our primary experimental results rely on an oracle setting where candidate event anchors are provided. As discussed in Appendix E, we found that even frontier LLMs (e.g., GPT-5, Gemini-Pro) currently struggle to perform valid end-to-end narrative reconstruction, primarily due to failures in low-level anchor identification and localization. While this highlights the difficulty of the proposed task, it also limits our current ability to evaluate fully autonomous narrative analysis systems without upstream assistance.

Domain and Language Specificity: LitVISTA is grounded in the LitBank corpus, which focuses on English literary texts from the public domain. While this choice ensures high-quality, expert-annotated narrative structures and avoids copyright issues, the findings may not fully generalize to other languages, modern internet fiction, or non-literary narrative forms where implicit structural cues might differ.

Annotation Scalability: To ensure topological consistency and theoretical depth, we employed a resource-intensive expert annotation process with consensus-based adjudication. This high standard for data quality inevitably constrains the scale of our dataset compared to automatically constructed corpora. Consequently, LitVISTA is designed as a high-precision evaluation benchmark rather than a large-scale training corpus.

Subjectivity of Literary Interpretation: Although we enforce strict axiomatic guidelines (Appendix B) to minimize ambiguity, literary boundaries and structural roles involve inherent interpretative subjectivity. Our "gold standard" represents a coherent, consensus-derived structural reading, but it may not capture every possible valid interpretation of a complex literary passage.

Participant Modeling: Our current formulation focuses on event-centric structure, modeling narrative orchestration through event anchors and their functional roles. It does not explicitly represent participants or character-level interactions, which are also central to narrative understanding. Extending VISTA to incorporate participant structure remains an important direction for future work.

10 Ethical Considerations

Data Source, Licensing, and Privacy: The LitVISTA benchmark builds upon the LitBank corpus, a dataset of 100 English-language fiction works sourced from Project Gutenberg. Since these texts belong to the public domain, the dataset contains no personally identifying information (PII) of living individuals. LitBank is licensed under a Creative Commons Attribution 4.0 International License (CC-BY 4.0), and we strictly adhere to these terms in distributing our derived artifacts.

Intended Use: Aligning with the scientific intent of Project Gutenberg and LitBank, we release LitVISTA to support research in natural language processing and computational humanities. The benchmark is intended solely for academic research to facilitate the study of narrative dynamics and evaluate the structural capabilities of large language models.

Annotator Compensation and Process: The annotation effort involved a small team of contributors; details of annotator background are provided in Section 3.2.1, while annotation details are described in Section 3.2.2. All participants were informed of the research goals and workload in advance, participated on a voluntary basis, and received no monetary compensation.

Use of AI Tools: We permitted annotators to use AI tools solely for summarizing broader literary contexts and clarifying plot backgrounds, mitigating the time cost of reading full novels. The core tasks of identifying narrative anchors, assigning topological roles, and resolving dependencies were performed entirely manually by human annotators. No AI-generated labels were used in the construction of the gold standard dataset.

Potential Risks and Subjectivity: Literary interpretation involves inherent subjectivity. To mitigate this, we established a multi-phase annotation strategy supported by a Theoretical Codebook (Appendix B) and consensus-based adjudication. While LitVISTA represents a cohesive structural interpretation, users should be aware of the subjective nature characterizing computational literary studies.

References

Mieke Bal. 1986. Narratology: Introduction to the theory of narrative.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in english literature. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 44–54. European Language Resources Association.

Roland Barthes and Lionel Duisit. 1975. An introduction to the structural analysis of narrative. *New Literary History*, 6:237.

William F. Brewer and Edward H. Lichtenstein. 1982. Stories are to entertain: A structural-affect theory of stories. technical report no. 265.

Jérôme Seymour Bruner. 1991. The narrative construction of reality. *Critical Inquiry*, 18:1 – 21.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Seymour Benjamin Chatman. 1979. Story and discourse: Narrative structure in fiction and film.

David L. Davidson. 2001. The logical form of action sentences.

Gérard Genette, Jeanne Ericsson Lewin, and Jonathan D. Culler. 1980. Narrative discourse : an essay in method. *Comparative Literature*, 32:413.

Hans Ole Hatzel, Haimo Stiemer, Chris Biemann, and Evelyn Gius. 2023. Machine learning in computational literary studies. *it Inf. Technol.*, 65(4-5):200–217.

David Herman. 2009. Basic elements of narrative.

George Lakoff and Srini Narayanan. 2010. Toward a computational model of narrative. In *Computational Models of Narrative, Papers from the 2010 AAAI Fall Symposium, Arlington, Virginia, USA, November 11-13, 2010*, AAAI Technical Report. AAAI.

Inderjeet Mani. 2012. *Computational Modeling of Narrative*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Kyeongman Park, Nakyeong Yang, and Kyomin Jung. 2024. Longstory: Coherent, complete and length controlled long story generation. In *Advances in Knowledge Discovery and Data Mining - 28th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2024, Taipei, Taiwan, May 7-10, 2024, Proceedings, Part II*, Lecture Notes in Computer Science, pages 184–196. Springer.

Andrew Piper. 2023. Computational narrative understanding: A big picture analysis. In *Proceedings of the Big Picture Workshop*, pages 28–39, Singapore. Association for Computational Linguistics.

Donald E. Polkinghorne. 2010. Narrative knowing and the human sciences.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering, Papers from 2003 AAAI Spring Symposium, Stanford University, Stanford, CA, USA*, pages 28–34. AAAI Press.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning - a Guide to Corpus-Building for Applications*. O'Reilly.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Meir Sternberg. 1992. Telling in time (ii): Chronology, teleology, narrativity. *Poetics Today*, 13:463.

Melanie Subbiah, Sean Zhang, Lydia B. Chilton, and Kathleen R. McKeown. 2024. Reading subtext: Evaluating large language models on short story summarization with writers. *Trans. Assoc. Comput. Linguistics*, 12:1290–1310.

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, Miami, Florida, USA. Association for Computational Linguistics.

Enrica Troiano and Piek T. J. M. Vossen. 2024. CLAUSE-ATLAS: A corpus of narrative information to scale up computational literary analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 3283–3296. ELRA and ICCL.

van Theo Dijk and Walter Kintsch. 1983. Strategies of discourse comprehension.

Prashanth Vijayaraghavan and Deb Roy. 2023. M-sense: Modeling narrative structure in short personal narratives using protagonist's mental representations. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13664–13672. AAAI Press.

Wenqing Wang, Mingqi Gao, Xinyu Hu, and Xiaojun Wan. 2025. Towards A "novel" benchmark: Evaluating literary fiction with large language models. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 21648–21673. Association for Computational Linguistics.

Ludwig Wittgenstein. 2014. Tractatus logico-philosophicus. *Nordic Wittgenstein Review*.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. 2025. Writingbench: A comprehensive benchmark for generative writing. *CoRR*, abs/2503.05244.

Haotian Xia, Hao Peng, Yunjia Qi, Bin Xu, Juanzi Li, Lei Hou, and Xiaozhi Wang. 2025. Storywriter: A multi-agent framework for long story generation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM 2025, Seoul, Republic of Korea, November 10-14, 2025*, pages 6559–6563. ACM.

Qiang Yi, Yangfan He, Jianhui Wang, Xinyuan Song, Shiyao Qian, Miao Zhang, Li Sun, and Tianyu Shi. 2025. SCORE: story coherence and retrieval enhancement for AI narratives. *CoRR*, abs/2503.23512.

Lixing Zhu, Runcong Zhao, Lin Gui, and Yulan He. 2023. Are NLP models good at tracing thoughts: An overview of narrative understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10098–10121, Singapore. Association for Computational Linguistics.

Lisa Zunshine. 2006. Why we read fiction: Theory of mind and the novel.

A Illustrating Narrative Configuration

This appendix provides concrete illustrations of *Narrative Configuration* as defined in Section 2.2. The goal is to clarify how different configurations of the same underlying events give rise to distinct narrative structures through the functional roles of $\mathcal{V}_I$, $\mathcal{V}_R$, and $\mathcal{V}_P$.

Across all examples, the underlying event content remains fixed. What varies is the structural organization imposed by narrative orchestration. These examples demonstrate how narrative meaning emerges from structural configuration rather than from the events themselves.

A.1 Structural Backbone

At the most basic level, a narrative can be represented as a minimal progression chain composed exclusively of Impulses ($\mathcal{V}_I$). This backbone encodes the irreversible advancement of the narrative state and preserves logical continuity between events.

Consider the following two variants, which share the same set of Impulse events but differ in their ordering:

Variation A (Chronological): ... Alice *poisons*$_{v_1}$ the coffee ... Bob *drinks*$_{v_2}$ it ... finally ... Bob *is saved*$_{v_3}$ by emergency treatment ...

Variation B (Reordered): ... Bob *drinks*$_{v_2}$ the coffee ... finally ... Bob *is saved*$_{v_3}$ after a rescue ... the cause is revealed ... Alice had *poisoned*$_{v_1}$ the cup ...

Both variants rely exclusively on $\mathcal{V}_I$ events and therefore encode the same narrative backbone. However, reordering the Impulses alters the distribution of information over narrative time, affecting reader expectation without introducing additional structural operations. This illustrates that even within $\mathcal{V}_I$, narrative effects can arise from configuration rather than content.

A.2 Lateral Expansion via Resonance

While the Impulse chain defines narrative progression, it offers limited expressive capacity. Structural richness emerges when Resonances ($\mathcal{V}_R$) are introduced to laterally expand the narrative state without advancing the progress index.

Using the same Impulse backbone (*poisons*$_{v_1}$, *drinks*$_{v_2}$, *is saved*$_{v_3}$), consider the following configuration:

Variation C (Resonant Expansion): Snow *falls*$_{v_R}$ outside while warm jazz *plays*$_{v_R}$. ... Bob *drinks*$_{v_2}$ the coffee ... finally ... Bob *is saved*$_{v_3}$ after a rescue ...

Here, the Resonance events attach to the Impulse *drinks*$_{v_2}$, enriching the narrative state without modifying the progression itself. Structurally, $\mathcal{V}_R$ introduces descriptive expansion that shapes reader interpretation while remaining subordinate to the backbone. The resulting narrative effect emerges from the accumulation of contextual information rather than from additional events.

A.3 Vertical Deepening via Pause

Pauses ($\mathcal{V}_P$) operate orthogonally to both progression and expansion. They suspend narrative advancement and concentrate representational density within a single narrative moment.

Consider the following configuration:

Variation D (Pause-Induced Density): ... Bob *drinks*$_{v_2}$ the coffee, the cup *clatters*$_{v_P}$ to the floor, a high-pitched ring *drowns*$_{v_P}$ out all sound, the ceiling light *stretches*$_{v_P}$ into a star, his heartbeat *slams*$_{v_P}$ to a halt ... finally ... Bob *is saved*$_{v_3}$ ...

This sequence of Pause events decomposes a single narrative instant into multiple micro-observations. Rather than advancing the narrative state, these events intensify local representation, producing high expressive density within a fixed temporal window. Structurally, this corresponds to movement along the Z-axis of VISTA Space.

A.4 Structural Choice and Global Interpretation

Although Resonances and Pauses are not required to preserve logical continuity, their inclusion determines how the narrative is globally interpreted. Different configurations over the same backbone yield systematically different narrative structures.

The following examples illustrate how discretionary structural choices shape global narrative interpretation:

Variation E (Internalization): ... Bob *drinks*$_{v_2}$ the coffee ... on the operating table, Bob *recalls*$_{v_P}$ his promise to a dying friend. This memory *ignites*$_{v_R}$ his will to survive ... finally ... Bob *is saved*$_{v_3}$ ...

Variation F (Externalization): ... Bob *drinks*$_{v_2}$ the coffee ... the camera *pans*$_{v_R}$ to a generic logo, then *zooms*$_{v_P}$ in on the brand of the life-support machine ... finally ... Bob *is saved*$_{v_3}$ ...

Although both variants preserve the same Impulse structure, their configurations emphasize different narrative dimensions. Variation E concentrates representational mass on internal state transitions, whereas Variation F allocates structural attention to external objects. These differences arise entirely from narrative configuration rather than from changes to event content.

A.5 Conclusion: Structural Implications for Computation

These examples demonstrate that narrative meaning is encoded in the structural configuration of events rather than in the events themselves. The Impulse backbone ensures logical progression, while Resonances and Pauses govern expansion and intensification within VISTA Space.

By formalizing these roles and their dependencies, VISTA provides a computationally explicit framework for modeling narrative structure. This framework supports systematic analysis of narrative organization and enables empirical evaluation of whether models construct integrated representations across narrative dimensions.

B Annotation Guidelines

We acknowledge the inherent dilemma between minimizing the cognitive load for annotators and maintaining the theoretical depth required for high-complexity tasks. Demanding extensive linguistic expertise is impractical, yet performing topological analysis without theoretical constraints inevitably leads to inconsistency. To resolve this trade-off, we adopted a **pragmatic tiered strategy**:

- The **Annotator Manual** is designed as the primary, accessible guide for standard workflow, prioritizing intuition over formalism.
- The **Theoretical Codebook** serves as the ultimate axiomatic constitution, intended to be consulted strictly for arbitration during ambiguous or borderline cases.

B.1 VISTA Annotator Manual

VISTA Annotator Manual

Note to Annotators: This document outlines the standard operating procedures. For any ambiguity or edge case not covered here, please refer to the Appendix B.2 for the final axiomatic ruling.

---

1. Task Objective

The goal is to reconstruct the linear text into a narrative topology. Annotators must identify **Narrative Anchors** (verbs) and classify them based on their manipulation of the **Narrative Progress Index** (τ).

2. Core Classifications

Refer to **Codebook Section 1 & 2** for formal definitions of τ and Anchors.

Impulse ($\mathcal{V}_I$)

- **Function: Transition** ($\tau \rightarrow \tau + 1$). The story turns the page.
- **The Necessity Test:** Try deleting the verb. If the preceding event cannot logically lead to the subsequent event (creating a causal gap), it is $\mathcal{V}_I$. *(See Codebook Axiom 2.2)*

Resonance ($\mathcal{V}_R$)

- **Function: Micro-shift** ($\tau + \epsilon$). The story scans the current page.
- **The Texture Test:** If deleting the event removes detail but leaves the logical skeleton intact, it is $\mathcal{V}_R$. *(See Codebook Axiom 3.2)*

Pause ($\mathcal{V}_P$)

- **Function: Bullet Time** ($\tau + 0$). The story freezes to gaze deeply.
- **The Density Test:** If a cluster of verbs decomposes a single split-second moment into high-resolution details, it is $\mathcal{V}_P$. *(See Codebook Axiom 4.2)*

3. General Principles

- **Structure First:** Ignore semantic intensity; focus only on structural function. *(See Codebook Axiom 1.2)*
- **Minimization:** The $\mathcal{V}_I$ chain must be the minimum set required to sustain the plot.

---

4. Case Study: The Western Duel

Text: ... The stranger **draws**$_{[2]}$ his gun. In a flash, he **pulls**$_{[3]}$ the trigger, the Sheriff **side-steps**$_{[4]}$, the bullet **grazes**$_{[5]}$ his hat, the window **shatters**$_{[6]}$... The Sheriff **returns**$_{[8]}$ fire...

Annotation Workflow Demonstration: Step 1: Keystone Identification

- **draws**$_{[2]}$ and **returns**$_{[8]}$ are identified as $\mathcal{V}_I$ because they are the minimal nodes required to advance the conflict. *(Refer to Codebook Axiom 6.1)*

Step 2: Inertial Filling

- **pulls**$_{[3]}$ and **side-steps**$_{[4]}$ follow the trigger event. By default, they are provisionally marked as $\mathcal{V}_R$ (Accompaniment). *(Refer to Codebook Axiom 6.2)*

Step 3: Density Correction

- **grazes**$_{[5]}$ and **shatters**$_{[6]}$ describe micro-physics in a frozen instant.
- **Verdict:** Correct to $\mathcal{V}_P$.
- **Reasoning:** These nodes represent a vertical information dive, not a horizontal progression. *(Refer to Codebook Axiom 4.1)*

5. Ambiguity Resolution (FAQ)

Q: How to handle psychological actions (thinking, recalling)?

- **Verdict:** $\mathcal{V}_P$ (Pause).
- **Reference: Codebook Axiom 4.1**. Internal thoughts are topologically isomorphic to external slow-motion shots; both are vertical dives.

Q: How to segment triggers vs. phenomena (e.g., "fired" vs. "sparks")?

- **Verdict:** "Fired" is $\mathcal{V}_I$; "Sparks" is $\mathcal{V}_P$.
- **Reference: Codebook Axiom 5.1**. Phenomena are visual residues that must depend on a structural trigger.

B.2 VISTA Theoretical Codebook

This codebook provides the formal foundation of the VISTA annotation scheme, specifying the axiomatic principles that govern role assignment and dependency decisions. It complements the annotator manual by making explicit the theoretical criteria underlying these decisions, particularly in ambiguous or borderline cases.

VISTA Theoretical Codebook (Axiomatic System)

1. The Basic Unit Proposition The atom of narrative analysis is the "Event Operator."

- **Axiom 1.1 (Symbolic Proxy):** Verbs are symbolic proxies for underlying semantic units.
- **Axiom 1.2 (The Operator Law):** The value of a verb depends strictly on its **transformational effect** on the narrative state (E), and is orthogonal to its lexical semantic intensity.

2. The Necessity Proposition ($\mathcal{V}_I$) Impulse is the sole logical carrier of narrative progression.

- **Axiom 2.1 (The Backbone):** $\mathcal{V}_I$ constitutes the irreversible timeline of the story.
- **Axiom 2.2 (Logical Continuity):** Any two adjacent impulses v_i, v_{i+1} must satisfy a direct logical sequence relationship. If v_i is removed, v_{i+1} loses its precondition.

3. The Extension Proposition ($\mathcal{V}_R$) Resonance is the lateral expansion of the narrative dimension.

- **Axiom 3.1 (Attachment):** $\mathcal{V}_R$ must attach to a backbone node, providing a state description increment (δ).
- **Axiom 3.2 (The Micro-shift):** If $\Delta\text{State} = 0$ (logical index is constant) but physical time flows ($\tau + \epsilon$), the node is $\mathcal{V}_R$.

4. The Depth Proposition ($\mathcal{V}_P$) Pause is the vertical collapse of the narrative dimension.

- **Axiom 4.1 (Verticality):** $\mathcal{V}_P$ represents a vertical dive into a single moment (Z-axis), characterized by high information density and zero narrative velocity ($\tau + 0$).
- **Axiom 4.2 (Super-Resolution):** Any cluster of verbs performing a microscopic decomposition of a single instantaneous frame is defined as $\mathcal{V}_P$.

5. The Structural Proposition

- **Axiom 5.1 (Asymmetric Dependency):** All discretionary nodes ($\mathcal{V}_R, \mathcal{V}_P$) must ultimately depend on a structural node ($\mathcal{V}_I$).
- **Axiom 5.1.1 (Ultimate Dependence):** Every discretionary node ultimately resolves to a backbone node ($\mathcal{V}_I$), regardless of local attachment.
- **Axiom 5.1.2 (Operational Serialization):** Intermediate references are operational artifacts, not additional backbone relations.

6. The Operational Proposition Principles for resolving ambiguity during the annotation process.

- **Axiom 6.1 (Keystone Priority):** The annotation process must prioritize establishing the $\mathcal{V}_I$ chain.
- **Axiom 6.2 (The Relativity Law):** The class of a fuzzy node is determined by its **axial relationship** relative to the preceding anchor:
  - Progression $\rightarrow \mathcal{V}_I$
  - Accompaniment $\rightarrow \mathcal{V}_R$
  - Deepening $\rightarrow \mathcal{V}_P$

C Concrete Annotated Example

Visual Representation Note: In the actual VISTA dataset, topological labels are encoded using inline HTML-style tags (e.g., `<span style="color:red">verb</span>`). This encoding scheme is a deliberate design choice, calculated to leverage the inherent proficiency of modern Large Language Models (LLMs) in handling structured formatting constraints (e.g., HTML/XML schemas), thereby enhancing topological consistency during generation.

For the sake of readability in this document, we have rendered these raw tags directly as colored text. The color coding and notation scheme are defined as follows:

- **Red**: Impulse ($\mathcal{V}_I$), denoting narrative progression.
- **Green**: Resonance ($\mathcal{V}_R$), denoting descriptive expansion.
- **Blue**: Pause ($\mathcal{V}_P$), denoting vertical deepening.
- **Indices (@n / #n)**: Indicate the topological dependency between a governing node (@) and its dependent (#).

Below is an excerpt of an annotation sample from *Alice's Adventures in Wonderland*.

Input (Raw Text)

Excerpt from *Alice's Adventures in Wonderland*

Chapter I. Down the Rabbit-Hole

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, "and what is the use of a book," thought Alice "without pictures or conversations?"

So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.

There was nothing so *very* remarkable in that; nor did Alice think it so *very* much out of the way to hear the Rabbit say to itself, "Oh dear! Oh dear! I shall be late!" (when she thought it over afterwards, it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural); but when the Rabbit actually TOOK A WATCH OUT OF ITS WAISTCOAT-POCKET, and looked at it, and then hurried on, Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, and burning with curiosity, she ran across the field after it, and fortunately was just in time to see it pop down a large rabbit-hole under the hedge.

In another moment down went Alice after it, never once considering how in the world she was to get out again. The rabbit-hole went straight on like a tunnel for some way, and then dipped suddenly down, so suddenly that Alice had not a moment to think about stopping herself before she found herself falling down a very deep well.

Either the well was very deep, or she fell very slowly, for she had plenty of time as she went down to look about her and to wonder what was going to happen next. First, she tried to look down and make out what she was coming to, but it was too dark to see anything; then she looked at the sides of the well, and noticed that they were filled with cupboards and book-shelves; here and there she saw maps and pictures hung upon pegs. She took down a jar from one of the shelves as she passed; it was labelled 'ORANGE MARMALADE', but to her great disappointment it was empty: she did not like to drop the jar for fear of killing somebody, so managed to put it into one of the cupboards as she fell past it.

"Well!" thought Alice to herself, "after such a fall as this, I shall think nothing of tumbling down stairs! How brave they'll all think me at home! Why, I wouldn't say anything about it, even if I fell off the top of the house!" (Which was very likely true.)

Output (Topological Annotation)

Down the Rabbit-Hole Alice was beginning to get very **tired@1** of sitting by her sister on the bank, and of having nothing to do: once or twice she had **peeped#1** into the book her sister was **reading**, but it had no pictures or conversations in it, "and what is the use of a book," **thought** Alice "without pictures or conversations?"
So she was **considering** in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes **ran** close by her.
There was nothing so VERY remarkable in that; nor did Alice think it so VERY much out of the way to **hear@2** the Rabbit **say** to itself, "Oh dear! Oh dear! I shall be late!" (when she **thought#2** it over afterwards, it **occurred** to her that she ought to have wondered at this, but at the time it all seemed quite natural); but when the Rabbit actually **TOOK** A WATCH OUT OF ITS WAISTCOAT-POCKET, and **looked** at it, and then **hurried** on, Alice **started** to her feet, for it **flashed** across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, and burning with curiosity, she **ran** across the field after it, and fortunately was just in time to **see** it **pop** down a large rabbit-hole under the hedge.
In another moment down **went** Alice after it, never once considering how in the world she was to get out again. The rabbit-hole went straight on like a tunnel for some way, and then dipped suddenly down, so suddenly that Alice had not a moment to think about stopping herself before she found herself **falling** down a very deep well.
Either the well was very deep, or she fell very slowly, for she had plenty of time as she **went** down to **look** about her and to **wonder** what was going to happen next. First, she **tried** to look down and make out what she was coming to, but it was too dark to see anything; then she **looked** at the sides of the well, and **noticed@3** that they were filled with cupboards and book-shelves; here and there she saw maps and pictures hung upon pegs.
She **took#3** down a jar from one of the shelves as she **passed**; it was labelled "ORANGE MARMALADE", but to her great **disappointment** it was empty: she did not like to drop the jar for fear of killing somebody, so managed to **put** it into one of the cupboards as she **fell** past it.
"Well!" **thought** Alice to herself, "after such a **fall** as this, I shall think nothing of tumbling down stairs! How brave they'll all think me at home! Why, I wouldn't say anything about it, even if I fell off the top of the house!" (Which was very likely true.)

D Details of Experimental Settings

This section summarizes the experimental settings and prompt designs. We set the temperature to `0.0` whenever applicable; otherwise, default settings are used.

We use two prompt variants. The Oracle Evaluation Prompt (Appendix D.1) takes the raw text together with event anchors and their character offsets, whereas the End-to-End Evaluation Prompt (Appendix D.2) takes only the raw text and therefore specifies anchor definitions more explicitly. Both prompts include one fully annotated example.

D.1 Oracle Evaluation Prompt

Prompt: Narrative Topology Classification (Pre-identified Anchors)

System Instruction: You are an expert Narrative Analyst. You are tasked with analyzing a text to construct a structured dependency graph.

CRITICAL CHANGE: You do NOT need to extract words from scratch. You will be provided with the **Input Text** and a list of **Pre-identified Anchors** (comprising ID, Offsets, and Word).
Your task is to assign the correct **Category** and **Head** (dependency) for each provided Anchor, strictly following the framework definitions below.

I. Foundational Definitions
The Narrative Anchor (v) An Anchor is a symbolic proxy for a semantic event or state change.

- **Context:** You are provided with these Anchors. They include Finite Verbs (e.g., "draws") and Event Nominals (e.g., "departure").
- **Your Job:** Do not add or remove anchors. Analyze only the ones provided in the list.

The Narrative Progress Index (τ) Narrative time is NOT chronological time. We track the Narrative Progress Index (τ), which represents the logical stage of the plot.

- **Rule:** τ only increments when the narrative state *must* change to enable the next event.
- **Constraint:** Mere descriptions or internal thoughts do not advance τ; they expand the current stage.

II. Task Definitions: The Topological Roles
For every provided **Anchor**, you must classify its operation on the Index (τ) using the following three roles:
Role A: IMPULSE (The Plot Driver)

- **Operation:** $\tau + 1$ (Advances the Index).
- **Definition:** These are the backbone events. They irreversibly change the state of the story.
- **The Necessity Test:** If you delete this anchor, does the logical chain break? If the next event loses its cause/precondition, this is an Impulse.

Role B: RESONANCE (The Lateral Expansion)

- **Operation:** τ (Same Index, Lateral shift).
- **Definition:** These events happen alongside the Impulse to provide atmosphere, manner, or context.
- **The Texture Test:** If you delete this anchor, is the plot skeleton preserved, losing only descriptive detail? If yes, it is a Resonance.

Role C: PAUSE (The Vertical Intensity)

- **Operation:** τ (Index Freeze).
- **Definition:** The narrative flow halts to load "Information Density" into a single moment.
- **The Density Test:** Does this anchor represent a split-second micro-action (physics) or a dive into internal psychology (thoughts)? If it dives "inward" instead of moving "forward," it is a Pause.

III. Dependency Logic (Determining the "Head")

- **If Impulse:** Points to the **previous Impulse** ID (or -1 if it is the first/root).
- **If Resonance/Pause:** Points to the ID of the **Impulse** that governs the current state (the Impulse being modified or described).

IV. Output Formatting Strategy
You must output a structured list (simulated table).

- **Format:** Tab-separated or fixed-width text.
- **Constraint:** The `ID`, `Offsets`, and `Word` columns must MATCH the Input Anchors exactly.

Columns Definition:

1. **ID**: The unique integer provided in the input.
2. **Category**: Your classification (Impulse, Resonance, or Pause).
3. **Offsets**: The offsets provided in the input (e.g., `331,334`).
4. **Word**: The word provided in the input.
5. **Head**: The ID of the parent node (calculated by you).

Output Template:

```
ID   Category     Offsets     Word     Head
0    Resonance    331,334     had      1
1    Impulse      796,803     imputes  -1
```

V. One-Shot Demonstration

Input Text: "CHAPTER I. Down the Rabbit-Hole Alice was beginning to get very tired of sitting by her sister on the bank , and of having nothing to do : once or twice she had peeped into the book her sister was reading , but it had no pictures or conversations in it , ' and what is the use of a book , ' thought Alice ' without pictures or conversations ? ' So she was considering in her own mind (as well as she could , for the hot day made her feel very sleepy and stupid) , whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies , when suddenly a White Rabbit with pink eyes ran close by her . There was nothing so VERY remarkable in that ; nor did Alice think it so VERY much out of the way to hear the Rabbit say to itself , ' Oh dear ! Oh dear ! I shall be late ! ' (when she thought it over afterwards , it occurred to her that she ought to have wondered at this , but at the time it all seemed quite natural) ; but when the Rabbit actually TOOK A WATCH OUT OF ITS WAISTCOAT-POCKET , and looked at it , and then hurried on , Alice started to her feet , for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket , or a watch to take out of it , and burning with curiosity , she ran across the field after it , and fortunately was just in time to see it pop down a large rabbit-hole under the hedge ."

Input Anchors:

```
0    64,69       tired
1    161,167     peeped
2    197,204     reading
3    291,298     thought
4    356,367     considering
5    622,625     ran
6    742,746     hear
7    758,761     say
8    827,834     thought
9    859,867     occurred
10   994,998     TOOK
11   1041,1047   looked
12   1065,1072   hurried
13   1084,1091   started
14   1113,1120   flashed
```

```
15   1274,1277   ran
16   1342,1345   see
17   1349,1352   pop
```

Reasoning:

1. tired (ID 0): State change (becoming tired). Advances narrative state. → Impulse. Head: -1.
2. peeped (ID 1): Minor action occurring alongside the main state. Does not advance plot stage. → Resonance. Head: 0.
3. reading (ID 2): Contextual activity of the sister. Expands the scene. → Resonance. Head: 1.
4. thought (ID 3): Internal mental process. Freezes time to load information. → Pause. Head: 0.

Output:

```
0    Impulse     64,69       tired       -1
1    Resonance   161,167     peeped      0
2    Resonance   197,204     reading     1
3    Pause       291,298     thought     0
4    Pause       356,367     considering 3
5    Impulse     622,625     ran         0
6    Impulse     742,746     hear        5
7    Impulse     758,761     say         6
8    Resonance   827,834     thought     6
9    Resonance   859,867     occurred    8
10   Resonance   994,998     TOOK        9
11   Resonance   1041,1047   looked      10
12   Resonance   1065,1072   hurried     11
13   Impulse     1084,1091   started     7
14   Impulse     1113,1120   flashed     13
15   Impulse     1274,1277   ran         14
16   Impulse     1342,1345   see         15
17   Impulse     1349,1352   pop         16
Any other text is prohibited from being output.
```

VI. Task
Input Text: [INSERT TEXT HERE]
Input Anchors: [INSERT ANCHOR LIST HERE (Format: ID Offsets Word)]

D.2 End-to-End Evaluation Prompt

System Instruction: You are an expert Narrative Analyst. You are tasked with deconstructing a text into a structured dependency graph. To do this, you must first understand the fundamental definitions of the framework provided below. Do not rely on outside knowledge; strictly follow these definitions.

I. Foundational Definitions
The Narrative Anchor (v) Before analyzing structure, you must identify the atomic units of the narrative, called Anchors.

- **Definition:** An Anchor is a symbolic proxy for a semantic event or state change.
- **Scope:** This includes **Finite Verbs** (e.g., “draws”, “ran”) AND **Event Nominals** (nouns that

imply an event structure, e.g., “departure”, “marriage”, “thought”).

- **Exclusion:** Do NOT tag auxiliary verbs (is, was, had) or functional connectors unless they are the sole carrier of meaning.

The Narrative Progress Index (τ) Narrative time is NOT chronological time. We track the Narrative Progress Index (τ), which represents the logical stage of the plot.

- **Rule:** τ only increments when the narrative state *must* change to enable the next event.
- **Constraint:** Mere descriptions or internal thoughts do not advance τ; they expand the current stage.

II. Task Definitions: The Topological Roles
For every identified **Anchor**, you must classify its operation on the Index (τ) using the following three roles:
Role A: IMPULSE (The Plot Driver)

- **Operation:** $\tau + 1$ (Advances the Index).
- **Definition:** These are the backbone events. They irreversibly change the state of the story.
- **The Necessity Test:** If you delete this anchor, does the logical chain break? If the next event loses its cause/precondition, this is an Impulse.

Role B: RESONANCE (The Lateral Expansion)

- **Operation:** τ (Same Index, Lateral shift).
- **Definition:** These events happen alongside the Impulse to provide atmosphere, manner, or context.
- **The Texture Test:** If you delete this anchor, is the plot skeleton preserved, losing only descriptive detail? If yes, it is a Resonance.

Role C: PAUSE (The Vertical Intensity)

- **Operation:** τ (Index Freeze).
- **Definition:** The narrative flow halts to load “Information Density” into a single moment.
- **The Density Test:** Does this anchor represent a split-second micro-action (physics) or a dive into internal psychology (thoughts)? If it dives “inward” instead of moving “forward,” it is a Pause.

III. Output Formatting Strategy
You must output the analysis as a structured list (simulated table) containing the following columns. Do NOT use HTML tags.
Columns Definition:

1. **ID**: A unique sequential integer (0, 1, 2...) for each Anchor found.
2. **Category**: The Role (Impulse, Resonance, or Pause).
3. **Offsets**: The start and end character position of the word in the input text (e.g., `331,334`). *Note: Estimate the offsets as accurately as possible based on the provided text.*
4. **Word**: The exact text of the Anchor.

5. **Head**: The ID of the parent node.
    - **If Impulse**: Points to the *previous* Impulse ID (or -1 if it is the first/root).
    - **If Resonance/Pause**: Points to the ID of the **Impulse** that governs the current state (the Impulse being modified).

Output Template:

```
ID   Category     Offsets     Word     Head
0    Resonance    331,334     had      1
1    Impulse      796,803     imputes  -1
```

IV. One-Shot Demonstration
Input Text: "CHAPTER I. Down the Rabbit-Hole Alice was beginning to get very tired of sitting by her sister on the bank , and of having nothing to do : once or twice she had peeped into the book her sister was reading , but it had no pictures or conversations in it , ' and what is the use of a book , ' thought Alice ' without pictures or conversations ? ' So she was considering in her own mind (as well as she could , for the hot day made her feel very sleepy and stupid) , whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies , when suddenly a White Rabbit with pink eyes ran close by her . There was nothing so VERY remarkable in that ; nor did Alice think it so VERY much out of the way to hear the Rabbit say to itself , ' Oh dear ! Oh dear ! I shall be late ! ' (when she thought it over afterwards , it occurred to her that she ought to have wondered at this , but at the time it all seemed quite natural) ; but when the Rabbit actually TOOK A WATCH OUT OF ITS WAISTCOAT-POCKET , and looked at it , and then hurried on , Alice started to her feet , for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket , or a watch to take out of it , and burning with curiosity , she ran across the field after it , and fortunately was just in time to see it pop down a large rabbit-hole under the hedge ."
Reasoning:

1. **tired** (ID 0): State change (becoming tired). Advances narrative state. → **Impulse**. Head: -1.
2. **peeped** (ID 1): Minor action occurring alongside the main state. Does not advance plot stage. → **Resonance**. Head: 0.
3. **reading** (ID 2): Contextual activity of the sister. Expands the scene. → **Resonance**. Head: 1.
4. **thought** (ID 3): Internal mental process. Freezes time to load information. → **Pause**. Head: 0.

Output:

```
0    Impulse     64,69       tired       -1
1    Resonance   161,167     peeped      0
2    Resonance   197,204     reading     1
3    Pause       291,298     thought     0
4    Pause       356,367     considering 3
5    Impulse     622,625     ran         0
6    Impulse     742,746     hear        5
7    Impulse     758,761     say         6
8    Resonance   827,834     thought     6
9    Resonance   859,867     occurred    8
10   Resonance   994,998     TOOK        9
11   Resonance   1041,1047   looked      10
12   Resonance   1065,1072   hurried     11
13   Impulse     1084,1091   started     7
```

```
14   Impulse     1113,1120   flashed     13
15   Impulse     1274,1277   ran         14
16   Impulse     1342,1345   see         15
17   Impulse     1349,1352   pop         16
```

V. Task
Analyze the following text strictly following the Definitions, Logical Tests, and Output Format above.
Input Text: [INSERT TEXT HERE]

E End-to-End Analysis

This appendix presents an end-to-end analysis to complement the oracle event-level experiments reported in the main paper. The goal of this analysis is to examine whether current large language models can perform narrative orchestration when provided only with raw text and a fully specified prompt, without access to gold event anchors.

We first summarize the overall findings and failure modes observed in the end-to-end setting (Section E.1). We then present representative model outputs alongside the corresponding ground-truth annotations to illustrate the observed errors in detail (Section E.2).

E.1 End-to-End Results and Analysis

We evaluate a representative set of frontier models in an end-to-end setting, including DeepSeek-v3.2, Gemini-3-Pro-Preview-Thinking, GPT-5, GPT-5-Thinking, Grok-4.1-Thinking, and Qwen3-235B-A22. In this setting, models are provided only with the raw narrative text and a fully specified prompt that defines narrative anchors, their functional roles, and the dependency structure, along with a concrete illustrative example.

Across all tested models, performance in the end-to-end setting is uniformly zero. Specifically, none of the models are able to produce a valid reconstruction of the LitVISTA graph that satisfies the evaluation criteria.

To diagnose the source of this failure, we analyze the raw model outputs in detail. Representative predictions are shown in Section E.2 alongside the corresponding ground-truth annotations. Two systematic failure modes consistently emerge:

- **Incomplete anchor identification**: Given a narrative with around one hundred events, a substantial fraction of anchors are consistently omitted. Models fail to exhaustively identify all event anchors in the text. For example, in the case of DeepSeek-v3.2, numerous event anchors like "CONTAINING" and "BIRTH" appear, but several key events like "lived" and "proceed" are omitted.

- **Misalignment of spans**: Even when an anchor is identified, models often mis-specify its exact token span or positional offset, leading to misaligned or invalid anchors. For instance, GPT-5 outputs anchors such as "CONDESCENDED" but misaligns spans (e.g., "2500,2511") that don't correspond to the actual ground-truth position.

These errors are characteristic of probabilistic, generative models. Exhaustive anchor extraction and precise span localization require strict coverage guarantees and exact alignment with the source text, properties that current autoregressive generation paradigms do not reliably provide. Because anchor identification and localization constitute the first step in the narrative reconstruction pipeline, errors at this stage prevent subsequent role assignment and dependency resolution from being meaningfully evaluated, resulting in zero scores under the end-to-end setting.

Taken together, these results indicate that the observed end-to-end failure reflects limitations in upstream anchor identification and localization rather than deficiencies in model capacity or dataset quality. As demonstrated in the main paper under the oracle setting, multiple models achieve strong performance when gold event anchors are provided. For example, Claude-sonnet-4-thinking attains a balanced Anchor F1 of 0.4914 and a Dependency F1 of 0.5624, while GPT-5.1-thinking reaches a Dependency Parsing F1 as high as 0.8135. These findings confirm that the downstream narrative orchestration task itself is well within the representational capacity of current models.

E.2 Representative Model Outputs

To qualitatively illustrate the failure modes discussed above, we present representative end-to-end predictions produced by different models on the same narrative input. The example is drawn from a single chapter of *The History of Tom Jones, a Foundling*, for which the LitVISTA annotation contains exactly fourteen event anchors.

For each model, we report its predicted anchors together with assigned roles, token offsets, and dependency heads. While the gold annotation consists of a compact and well-defined set of anchors, model predictions typically contain substantially more entries, along with omissions, misaligned spans, and structural inconsistencies. Ellipses indicate omitted portions of the prediction.

Ground Truth (LitVISTA Annotation). The gold annotation contains exactly fourteen event anchors. All anchors are shown in full below.

ID	Category	Offsets	Word	Head
0	Impulse	2500,2511	condescended	-1
1	Resonance	2650,2655	prefix	0
2	Resonance	2750,2753	give	0
3	Resonance	2900,2903	made	0
4	Pause	3200,3203	fear	0
5	Resonance	3600,3608	represent	0
6	Resonance	3720,3723	hash	0
7	Pause	3950,3954	doubt	0
8	Resonance	4250,4255	detain	0
9	Impulse	4330,4336	proceed	0
10	Impulse	4700,4704	lived	9
11	Resonance	4780,4785	called	10
12	Resonance	4920,4928	contended	10
13	Resonance	5070,5077	bestowed	10
14	Resonance	5420,5426	decreed	10

GPT-5. GPT-5 generates fewer anchors than DeepSeek-v3.2, but still exceeds the gold count and fails to recover the complete gold structure.

ID	Category	Offsets	Word	Head
0	Impulse	1150,1162	condescended	-1
1	Resonance	1177,1183	prefix	0
2	Resonance	1217,1220	give	0
3	Resonance	1310,1314	made	0
4	Impulse	2045,2052	proceed	0
5	Impulse	2210,2215	lived	4
6	Resonance	2635,2642	married	5
7	Impulse	2700,2704	died	5
8	Impulse	2748,2755	burying	7
9	Pause	2860,2863	said	8
10	Pause	2888,2897	considered	8
11	Impulse	2938,2943	lived	8
12	Resonance	3245,3251	intend	11
13	Resonance	3380,3383	told	11
14	Resonance	3412,3421	inherited	11
15	Resonance	3610,3615	built	11

GPT-5-Thinking. GPT-5-Thinking generates a sequence of event anchors, though it still produces errors in coverage, span localization, and anchor alignment. Below, we show the full output for the first 17 predicted anchors, followed by the last two anchors.

ID	Category	Offsets	Word	Head
0	Impulse	131,144	ought	-1
1	Resonance	157,162	consider	0
2	Resonance	253,263	provides	0
3	Resonance	331,345	is	0
4	Impulse	347,353	may	0
5	Resonance	490,502	known	4
6	Resonance	550,566	have	0
7	Impulse	573,586	ensure	4
		...		
18	Impulse	1240,1246	fear	16
19	Resonance	1315,1326	very	17

DeepSeek-v3.2. DeepSeek-v3.2 produces a long sequence of predicted anchors that substantially exceeds the fourteen gold events. Below we show the beginning of the prediction in full, followed by selected later segments.

ID	Category	Offsets	Word	Head
0	Impulse	7,15	CONTAINING	-1
1	Resonance	27,34	BIRTH	0
2	Resonance	81,90	ACQUAINT	0
3	Impulse	212,224	consider	0
4	Resonance	282,287	gives	3
5	Resonance	373,379	keeps	3
6	Resonance	455,464	provides	5
7	Resonance	502,511	disagreeable	6
		...		
122	Impulse	9737,9742	plead	118
123	Resonance	9983,9995	inherited	122

Gemini-3-Pro-Preview-Thinking. Gemini-3-Pro-Preview-Thinking produces the sparsest output among the models shown, yet still fails to recover all fourteen gold anchors.

ID	Category	Offsets	Word	Head
0	Impulse	183,191	consider	-1
1	Resonance	227,232	gives	0
2	Resonance	283,288	keeps	0
3	Impulse	612,619	happens	0
4	Resonance	678,684	insist	3
5	Impulse	868,875	prevent	3
6	Impulse	1317,1329	condescended	5
7	Impulse	1392,1398	prefix	6
		...		
28	Impulse	6692,6696	told	27
29	Impulse	6813,6822	concluded	28