Critic's Eye: Review-Grounded Graph Reasoning for Long-Form Motif Discovery

Anonymous ACL submission

Abstract

Deep narrative understanding requires distinguishing *what* happens from *how* it is presented, moving beyond the chronological *fabula* to decode the *sužet* , the discursive organization through which themes are staged. We study this through motif recurrence, where symbolic significance emerges not from local cues, but through repetition-with-variation across the whole work. We operationalize this setting as Computational Motif Discovery, a transductive task that predicts missing *line→motif* links using the full narrative structure of a play. We propose CRITIC'S EYE , which models the work as a heterogeneous Narrative Topology Graph and performs discriminative inference over global structural evidence. CRITIC'S EYE achieves 84.8% Hit@5 on our benchmark, nearly quadrupling state-of-the-art proprietary foundation models (Gemini-3-pro, ∼22.6%). Our analysis reveals that despite massive scaling, sequence-based models hit a "semantic ceiling", struggling to resolve dispersed dependencies. These findings suggest that explicit structural priors are a far more effective inductive bias than parameter scaling for decoding the complex architecture of long-form literature.

1 Introduction

A fundamental challenge in narrative intelligence is distinguishing the raw material of a story from its artistic presentation. This distinction is formalized by Russian Formalism as the duality between the ***fabula*** (the chronological sequence of events) and the ***sužet*** (the discursive organization of the text) (Propp, 1968; Bal and Boheemen, 2009). While the *fabula* follows the linear causality of "what happened," the *sužet* encodes the topological logic of "how it is presented," employing techniques such as recurrence, focalization, and anachrony to stage meaning.

In computational terms, the two correspond to different supervision regimes: *fabula* -level signals tend to be locally explicit and extractable, whereas ***sužet*** -level meaning is often distributed and relational, emerging only from how elements recur and interact across a work. A particularly central form of *sužet* -level evidence is **motif recurrence**: recurring symbolic elements whose interpretive force accrues through repetition with variation across a work.

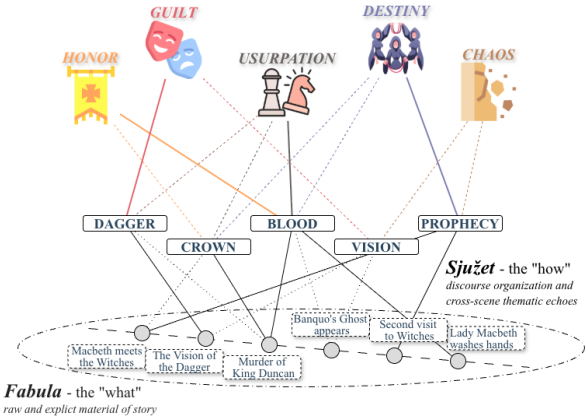

Figure 1: **The Process of Sensemaking**, Events constitute the ***fabula*** as the chronological foundation. The ***sužet*** overlays this foundation, linking disparate events through shared imagery to construct a discourse.

This conception aligns with a longstanding aspiration in NLP: to move beyond surface processing toward models that can resolve ambiguity and construct coherent meaning. However, dominant task formulations have largely emphasized *fabula* -level inference, favoring explicit, locally supervised content such as event extraction (Sun et al., 2024), plot summarization, or attribute summarization (Yuan et al., 2024). While effective for factual reporting, such objectives inherently flatten the rich, non-linear semantics of literature. By stripping away the discursive organization, current models neglect the *sužet* , missing the critical interpretative cues required to decode a work's deeper intent.

We operationalize this missing component through **Computational Motif Discovery**. Un-

like explicit events that follow linear time, motifs are structural anchors determined by their distribution across the whole work: they induce long-range bindings that connect distant moments into a thematic configuration. As Spurgeon (1935) observes, motifs consist of ideas that are "interwoven," forming a structure "subtle and complex" that defies simple plot mechanics. Their meanings are often non-stationary and context-dependent (e.g., `blood` shifting from *honor* to *guilt* across acts), making motif resolution poorly matched to surface-level retrieval or inductive classification. This makes motif resolution a particularly clean testbed for *sjužet* -level inference: the target is not to summarize content, but to recover non-local bindings induced by recurrence-with-variation.

This setting exposes the intrinsic limits of sequence-based modeling. Despite their surface fluency, Large Language Models (LLMs) fundamentally process text as flat streams. Recent empirical studies indicate that LLMs struggle with long-horizon coherence (Tian et al., 2024) and planning (Senanayake and Ware, 2025), often failing to integrate evidence that is distributed across thousands of lines ("lost-in-the-middle") (Levy et al., 2024). In our experiments, even state-of-the-art proprietary foundation models (e.g., Claude, GPT-series, Gemini) hit a distinct “ceiling” on motif resolution. Despite massive parameter scaling and likely pre-training exposure to the source text, they achieve a Hit@5 of only $\sim$**23%**, lagging far behind our approach (**84.8%**). This suggests that the bottleneck is not computational capacity or local semantics, but the lack of explicit topological priors required to model global recurrence.

To address this, we propose CRITIC’S EYE , a framework that models narrative not as a sequence, but as a heterogeneous **Narrative Topology Graph**. By explicitly encoding the interplay of lines, characters, and scenes, we formulate motif resolution as "transductive link discovery", enabling the model to propagate weak thematic signals across the global structure to resolve interpretive ambiguities.

2 Related Work

Modeling the *Fabula*: Entities and Events. A robust literature in NLP models explicit narrative content, the "what happened", through entities, character representations, and event structures. On the character side, prior work modeled discrete roles and archetypes (Bamman et al., 2014; Stammbach et al., 2022), while recent efforts expand toward character profiling with LLMs, summarizing attributes and relationships from broader context (Yuan et al., 2024). Related directions include persona-driven decision modeling (Xu et al., 2025) and goal or intent induction from semantic encodings (Rahimtoroghi et al., 2017). On the event side, causality has been identified as central to story understanding: Sun et al. (2024) study open-world event causality, building on earlier paradigms for event chain extraction and temporal dynamics (Chambers and Jurafsky, 2009; Fang et al., 2024). Recent benchmarks continue to push for narrative-driven understanding in drama series (Zhang et al.). Across these lines, supervision and evaluation typically favor locally explicit evidence (e.g., event mentions, pairwise relations, span-level answers), emphasizing plot-level reconstruction over interpretive meaning.

Modeling the *sjužet* : Discourse and Long-Form Coherence. Beyond fabula, other work attempts to capture narrative discourse, how a story is presented, by operationalizing discursive features. Piper and Bagga (2024) employ LLMs to annotate deictic dimensions such as time, setting, and perspective, and Zhao et al. (2024) leverage narrative structures to guide interactive role-play. In generation and evaluation, several studies analyze the discourse-level properties of LLM stories and identify deficits in narrative planning and arc control (Tian et al., 2024; Senanayake and Ware, 2025), or struggles in decomposing complex narratives into structural scene graphs (Yang et al., 2025). Metrics for long-form coherence, such as BooookScore (Chang et al., 2024), provide useful signals but still face fundamental challenges when models process text as flat sequences under finite context windows, These include “lost-in-the-middle” effects (Levy et al., 2024; Baker et al., 2024), and inverse scaling phenomena on complex reasoning tasks (McKenzie et al., 2023). More broadly, widely used NLU benchmarks tend to privilege fact-centric objectives (e.g., extractive or multi-hop QA), rewarding locally explicit answers over global interpretive coherence (Rajpurkar et al., 2016; Yang et al., 2018).

From Extraction to Topological Discovery. We move beyond detecting isolated elements (characters, events, discursive tags) to model the latent topology that binds them across a whole

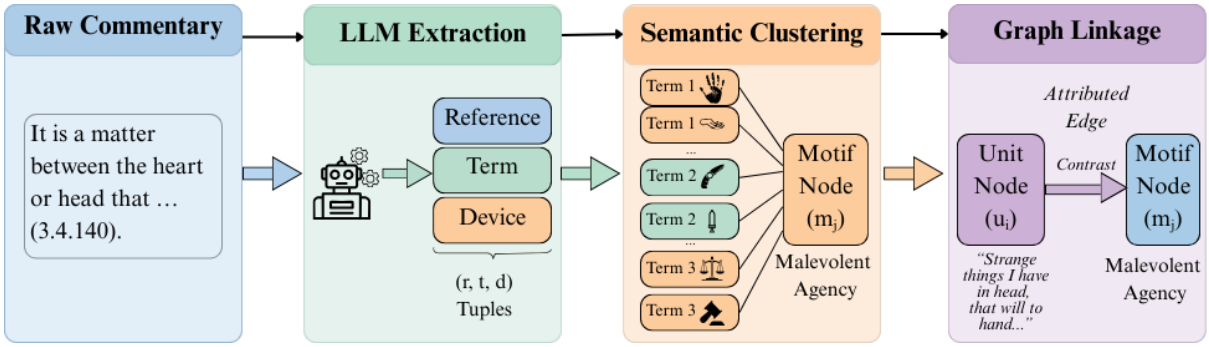

Figure 2: **The CRITIC'S EYE Graph Construction Pipeline.** We fuse the explicit **Narrative Backbone** (parsed from TEI XML) with a latent **Thematic Layer** derived from literary criticism. (1) An LLM extracts structured tuples (Reference, Theme, Device) from unstructured commentary. (2) These mentions are semantically clustered into canonical motif nodes. (3) The resulting heterogeneous graph encodes both the linear ***fabula*** and the critical *sužet* , serving as the substrate for transductive inference. Detailed prompt for extraction is provided in Appendix B

work. While prior work has utilized graphs for explicit event relations (Fang et al., 2024), temporal forecasting (Cai et al., 2025), or logic-constrained completion (Huang et al., 2025), these methods often rely on relational assumptions (e.g., decay, transitivity) that do not transfer to literary analysis. We focus on **Computational Motif Discovery**: inferring motif associations whose evidence is "interwoven" (Spurgeon, 1935) and distributed through recurrence-with-variation rather than localized mentions. This task targets narrative *linking functions* (Herman, 2003) and requires resolving non-stationary, structurally mediated dependencies: a setting naturally cast as whole-work, transductive link discovery rather than local sequence prediction.

3 Method

3.1 Formalizing the Narrative Space

A literary narrative is not merely a linear sequence of tokens, but a hierarchical system composed of interacting entities. To capture this, we adopt a formalization grounded in structural narratology (Bal and Boheemen, 2009; Genette, 1980), mapping the components of narrative discourse to an observable backbone tuple $\mathcal{N} = (\mathcal{U}, \mathcal{A}, \mathcal{S})$:

- **Atomic Units ($\mathcal{U}$): The Micro-Level Voice.** Let $\mathcal{U} = \{u_1, ..., u_n\}$ denote atomic textual units (*e.g. lines in drama, sentences in novels*), ordered by their presentation in the text. These correspond to Genette's concept of *voice*, the immediate textual surface where motifs are realized. These units are strictly ordered by their linear presentation ($t = 1 \dots n$).
- **Agents ($\mathcal{A}$): The Meso-Level Persona.** Let $\mathcal{A} = \{a_1, ..., a_k\}$ denote narrative agents (*e.g. characters, narrators*). Following Bamman et al. (2014), we model characters not as static attributes but as *bundles of habits* defined by their utterances. We define a partial speaker mapping $\pi : \mathcal{U} \to \mathcal{A} \cup \{\perp\}$, connecting each atomic unit to its originating persona.
- **Structural Containers ($\mathcal{S}$): The Macro-Level Mood.** $\mathcal{S}$ represents the spatiotemporal hierarchy (*e.g. acts, scenes*). Analogous to narrative *mood* or setting, these containers impose boundaries on interaction. Formally, they form a rooted tree via a parent function $p : \mathcal{S} \to \mathcal{S} \cup \{\varnothing\}$, with an assignment $\rho : \mathcal{U} \to \mathcal{S}$ anchoring each unit to its minimal container.

All components in $(\mathcal{U}, \mathcal{A}, \mathcal{S})$ are explicitly observable from the source text, forming the topological scaffold upon which thematic reasoning operates.

3.2 Constructing the Narrative Topology

We operationalize $\mathcal{N}$ as a heterogeneous narrative topology graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, transforming the theoretical schema into a computable structure. Let $\mathcal{V} = \mathcal{U} \cup \mathcal{A} \cup \mathcal{S}$.

1. **Textual flow ($\mathcal{E}_{\text{SEQ}}$):** $(u_t \xrightarrow{\text{SEQ}} u_{t+1})$ captures the *fabula* -level continuity or local discourse flow. We initialize node features using a frozen sentence encoder to preserve local semantics.

2. **Hierarchical composition ($\mathcal{E}_{\text{INC}}$):**

   Edges $(u \xrightarrow{\text{INC}} \rho(u))$ and $(s \xrightarrow{\text{INC}} p(s))$ encode the vertical organization of the *sužet* ,

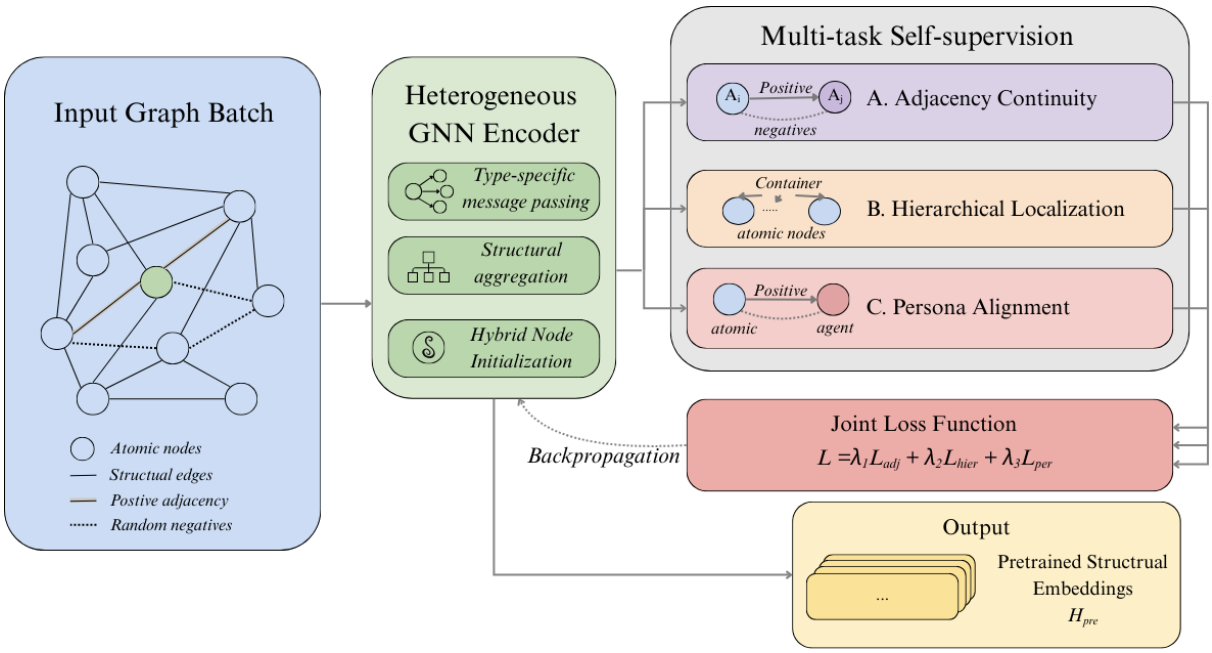

Figure 3: **Stage 1: Multi-Task Structural Pre-training.** Before accessing motif labels, the encoder is forced to internalize the narrative "anatomy" via three intrinsic objectives: **(A) Adjacency Continuity** ($u_t \rightarrow u_{t+1}$) captures local discourse flow; **(B) Hierarchical Localization** ($u \rightarrow s$) recovers macro-structural positioning; and **(C) Persona Alignment** ($u \rightarrow a$) learns character-specific stylistic features. This *structure-first* strategy aligns the embedding space with the narrative topology.

creating topological shortcuts that allow information to propagate across long temporal distances within the same scene.

3. **Agentive interaction ($\mathcal{E}_{\text{INT}}$):** $(\pi(u) \xrightarrow{\text{INT}} u)$ model the character-driven nature of drama, allowing the model to learn persona-specific stylistic features.

3.3 Motif Instantiation via Critical Clustering

To instantiate motif nodes and sparse positive links, we leverage literary criticism as a source of expert supervision. Since critics describe identical themes using varied terminology (e.g., *blood*, *gore*, *sanguine imagery*), we employ a **Schema-Constrained Extraction** pipeline.

1. **Extraction:** We use an LLM to extract tuples (r, t, d) from commentary, where r is a resolvable reference, t is a thematic descriptor, and d is a rhetorical device tag (*e.g.* Metaphor).

2. **Clustering:** We cluster descriptors t into canonical motif nodes $\mathcal{M}$ using agglomerative clustering on semantic embeddings.

3. **Attributed Linkage:** We instantiate observed edges $\mathcal{E}_{obs}$ between units u and motifs m. Crucially, we maintain parallel edges for recurring mentions, using citation frequency as a proxy for evidence strength.

3.4 Two-Stage Structure-to-Semantics Transfer

Our training pipeline adopts a Structure-First strategy, reflecting the hypothesis that literary meaning is constrained by topological position before it is filled with semantic content.

Stage 1: Multi-Task Structural Pre-training. We force the encoder to internalize the narrative "anatomy" via three self-supervised objectives (Fig. 3), without accessing motif labels:

- **Adjacency Continuity ($\mathcal{L}_{seq}$):** Capturing local flow ($u_t \rightarrow u_{t+1}$).
- **Hierarchical Localization ($\mathcal{L}_{hier}$):** Recovering macro-structural position ($u \rightarrow s$).
- **Persona Alignment ($\mathcal{L}_{agent}$):** Inferring the speaker from structural context.

The encoder f_θ minimizes $\mathcal{L}_{pre} = \lambda_1 \mathcal{L}_{seq} + \lambda_2 \mathcal{L}_{hier} + \lambda_3 \mathcal{L}_{agent}$, mapping nodes to a structure-aware topological space.

Stage 2: Semantic Alignment via Contrastive Learning. We then align this topological space with the semantic space of motifs. Using a frozen Motif Encoder g_ϕ (e.g., BGE), we treat motif discovery as a dense retrieval task. We opti-

Node Type	Count	Edge Type	Count
Act / Scene	5 / 28	Hierarchical ($\mathcal{E}_{\text{HIER}}$)	3,397
Speech Section	865	Sequential ($\mathcal{E}_{\text{SEQ}}$)	2,602
Atomic Line	2,456	Metadata ($\mathcal{E}_{\text{AGENT}}$)	1,058
Character	44	**Total Explicit**	**7,057**
Latent Motif	283	*Thematic Link*	923

Table 1: Statistics of the *Macbeth* Narrative Graph.

mize an InfoNCE loss to maximize the similarity between the structure-aware unit embedding $\mathbf{h}_{atomic} = f_\theta(u)$ and its ground-truth motif embedding $\mathbf{h}_{motif}$:

$$\mathcal{L}_{\text{align}} = -\log \frac{\exp(s(\mathbf{h}_{\text{atomic}}, \mathbf{h}^{+}_{\text{motif}})/\tau)}{\sum_{j \in \mathcal{N}} \exp(s(\mathbf{h}_{\text{atomic}}, \mathbf{h}^{j}_{\text{motif}})/\tau)} \quad (1)$$

This effectively "anchors" the floating semantic concepts to the rigid narrative topology.

4 The MOTIFBENCH Dataset

To evaluate CRITIC'S EYE , we constructed MOTIFBENCH, a specialized graph dataset centered on Shakespeare's *Macbeth*. Unlike crowd-sourced benchmarks, our ground truth is derived from authoritative literary criticism. The dataset comprises two distinct layers: the explicit **Narrative Backbone** and the latent **Critical Layer**.

4.1 The Narrative Backbone

We parse the TEI-encoded XML of *Macbeth* from the Folger Shakespeare Library (Shakespeare, n.d.) into a heterogeneous graph $\mathcal{G}$.

The node schema corresponds to the play's explicit structure (acts, scenes, speeches, lines) and its social layer (characters). Edges encode both hierarchical inclusion (e.g., $Line \rightarrow Speech \rightarrow Scene$) and sequential flow. As shown in Table 1, this topology provides the scaffold onto which latent motifs are attached.

4.2 The Critical Corpus (Weak Supervision)

To instantiate the motif extraction pipeline, we curated a corpus of 26 documents spanning from 1765 to 2020. Crucially, rather than random selection, we stratified sources into three interpretive lenses (Table 2) to ensure the narrative graph captures diverse semantic dimensions:

Crucially, we adopted an *accessibility-first* policy to facilitate reproducibility. The majority of our corpus is sourced from **Project Gutenberg** (Hart, 2010) and is in the public domain. To strictly adhere to intellectual property rights, our public dataset release follows a split-distribution protocol: we provide the full digitized text for public domain works, while copyrighted sources (mostly recent phenomenological criticism) are released as structured metadata (*e.g. D*OI, Chapter Title) to enable researchers to retrieve the source text independently while preserving the integrity of the graph structure.

Stratum (Type)	Focus & Role in Narrative Graph
1. Humanist	Focus: Character agency, ethical conflicts, and tragedy.
(12 Monographs	*Role*: **High-Density Supervision.** Comprising foundational book-length studies and major critical volumes.
& Volumes)	*Key Sources:* Bradley (1904), Bloom et al. (1987), Fletcher (1844), Hazlitt (1903), Weiss (1876).
2. Socio-Cultural	Focus: Folklore, politics, history, and theatre practice.
(8 Mixed Sources)	*Role*: **Contextual Anchoring.** Captures world-building motifs (e.g., *Witchcraft*, *Kingship*) missed by pure text analysis.
(Books/Essays)	*Key Sources:* Thiselton-Dyer (1883) (Folklore), Adams (1917) (Theatre), Tolstoy (1906) (Counter-criticism), Moschovakis (2008).
3. Phenomenological	Focus: Sensory imagery, objects, and embodied cognition.
(6 Articles)	*Role*: **Latent Discovery.** Forces the model to recover non-explicit associations (e.g., *Smell*, *Dread*).
(Specific Studies)	*Key Sources:* Harris (2007), Spolsky (2011), Cheung (1984), Sachon (2020).

Table 2: **Stratification of the Critical Corpus.** Sources are categorized by interpretive lens. Note that the *Humanist* stratum comprises of extensive monographs and volumes, providing a dense backbone of character analysis, while other strata introduce targeted, fine-grained perspectives. Full bibliographic details are in Appendix A.

4.3 Challenge: The Long-Tail Topology

A quantitative analysis reveals that literary motifs follow a severe long-tail distribution (Figure 4). The dataset exhibits an Imbalance Ratio (IR), defined as $\frac{freq_{max}}{freq_{median}}$, of 35.0. While the mean frequency of active motifs is 1.82, the P90 frequency is only 3.0, indicating that 90% of active thematic concepts are supported by three or fewer examples. Crucially, the tail classes (frequency ≤ 5) comprise **96.9%** (187/193) of the active vocabulary.

Method	Hit				nDCG				MRR
	@1	@3	@5	@10	@1	@3	@5	@10	
Embedding SOTA *(Best of dense retrievers)*									
BGE-Large-v1.5	0.039	0.071	0.081	0.131	0.037	0.049	0.052	0.067	0.075
Open-Source LLMs *(Highlighting Inverse Scaling)*									
GPT-OSS-20B (+RAG)	0.042	0.088	0.117*	0.120	0.042	0.070	0.081	0.082	0.070
GPT-OSS-120B (+RAG) †	0.032	0.064	0.085*	0.085	0.032	0.049	0.058	0.058	0.050
DeepSeek-R1-70B (Zero-shot)	0.081	0.124	0.134	0.159	0.081	0.106	0.110	0.118	0.105
Proprietary Foundation Models *(Highlighting Semantic Ceiling)*									
GPT-5.2-Chat (Zero-shot)	0.064	0.127	0.170	0.184	0.064	0.099	0.116	0.120	0.100
Claude-Opus-4.5 (Zero-shot)	0.078	0.148	0.198	0.276	0.078	0.118	0.138	0.163	0.129
Gemini-3-preview (+RAG)	0.088	0.156	0.226	0.332	0.088	0.088	0.156	0.191	0.125
CRITIC'S EYE (Ours)	**0.548**	**0.774**	**0.848**	**0.922**	**0.548**	**0.646**	**0.682**	**0.716**	**0.676**

Note: We report the **top-performing model** for each category. † **GPT-OSS-120B** is included to illustrate the *inverse scaling* phenomenon (underperforming the 20B model due to instruction drift). Underlines denote the best baseline performance. * indicates the main metric we focus on (Hit@5). Full results for all 12 evaluated models are provided in Appendix D.

Table 3: **Performance Comparison with Frontier Models.** Even the strongest proprietary model (Gemini-3) hits a **ceiling** (∼23% Hit@5), lagging significantly behind CRITIC'S EYE 's structural reasoning.

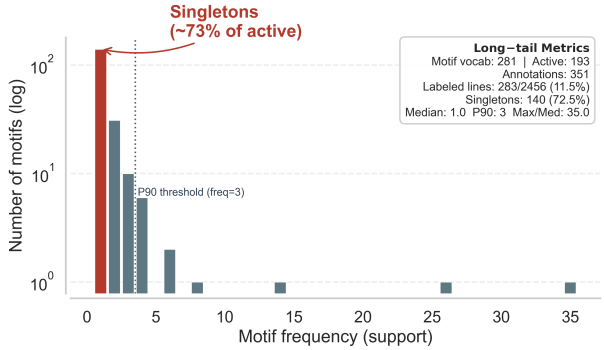

Figure 4: **The Severe Long-Tail Distribution of Literary Motifs.** MOTIFBENCH follows a sharp power law (Imbalance Ratio = 35.0), with over **96%** of motifs in the sparse "tail" (freq ≤ 5). This extreme sparsity renders standard random splits ineffective, motivating our transductive formulation to capture these rare signals.

This extreme sparsity validates our transductive formulation: in a standard random inductive split, ∼67% of motifs would likely have zero training examples in the labeled set, resulting in topological isolation. By treating the problem as transductive link prediction, CRITIC'S EYE enables the propagation of weak signals from the few "head" motifs to the sparse "tail" via the dense narrative backbone.

5 Experiments

We evaluate CRITIC'S EYE against state-of-the-art LLMs and embedding models on the *Macbeth* motif discovery benchmark. The task is formulated as *inventory-constrained retrieval*: given a query unit u, the model must rank the fixed motif inventory $\mathcal{M}$ ($|\mathcal{M}| = 283$). We adopt **Hit@**k and **nDCG@**k, and **MRR**. as primary metrics. Given the sparse, positive-only nature of literary criticism, where unannotated lines often represent overlooked themes rather than definitive negatives. These recall-oriented metrics best reflect the system's utility as a discovery aid. Detailed experimental settings are provided in Appendix C.

5.1 Main Results

Table 3 presents the main results. We compare CRITIC'S EYE against three baseline families: (1) **Embedding Retrievers** (Sentence-BERT, BGE); (2) **Open-Source LLMs** (Llama-3, Qwen, DeepSeek-R1); and (3) **Proprietary Foundation Models** (GPT-series, Claude, Gemini-3), representing the upper bound of current general-purpose reasoning.

CRITIC'S EYE establishes a new state-of-the-art. As shown in Table 3, CRITIC'S EYE achieves decisive improvements across all metrics. Our model achieves a Hit@5 of **84.8%**, nearly quadrupling the performance of the leading proprietary baseline, Gemini-3-pro (22.6%). Crucially, while proprietary models outperform open weights, they hit a distinct "semantic ceiling" in the low-20s range (Hit@5). This suggests that scaling model parameters or context windows alone cannot resolve the structural dependencies of the *sjužet* ; the gap remains topological, not computational.

Retrieval augmentation yields diminishing returns. We experimented with Self-RAG, providing models with semantically similar lines. The

Model Configuration	Hit				nDCG				MRR
	@1	@3	@5	@10	@1	@3	@5	@10	
w/o pre-training (frozen random)	0.1237	0.2544	0.3145	0.3746	0.1237	0.1710	0.1944	0.2134	0.2189
CRITIC'S EYE (*hetero*)	0.2792	0.4947	0.5901	0.7208	0.2792	0.3742	0.4165	0.4657	0.4245
CRITIC'S EYE (*seq*)	0.4876	0.6608	0.7562	0.8198	0.4876	0.5543	0.5960	0.6223	0.6019
CRITIC'S EYE (*full*)	**0.5654**	**0.7809**	**0.8481**	**0.9293**	**0.5654**	**0.6618**	**0.6979**	**0.7321**	**0.6942**

Table 4: **Structural pre-training and topology ablation** (N=283 query lines; $|\mathcal{M}|$=281 motifs). We evaluate with a frozen-encoder probing protocol to measure representation quality. Pre-training is necessary (random vs. pre-trained), and integrating richer structure yields monotonic gains (*hetero* < *seq* < *full*).

utility of RAG proves inconsistent: while it aids smaller models and boosts Recall for Gemini-3 (Hit@10 rises to 33%), it frequently degrades the ranking quality (nDCG) for reasoners like Claude and GPT-5.2. Even for the strongest model, RAG only provides marginal gains ($\sim$1% boost in Hit@5), failing to close the structural gap. This confirms that **lexical similarity is a poor proxy for thematic recurrence**, often introducing competing contexts rather than clarifying the motif.

5.2 Analysis

Our results reveal that model scale alone offers an inefficient path to structural understanding, manifesting in two distinct failure modes.

Inverse Scaling in Instruction Following. Counter-intuitively, the largest open model, *GPT-OSS-120B*, underperforms its smaller 20B counterpart (Hit@5: 0.099 vs. 0.106). Error analysis links this to **schema non-compliance**: the 120B model prioritizes generative fluency over strict output constraints. This confirms the "inverse scaling" phenomenon (McKenzie et al., 2023), where strong generative priors actively hinder rigid classification tasks.

The "Semantic Ceiling" and Efficiency Paradox. Comparing open weights to proprietary giants (e.g., Gemini-3), we observe that while scaling provides gains (Hit@5 rising from $\sim$13% to $\sim$23%), it hits sharp diminishing returns. Despite massive compute and likely **positive data contamination** (exposure to critical texts), general-purpose models hit a distinct "Ceiling" in the low-20s range. In contrast, CRITIC'S EYE demonstrates a nearly $4\times$ performance improvement (84.8%) with only a fraction of the parameter count. This proves that for narrative topology, **explicit structural priors** are a far more efficient inductive bias than brute-force scaling, which can only asymptotically approximate the *sužet* .

5.3 Ablation Study

To rigorously disentangle the contribution of structural priors from the capacity of the GNN backbone, we adopt a frozen-encoder probing protocol. Instead of full fine-tuning, we freeze the pre-trained encoder and optimize only a lightweight matching head. This strict setup measures how easily motif relevance is *recoverable* from the representation space itself.

Structure-First Pre-training is Essential. Table 4 reveals a stark separation in representation quality. With a randomly initialized encoder, performance is poor (Hit@5 $\approx$ 0.31), implying that motif relevance is not intrinsic to the raw graph. In contrast, our pre-trained encoder achieves Hit@5 $\approx$ 0.85. This absolute gain of **+0.53** confirms that the self-supervised objectives (Section 3.4) effectively transform raw narrative connectivity into a semantically aligned space, making motif discovery a linear problem rather than a complex non-linear one.

Synergy of Heterogeneous Topology. We further ablate the topological components (Table 4). A clear hierarchy emerges:

1. *Hetero-only* (Hit@5 0.59) provides useful global anchors (e.g., character/scene associations) but lacks local context.

2. *Seq-only* (Hit@5 0.76) captures the local discourse flow, which is information-rich but locally ambiguous.

3. ***Full Topology*** (Hit@5 0.85) achieves the best performance.

This result confirms a complementary effect: heterogeneous anchors (e.g., *SpeakerOf*) help disambiguate locally similar utterances that play different roles in the dramatic scaffold.

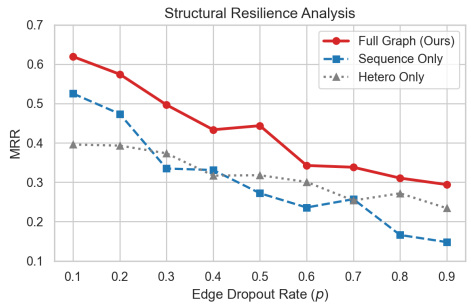

Figure 5: **Structural resilience under edge dropout.** Mean Reciprocal Rank (MRR) as a function of inference-time edge dropout rate p. The Full topology degrades more gracefully than Sequence-only across a wide range of dropout rates, indicating higher robustness to partial graph observations.

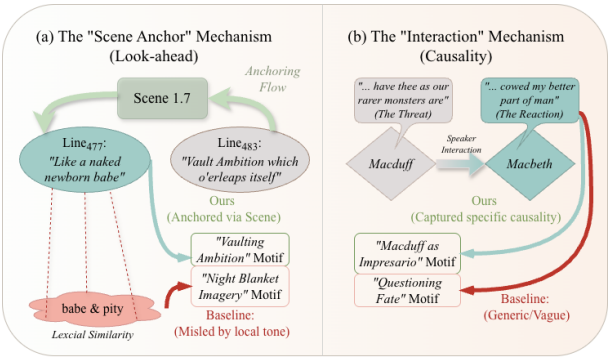

Figure 6: **Structural reasoning beyond local cues.** (a) CRITIC'S EYE anchors a metaphor to its scene context, avoiding a lexical-similarity trap. (b) CRITIC'S EYE propagates cross-speaker causality via interaction structure, resolving a specific motif where a sequential baseline predicts a generic one.

Structural Resilience and Redundancy. Beyond peak accuracy, we stress-test robustness by applying random edge dropout at inference time (Figure 5). As the dropout rate p increases, the *Sequence-only* model degrades sharply, reflecting its fragility to broken local contexts. In contrast, the *Full* topology maintains consistently higher MRR and exhibits a smoother decline. This demonstrates **structural redundancy**: when the local sequential path is severed, the heterogeneous edges ($\mathcal{E}_{\text{HIER}}, \mathcal{E}_{\text{AGENT}}$) provide alternative message-passing routes, preserving thematic signals through the "dilatory space" of the narrative.

5.4 Qualitative Analysis: Structural Disambiguation

Figure 6 demonstrates how topological priors resolve ambiguities that mislead sequence-only models.

Case (a): Scene Anchoring. The line *"pity, like a naked newborn babe"* (Act 1, Scene 7) contains salient lexical imagery that distracts sequence baselines toward *Night Imagery*. CRITIC'S EYE instead routes evidence through the Scene node, aggregating thematically consistent cues within the same dramatic unit. In particular, Act 1, Scene 7 also contains the explicit ambition framing (e.g., "vaulting ambition", Line #483), making ***Ambition*** topologically supported even when the target span itself is metaphorical.

Case (b): Interactional Causality. For Macbeth's reaction *"it hath cowed my better part of man"*, baselines default to an underspecified *Questioning Fate* label when encoding the line in isolation. Our graph makes the immediate antecedent accessible via the spoke_by edge from Macduff's threat to exhibit Macbeth as a "monster", allowing CRITIC'S EYE to recover the specific ***Impresario*** motif grounded in the cross-speaker exchange.

Why topology helps. Both cases illustrate *structural under-conditioning* in sequence encoders: the decisive evidence is not in the target span but in its structural neighborhood (within the same scene or an adjacent turn). In our graph, such cues become reachable via short typed paths (Line→ Scene→Line; Line→ spoke_by →Character), imposing an explicit *evidence-routing prior*. As a result, CRITIC'S EYE prefers motifs supported by coherent structure over those driven by local lexical salience or underspecified global labels. Detailed analysis is provided in Appendix E.

6 Conclusion

We presented CRITIC'S EYE , a graph-based framework that bridges the gap between the linear *fabula* and the topological *sjužet* in narrative understanding. While standard sequential models struggle with the non-chronological distortions of literary discourse, our framework operationalizes these structural dependencies into a heterogeneous graph. By employing a transductive learning objective, our approach overcomes the severe long-tail sparsity of thematic motifs, significantly outperforming LLM baselines. This indicates that the introduced structural inductive bias provides necessary scaffolding for reasoning about dispersed narrative clues. Consequently, our work suggests that for deep literary analysis, models must move beyond sequence modeling to internalize the complex, non-linear architecture of storytelling.

Limitations

While our results demonstrate the efficacy of structure-aware pre-training, we acknowledge inherent constraints regarding granularity, data density, and the trade-off between precision and scalability.

Genre Adaptation and Granularity. Our experiments focus on drama, where topological markers (Acts/Scenes/Turns) are explicit. Generalizing to prose fiction requires fundamentally different, finer-grained annotation strategies. In novels, the *sjužet* organization is often implicit (e.g., free indirect discourse, stream of consciousness) and lacks clear segmentation. Adapting CRITIC'S EYE to such forms would necessitate computationally expensive discourse parsing to recover the latent scaffold, imposing a trade-off between structural precision and the breadth of compatible literary forms.

The "Well-Studied" Assumption. Our pipeline operates under a **canonical work assumption**: it relies on the availability of a rich scholarly infrastructure— specifically, high-fidelity structural backbones (TEI) and dense expert commentary. This restricts our method to "well-studied" texts (e.g., Shakespeare, Joyce) where such dual-layer supervision is available. Applying this framework to "in-the-wild" web fiction or under-resourced languages would require robust upstream parsing, where noise could propagate to the GNN and degrade representation quality.

Closed-World Constraint vs. Open Discovery. We currently frame motif identification as retrieval over a fixed, expert-curated inventory ($|\mathcal{M}| = 281$). This closed-world assumption allows for high-precision alignment with scholarly consensus but precludes *ab initio* discovery of novel or evolving themes. Unlike self-supervised LLMs that consume raw text to maximize generality, CRITIC'S EYE serves as a **specialized, high-precision instrument** for Deep Digital Humanities. While this design ensures rigorous modeling of complex narrative architectures, it inherently limits the system's ability to scale as a general-purpose foundation model for arbitrary text analysis.

Linguistic Domain Shift. Our node initialization relies on off-the-shelf sentence encoders (e.g., BGE) trained primarily on modern English. This introduces a **domain gap** when processing the Early Modern English of Shakespeare (e.g., archaic syntax and semantic drift). While our structural pre-training objectives partially mitigate this by contextualizing embeddings within the narrative topology, the underlying semantic representations remain suboptimal compared to a model linguistically adapted to the specific historical period.

Downstream Integration with Generative Models. Our current framework focuses on the *discriminative* task of motif discovery. While we establish that structural priors are essential for accurate retrieval, we leave the question of **downstream utilization** under-explored. Specifically, how these structure-aware topological embeddings can be injected back into LLMs— for instance, to guide controllable text generation, automate literary essay writing, or enhance Graph-RAG pipelines— remains an open research direction. We envision CRITIC'S EYE not as a replacement for foundation models, but as a modular structural adapter that bridges the gap between raw text processing and deep thematic reasoning.

Ethics Statement

Data Usage and Intellectual Property. The primary textual data used in this study, William Shakespeare's *Macbeth*, is in the public domain. The critical essays and secondary literature used for motif inventory construction were accessed through academic repositories consistent with fair use principles for research purposes. We have ensured that our released dataset contains no personally identifiable information (PII) or copyright-infringing material.

Content Safety and Literary Context. While the source text (*Macbeth*) inherently contains themes of violence, regicide, and psychological distress, these are analyzed strictly within a literary and historical context. In our use of Large Language Models (LLMs) for weak supervision and comparative evaluation, we implemented prompt constraints to ensure the models remained focused on literary analysis and did not generate gratuitous, harmful, or non-consensual content. We manually reviewed a subset of model outputs to verify adherence to these safety guidelines.

Model Usage and Compute. We utilized both open-weights models (e.g., Llama-3, Qwen) and proprietary APIs (e.g., GPT, Claude). We adhered

to the respective Terms of Use and Acceptable Use Policies of all service providers. Furthermore, our proposed method (CRITIC'S EYE) relies on lightweight Graph Neural Networks, which incur significantly lower computational costs and carbon footprint compared to fine-tuning Large Language Models, aligning with Green AI principles.

Human Subjects. This research did not involve human subjects or crowdsourced annotators (e.g., Amazon Mechanical Turk). All "ground truth" labels were derived from established literary criticism or expert verification by the authors. Therefore, no Institutional Review Board (IRB) approval was required.

References

Joseph Quincy Adams. 1917. *Shakespearean Playhouses: A History of English Theatres from the Beginnings to the Restoration*. Houghton Mifflin Company, Boston. Project Gutenberg eBook #22397.

George Arthur Baker, Ankush Raut, Sagi Shaier, Lawrence E. Hunter, and prefix=von der useprefix=false family=Wense, given=Katharina. 2024. Lost in the middle, and In-between: Enhancing language models' ability to reason over long contexts in multi-hop QA. *Preprint*, arXiv:2412.10079.

Mieke Bal and van Christine Boheemen. 2009. *Narratology: Introduction to the Theory of Narrative*. University of Toronto Press.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379. Association for Computational Linguistics.

H. Bloom, S.P.H.H. Bloom, and W. Shakespeare. 1987. *William Shakespeare's Macbeth*. Modern Critical Interpretations. Chelsea House.

Harold Bloom and Janyce Marson, editors. 2008. *Macbeth*. Bloom's Shakespeare Through the Ages. Infobase Publishing, New York. General Editor: Harold Bloom; Volume Editor: Janyce Marson.

A. C. Bradley. 1904. *Shakespearean Tragedy: Lectures on Hamlet, Othello, King Lear, Macbeth*. Macmillan and Co., London. Project Gutenberg eBook #16966.

Wenyu Cai, Mengfan Li, Xuanhua Shi, Yuanxin Fan, Quntao Zhu, and Hai Jin. 2025. RE-SEGNN: Recurrent semantic evidence-aware graph neural network for temporal knowledge graph forecasting. 68(2):122104.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. BooookScore: A systematic exploration of book-length summarization in the era of LLMs. *Preprint*, arXiv:2310.00785.

King-Kok Cheung. 1984. Shakespeare and kierkegaard: "dread" in macbeth. 35(4):430–439.

John D. Cox. 2013. Religion and suffering in Macbeth. *Christianity & Literature*, 62(2):225–240.

Benedetto Croce. 1920. *Ariosto, Shakespeare and Corneille*. Henry Holt and Company, New York. Project Gutenberg eBook #54165.

Zhiyu Fang, Shuai-Long Lei, Xiaobin Zhu, Chun Yang, Shi-Xue Zhang, Xu-Cheng Yin, and Jingyan Qin. 2024. Transformer-based reasoning for learning evolutionary chain of events on temporal knowledge graph. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 70–79, New York, NY, USA. Association for Computing Machinery.

George Fletcher. 1844. Macbeth: Shakespearean criticism and acting. *The Westminster Review*, pages 1–72. Art. I. Review of Knight's Cabinet Edition of Shakspere.

Gérard Genette. 1980. *Narrative Discourse: An Essay in Method*. Cornell University Press.

John Gerlach. 1973. Shakespeare, kurosawa, and "macbeth": A response to J . Blumenthal. 1(4):352–359.

Jonathan Gil Harris. 2007. The smell of Macbeth. *Shakespeare Quarterly*, 58(4):465–486.

Michael Hart. 2010. The project gutenberg. *Unesco courier*.

William Hazlitt. 1903. *Characters of Shakespeare's Plays*, chapter Macbeth. The Macmillan Company, London. Originally published in 1817.

C. H. Herford. 1921. *Shakespeare's Treatment of Love & Marriage, and Other Essays*. T. Fisher Unwin, London. Project Gutenberg eBook #69468.

David Herman. 2003. *Narrative Theory and the Cognitive Sciences*. CSLI Publications.

Peixin Huang, Xiang Zhao, Minghao Hu, Zhen Tan, and Weidong Xiao. 2025. Logic induced high-order reasoning network for event-event relation extraction. 39(23):24141–24149.

Indira Gandhi National Open University. n.d. *Unit 4: Macbeth: Critical Responses*. School of Humanities, IGNOU, New Delhi. Course material for MEG-02: British Drama.

Samuel Johnson. 1765. *Preface to Shakespeare*. J. & R. Tonson, London. Project Gutenberg eBook #5429.

J. Gregory Keller. 1995. The moral thinking of Macbeth. *Philosophy and Literature*, 19(1):41–56.

Tetsuo Kishi, Roger Pringle, and Stanley Wells, editors. 1994. *Shakespeare and Cultural Traditions: The Selected Proceedings of the International Shakespeare Association World Congress, Tokyo, 1991*. University of Delaware Press, Newark. Proceedings of the International Shakespeare Association World Congress.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: The impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353. Association for Computational Linguistics.

Ian R. McKenzie, Alexander Lyzhov, Michael Martin Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Xudong Shen, Joe Cavanagh, Andrew George Gritsevskiy, Derik Kauffman, Aaron T. Kirtland, Zhengping Zhou, Yuhui Zhang, Sicong Huang, Daniel Wurgaft, Max Weiss, Alexis Ross, Gabriel Recchia, and 7 others. 2023. Inverse Scaling: When Bigger Isn't Better.

Nick Moschovakis, editor. 2008. *Macbeth: New Critical Essays*. Shakespeare Criticism. Routledge, London.

Fuad Abdul Muttaleb and Mohammad Khair Mohammad Rawashdeh. 2019. Macbeth's political imagination: The struggle for kingship in Macbeth. *Jerash for Research and Studies*, 20(2):671–683.

Masahiko Omura. 2018. The problem of sympathy in Macbeth. *Hitotsubashi Kenkyu*, 43(1):1–18.

Andrew Piper and Sunyam Bagga. 2024. Using large language models for understanding narrative discourse. In *Proceedings of the 6th Workshop on Narrative Understanding*, pages 37–46. Association for Computational Linguistics.

V. Propp. 1968. *Morphology of the Folktale: Second Edition*. University of Texas Press.

Sigit Purnomo. 2013. Tragedy and moral values in william shakespeare's macbeth: A structural analysis. *Register Journal*, 6:125.

Elahe Rahimtoroghi, Jiaqi Wu, Ruimin Wang, Pranav Anand, and Marilyn Walker. 2017. Modelling protagonist goals and desires in first-person narrative. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–369. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Susan Sachon. 2020. *Shakespeare, Objects and Phenomenology: Daggers of the Mind*. Palgrave Macmillan, Cham.

Lasantha Senanayake and Stephen G. Ware. 2025. Language models as narrative planning heuristics. In *Proceedings of the 20th International Conference on the Foundations of Digital Games*, FDG '25, pages 1–9. Association for Computing Machinery.

William Shakespeare. n.d. Macbeth. Folger Digital Texts. Accessed: 2025-12-17.

D.N. Smith. 1961. *Shakespeare Criticism: A Selection, 1623-1840*. The World's Classics. Oxford University Press.

Ellen Spolsky. 2011. An embodied view of misunderstanding in macbeth. 32(3):489–520.

Caroline F. E. Spurgeon. 1935. *Shakespeare's Imagery and What It Tells Us*. Cambridge University Press.

Dominik Stammbach, Maria Antoniak, and Elliott Ash. 2022. Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 47–56. Association for Computational Linguistics.

Yidan Sun, Qin Chao, and Boyang Li. 2024. Event causality is key to computational story understanding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3493–3511. Association for Computational Linguistics.

Thomas Firminger Thiselton-Dyer. 1883. *Folk-lore of Shakespeare*. Griffith & Farran, London. Project Gutenberg eBook #32183.

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.

Leo Tolstoy. 1906. *Tolstoy on Shakespeare: A Critical Essay on Shakespeare*. Funk & Wagnalls Company, New York. Includes essays by Ernest Crosby and a letter by Bernard Shaw. Project Gutenberg eBook #27726.

John Weiss. 1876. *Wit, Humor, and Shakspeare: Twelve Essays*. Roberts Brothers, Boston. Project Gutenberg eBook #65060.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqingdong, and Yanghua Xiao. 2025. Character is destiny: Can persona-assigned language models make personal choices? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15038–15059. Association for Computational Linguistics.

Dongil Yang, Minjin Kim, Sunghwan Kim, Beongwoo Kwak, Minjun Park, Jinseok Hong, Woontack Woo, and Jinyoung Yeo. 2025. LLM meets scene graph: Can large language models understand and generate scene graphs? a benchmark and empirical study. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21335–21360, Vienna, Austria. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8015–8036. Association for Computational Linguistics.

Chenkai Zhang, Yiming Lei, Zeming Liu, Haitao Leng, ShaoGuo Liu, Tingting Gao, Qingjie Liu, and Yunhong Wang. SeriesBench: A benchmark for narrative-driven drama series understanding. pages 28995–29004.

Runcong Zhao, Wenjia Zhang, Jiazheng Li, Lixing Zhu, Yanran Li, Yulan He, and Lin Gui. 2024. NarrativePlay: Interactive narrative understanding. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 82–93. Association for Computational Linguistics.

A Detailed List of the Critical Corpus

The detailed list is shown in Table 5 on page 13.

B Motif Extraction Prompts

To operationalize the weak supervision step (Section 3.3), we employed a schema-constrained prompting strategy using LLM. The interaction consists of a system instruction defining the critical persona, a user query containing the target text, and an iterative repair mechanism to ensure valid JSON output.

B.1 Prompt Templates

System Prompt

You are an expert literary critic analyzing Shakespeare's *Macbeth*. Your task is to identify the **Top-10 most relevant literary motif IDs** for the target line.

Motif Candidates (ID: Description):
`[Full Inventory of 283 Motifs with IDs and Definitions inserted here]`

Output Format Instructions:
The output must be a valid JSON object conforming to the specified schema.

User Prompt

Target Line: "`[Target Line Text]`"

Context (Preceding/Succeeding Lines):
`[Context Window Text]`

Please extract the top 10 motif IDs based on the provided candidates.

Error Correction Prompt (Triggered on Parse Fail)

The last output was not a valid JSON or did not match the schema.
Error: `[Python Exception Message]`
Please fix the JSON output. Return ONLY the JSON.

Figure 7: Prompt templates used for weak supervision signal extraction. We inject the full definition of $|\mathcal{M}| = 283$ motifs into the system context.

B.2 Output Schema

We enforced structured generation using a Pydantic definition to guarantee that the output is parsable. The schema requires a ranked list of integers.

```
{
  "motif_ids": [
    integer,  // Rank 1 Motif ID
    integer,  // Rank 2 Motif ID
    ...
    integer   // Rank 10 Motif ID
  ]
}
```

During extraction, we set `temperature=0.0` to maximize determinism.

Table 5: The Critical Corpus for *Macbeth* Motif Extraction (Weak Supervision Sources).

Author / Source	Title
Johnson (1765)	*Preface to Shakespeare*
Fletcher (1844)	*Macbeth: Shakespearean Criticism and Acting*
Weiss (1876)	*Wit, Humor, and Shakspeare: Twelve Essays*
Thiselton-Dyer (1883)	*Folk-lore of Shakespeare*
Hazlitt (1903)	*Eighteenth Century Essays on Shakespeare*
Bradley (1904)	*Shakespearean Tragedy: Lectures on Hamlet, Othello, King Lear, Macbeth*
Tolstoy (1906)	*Tolstoy on Shakespeare: A Critical Essay on Shakespeare*
Adams (1917)	*Shakespearean Playhouses: A History of English Theatres from the Beginnings to the Restoration*
Croce (1920)	*Ariosto, Shakespeare and Corneille*
Herford (1921)	*Shakespeare's Treatment of Love & Marriage, and Other Essays*
Smith (1961)	*Shakespeare Criticism: A Selection, 1623-1840*
Gerlach (1973)	*Shakespeare, Kurosawa, and "Macbeth": A Response to J. Blumenthal*
Cheung (1984)	*Shakespeare and Kierkegaard: "Dread" in Macbeth*
Bloom et al. (1987)	*William Shakespeare's Macbeth*
Kishi et al. (1994)	*Shakespeare and Cultural Traditions: The Selected Proceedings of the International Shakespeare Association World Congress*
Keller (1995)	*The Moral Thinking of Macbeth*
Harris (2007)	*The Smell of Macbeth*
Bloom and Marson (2008)	*Macbeth (Bloom's Shakespeare Through the Ages)*
Moschovakis (2008)	*Macbeth: New Critical Essays*
Spolsky (2011)	*An Embodied View of Misunderstanding in Macbeth*
Cox (2013)	*Religion and Suffering in Macbeth*
Purnomo (2013)	*Tragedy and Moral Values in William Shakespeare's Macbeth: A Structural Analysis*
Omura (2018)	*The Problem of Sympathy in Macbeth*
Muttaleb and Rawashdeh (2019)	*Macbeth's Political Imagination: The Struggle for Kingship in Macbeth*
Sachon (2020)	*Shakespeare, Objects and Phenomenology: Daggers of the Mind*
Indira Gandhi National Open University (n.d.)	*Unit 4: Macbeth: Critical Responses*

C Reproducibility Details

We detail the hyperparameters and hardware configurations to ensure reproducibility.

C.1 Hyperparameters

- **Encoder Architecture:** 2-layer Heterogeneous Graph Transformer (HGT).
- **Hidden Dimension:** 768.
- **Optimization:** AdamW optimizer with learning rate 2e-5, weight decay 0.01.
- **Model Size:** batch size 32, 50 epochs on a single Apple M2 (macOS; PyTorch MPS).
- **Compute Budget:** training completed within < 1 day wall-clock time on the above machine.

C.2 Baseline Configurations

For closed-source models (e.g., GPT-5.2-Chat, Claude-Opus-4.5), we utilized OpenRouter's respective APIs with temperature set to 0.0 to maximize determinism.

D Full Benchmark Results

Table 6 presents the comprehensive performance metrics for all evaluated models, including smaller open-source baselines (e.g., Qwen-4B) and random guess baselines omitted from the main text for brevity.

E Qualitative Case Studies

We provide the full textual context and model outputs for the case studies discussed in Section 5.4.

In this section, we expand the two case studies in Section 5.4 by making explicit (i) what the baselines attend to, (ii) what structural evidence CRITIC'S EYE has access to, and (iii) how this changes the predicted motif.

E.1 Case Study 1: Metaphor disambiguation via scene-level structure

Target Line: *"And pity, like a naked newborn babe"* (Line #477, Act 1, Scene 7).

Method	Hit				nDCG				MRR
	@1	@3	@5	@10	@1	@3	@5	@10	
Embedding Baselines									
Random Guess	0.0035	0.0106	0.0177	0.0353	0.0035	0.0075	0.0104	0.0161	0.0220
Sentence-BERT	0.0424	0.0742	0.1060	0.1060	0.0406	0.0531	0.0637	0.0736	0.0812
BGE-Large-v1.5	0.0389	0.0707	0.0813	0.1307	0.0371	0.0488	0.0521	0.0665	0.0748
Open-Source LLMs									
Qwen-3-4B									
Zero-shot	0.0247	0.0353	0.0459	0.0565	0.0247	0.0305	0.0349	0.0384	0.0328
+ RAG	0.0212	0.0314	0.0636	0.0742	0.0212	0.0413	0.0413	0.0446	0.0354
Llama-3-8B									
Zero-shot	0.0318	0.0707	0.0777	0.0989	0.0318	0.0545	0.0574	0.0641	0.0531
+ RAG	0.0283	0.0459	0.0459	0.0530	0.0283	0.0380	0.0380	0.0405	0.0364
GPT-OSS-20B									
Zero-shot	0.0424	0.0707	0.1025	0.1060	0.0424	0.0598	0.0596	0.0742	0.0639
+ RAG	0.0424	0.0883	0.1166	0.1201	0.0424	0.0700	0.0811	0.0822	0.0699
DeepSeek-R1-70B									
Zero-shot	0.0813	0.1237	0.1343	0.1590	0.0813	0.1057	0.1101	0.1180	0.1052
+ RAG	0.0742	0.1166	0.1343	0.1802	0.0742	0.0991	0.1063	0.1213	0.1032
GPT-OSS-120B									
Zero-shot	0.0247	0.0495	0.0989	0.0989	0.0247	0.0394	0.0596	0.0596	0.0471
+ RAG	0.0318	0.0636	0.0848	0.0848	0.0318	0.0491	0.0582	0.0582	0.0495
Closed-Source LLMs (Proprietary)									
GPT-5.2-Chat									
Zero-shot	0.0636	0.1272	0.1696	0.1837	0.0636	0.0986	0.1155	0.1201	0.0998
+ RAG	0.0495	0.1166	0.1413	0.2120	0.0495	0.0872	0.0974	0.2082	0.0928
Claude-Opus-4.5									
Zero-shot	0.0777	0.1484	0.1979	0.2756	0.0777	0.1177	0.1381	0.1633	0.1290
+ RAG	0.0671	0.1343	0.1661	0.2615	0.0671	0.1058	0.1193	0.1499	0.1162
Grok-4.1-Fast									
Zero-shot	0.0636	0.1378	0.1731	0.2509	0.0636	0.1049	0.1190	0.1439	0.1114
+ RAG	0.0777	0.1449	0.1802	0.2650	0.0777	0.1164	0.1304	0.1572	0.1248
Gemini-3-pro-preview									
Zero-shot	0.0707	<u>0.1625</u>	0.2155	0.2862	0.0707	<u>0.1231</u>	0.1453	0.1683	<u>0.1317</u>
+ RAG	<u>0.0883</u>	0.1555	<u>0.2261</u>	<u>0.3322</u>	<u>0.0883</u>	0.0883	<u>0.1562</u>	<u>0.1909</u>	0.1248
CRITIC'S EYE (Ours)	**0.5477**	**0.7739**	**0.8481**	**0.9223**	**0.5477**	**0.6462**	**0.6820**	**0.7156**	**0.6762**

Note: nDCG@1 is equivalent to Hit@1 under binary relevance. **Baselines:** We underline the second-best performance (best among baselines). Newer closed-source models (e.g., Claude-Opus-4.5) outperform open weights but still lag significantly behind CRITIC'S EYE .

Table 6: Performance on motif identification ($N = 283$). Comparison against Open and Closed SOTA LLMs.

Context. The line occurs in Macbeth's soliloquy weighing the assassination of Duncan. Although the local surface contains innocence-related imagery ("babe") and affect ("pity"), the surrounding speech in Act 1, Scene 7 foregrounds Macbeth's motive of *ambition*.

Baseline Failures.

- The *Sequence-Only* model misclassifies this as *"Night Blanket Imagery"*, misled by the somber, metaphorical tone of the immediate text.
- The *Heterogeneous-Only* model predicts the generic *"Double Trust in Messengers"*, relying on coarse character correlations with weak textual grounding.

CRITIC'S EYE prediction and supporting structure. CRITIC'S EYE ranks ***Vaulting Ambition*** as the top motif. Notably, the explicit phrase "vaulting ambition" appears a few lines later (L#1003), but both lines are located in the same structural unit (Act 1, Scene 7). In our graph, the target line is connected to the corresponding `Scene` node, which in turn links to multiple ambition-related cues within the same scene. This scene-level connectivity provides a short topological path from the target metaphor to ambition-centered evidence, reducing the tendency to overfit to local lexical similarity.

E.2 Case Study 2: Causality across speakers via interaction structure

Target Line: *"For it hath cowed my better part of man!"* (Line #2387, Act 5, Scene 8).

Context. Macbeth's utterance immediately follows Macduff's threat to exhibit him publicly as a "monster" if Macbeth refuses to fight, i.e., the reaction is locally triggered by a preceding cross-speaker exchange.

Baseline prediction. The sequence baseline predicts *Questioning Fate and Identity*. While globally compatible with Macbeth's late-stage realization, it under-specifies the immediate trigger in this exchange.

CRITIC'S EYE prediction and supporting structure. CRITIC'S EYE ranks ***Macduff as Impresario*** as the top motif, capturing the humiliation/showman framing in Macduff's preceding threat. In the graph, the target line is linked to (i) the local dialogue context and (ii) the speaker-interaction structure connecting Macduff's antecedent threat to Macbeth's response. This additional cross-speaker connectivity makes the causal antecedent topologically accessible at inference time, favoring a fine-grained motif grounded in the exchange rather than a generic fate-related label.