



MASTERARBEIT | MASTER'S THESIS

Titel | Title

Assessing Story Coherence of Long Narratives using Large
Language Models and Retrieval-Augmented Generation

verfasst von | submitted by
Enno Peter Stubler B.Sc.

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of
Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree
programme code as it appears on the
student record sheet:

UA 066 921

Studienrichtung lt. Studienblatt | Degree
programme as it appears on the student
record sheet:

Masterstudium Informatik

Betreut von | Supervisor:

Univ.-Prof. Dipl.-Ing. Dr. Helmut Hlavacs

Acknowledgements

First and foremost, I would like to express gratitude to my supervisor, Dr. Helmut Hlavacs, for his guidance and support throughout my research journey. His feedback has been instrumental for this thesis.

I would also like to extend my thanks to my friends, not only for their continuous encouragement and patience, but also for taking the time to read this work and provide thoughtful feedback. Your comments and suggestions contributed to improving the quality of this thesis.

Abstract

Long-form story narratives present unique challenges in automated coherence evaluation due to their scale, complexity, and diverse structures. In this work, the assessment of story coherence in extended texts is explored using large language models (LLMs) enhanced with retrieval-augmented generation (RAG). The work begins by surveying the key dimensions underpinning this research area. Building on these foundations, four validation systems are introduced, designed to score story coherence within long texts. The evaluation leverages a dataset of 120 full-length books from Project Gutenberg, each exceeding 150,000 tokens and spanning multiple literary genres. An in-depth analysis of model performance is conducted, including detailed case studies on selected titles. Additionally, the systems are benchmarked in terms of validation accuracy and computational cost, offering insights into their practical trade-offs for narrative assessment. The results show that the proposed systems can reliably distinguish between coherent and incoherent (positive and negative) story samples—whereas a baseline validation method fails to do so—demonstrating the effectiveness of the approach. Finally, specific incoherent scenarios are systematically simulated by applying perturbations to the validation examples. It was shown that the proposed validation methods respond more effectively to negatively altered text segments and are able to identify errors more reliably.

Kurzfassung

Große Texte stellen für die automatische Bewertung der Handlungskohärenz eine besondere Herausforderung dar – nicht zuletzt wegen ihrer schiereren Länge, strukturellen Vielfalt und narrativen Tiefe. In der vorliegenden Arbeit nähern wir uns dieser Problematik mit einem neuartigen Ansatz: Es werden große Sprachmodelle mit Retrieval Augmented Generation Systemen kombiniert, um die Kohärenz ausgedehnter Texte systematisch zu analysieren. Zuerst werden die zentralen Dimensionen, die diesem Forschungsbereich zugrunde liegen untersucht. Auf dieser Grundlage werden vier Validierungssysteme vorgestellt. Die Evaluation basiert auf einem Datensatz von 120 Büchern aus der Project Gutenberg Datenbank, die jeweils über 150.000 Tokens umfassen und mehrere literarische Genres abdecken. Es wird eine detaillierte Analyse der Methoden durchgeführt, einschließlich eingehender Fallstudien zu ausgewählten Titeln. Darüber hinaus werden die Systeme hinsichtlich Validierungsgenauigkeit und Rechenaufwand verglichen, um Einblicke in die praktischen Abwägungen bei der Bewertung von Narrativen zu geben. Die Ergebnisse zeigen, dass die vorgeschlagenen Systeme zuverlässig zwischen kohärenten und inkohärenten (positiven und negativen) Erzählproben unterscheiden können – im Gegensatz zu einer Baseline-Methode, die hierzu nicht in der Lage ist – und belegen so die Wirksamkeit des Ansatzes. Letztlich werden mithilfe von Perturbationen der Validierungsbeispiele gezielt bestimmte inkohärente Szenarien nachgestellt. Dabei konnte gezeigt werden, dass die Validierungsmethoden besser auf negativ veränderte Textabschnitte reagieren und Fehler zuverlässiger aufzeigen.

Contents

Acknowledgements	i
Abstract	iii
Kurzfassung	v
List of Figures	ix
1. Introduction	1
2. Related Works	3
3. Fundamentals	9
3.1. Principles of Narrative Coherence	9
3.2. Evaluating Narrative Quality with LLMs	11
3.3. Retrieval-augmented Generation Systems	14
3.4. Knowledge Graphs in Story Evaluation	15
4. Assessing Story Coherence	17
4.1. Baseline Validation	18
4.2. GRAG Simple Validation	18
4.3. GRAG Context-Aware Validation	22
4.4. GRAG Complex Question Validation	25
4.5. GRAG Relationship Score Validation	29
5. Evaluation	35
5.1. Dataset Preparation	35
5.2. Validation Results	40
5.3. Parameters	48
5.4. Case Study	50
5.5. Analysis of Validation Results	55
5.6. Discussion	59
6. Conclusion	61
Bibliography	65
A. Appendix	75

List of Figures

3.1. Comparison of LLM optimization methods by external knowledge requirements and model adaptation, Reprinted from [GXG ⁺ 24]	13
4.1. Basic validation System	17
4.2. An example ZS-prompt that instruct the evaluation of coherence	19
4.3. RAG analysis method Design	20
4.4. RAG graph of the book “The Slave Planet”	21
4.5. An example prompt that instruct the evaluation of the thematic coherence	22
4.6. RAG graph context extraction method design	23
4.7. An example query prompt for the GRAG to retrieve context	23
4.8. RAG graph complex question method design	25
4.9. System prompt to extract coherence precondition questions	27
4.10. GRAG output instruction prompt	28
4.11. RAG graph comparison method design	29
4.12. The GRAG result of a TTV in the graph comparison validation method	31
4.13. GRAG edge evaluation prompt	32
4.14. GRAG edge statement analysis result	33
5.1. Word cloud of themes and genres present in the 120 book corpus.	36
5.2. Experiment setup and chapter comparison validation	37
5.3. Experiment setup and perturbed chapter validation	38
5.4. A dynamic prompt to create PTPCs	39
5.5. Chapter comparison mean of all validation results by coherence type and validation method	41
5.6. One sided wilcoxon test comparing each measurement of the baselline validation method and the context-aware validation method	42
5.7. Perturbed validation mean of all validation results by coherence type and validation method	43
5.8. Local but not global coherence PTPC prompt	44
5.9. Mean comparison of naive local and global perturbed samples	45
5.10. Mean comparison of local and global perturbed samples with refined perturbations	46
5.11. Time measurement comparison of validation methods	47
5.12. Validation result comparison off two TTV sample sizes 500 words(left) and 1000 words (right)	50
5.13. Local and global search parameter comparison at the context-aware validation method	51

List of Figures

5.14. Baseline validation (top) vs context-aware validation (bottom) results in focus of non-coherent story samples	53
5.15. Mean validation results on non-coherent data samples only	55
5.16. The summary of the positive validation sample from the book: The Castle of Otranto	56

1. Introduction

By 2025, large language models (LLMs) have become deeply integrated into various aspects of daily life, significantly influencing how individuals engage with information and perform creative tasks. Their widespread adoption underscores a transformative shift in human-computer interaction and information practices. This influence of LLMs is also markedly evident in creative domains. In the field of creative writing, LLMs are employed to generate ideas, compose dialogue, and structure narratives. These tools have demonstrated an ability to produce outputs with a diversity and quality comparable to human-authored content [CPR⁺25]. Beyond creative writing LLMs are increasingly used to generate narratives and facilitate user collaboration in story creation, enhancing immersion and personalization [THL⁺24, TNG⁺24, NGG20]. These advancements highlight LLMs' potential in shaping narrative-driven applications in both entertainment and educational contexts.

On the other hand a study by [YJ24] highlights that LLMs struggle with maintaining narrative consistency in storytelling, particularly in aspects such as character development and coherence. The authors note that, although LLMs can generate locally coherent text, they frequently fail to preserve global narrative structures, leading to disjointed and less engaging stories. However this claim holds for medium-length texts under 10,000 words, where evaluation methods are well established. [QDH⁺24] highlight that critical gap in current evaluation methods for LLMs. Widely used metrics such as BLEU [PRWZ02] and ROUGE [Lin04] primarily measure surface-level lexical overlap, failing to assess deeper narrative qualities like coherence, character development, and thematic progression.

In light of these recent developments, automatic story generation (ASG) emerge as a valuable tool for the entertainment industry, particularly in applications such as interactive storytelling, video game quest generation, and digital media production. The integration of robust validation methods plays a crucial role in ensuring the quality of generated narratives, as these techniques detect and correct inconsistencies, improving the overall coherence and creative depth of ASG outputs [SPS⁺19, TDM⁺19]. By embedding feedback loops into the generation process, ASG systems iteratively refine their narratives, resulting in better content [GHZ⁺20, PA20]. Importantly, recent studies highlight the potential financial benefits of ASG technologies, as they contribute to reducing the costs and time associated with manual content creation while enabling scalable and personalized storytelling experiences [ALP21]. This findings calls for novel evaluation frameworks of long-form storytelling, underscoring the need for more robust and comprehensive tools to meaningfully assess extended text sizes.

At the same time, a technological infrastructure has developed to support and extend the capabilities of LLMs. Central to this are semantic embedding models, which have shown improvements in capturing contextual meaning and enabling semantic search

1. Introduction

[TSG⁺24]. Vector databases now play a key role in managing high-dimensional embedding spaces, allowing efficient similarity search and enabling real-time retrieval in LLM-driven applications [JSH⁺25]. Additionally, retrieval-augmented generation (RAG) systems have emerged as effective tools for incorporating external, verifiable knowledge into LLMs outputs, helping to reduce hallucinations and factual inaccuracies [LPP⁺21, GXG⁺24]. Alongside these advancements, techniques such as fine-tuning, prompt engineering, and inference optimization have been developed to adapt LLMs for specific domains while improving computational efficiency [ZLCF24, WFH⁺23]. These innovations collectively reflect a trend toward modular, agent-based Artificial Intelligence (AI) systems capable of executing complex, multi-step reasoning processes [Kan25]. Building on this foundation, the present study introduces novel validation methods for ultra-long narrative texts (ULNT) that combine LLMs reasoning with the contextual accuracy of RAG-based retrieval systems [JYHA24].

Finally, evaluating in ASG remains methodologically unclear, with inconsistent and scattered metrics across the literature. Common evaluation dimensions include coherence, plot consistency, character development, linguistic fluency, clarity, stylistic consistency, creativity, and emotional impact—each involving multiple sub-aspects. Despite their frequent use, these metrics are conceptually complex, interdependent, and lack standardized definitions or benchmarks [YJ24, CCSC22]. This work focuses on plot consistency, also known as narrative coherence, which is regarded as a core dimension for assessing the logical and causal structure of stories. It defines the works’s scope and exemplifies the broader challenges in automatic story evaluation (ASE) quality in generative systems. The following theoretical framework outlines the foundations of this approach by reviewing relevant strands of research: narrative theory and narratology, recent advancements of LLMs in storytelling and evaluation, the role of RAG, and efforts aimed at evaluating ultra-long text generation (ULTG).

Research Questions

The preceding part outlined diverse strands of research that reveal both the growing potential and the current limitations of automated narrative evaluation—especially for long-form and complex narratives. Building on this background, central research questions arise, aimed at advancing the interpretability, and reliability of narrative evaluation systems in light of the current technology stack:

RQ1: How can LLMs and RAG systems effectively validate story coherence of long-form text narratives?

RQ2: How can the results of a validation method for ULNT be meaningfully interpreted in the absence of annotated benchmarks or standardized evaluation datasets?

RQ3: Can the proposed validation system reliably detect narrative inconsistencies and provide both a meaningful coherence score and useful feedback to guide revision or analysis?

2. Related Works

In the academic context of this work, two main categories are of central importance: first, ASG, which encompasses the task of generating a story based on a given input; and second, ASE, assessing story performance. Traditionally, methods in both story generation and evaluation have often relied on statistical or lexical approaches. Among the most widely used systems are BLEU [PRWZ02], originally developed for evaluating machine translation based on lexical similarity, and ROUGE [Lin04], which measures lexical overlap between generated and reference texts. The breakthrough advancements in AI and natural language processing (NLP) in recent years have opened entirely new approaches to these tasks. Recent studies have demonstrated that current LLMs models already outperform traditional lexical metrics, which lack the capacity to assess complex aspects of language, such as coherence, creativity, or narrative consistency. [NGG20] present a collaborative storytelling system where humans and a large language model co-author stories. It showcases the already strong performance of LLM methods in the year of 2020. Furthermore, a more recent paper [CSC24] investigates using generative methods for human annotators, analyzing their correlation with human ratings and other automatic metrics. The study finds that LLMs outperform existing automatic measures. While the method surpasses traditional metrics, the prompting approach used still falls short of matching human performance. As a result, the scope of this thesis is further narrowed to methods based on NLP or LLMs that focus specifically on story generation and evaluation.

Furthermore, this thesis concentrates on open-ended text generation, which necessitates a further narrowing to so-called reference-based evaluation methods. These methods are particularly necessary for open-ended generation tasks, as they evaluate text continuations based on a given story context. A detailed examination of various ASG approaches reveals that corpus sizes of modern methods for open-ended story generation are growing rapidly in sample size. This is primarily because such systems are often employed in an iterative manner, continuously generating text continuations that serve as the basis for subsequent story continuations. Therefore, modern ASG methods like [WHL⁺24] typically employ some form of context or memory mechanism to meet these requirements, which demonstrates how memory-enhanced architectures can effectively manage long-term dependencies in story generation. A similar approach is used in [PERPL24] which employs an action-selection model to iteratively guide story generation toward narratives through a planning-writing loop.

Reviewing these papers and other recent works with modern approaches [MQL24] and [HCL⁺24], another aspect of the research landscape becomes clear: most ASG evaluations rely on human annotators or automatic methods designed for short texts. Moreover, many evaluations focus on short stories, even when the systems are intended for long-form

2. Related Works

generation. This reveals a possible research gap—namely, the lack of automatic evaluation methods for long-form, open-ended story generation. This observation is further supported by the paper “What Makes a Good Story and How Can We Measure It? A Comprehensive Survey of Story Evaluation” [YJ24], which investigates precisely our research areas and presents current methods for ASG and ASE, while also providing a concise overview of the field in general. The paper highlights several important conclusions, including that there is a need for ultra-long story evaluation methods and that there are currently no long-story benchmarks available for reference-based story evaluation methods.

Based on that, this paper navigates within the concise realm of **generative and reference-based textual story evaluation**. In this realm, several interesting methods introduce notable experiments and concepts. Firstly, [FNJL23] explores the capabilities of large language models to assess generated texts through natural language instructions. It evaluates 19 models, ranging from 80 million to 175 billion parameters, across four text generation tasks, 22 evaluation aspects, and 37 datasets. The results demonstrate that GPTScore enables evaluation without requiring annotated data, effectively addressing key challenges in text evaluation. An advancement related to GPTScore is the RECURRENT-GPT system [ZJC⁺23], which simulates the recurrence mechanism of RNNs using large language models. It generates text in steps, storing long-term and short-term memory within prompts, allowing for the generation of long texts without forgetting earlier content. The paper also addresses context loss and limited context size by using a sliding window approach to manage and retrieve relevant past information.

Another relevant work is EIPE-text [YWL⁺23]. A framework for long-form narrative generation that iteratively extracts and refines narrative plans from existing stories. In experiments on novel and storytelling tasks, EIPE-text outperforms baseline methods—including the RECURRENTGPT and GPTScore-like methods—in human evaluations. It is particularly notable for introducing a Question Answering (QA)-based validation mechanism to evaluate and refine narrative plans—a concept that is becoming increasingly relevant with the advancement of systems like RAG, which have demonstrated good performance on QA tasks.

[XLCL23] presents an evaluation method that intentionally introduces perturbations (or blunders) into story texts and measures the differences between the original and perturbed versions using large language models. By comparing these differences, DeltaScore evaluates aspects such as fluency, coherence, relatedness, logicity, and interestingness, showing strong performance on storytelling datasets. CoheSentia [MT23] introduces a novel benchmark for evaluating human-perceived coherence in automatically generated texts, using both global and incremental (local, sentence-by-sentence) annotation methods to provide detailed coherence assessments. Notably, CoheSentia is the first to explicitly distinguish between local and global coherence scores—an important aspect overlooked by NLP methods like GPTScore.

Finally SCORE [YHW⁺25] is a framework designed to improve narrative coherence in AI-generated stories by detecting and resolving inconsistencies through key item tracking and episode summaries, using a RAG approach. Experiments show that SCORE enhances story coherence and consistency, outperforming baseline GPT models. This

paper is particularly interesting as it actively addresses the context limitations problem by integrating a RAG system into its evaluation method, and it also uses a similar approach to GPTScore as its baseline evaluation metric.

Another interesting and rapidly growing approach in the field is agentic, automatic story evaluation and generation. A recent survey on AI agents in large language model systems [FTD25], covering developments between the years 2019 and 2025, examined the use of agents for evaluating, and enhancing large language model performance across domains, like reasoning, code generation, information retrieval, multimodal tasks, and interactive assessments. The survey highlighted several successful applications of agent-based methods in the area of multimedia applications, such as drama script generation and evaluation agents. These applications are closely tied to the concept of using large language models as autonomous evaluators, referred to as “LLM-as-a-judge.” In the domain of generative and reference-based textual story evaluation, [XPQ⁺25] represents a notable example. STORYWRITER is a multi-agent story generation framework composed of three agents: an outline agent, a planning agent, and a writing agent, which collaboratively produce coherent, medium-length narratives. The system was evaluated by graduate students in English programs, and automated methods using the Hanna [CCSC22] dataset. In both evaluation settings, STORYWRITER outperformed existing baseline systems.

While not directly introducing a new evaluation method or benchmark, [SDS⁺21] highlights a issue in NLG evaluation: human evaluation scores for metrics like coherence are often uncorrelated to automatic evaluation results. To address this, the authors propose a perturbation checklist, designed to test specific evaluation criteria—which reveal that existing automatic metrics frequently fail to align with human judgments when targeted quality aspects are altered. The paper presents several perturbation methods and concepts that also address capturing differences in local and global coherence. Another interesting source is [GLSM⁺21], which adapts the concept of injecting perturbations into ground-truth story plots to evaluate the performance of their system. They distinguish between plot-level perturbations and local perturbations. For plot-level perturbations, they apply techniques such as sentence reordering, repetition injection, negation, and keyword renaming—such as changing character names—which is particularly effective at disrupting plot coherence. For local perturbations, they use methods like word jumbles and other lexical alterations. The paper presents each perturbation method in detail, discussing the advantages and disadvantages of each. This approach is especially helpful for developing effective validation methods for non-annotated text data.

Validation methods rely on benchmarks and datasets for comparison. Currently, the Storium [AWW⁺20] dataset holds the largest annotated corpus of fiction stories, with an average of 19,278 tokens per story, but unfortunately, it is not publicly available. Besides Storium, PG-19 [RPJL19] is a large-scale, long-form language dataset designed for training and evaluating models on long-range context understanding and story generation. It contains full-length novels, making it ideal for ultra-long story generation tasks; however, unlike Storium, PG-19 is not annotated. This highlights the clear limitation of the absence of publicly available, ultra-long annotated datasets. Interestingly, the PERSE [WYZ⁺24] evaluation framework introduces a creative workaround by using IMDb movie

2. Related Works

ratings to indirectly annotate film scripts, enabling large-scale evaluation. Their method shows a better alignment with human judgments compared to GPT-4, demonstrating strong performance in story generation and other open-ended tasks. Notable datasets like ROCStories [MCH⁺16] and WritingPrompts [WMY⁺25] are among the most widely used benchmarks for ASE and ASG) tasks. However, both focus on short to medium-length stories and cannot capture contextual dependencies present in ultra-long stories. ROCStories contains samples averaging around 88 tokens per story, while WritingPrompts has longer samples with an average of 735 tokens per story, still far from covering long-form narrative structures.

While context plays a crucial role in the evaluation of stories, modern ASE methods increasingly focus on implementing context memory mechanisms. RecurrentGPT uses a sliding LSTM-inspired approach, storing long-term information externally and short-term memory within prompts to maintain context over long generations. Other methods adopt knowledge graph-based retrieval to capture and utilize contextual information during evaluation. For example, [WHL⁺24] employs entity extraction to build a knowledge graph, and through a query module, retrieves relevant information from the graph to produce a grounded and context-aware assessment of the generated text compared to the reference. The concept behind these approaches was showcased in [ACJR12], where the authors analyzed *Alice in Wonderland* using a knowledge graph. By applying metrics such as graph degree, centrality and clustering, they identified key characters and examined the structure of relationships within the story. This demonstrated that knowledge graphs can effectively capture and analyze narrative structures, making them useful tools for assessing plot dynamics and context in complex stories.

RAG systems have proven to be powerful tools for knowledge retrieval and question answering tasks. the survey paper [GXG⁺24] reviewed RAG methods for large language models, focusing on reducing hallucinations and improving factual accuracy. They categorized RAG systems based on retrieval strategies, generation models, and integration techniques, highlighting key trends and open challenges. Their findings showed that RAG systems consistently outperform LLM approaches on QA benchmarks. This suggests that RAG methods could be effectively applied in ASE, particularly for context-aware story evaluation—especially since Paper [YWL⁺23] demonstrated that QA-based approaches can perform strongly in analyzing story content.

The most recent advancement in this field is the combination of knowledge graph (KG)s and RAG systems, namely graph retrieval-augmented generation (GRAG) systems. [ETC⁺25] proposed a graph-based RAG approach for query-focused summarization across large corpora. Their method constructs knowledge graphs with hierarchical summaries, enabling global, comprehensive answers from large document collections. The resulting framework can retrieve summaries linked to entities in the knowledge graph, and through graph traversal, it can analyze neighboring nodes for deeper context and reasoning. The paper challenges the system with text corpora exceeding one million tokens and with questions that require a global understanding of the text. They demonstrate that their approach can outperform traditional RAG systems in this setting. This is particularly noteworthy, as global coherence in narratives also requires a comprehensive understanding

of the overall plot. Therefore, such systems could contribute to the goal of evaluating ultra-long stories.

Finally, another dimension of research contributing to this field is the analysis of prompting strategies. Prompt-based learning, also known as Prompt engineering, has become a recognized paradigm wherein task-specific instructions guide LLMs to perform advanced tasks without modifying their core parameters. [SSS⁺25] presents a survey that categorizes prompting techniques by application areas, detailing the methodologies, types of prompts (manual vs. learned), associated models, datasets, and use cases. The paper highlights the strengths and limitations of each approach, offering a taxonomy and visual summaries. While the survey provides an insightful and well-organized overview, it does not empirically compare the performance of different prompting techniques across tasks, leaving a gap in direct evaluation. Nevertheless, it contributes significantly by introducing and contextualizing concepts central to this work, such as Knowledge-Based Reasoning. It also addresses prompting methods to mitigate hallucination in language models—particularly in the context of RAG techniques—and explores advanced reasoning strategies like Chain-of-Thought prompting.

[WWS⁺23] introduces the concept of Chain-of-Thought Prompting, a simple yet powerful method to enhance the reasoning abilities of LLMs. The core idea is to provide a few exemplars in the prompt that include intermediate reasoning steps, effectively modeling how humans think through problems step-by-step. The authors conduct experiments on three large-scale LLMs of varying sizes and show that CoT prompting leads to significant improvements across reasoning tasks. They showed that smaller, not fine-tuned models, when prompted with chain-of-thought examples, achieve a better performance on the GSM8K math word problem benchmark than bigger fine-tuned models, which indicates a strong performance across other tasks as well.

Building on this innovation [PLS⁺23] introduce a framework that enhances LLMs reasoning by automatically generating intermediate steps and integrating external tools. Unlike traditional Chain-of-Thought approaches that rely on hand-crafted prompts and scripted interactions, ART selects demonstrations from a specialized task library and dynamically incorporates tool outputs during inference. Their task library is specifically designed to cover key areas: Arithmetic, Code, Search and Question Decomposition, Free-form Reasoning, and String Operations(formatting, entailment, etc.). They demonstrates substantial improvements over few-shot and CoT prompting on tasks in BigBench and MMLU benchmarks and matches the performance of hand-crafted solutions on most tasks. While, e.g, BigBench does not include tasks related to story evaluation or narrative reasoning, the ART framework is nonetheless highly innovative. Extending the task library could extend its abilities in automatic prompt-step-generation in specialized domains, such as narrative evaluation, that benefit from reasoning about plot coherence or character development.

The concept of automatic prompt generation also presents a viable solution within the indexing process of GRAG-based knowledge graphs. GRAG constructs knowledge graphs through an indexing procedure in which LLMs are guided by a set of prompts to extract entities and relationships from a source content. Since different text genres contain

2. *Related Works*

distinct entity types—for instance, “use-cases” in software documentation and “characters” in narrative fiction—the default indexing prompts are intentionally designed to be vague and broadly applicable. However, this generality often limits precision, particularly in specialized domains where more targeted extraction is desirable. To address this, [Res24] introduced an automated prompt tuning system for GRAG that enhances domain adaptability through automatic few-shot prompt generation. The system samples a small subset of domain-relevant text and instructs a LLM to generate tailored few-shot examples, which are then used to refine prompts for a given domain. This innovative approach has shown promise in improving knowledge graph construction across various technical and scientific domains and could similarly benefit the generation of retrieval-augmented graphs in narrative-focused applications. In particular, automatic prompt tuning may improve entity and relationship extraction in genre-specific storytelling by aligning prompts with narrative structures unique to those texts.

Finally [PSB21], find that much of the work has remained largely disconnected from longstanding theoretical frameworks developed in the humanities, social sciences, and cognitive sciences. Therefore, they address this by introducing dominant narratological theories to the NLP community and positioning current computational research within the narratology domain. The authors argue that grounding NLP research in narrative theory not only deepens our empirical understanding of narrative but also opens new avenues for practical applications. In doing so, the paper proposes a set of key narrative elements essential for computational modeling—such as narrative sequence, structure, setting, temporality, perspective, and narrative economies—highlighting their significance for developing more theoretically informed and robust narrative processing systems.

Based on the collected information, a clear research gap becomes evident. First, current evaluation methods do not yet cover the latest promising technologies, namely GRAG systems. Second, existing validation approaches do not adequately test ultra-long text samples exceeding 100,000 tokens, leaving a critical gap in assessing global understanding and coherence. Third, there is a lack of annotated benchmark datasets specifically designed for such ultra-long text scenarios, which hinders evaluation and comparison of models in this emerging area.

3. Fundamentals

This section establishes the theoretical foundation for the methods used in this work. It begins by introducing the principles of narrative coherence, which define key aspects of how stories are structured and interpreted. Building on this, this section reviews current approaches to evaluating narrative quality with LLMs, explores the potential of retrieval-augmented generation systems, and examines the role of knowledge graphs in story evaluation.

3.1. Principles of Narrative Coherence

The examination and validation of narratives present a complex challenge, as narratives are multifaceted and context-dependent. [Mca06] emphasizes that narrative coherence should not be understood as a fixed or objectively measurable quality. Instead, coherence is a socially and culturally mediated judgment that spans several dimensions, including temporal sequencing, causal logic, thematic integration, and perceived realism. Notably, what qualifies as “coherent” can differ significantly across cultural, historical, and psychological settings. As such, coherence emerges as a dynamic, holistic property that complicates empirical validation, posing implications for both psychological assessment and narrative evaluation. Narrative coherence is commonly analyzed across distinct dimensions—particularly thematic, cultural, and causal coherence. [HB00] proposed a structured framework for understanding narratives that incorporates these three key aspects: Thematic coherence refers to the presence of an overarching theme or consistent meaning throughout the narrative. Cultural coherence denotes the extent to which the narrative aligns with shared sociocultural expectations regarding the structure and content of a story. Causal coherence, in turn, captures the logical sequencing of events through cause-and-effect relationships, ensuring that the narrative unfolds in a way that is plausible within its context. Given this complexity, narrative coherence cannot be fully understood outside its context. Consequently, the methodological framework of this thesis explicitly used context as a foundational principle in narrative analysis. Also rather than applying a single fixed coherence metric, this work assumes that coherence must be interpreted through the lens of the three metrics explained by [CCSC22] in which narratives are understood.

Causal coherence refers to the degree to which events within a narrative are logically linked by cause, motive, or consequence. A causally coherent narrative establishes clear connections between events, ensuring that earlier occurrences plausibly lead to later developments. This structure supports character development and narrative progression. The importance of causal structure was formalized in early narratology; for example,

3. Fundamentals

[SG79] defined a well-formed story as one comprising an initiating problem and a resolution, bridged by intervening events that follow a logical sequence. This notion has been further supported by discourse psychologists such as [Tv85], who demonstrated that readers tend to search for and remember causally connected events more readily than isolated ones. Recent projects in ASG incorporate causal coherence as an evaluation metric. For instance, [CFB⁺21] introduced a question-answering based method to measure causal coherence, finding their method substantially improved coherence relative to neural baselines. Similarly, [YCSR22] employed neural planning to explicitly enforce causal consistency, demonstrating superior narrative coherence via automated question-answering evaluation. Furthermore, [ACBR20] used commonsense causal reasoning to generate coherent story sequences, validated by human judges as significantly more coherent compared to baseline models. Collectively, these studies confirm that systematically ensuring causal coherence enhances narrative quality.

Cultural coherence captures how well a narrative aligns with culturally shared scripts or expectations for how stories should unfold. It involves conformity to societal norms, roles, and values. Narratives that lack this alignment may be perceived as disjointed or unintelligible within a given cultural framework, even if they maintain internal consistency. This dimension has particular relevance in cross-cultural psychology and AI-based narrative systems, where models trained on culturally homogeneous data may fail to recognize or appropriately evaluate culturally diverse storytelling patterns.

[BD24] addressed this issue in their study. The study investigated how LLMs perform in culturally sensitive tasks such as open-ended question answering and ASG. To test the models' responsiveness to cultural context, the researchers embedded explicit cultural cues—specifically nationality—into the prompts and analyzed the resulting outputs. Their findings indicated that these adaptations only weakly reflected the actual cultural values associated with the respective countries.

Thematic coherence centers on the overall meaning a narrative. According [LW67] foundational work, story coherence requires a unifying perspective or moral defined as thematic coherence. Therefore, even episodic narratives lacking strong causal links can achieve high thematic coherence if they consistently reflect an overarching interpretive frame or theme. Recent research on ASE increasingly emphasizes thematic coherence as a central metric for assessing narrative quality. [CH11], for example, proposed a narrative cohesion framework that incorporates thematic coherence by using lexical networks to identify thematic unity. Narratives lacking a clear thematic structure were rated as less coherent. [PyP24] introduced the Model of Thematic Consistency, which flags narrative events that diverge from the central conflict as thematically incoherent. In a related approach, [TSD24] used topic modeling to evaluate thematic coherence, observing that AI-generated stories tend to exhibit less thematic variation than human-written ones.

Understanding narrative structure also involves analyzing how narrative elements are organized and abstracted into higher-level units across a text. At the passage level, this includes identifying narrative levels—such as embedded stories within stories—and detecting scene boundaries based on spatial, temporal, and character cues. These scene-level distinctions are often referred to using terms like “text worlds” or “narremes,” but

the term “scene” is preferred to clearly separate horizontal segmentation (scene detection) from vertical layering (narrative levels) and broader structural groupings like plotlines. A useful conceptual hierarchy for this purpose is: event \rightarrow scene \rightarrow level \rightarrow plotline \rightarrow plot. At the document level, various theoretical frameworks have been applied to group narrative segments into more abstract categories. One such approach focuses on identifying the “most reportable event” to capture a narrative’s core meaning, although this is more effective for short, self-contained stories. For longer and more complex narratives, alternative models emphasize turning points, moments of disequilibrium or change of state, and rising or falling patterns of fortune—often measured through sentiment analysis.

Narrative sequence refers to the way events are selected and ordered by a narrator to form the structure of narrative discourse. This sequencing is central to storytelling, as it determines how the story unfolds for the reader or audience. Importantly, the order in which events are narrated does not always match the chronological order in which they occur within the storyworld. Narrators often exercise selectivity—choosing certain events to highlight while omitting others—and may present these events out of temporal sequence to achieve specific narrative effects, such as building suspense or emphasizing causality. This creates a distinction between the story (the chronological sequence of events as they would have occurred) and the discourse (the order in which those events are presented). Temporality, on the other hand, refers to the pacing of the narrative—what is often described as the relationship between narrative time and narrated time. It reflects how fast or slow a story is told. For example, some narratives feel almost like real-time experiences, with events unfolding at a one-to-one pace, while others may summarize years or even decades within a few paragraphs. This manipulation of narrative time not only affects the rhythm of the story but also influences how the audience experiences character development, thematic progression, and emotional impact.

3.2. Evaluating Narrative Quality with LLMs

The integration of LLMs into automated narrative analysis has fundamentally reshaped how computational systems interpret and evaluate storytelling elements. Recent models like Openai GPT-4 [Ope24], or DeepseekV3 [DA25] have shown increasingly advanced capabilities in both narrative generation and evaluation. This progress has created growing interest in using LLMs to make judgments traditionally requiring literary or dramaturgical insight [CSC24, XCL23]. LLMs are deep learning architectures, typically based on the Transformer framework, designed to model the statistical properties of natural language through the unsupervised learning of massive text corpora. These models are trained to predict the probability distribution of the next token in a sequence, enabling them to generate coherent and contextually appropriate text across a wide range of domains. LLMs leverage self-attention mechanisms to capture complex dependencies and relationships between words, both within and across sentences. Their large parameter count, in the range of billions, allows them to store vast amounts of linguistic knowledge and world information implicitly within their weights.

The concept of that evaluates text is also known as the “LLMs-as-a-judge” paradigm. It

3. Fundamentals

showed that large language models such as GPT-4 [Ope24] can closely approximate human judgments on criteria including coherence, relevance, creativity, and factual accuracy. These models are typically prompted with evaluative instructions or trained on preference datasets to simulate human-like assessment. Despite their effectiveness, several limitations have been identified. LLM judges are prone to overvaluing surface-level features such as fluency, verbosity, or confidence, often rating AI-generated stories as more creative than they are—especially when compared to expert human assessments. Additionally, recent studies have shown that LLMs tend to penalize uncertainty in responses, preferring confident but potentially incorrect answers, a bias not shared by human judges. This sensitivity to phrasing and stylistic cues introduces a risk of misaligned or misleading evaluations. As a result, researchers stress the importance of carefully calibrating LLM-based evaluators and combining them with human oversight to ensure that evaluation metrics reflect true quality, rather than artifacts of language model behavior.

Optimizing the reasoning performance of LLMs encompasses several dimensions, one of the most impactful being prompt engineering. This approach involves the strategic formulation of task-specific instructions—known as prompts—to direct model outputs without modifying the model’s underlying parameters. By enabling model behavior to be shaped through carefully constructed input prompts, this technique allows for improved performance across a wide range of tasks and domains. In recent years, numerous effective prompt engineering strategies have emerged, including Chain-of-Thought prompting, automatic reflection and many more. These methods have become integral to contemporary LLM-based systems, significantly influencing frameworks such as RAG or agentic architectures. As such, prompt engineering is a foundational component in optimizing the reasoning capabilities and practical deployment of modern LLM utilizing systems.

Another key dimension in enhancing the reasoning performance of LLMs is fine-tuning. Unlike prompt engineering, which operates at the input level, fine-tuning involves modifying a model’s internal parameters by training it on additional data tailored to specific tasks, domains, or reasoning styles. Fine-tuning has proven especially valuable in scenarios where high precision or domain adaptation is required, such as legal, medical, or scientific applications. On the other hand, fine-tuning presents several challenges. It is often computationally expensive, requiring substantial time and high-quality curated data. Moreover, the process is technically complex, demanding expertise in model architecture, training dynamics, and evaluation metrics. Poorly executed fine-tuning can lead to issues such as overfitting, catastrophic forgetting, or unintended bias amplification. Frequent retraining may also be necessary to keep up with domain changes, further increasing the operational burden. Despite these limitations, fine-tuning remains a powerful and often necessary approach for an LLM optimization.

Paper [GXG⁺24] explains that the major methods for optimizing LLMs involve a combination of prompt engineering, fine-tuning, and external knowledge integration. Figure 3.1 provides an overview of these optimization strategies and illustrates how they can be understood in relation to one another, highlighting their complementary roles in enhancing LLM performance across various tasks and domains.

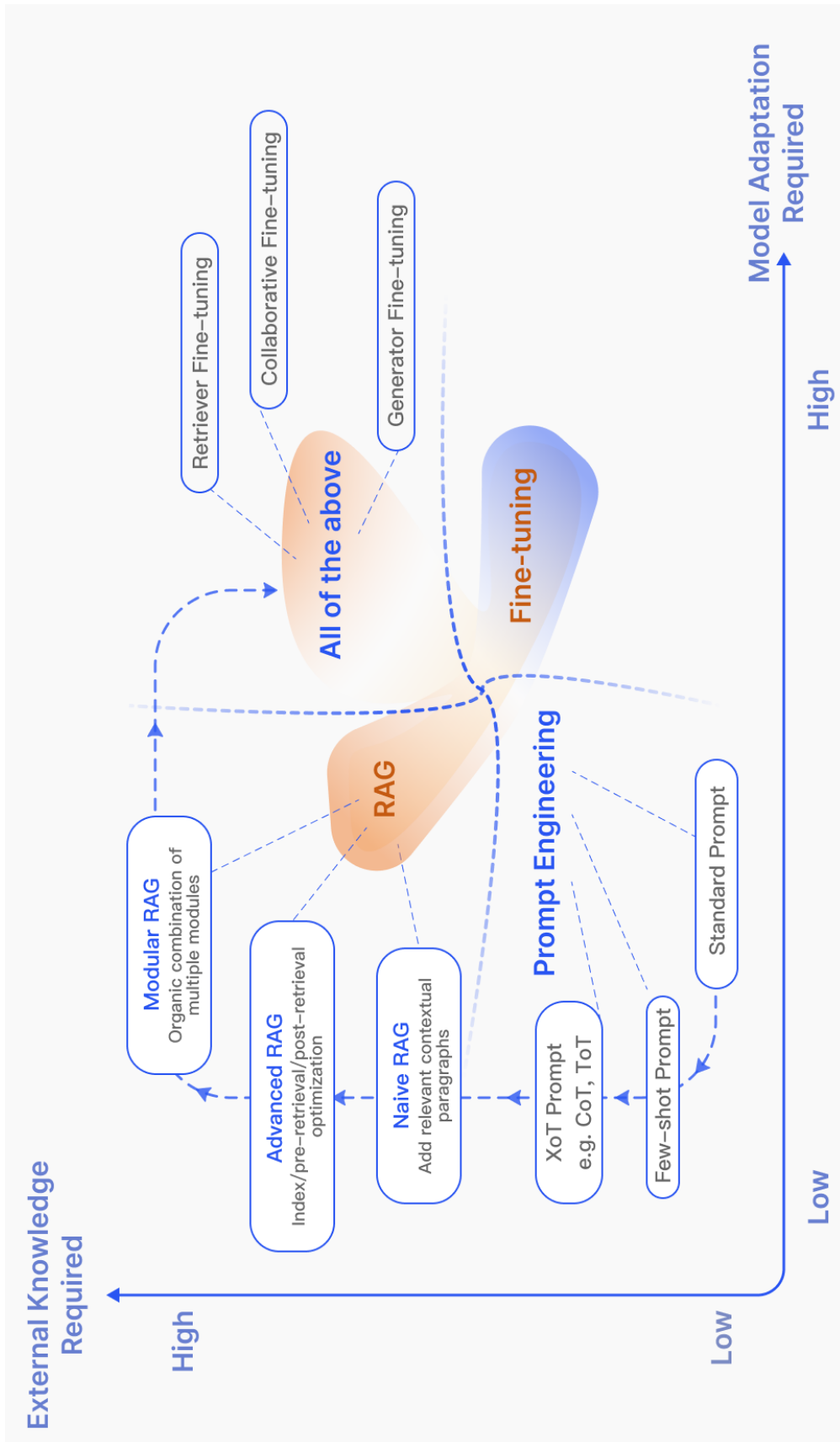


Figure 3.1.: Comparison of LLM optimization methods by external knowledge requirements and model adaptation, Reprinted from [GXG⁺24]

3.3. Retrieval-augmented Generation Systems

Modern LLMs such as GPT-4 [Ope24] and LLaMA [Met24] support context windows spanning tens of thousands of tokens. However, their effective usage of this extended context remains significantly shorter in practice. Empirical results indicate that LLMs typically utilize less than half of their nominal context capacity. For instance, on the RULER long-context benchmark [HSK⁺24], LLaMA 3.1-70B—despite its claimed 128K token window—effectively handles only about 64K tokens, falling below 50% utilization.

This discrepancy is largely attributed to pretraining data biases. Specifically, tokens that appear far from the current position are relatively rare during training, which limits the model’s ability to learn long-range dependencies. Consequently, as the input context increases, model performance degrades. In LongICLBench [LZD⁺24], open-source LLMs exhibit steep accuracy declines on tasks requiring very long demonstrations (20–50K tokens). A similar pattern is observed in RULER’s [HSK⁺24] synthetic tasks, where model accuracy drops sharply once the context size exceeds the model’s effective processing range.

RAG is a development in natural language processing aimed at addressing the limitations of LLMs concerning their fixed context windows. RAG systems overcome the limitation described above by integrating a retrieval component that dynamically sources relevant external information and provides it to the LLMs during inference [LPP⁺21, IG21]. The architecture typically combines a dense retriever—often based on dual encoder models—with a generative model such as a Transformer-based language model. During inference, the retriever component first encodes the input query and retrieves the most semantically relevant documents from a large corpus. These retrieved documents serve as dynamic external memory and are concatenated with the original query or used to condition the generation process. The generative model then processes this enriched input to produce coherent and contextually informed text. By decoupling knowledge retrieval from generation, RAG enables the model to leverage up-to-date external information, mitigate hallucinations, and improve factual grounding, offering a scalable solution for open-domain question answering and other knowledge-intensive natural language processing tasks.

Studies confirm that RAG systems significantly enhance performance on knowledge-intensive tasks. In biomedical question answering, for instance, methods like BriefContext, which uses a map-reduce-style retrieval mechanism, have outperformed standard prompting techniques [YRB⁺22]. In general-domain applications, RAG-based systems such as Atlas and REALM have matched or surpassed closed-book LLMs, mainly by reducing dependence on outdated training data and lowering the incidence of hallucinated outputs [GLT⁺20, ILL⁺22].

Nevertheless, RAG systems also introduce new challenges. Their effectiveness relies heavily on the quality of the retrieved content. Irrelevant, incomplete, or noisy documents can mislead the language model, resulting in degraded output quality. Additionally, pre-processing large documents into retrievable chunks presents trade-offs. Fixed-length chunking can disrupt coherence, while overly granular or variable chunking risks losing critical context [MS25]. Moreover, the RAG process remains constrained by the LLMs

input length. Only a limited number of documents can be included in the context window, requiring strategies such as query-focused chunking, iterative retrieval, or summarization to optimize context usage. These methods are more complex and slower because they require different systems—like retrieval engines, embedding databases, and ranking algorithms—to work together.[LPP⁺21].

3.4. Knowledge Graphs in Story Evaluation

Another line of research focused on the role of knowledge graph (KG)s in narrative analysis. KGs are structured representations of information where entities (nodes) are connected by relationships (edges), typically expressed as triples in the form of (subject, predicate, object). These graphs encode knowledge by capturing semantic relationships between concepts, enabling machines to interpret and reason over complex data. [ACJR12] examined the social network within *Alice in Wonderland* to demonstrate how character interaction networks can be used to assess narrative coherence. By applying social network metrics such as degree centrality and clustering, they identified central characters and analyzed the structure of relationships. Their findings showed that deviations from expected interaction patterns may indicate inconsistencies or structural flaws in the story. This work suggests that KGs, especially in the form of social network graphs (SNG)s, can function as a diagnostic tool for evaluating whether a narrative maintains coherent and logically consistent relational structures. SNGs model the relationships and interactions between characters in a story, focusing specifically on social dynamics such as alliances, conflicts, or communication. They provide a structural view of character networks, showing patterns of interaction and revealing the marginality of individual characters within the narrative. At the same time researchers highlighted the effectiveness of these Graphs as structured representations for assessing narrative coherence in both human-authored and machine-generated stories. KGs support the modeling of entities, events, and their interrelations over time, offering a semantic structure to monitor consistency and causal dependencies throughout a narrative. For example, [YHW⁺25] propose a dynamic KG-based memory mechanism that evaluates coherence by retrieving and comparing semantically similar past events, achieving a 23.6% improvement in coherence metrics compared to baseline language models.

To close this bridge between KGs and RAG systems, recent research has advanced the integration of structured and unstructured information within language models. other than the flat retrieval of RAG systems, GRAG incorporate structured data from graph-based knowledge representations. This shift enables models to reason over complex relational structures, beyond surface-level retrieval. In contexts like KGs and SNGs, where information is naturally organized as nodes (e.g., entities, individuals) and edges (e.g., relationships, interactions), this integration supports multi-hop inference, entity disambiguation, and semantically enriched contextualization. multi-hop inference allows reasoning over several connected nodes in a knowledge graph to answer complex queries. Instead of retrieving only directly related information, the system traverses multiple edges to collect facts from distant but relevant parts of the graph. This enables reasoning and supports

3. *Fundamentals*

answering questions that require a global understanding, such as tracking storylines across large documents. Entity disambiguation ensures that ambiguous terms—such as names or places—are correctly linked to the right entities in the graph, avoiding confusion. Semantically enriched contextualization then integrates structured relationships from the graph into the retrieved information, enriching the context for generation. This results in more coherent and precise outputs. By enabling LLMs to interact with graph-structured data, GRAG forms a hybrid architecture that merges the generative strengths of LLMs with the relational precision of graph analytics [HWS⁺25, PZL⁺24].

4. Assessing Story Coherence

In this section, novel methods are proposed that build on the research presented in the previous section. These methods are designed to evaluate the coherence of narrative extensions based on three key metrics: **causal**, **cultural**, and **thematic** coherence. Each approach incorporates the following foundational elements, as illustrated in Figure 4.1.

1. A pre-existing, coherent reference story
2. A story segment that requires validation.
3. A validation routine that compares the new segment to the ground truth.
4. An output format that provides a coherence score for each of the three metrics on a scale from 1 to 10 and a brief description.

This setup functions as an evaluation system for open-ended text generation. Its primary objective is to assess whether a newly created paragraph or chapter maintains narrative coherence when compared to the reference story, which serves as the source material storyline. These three metrics are chosen not only because of their importance in the field of narratology, but also because LLMs can understand and reason about them. This is important, as each validation method—especially the baseline method—relies on prompting LLMs. Since it has been observed that shorter input prompts lead to better results, the aim is to avoid explaining the concept of each coherence type in every prompt. Instead, brief descriptions such as “*Causal Coherence (cause-effect logic)*” should be sufficient within the prompt. While there are no formal studies that verify this statement, the following assumption is formulated: “*LLMs can understand and reason about causal,*

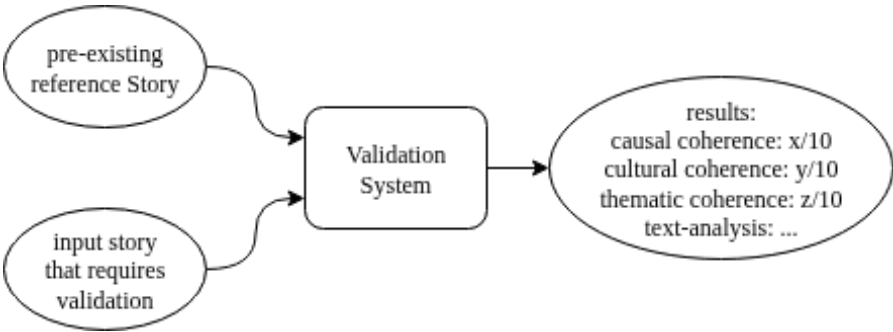


Figure 4.1.: Basic validation System

4. Assessing Story Coherence

cultural, and thematic coherence in a manner that enables them to perform the analytical tasks required.” Experiments (see 5.2) demonstrate that modern or state-of-the-art LLMs indeed understand these easy concepts and are capable of generating reasonable coherence scoring.

4.1. Baseline Validation

The baseline method was created to provide a point of comparison for developing other methods aligned with the findings in GPT-Score [FNJL23]. For its validation, a few-shot (FS) prompt was constructed, which included the task description, the source material book (SMB), and the text to validate (TTV), see Figure 4.2. FS prompting refers to giving a language model a small number of examples of how a task should be performed within the prompt. This approach is necessary because computational prompting often requires the model to produce outputs in exact data types or predefined JSON structures. In zero-shot (ZS) prompting scenarios—where no examples are provided—this frequently leads to inconsistent or malformed outputs. To address this issue, the system incorporates a retry mechanism for all prompting attempts. This safeguard is important even when employing few-shot prompting, as structural inconsistencies can still arise despite the inclusion of example outputs. To keep the setup as simple as possible, no summarization was applied. This led to extremely large prompts, nearly exhausting the maximum input token window of the DeepSeek LLMs that are used for the experiments later on. As a result, this constraint limited the later experiments, since the maximum input size is lower than the length of the SMB and the TTV combined, even though later methods could potentially handle larger texts. Given the context size limitations and the efficient context usage problem discussed in the introduction, the results from the baseline are expected to be imperfect.

One major constraint of this method is that incoherent source material—rather than incoherent validation text—may also be validated, meaning that the resulting coherence score might be influenced by the quality of the source material. This occurs because, unlike in the other methods, there is no strict separation between the data components. As a result, this factor must be taken into account when interpreting the test results.

4.2. GRAG Simple Validation

The first and simplest evaluation method demonstrates how effectively a modern GRAG framework can conduct coherence analysis independently. The GRAG framework used in this study is capable of answering complex questions and executing instructions based on the provided SMB. The validation process consists of the following design in Figure 4.3

Indexing the RAG Graph with the SMB

In the beginning, the indexing routine of the GRAG library used is triggered with the SMB [ETC⁺25]. In the indexing process, documents are first chunked and embedded

System prompt:

You are a precise and insightful literary evaluator. You assess how well a derived story (Story B) aligns with a source text (Book A) across three dimensions:

Causal Coherence (cause-effect logic)

Thematic Coherence (main ideas/themes)

Cultural Coherence (e.g., redemption, success, personal growth).

For each dimension, assign a score from 0 (incoherent) to 10 (fully coherent), and provide a brief but thoughtful analysis.

Always return your evaluation in the following JSON format:

```
{
  "causal_coherence": {
    "score": X,
    "analysis": "...",
  },
  "thematic_coherence": {
    "score": Y,
    "analysis": "...",
  },
  "cultural_coherence": {
    "score": Z,
    "analysis": "...",
  }
}
```

Book A:

...

Story B:

...

Figure 4.2.: An example ZS-prompt that instruct the evaluation of coherence

4. Assessing Story Coherence

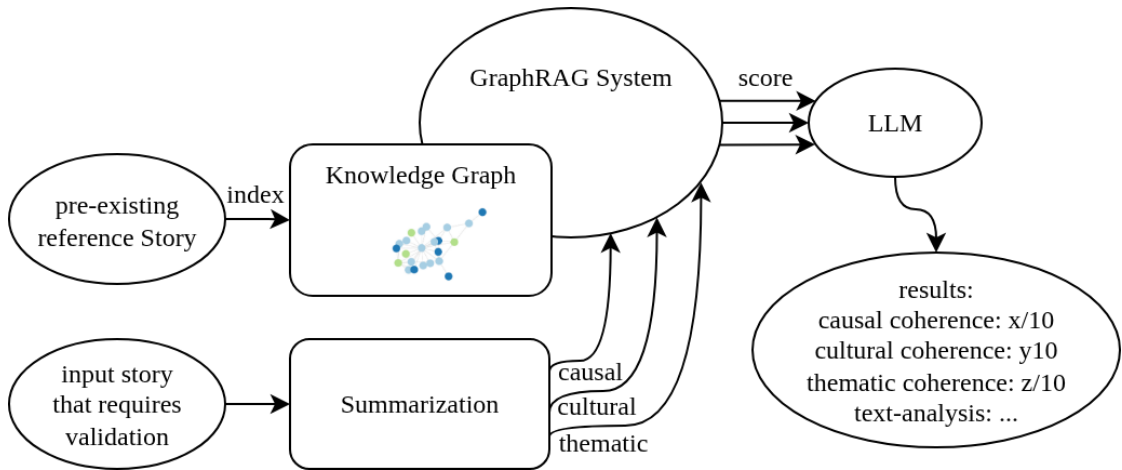


Figure 4.3.: RAG analysis method Design

using LLMs. Then, GRAG constructs a knowledge graph where each node represents a document chunk and edges encode semantic or logical relationships (e.g., citations, co-occurrence, or topical similarity). This graph is also enriched with metadata and relations to provide more information. During queries, relevant nodes and their neighborhoods in the graph are retrieved, allowing LLMs to generate more grounded and connected answers. The standard pipeline includes the following steps:

1. Extracting entities, relationships, and claims from raw text.
2. Performing community detection on extracted entities.
3. Generating summaries and reports at multiple levels of granularity.
4. Embedding entities into a graph-based vector space.
5. Embedding text chunks into a textual vector space.

The output of this pipeline is saved as Parquet tables by default [D⁺13], and embeddings are stored in the configured vector database. This data can be visualized shown in Figure 4.4.

The data visualizes the KG of the source material, with nodes and edges representing key narrative elements and their relationships. Green nodes indicate entities of type geographic location, red nodes represent events, dark green nodes denote undefined entities—used for important elements that cannot be clearly assigned to an abstract type—light blue nodes correspond to organizations, and dark blue nodes represent persons. The edges reflect textual events that associate the connected entities within the narrative. The displayed graph depicts a sub-portion of the SNG from the story *The Slave Planet*, highlighting the main character, Dodd, positioned at the center of the graph with numerous relational edges connecting him to various other entities. This graph was created with the GRAG visualization library [now25].

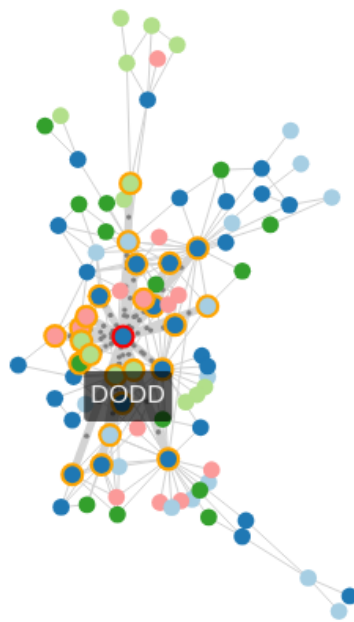


Figure 4.4.: RAG graph of the book “The Slave Planet”

4. Assessing Story Coherence

```
User prompt:

compare the provided story below with the source material and
evaluate the thematic coherence context related to the source
material. Thematic coherence are consistent themes or motifs
emerge across the story (e.g., overcoming adversity).

Provide the context and provide where the provided story does not fit
in the thematic themes and motifs.

The results must be in form of a pro and contra analysis
provided story:
...
```

Figure 4.5.: An example prompt that instruct the evaluation of the thematic coherence

Summarizing the TTV

To improve response quality and reduce prompt size, the validation text is summarized. A domain-specific summarization prompt is used to preserve coherence and retain relevant aspects of the plot. While studies suggest that information loss from summarization is generally minimal, it has not been explicitly measured or proven in the context of plot coherence. Based on that, the risk of partial information loss is acknowledged.

Creating GRAG prompts for each coherence metric

The GRAG interface is used similarly to a traditional LLM chatbot. Prior experiments have shown that the GRAG system struggles to assign coherence scores directly and to return consistent numerical results. Therefore, in this step, the framework is instructed to analyze and return only a pro/con list for each metric.

Each coherence metric is queried separately to reduce prompt complexity. This approach is based on the assumption that, due to its structured knowledge of entities, communities, events, and linked text snippets, the GRAG system should be capable of understanding thematic, cultural, and causal relationships represented in the graph. An example of such a prompt can be seen in Figure 4.5. After analyzing each coherence metric using the GRAG system, a FS prompt is created to transform the GRAG generated pro- and contralist into coherence scores.

4.3. GRAG Context-Aware Validation

The next validation method, called Context-Aware Validation, uses the GRAG system as a context retrieval system. The method requests relevant contextual material from a KG,

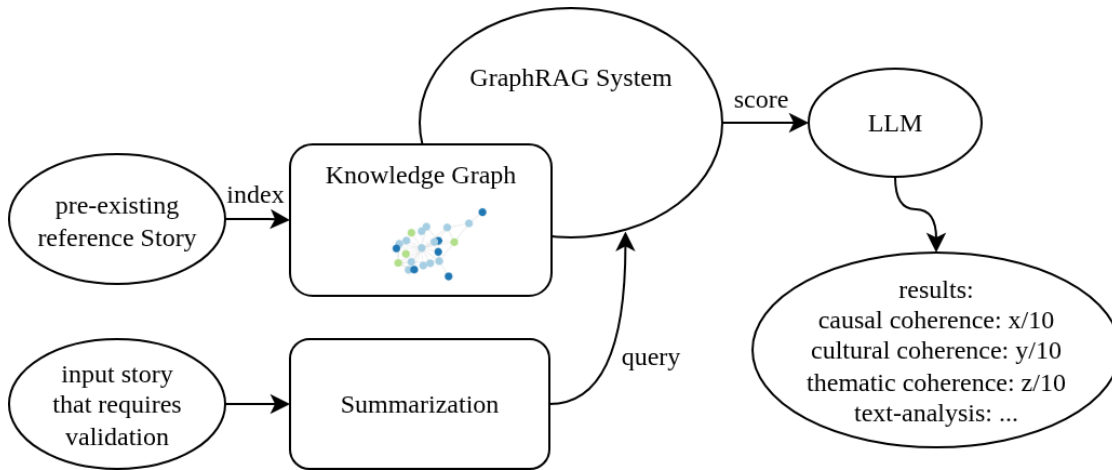


Figure 4.6.: RAG graph context extraction method design

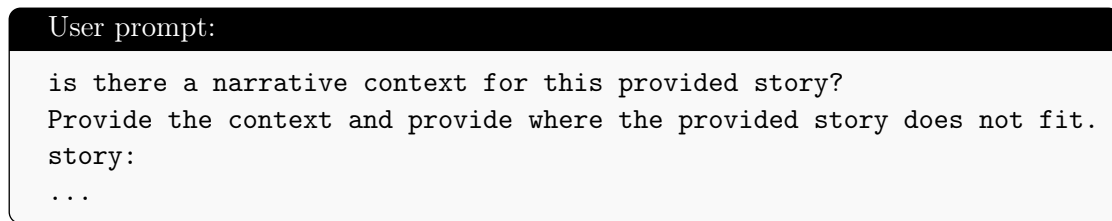


Figure 4.7.: An example query prompt for the GRAG to retrieve context

which is then used as curated input for a LLM that carry out the analysis. The LLM scores the context in comparison to the text being validated. This method operates on the following assumptions:

1. *“If a text A is coherent in comparison to SMB B , then B must have a certain degree of contextual relevance to the attributes mentioned in A .“*
2. *“The context provided by SMB B holds the information necessary to make an educated validation of text A .“*

The GRAG query prompt, see Figure 4.7 is designed as a straightforward instruction that requests both the contextual placement of story A within the narrative framework of story B , and an analysis identifying areas where the Graphrag system detects no narrative overlap between the two stories.

Local and global search parameter

The instruction against the GRAG library can be performed via two different query types. Both are promising for the use case, as coherence is both entity-based and global at the

4. Assessing Story Coherence

same time. Therefore, the tests were executed with both query versions. These query types are the local search and the global search. Each method is suited to different types of user questions.

The local search is optimized for precise, entity-focused inquiries. It operates by first identifying the key entities contained within the user’s question. Following this, it navigates the KG in order to collect relevant information, including associated entities, their relationships, as well as pertinent text extracted from the source documents. This method proves particularly effective for questions that demand detailed knowledge regarding specific entities, such as: “In which university did Victor Frankenstein learned about...”

In contrast, the global search is intended for broad, theme-oriented questions that require an overarching understanding of the complete dataset. It makes use of precomputed summaries of graph communities and applies a map-reduce strategy to generate comprehensive answers. This approach is suited for queries such as: “What is the overall theme of the source material?”. Although more demanding in terms of computational resources, the global search yields a perspective by integrating information across different communities.

Coherence validation

The actual validation process involves evaluating the generated text A against the SMB B , which has been adapted to the context of A in B through the GRAG system. This validation is carried out using an instruction format closely aligned with the baseline model. The key distinction lies in the way the prompt is constructed: Instead of presenting both full narratives in their entirety, the validation is focused exclusively on plot-relevant content from the generated story. Irrelevant segments are filtered out in advance by the GRAG system.

This approach introduces certain challenges with respect to assessing causal coherence. Specifically, if essential plot elements are introduced early within the generated TTV A , but do not exist in the SMB B — excluding duplicated descriptions — this results in incoherence. Consequently, a story can only achieve the highest coherence score if *all* of its contextual foundation is present in the SMB, and *none* of that foundational context is introduced for the first time within the generated text itself.

This poses a problem because in practical narrative structures, chapters typically exhibit strong local coherence: Much of the plot is established and developed within the same chapter. Chapters tend to have a self-contained narrative arc—consisting of a beginning, middle, and end—rather than functioning as mere segments of a continuous story. To address this limitation, one proposed direction for future research is to leverage existing small-scale text validation models to evaluate the *locality* of text A . The outcomes of such local validation could then be integrated into the broader context analysis, offering a more accurate assessment of coherence. This strategy may serve as a valuable foundation for improving validation methods in ULTG.

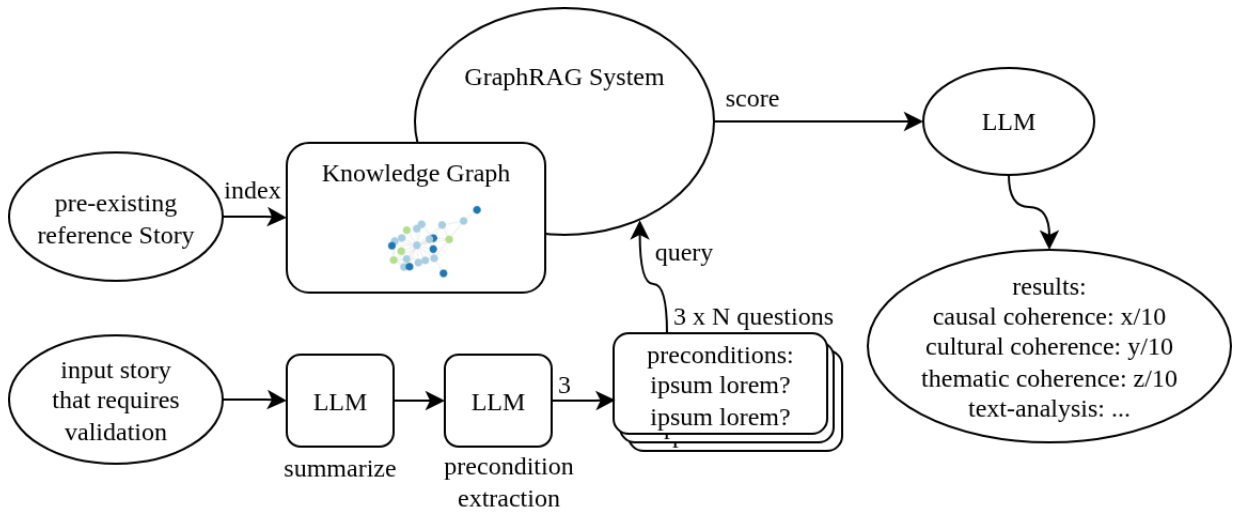


Figure 4.8.: RAG graph complex question method design

Parameter considerations

Another important aspect to consider are the parameters influencing the summarization process. In addition to the local and global context distinctions, the length of the summarization itself plays a role. Summarization inherently involves information loss; therefore, longer or more detailed summaries are expected to retain more context and potentially lead to improved outcomes in downstream tasks.

However, it has been observed that the reasoning capabilities of LLMs tend to degrade when processing longer instruction prompts. Since the GRAG framework is entirely LLM-driven, this limitation equally affects the summarization and instruction steps within its pipeline. As a result, determining the optimal length of a summary becomes non-trivial and should ideally be informed by empirical performance evaluations. In preliminary, sample-based investigations, it was found that summaries of approximately 300 words strike a practical balance: they preserve a substantial amount of context while avoiding the pitfalls of excessive prompt length. This observation suggests a potential benchmark for future experiments in fine tuning the summarization size.

4.4. GRAG Complex Question Validation

The next validation process is grounded in the premise that for a narrative to be coherent and meaningful, certain factual elements in the SMB—referred to as *preconditions*— must be present/true. Initially, the system is initialized and indexed with the SMB to construct the RAG graph. Following this, the story intended for validation is summarized. From this summary, the system extracts a set of precondition questions, which are essential to assess whether the story aligns with the established SMB. These questions are designed to probe various dimensions of coherence, such as thematic consistency, causal relationships,

4. Assessing Story Coherence

and character motivations. For instance, a question addressing thematic coherence might be:

Is the theme of psychological tension and moral ambiguity present in the SMB?

The system automatically generates a set number N of such questions for each category of coherence. These questions are then submitted to the GraphRAG system, which evaluates whether each precondition can be affirmed based on the indexed source content.

A validation score is computed based on the proportion of questions that received (i.e., “YES”) responses, reflecting the degree of alignment between the target narrative and the SMB. Finally, the outcome of the analysis is presented as a combination of the precondition questions and their corresponding answers, offering an interpretable explanation of the validation results.

Precondition extraction and prompt considerations

Precondition questions are generated using a separate LLM instruction, with a specific prompt tailored to each coherence type, see Figure 4.9. This prompt is designed to create a list of questions in JSON format. The story extension, referred to as Appendix A, is inserted at runtime during the evaluation process. The extraction prompt used to generate precondition questions consists of several carefully designed instructions intended to mitigate issues in this approach:

“Each question should be phrased so that an expert, if the extension is coherent, would answer ‘YES’ ”. This ensures that the questions are aligned with a positive boolean orientation. “YES” answers indicate coherence between the story and the SMB.

“The questions must be answerable based only on the summary and the beginning of the story, without needing to know the full text.”. This constraint guarantees that the questions rely solely on the indexed SMB and do not reference elements introduced exclusively in the story being validated.

“Explicitly name all characters and entities in every question; do not use pronouns.”. This guideline addresses a common failure mode in LLMs, where vague or ambiguous references can make questions unanswerable unless the full narrative is known. Naming entities explicitly improves grounding and ensures that the questions remain resolvable within the indexed content.

Finally *“Questions must be very concise but retain all essential information. Remove any words that do not carry direct informational value.”* ensures that the questions are sufficiently informative yet compact, optimizing their suitability for vector-based retrieval in the RAG system. Concise questions with focused context improve the chances that relevant information will be retrieved, thereby increasing the accuracy of the coherence validation. Additionally, the output format of the GRAG framework can be modified through textual instructions. As previously mentioned, the framework is fully LLM-driven, and its output format is specified by appending a formatting directive to the end of the GRAG prompt that generates the results. Unlike standard LLM prompting, where the output format is typically embedded within the input instruction, GRAG separates the

```
System prompt:
you are a literary expert. Given the following summary of a story
extension, generate 3 questions that can be answered with YES
by another experts source material. These questions should move
beyond surface-level details and deeply explore
the Appendix:

<appendixA>

your instructions:

The questions must be answerable based only on the summary and the
beginning of the story, without needing to know the full text.

Each question should be phrased so that an expert, if the extension
is coherent, would answer "YES."

Explicitly name all characters and entities in every question;
do not use pronouns.

Questions must be very concise but retain all essential information.
Remove any words that do not carry direct informational value.

Only generate questions; do not answer them.

Return the output as a JSON list of strings like this
{
  "result": [ "Question 1", "Question 2", "Question 3" ]
}
```

Figure 4.9.: System prompt to extract coherence precondition questions

4. Assessing Story Coherence

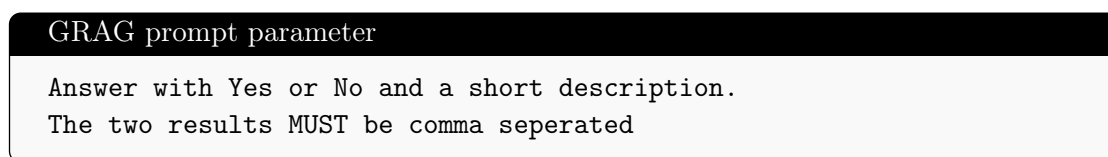


Figure 4.10.: GRAG output instruction prompt

input (i.e., the content to be analyzed) from the output instruction during its execution process. This distinction makes it necessary to explicitly manipulate the output format to ensure that the results are machine-readable and parsable. To achieve this, the output format instruction was modified in Figure 4.10.

First person problematic

A major challenge in this context arises from the use of first-person writing styles in the SMB and/or the TTV. This issue manifests at two levels.

The first level concerns the construction of the KG. While the graph can reliably identify relationships involving the “narrator” , it struggles when the text includes letters, journal entries, or other embedded perspectives that shift away from the narrator’s viewpoint. These perspective changes introduce ambiguity into entity and relationship extraction.

The second level involves the generation of precondition questions. First-person narration complicates this process because, without clear referential links between entities in the text and those in the KG, it becomes impossible to formulate answerable precondition questions—even when the TTV is coherent. For example, a question such as “Is the behavior of the characters consistent with their established ... ?” is highly ineffective, as neither the language model nor the knowledge graph can reliably infer which characters are being referenced.

To address this issue, the SMB is translated into third-person narration by substituting first-person references with the narrator’s name or equivalent identifiers. This approach significantly improves both the clarity of the KG and the quality of generated coherence questions. However, the translation process used in this study is naive and does not account for more complex perspective shifts, which may in turn introduce new inconsistencies.

Parameter considerations

This approach also involves several parameter considerations, including summarization length, as well as the use of global versus local search modes described in earlier sections of this work. One important parameter is the number of precondition questions generated per coherence metric. Ideally, this value would be dynamic, adapting to the semantic density and narrative structure of the text—some passages may contain ten relevant thematic preconditions, while others may contain none.

However, setting this as a fixed parameter presents challenges. Observations revealed that the LLMs struggles with generating either zero or a large number (more than five)

4.5. GRAG Relationship Score Validation

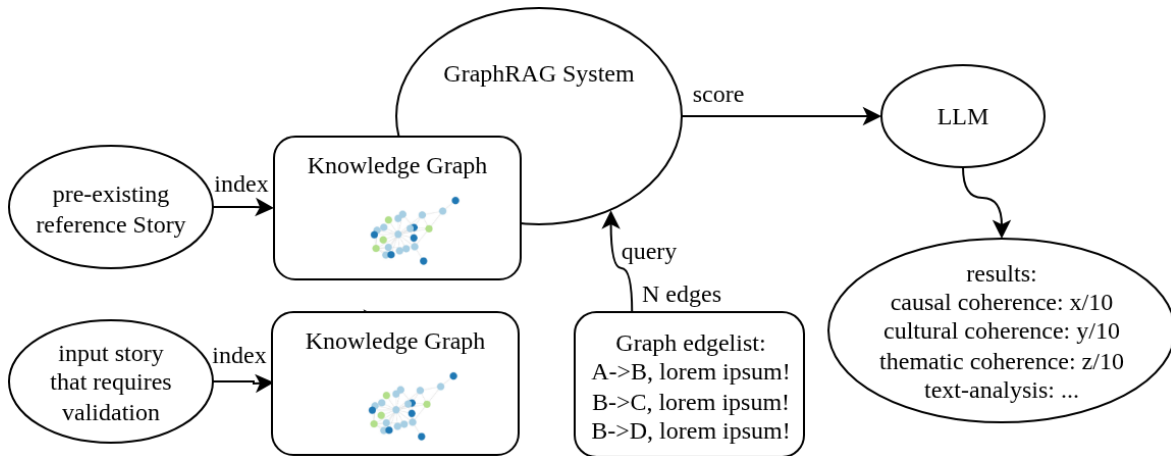


Figure 4.11.: RAG graph comparison method design

of questions. Generating no questions often violates the expectations set by the prompt, while producing too many questions exceeds what the model treats as a “normal” or expected response. In an ideal scenario, the LLMs would be able to infer both the relevance and the appropriate number of questions autonomously.

Furthermore, experiments demonstrated that when instructed to generate many questions, the LLMs tends to produce similar, redundant questions, reducing their overall effectiveness. To address this and to control evaluation costs, a fixed number of three questions per coherence metric was selected as the default.

4.5. GRAG Relationship Score Validation

This validation approach aims to combine the strengths of contextual retrieval from RAG systems with the structured reasoning capabilities of KG. Unlike earlier methods that only index the SMB, this method constructs a KG not only for the SMB but also for the text to be validated.

In this framework, see Figure 4.11, each relationship (i.e., edge) in the KG derived from the story to be validated is individually evaluated against the SMB’s KG. A Edge consist of a source and target, and also a statement. For each edge, the system checks whether relevant contextual support exists and whether the statement encoded by the edge is coherent with the SMB. This evaluation is performed with respect to three coherence metrics. The GRAG system conducts this analysis based on the SMB. Finally, an LLM is employed to score each edge according to the three coherence metrics. These individual scores are aggregated to produce a total coherence score, which is calculated as the average of all edge-level scores.

Knowledge Graph

First, the validation method indexes the TTV using the graphrag library, which generates a KG of the TTV. Through this graph indexing process, the KG is constructed so that all findings related to the story are now encapsulated within the graph. The SNG presented in Figure 4.12 of the TTV within the graph comparison validation method illustrates a smaller Knowledge Graph derived from the data sample, centered around the main character Raffles. In total, the validated TTV reveals 18 relationships that can be confirmed through this validation process.

One particularly interesting feature of the knowledge graph is the isolated light green node in the top left of the graph shown in Figure 4.12. If the TTV aligns with the story and the entities such as “ruffles” and the companions are introduced, coherence can be detected and result in a high score. However, for this isolated node, which has no connecting edges, no validation takes place as this validation method focuses on the edges only. For such nodes, an isolated validation approach could potentially improve the results by checking whether the entity fits the source material independently of the graph structure.

Furthermore, this isolated node may have been introduced in the TTV but not yet incorporated into the main story. It might represent a parallel or disconnected storyline that remains unconnected in the current graph but could be linked in future text samples beyond the scope of the current TTV. In such cases, other methods might assign a low coherence score due to the TTVs incomplete perspective on this aspect of the narrative. In contrast, the graph comparison system simply ignores the node, as it lacks edges and therefore does not influence the overall graph-based coherence analysis.

This behavior has a conceptually significant impact on the validation results. Negative samples or highly perturbed samples that contain multiple isolated, incoherent nodes would be ignored by the graph comparison method, as these nodes lack edges and thus do not contribute to the calculated coherence. As a result, they are not reflected as negative coherence scores. In contrast, other validation methods—such as the context-aware validation approach—would treat these nodes (or, more precisely, the corresponding text segments) as insufficient context and would consequently assign a low coherence score. This difference in handling leads to generally lower scores in the context-aware validation method, while the graph comparison method tends to score higher, particularly for negative or perturbed samples.

Edge extraction

Once the TTV is indexed by the GRAG system, an edge list is generated that captures all relationships among the introduced characters, communities, and events. In the validation method described above, our focus lies on this edge list, as it effectively represents the narrative structure derived from the text. Each edge in the list encapsulates information about the connected entities and includes a description or statement that is subject to validation against the SMB. For example, one such statement might be:

“Bunny is involved in various robberies with Raffles, though often in a secondary

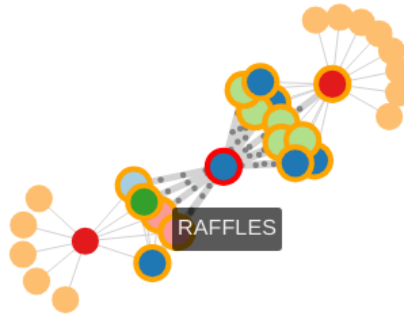


Figure 4.12.: The GRAG result of a TTV in the graph comparison validation method

role”

This statement then serves as the input for the source-material GRAG system’s instruction module, which generates an analysis based on the three dimensions of coherence. The instruction prompt, as shown in Figure 4.13, includes both the instruction and the definitions of the different coherence types.

It is important to note that the GRAG system performs the analysis on all three coherence dimensions simultaneously. This design choice opens a potential avenue for future research: investigating whether conducting individual analyses for each coherence metric yields more accurate or insightful results. This work adopts an all-in-one analysis strategy for the sake of simplicity. Experimental results have demonstrated that the outcomes for each coherence type, even when using a single-instruction query, are both detailed and well-articulated. This supports the decision to use the combined method as the default approach. However, this choice leads to increased computational costs—specifically, a threefold increase—since each edge requires three separate queries. A sample analysis output of the statement above is presented in Figure 4.14.

The output reveals a detailed analysis of the statement in comparison with the SMB, providing impressive insights along with the actual linked data, which could be leveraged for further analysis. Finally, this result is scored using an additional LLM instruction.

4. Assessing Story Coherence

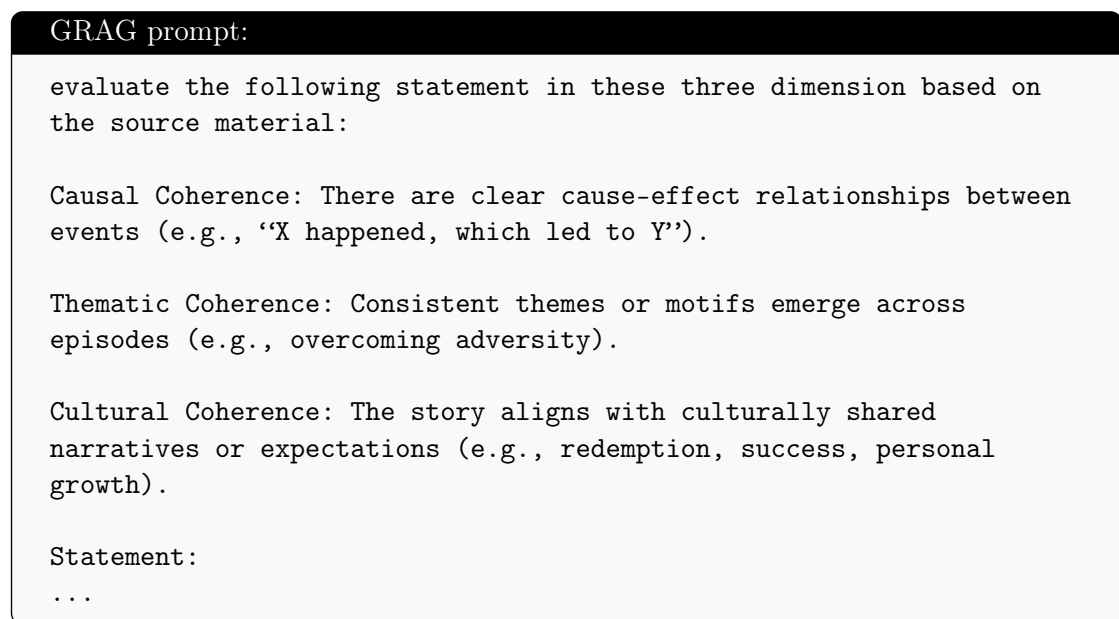


Figure 4.13.: GRAG edge evaluation prompt

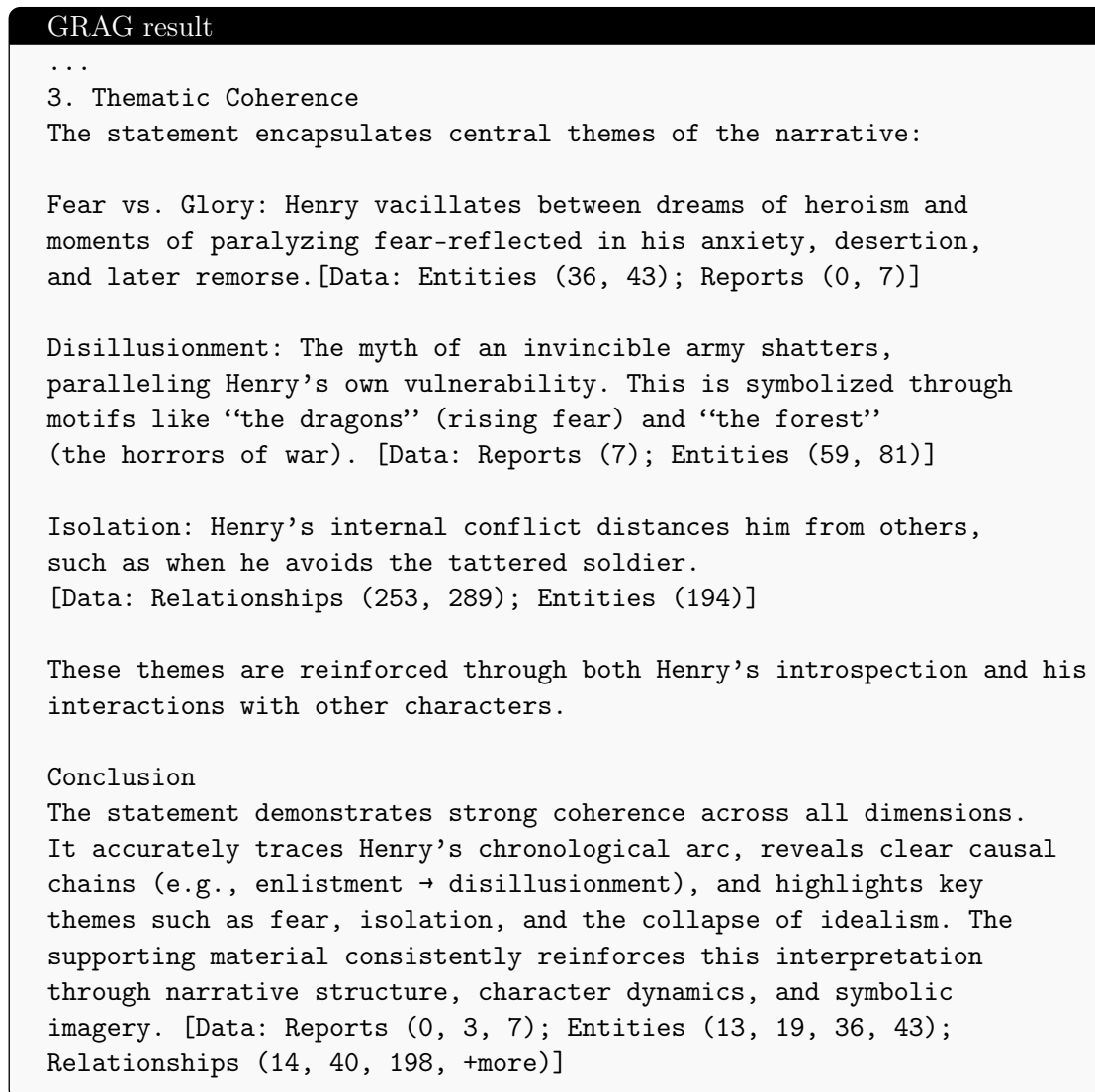


Figure 4.14.: GRAG edge statement analysis result

5. Evaluation

To evaluate the performance and conduct a detailed analysis of the four validation methods in comparison to the baseline model, a series of experiments was designed. These experiments aim to demonstrate overall performance and to provide insights through selected case studies. As discussed in previous sections, there is no available baseline dataset with gold-standard annotations that satisfies the requirements of our experiments. In particular, existing datasets lack sufficient token lengths (10,000+ tokens) and therefore do not support validation on ULNT.

As a result, evaluation strategies were developed that do not depend on gold-standard data but still produce meaningful and reliable insights. The following sections details the dataset preparation process, the metrics used to evaluate the results, the parameters influencing performance, and a case study that illustrates specific evaluations and observations.

5.1. Dataset Preparation

The experiments were conducted using texts from Project Gutenberg and a curated subset of English-language books available via the `sedthh/gutenberg_english` [sed21] project. For the purposes of this study, 120 fictional, plot-driven books were selected from the dataset. Non-narrative texts such as lexicons or encyclopedic entries, which lack coherent storylines, were excluded from the experiments.

The selected books range in length from 60,000 to 100,000 tokens and represent a variety of themes and genres, as illustrated in Figure 5.1. Notably, the dataset also includes books with varying degrees of narrative coherence—for instance, “Tales of Daring and Danger” [Hen90]—which provides valuable insights, particularly when evaluating the performance of validation methods on texts with less structured plots.

As noted in the Precondition Extraction subsection, the books included in the dataset are not limited to third-person narratives, nor are they curated based on writing style. The dataset contains a diverse range of narrative perspectives, though the exact distribution is unknown. This has direct implications for the evaluation of the Question Validation method, where the presence of first-person narratives introduces particular challenges. These challenges must be taken into account when interpreting the results.

Moreover, first-person narration also affects the performance of all the GRAG KG based methods. In such narratives, the narrator is often implicitly referenced using pronouns or generic terms, and the connection between the narrator and their actual name is not always explicitly stated or easily inferred. This can lead to incorrect or incomplete entity resolution, affecting the accuracy of the results. Consequently, stories written in

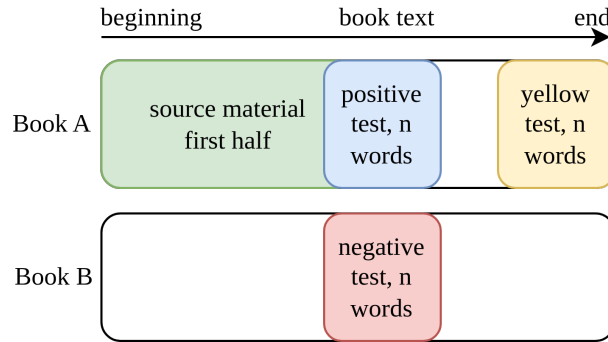


Figure 5.2.: Experiment setup and chapter comparison validation

Finally, a validation scenario is included in which the target text is not from the same book at all. In this case, the coherence score is expected to be significantly lower, reflecting the absence of shared narrative context. It is important to note that some degree of inference between texts may still be possible, even in negative samples. For this purpose, the negative sample in this setup uses an excerpt from *The Willows by Algernon Blackwood*. This particular text was chosen due to its relatively unique genre and theme—supernatural horror—which sets it apart from the majority of texts in the dataset. The expectation is that such content would have minimal, if any, overlap with the SMB used in validation, thereby serving as a strong negative example.

However, given the diversity of the 120 selected stories, it cannot be categorically assumed that no thematic or cultural overlap exists. Since *The Willows* is written from a Central European narrative perspective, some implicit cultural or thematic connections with other texts in the corpus may arise. Another noteworthy aspect of this setup is that each story segment is inherently coherent in itself. This means that the passages selected for validation are locally coherent, as they are drawn from professionally authored and widely recognized literary works. As such, even isolated text segments demonstrate a baseline level of readability and internal consistency.

This observation highlights an important distinction: a narrative can be locally incoherent while remaining globally consistent. In this case, the local coherence of the text should not be misinterpreted as an indicator of narrative continuity or global coherence. Therefore, this dimension of coherence is not the target of validation and should ideally be disregarded by the validation methods. Local coherence is an addressed problem with good results and could serve as an additional metric to enrich the interpretation of results. However, in this experimental setup, local coherence is not explicitly measured or evaluated.

Perturbed text experiment setup

The second validation setup is inspired by the concept introduced in the Deltascore framework [XLCL23], which proposes a validation mechanism capable of quantifying the extent to which a text B deviates or is perturbed from an original source text A

5. Evaluation

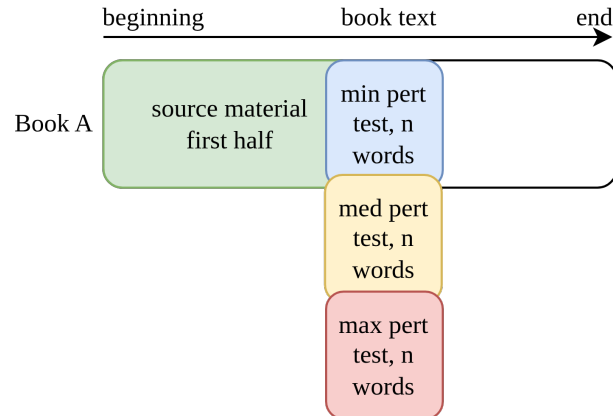


Figure 5.3.: Experiment setup and perturbed chapter validation

In this context, a similar perspective is adopted by evaluating how varying levels of perturbation affect the coherence score as determined by the validation methods. The underlying assumption is that a minimally perturbed version of the original text should receive a lower coherence score than the unaltered text, while a maximally perturbed version should result in an even lower score. This establishes a gradient from high to low coherence, aligned with the degree of textual perturbation.

This experimental setup aims to evaluate whether the validation methods are capable of reliably detecting this pattern—namely, whether they can differentiate between various levels of perturbation and reflect that difference in a consistent downward trend in coherence scores.

Additionally, in consideration of local coherence, the perturbation experiment is designed to reveal whether the validation method evaluates coherence in relation to the SMB or whether it is influenced by local incoherence within the perturbed text. Since the perturbations applied in this setup not only change the connection to the source but also disrupt the internal structure of the text, the results can indicate what type of coherence the method primarily responds to. If the validation method reflects a drop in coherence mainly due to changes from the source, it suggests that the method is sensitive to source-based coherence. However, if the coherence score is also significantly affected by local inconsistencies introduced through perturbation, this may indicate that the method is influenced by local coherence as well.

Perturbed text generation

The perturbations of original text segments are generated using a ZS prompting approach against a LLM. For each perturbation level—minimal, medium, and maximal—a corresponding prompt is formulated to transform the original text accordingly. The resulting output serves as the new text to be validated.

Each prompt is designed to produce a general perturbation that affects all aspects

```

User prompt:
Revise the following story and changing some of the following aspects:

Causal Coherence (break natural cause-effect flow)

Thematic Coherence (introduce clashing or trivial themes)

Cultural Coherence (deviate from familiar narrative expectations).

RULES: Just response the perturbed text, nothing else.
Dont write any other response other than the changed story.

The revision should be <severance>. “story”:

```

Figure 5.4.: A dynamic prompt to create PTPCs

of coherence simultaneously. That is, specialized prompts tailored to target specific coherence dimensions (e.g., causal, cultural, or thematic coherence) are not used. Instead, the same prompt is applied for simplicity reasons, but still allowing the perturbation to influence the text across multiple coherence aspects, see Figure 5.4.

Since the perturbations are generated through the reasoning capabilities of LLMs, there is inherent variability in the quality and nature of the resulting outputs. Notably, there is no validation step to ensure that a “minimal” perturbation is, in fact, minimal in terms of its narrative impact. This introduces a degree of uncertainty into the experiment.

A clear example of this issue arises with character name changes. Depending on a character’s role and relevance within the broader context of the story, altering their name could constitute either a minimal or a substantial perturbation. This underscores a central challenge: evaluating the coherence of perturbed texts may itself require access to global story context—precisely the type of reasoning that the validation methods under evaluation are intended to perform. As a result, the perturbation outcomes can vary unpredictably, depending on how the LLM interprets the original content. Nevertheless, across the full experiment with 120 books, the overall trend remains informative. Minimal perturbation prompts are more likely to produce small textual changes, while maximal prompts generally lead to more substantial alterations. Thus, even with variation in individual cases, the aggregate results still offer meaningful insight into whether the validation methods can detect and reflect a gradient of coherence degradation corresponding to increasing levels of perturbation.

5.2. Validation Results

In this section, the validation results of this work are presented. First, the overall mean scores of the different validation methods across both experimental setups — the chapter shuffle and the perturbed setup — are examined. The results are analyzed in comparison with one another, and correlation coefficients are calculated and a Wilcoxon test is performed to demonstrate that the outcomes are not due to random variation. Subsequently, false positive samples — specifically, stories without a consistent plot line, such as *Tales of Daring and Danger* — are investigated to assess how many of these incoherent stories were correctly identified by the various validation methods.

Validation results of chapter comparison

The following heatmap shows the average measured coherence scores across all 120 books for each validation method and coherence type. It includes positive, yellow, perturbed, and negative text samples. The heatmap visualizes the mean validation scores across all evaluated chapters, categorized by coherence type (causal, thematic, cultural) and validation method (baseline, local context-aware, global context-aware, complex question, RAG-Analysis, graph comparison). Each row represents a specific type of coherence, further split into subsets: full samples (e.g., `causal_coherence`), yellow samples (`*_y`), perturbed samples (`*_p`), and negative samples (`*_n`). The columns correspond to the six different validation methods. 5.5

According to the assumption stated in the experimental setup, where the coherence score is expected to be lowest for negative samples, slightly higher for perturbed samples, medium for yellow samples, and highest for positive samples, a clear trend is visible: For the validation methods local, global, complexQ, and graphComp, the coherence scores increase linearly from negative to positive samples. This result aligns with the theoretical expectation.

Since the perturbed samples come from the same book as the reference texts, it is expected that they show higher coherence scores than the negative samples, despite being partially disrupted. Most validation methods correctly reflect this behavior. However, the simple RAG analysis and the baseline method do not. The reason is that these methods do not strictly separate the validation text from the SMB. As a result, they measure both global coherence (i.e., within the dataset) and local coherence (within the individual sample). Because the perturbed samples are locally disrupted, they appear less coherent in these methods than the negative samples, which, although thematically unrelated, remain structurally intact on a local level.

Finally it can also be observed that the negative samples are most accurately assessed by the context-aware validation methods. Since the negative samples originate from a different book and should have no coherence relation to the SMB, the expected coherence scores for these samples are low. While there remains a minimal chance that certain aspects of the negative samples may coincidentally align with content in the SMB, this likelihood is negligible due to the selection of a highly specific genre for the negative samples. Consequently, coherence values close to zero are expected, with only a minor

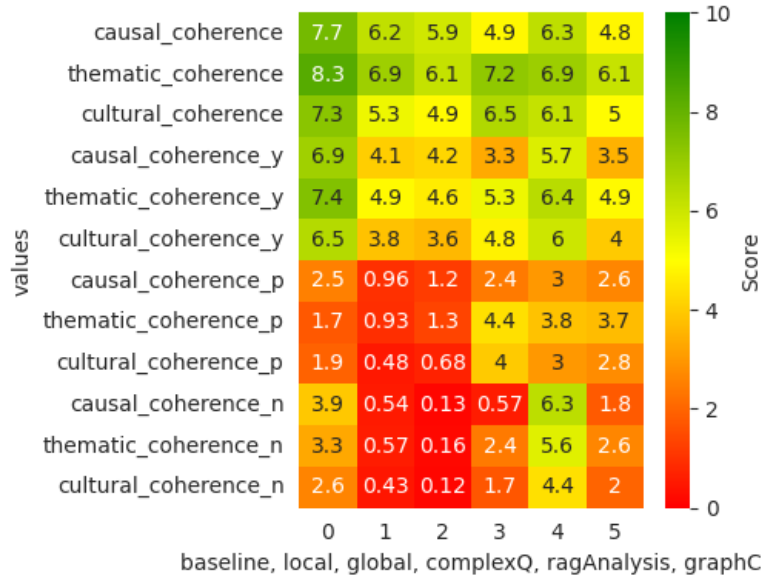


Figure 5.5.: Chapter comparison mean of all validation results by coherence type and validation method

margin for incidental overlap.

Another dimension of the analysis involves the Wilcoxon signed-rank test, applied to compare each data point against the baseline. This statistical test is used to demonstrate that the observed differences are not due to random chance. Specifically, it is observed that the scores for negative and perturbed samples are consistently lower than the baseline, while those for positive samples are consistently higher. To formally validate this, this test was conducted comparing the baseline and the context-aware validation scores as seen in Figure 5.6:

The bar chart displays the W -statistics (left y-axis, blue bars) and corresponding p -values (right y-axis, red line) from one-sided Wilcoxon signed-rank tests, applied across different types of coherence dimensions. Each x-axis label represents a coherence measure under three sample conditions: positive (no suffix), yellow ($_y$), perturbed ($_p$), and negative ($_n$) samples.

The results indicate that the context-aware validation (local) method consistently produces different scores overall, though some noise is observed at both ends for positive and negative samples. This suggests a generally robust difference over the baseline, with slight variability in edge cases. These differences can be attributed to the presence of false positives and false negatives within the test dataset. As discussed earlier, not all books exhibit a consistent or coherent storyline. Consequently, some positive samples may include narratives that lack strong coherence, while some negative or perturbed samples might coincidentally maintain a coherent flow. This variability is reflected in the observed noise, particularly at the extremes. In such cases, the context-aware validation method is

5. Evaluation

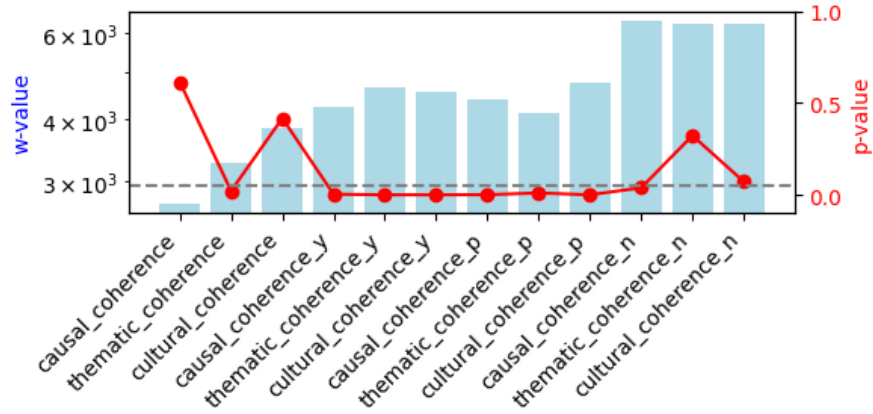


Figure 5.6.: One sided wilcoxon test comparing each measurement of the baseline validation method and the context-aware validation method

capable of identifying inconsistencies more effectively than the baseline. This behavior is further analyzed in the case study section, where selected stories are interpreted in depth to illustrate how context-aware validation captures narrative coherence anomalies more reliably.

Validation results of perturbation comparison

As demonstrated in the experiment presented in the previous, utilizing perturbed variations of a text can provide insight into the effectiveness of validation methods. For example, the results show that perturbed texts from positive text continuations (PTPC)s tend to receive lower ratings from the baseline model compared to unperturbed negative texts originating from entirely different stories. This observation suggests that story evaluators respond differently depending on whether a text has been altered, which raises important questions about the interpretability of such evaluators. In particular, it leads into how accurately validation models can assess PTPCs and whether perturbations can be used to uncover behavioral patterns in evaluation methods.

As an initial experimental assumption focusing specifically on perturbations, the following is proposed: *A validator should be able to assess narrative coherence in a consistent manner, ranking texts from lowest to highest coherence according to three predefined levels of perturbation—maximally PTPCs, moderately PTPCs, and minimally PTPCs.* Based on this assumption, coherence scores are computed using the chosen evaluation metric for each book across all three perturbation levels.

For this experiment, a prompting approach was used to generate PTPCs from each book. A PTPC is defined as a continuation that follows directly after the end of the SMB, maintaining narrative consistency. As introduced in section 5.1, each perturbation was produced using a ZS LLM prompt, with the degree of perturbation controlled via a specified “*severance*” parameter. This parameter aims to influence how strongly the

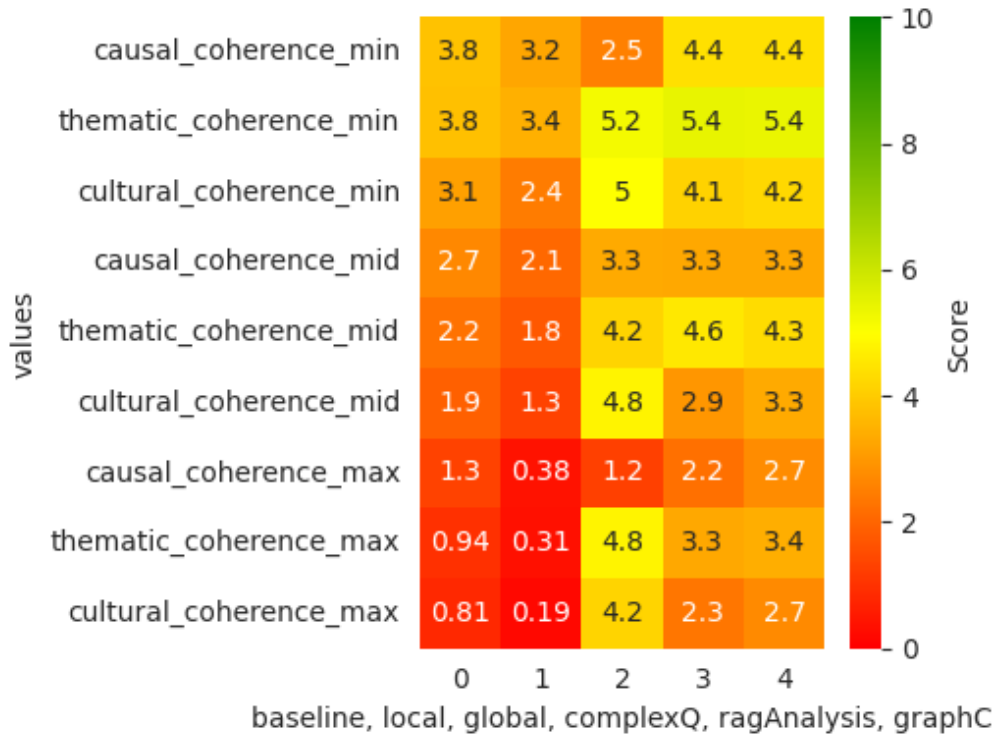


Figure 5.7.: Perturbed validation mean of all validation results by coherence type and validation method

continuation should diverge from the original text.

Figure 5.10 presents the results of this experiment, displaying the mean coherence scores for each evaluation metric across the three levels of perturbation. The experiment was conducted on the first 30 books from our 120-book corpus. For each of these books, story continuations were generated at all three perturbation levels—minimal, moderate, and maximal—and subsequently evaluated using each validation method. Each row corresponds to a specific type and level of coherence (causal, thematic, or cultural), while each column represents a different validation method. The color gradient reflects the coherence score, ranging from low (red) to high (green).

The figure supports the assumption: All evaluation methods exhibit a roughly linear trend in coherence scores across the three PTPC levels, with scores decreasing as perturbation severity increases. This indicates that all validation methods are, in principle, capable of detecting perturbations in the text.

However, results from the previous experiment shows, that at the baseline model, PTPC receive lower coherence scores than unrelated negative samples. This false behavior is not visible anymore in this experiment. This suggests that simple perturbation strategies are insufficient to capture differences in how various evaluation methods behave. The likely cause is that different validators are sensitive to different types of coherence—some

5. Evaluation

```
User prompt:

You are given a coherent narrative. Your task is to create a
perturbed version of the story that maintains global coherence -
the overall storyline must still make sense and follow a consistent
arc - but introduces local incoherence at the provided text level.

Constraints:
Do not change any character names.
You may change character motivations, beliefs, or roles
Introduce unclear logic, or jarring shifts in tone or style
within localized portions of the text.
characters may misinterpret events, act irrationally, or
misunderstand each other in ways that seem confusing on
the sentence level.
```

Figure 5.8.: Local but not global coherence PTPC prompt

more to local coherence (paragraph level coherence), others to global coherence (overall narrative consistency). Since the LLM-generated perturbations affect both local and global aspects, the resulting coherence scores across methods tend to converge, reducing the ability to see differences in coherence scores across validation methods.

To address the limitations identified in the first experiment, this work introduces a second experimental setup designed to distinguish between local and global coherence. This approach is based on the following assumption: *When a continuation preserves local coherence but disrupts global coherence, the resulting coherence score should be low. Conversely, if the continuation perturbs local coherence while maintaining global coherence with the SMB, the coherence score is expected to be relatively higher.*

To test this, the perturbation strategy was adjusted. Instead of varying the severity of perturbations, the new setup explicitly distinguishes between locally perturbed but globally coherent and globally perturbed but locally coherent samples. To create those perturbed samples from the positive text continuations, the prompt, see Figure 5.8, was used to perturb on a local level, but not on a global level.

The second perturbation experiment followed a similar methodology. Figure 5.10 presents the results of this experiment, conducted on the same sample of 30 books. Contrary to the assumption, the results do not align with expectations. In all cases, the globally perturbed but locally coherent PTPCs received lower coherence scores than the locally perturbed but globally coherent PTPCs. This suggests that the validation methods may be more sensitive to disruptions in local coherence than to breaks in global narrative consistency, or that the perturbation prompts did not succeed in cleanly isolating the two dimensions. A detailed analysis of the generated PTPCs revealed that many of them were, in fact, both locally and globally perturbed. A likely explanation is that the

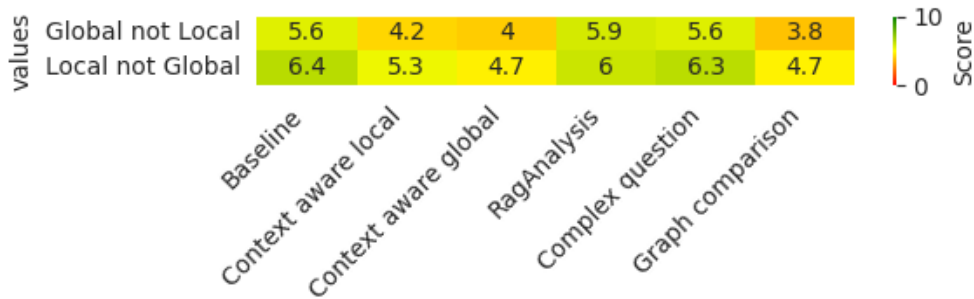


Figure 5.9.: Mean comparison of naive local and global perturbed samples

prompts used to guide the perturbation process were too complex, exceeding the language model’s ability to distinguish and reason between local and global coherence dimensions reliably. As a result, the model often introduced perturbations that affected both levels simultaneously.

To mitigate the limitations observed in the previous perturbation setup, the generation prompts for generating PTPCs were further optimized. Drawing inspiration from the refined perturbation strategies detailed in “Perturbation CheckLists for Evaluating NLG Evaluation Metrics” [SDS⁺21], a list of controlled perturbations was adopted to explicitly target these problems. For local coherence, sentence swapping, reordering of local events, and insertion of confusing or irrational character actions are applied. To perturb global coherence, changes are introduced to key narrative elements such as character names, locations, and major events. Meanwhile, perturbation of coherence aspects outside the intended level is explicitly disallowed—ensuring that when local coherence is perturbed, global coherence remains intact, and vice versa. The goal is to generate perturbed samples that are only locally or only globally disrupted, allowing for clearer isolation of each coherence type in evaluations.

Therefore, a third experiment was conducted using a methodology similar to that of Experiment 2, but utilizing the newly refined PTPCs. In the corresponding results figure, the y-axis represents the different perturbation types, while the x-axis displays the validation methods. Each cell in the matrix shows the mean coherence score for a given combination of perturbation type and validation method, averaged over all 20 book samples.

The results of this setup reveal that the baseline and RAG analysis validation methods yield higher coherence scores for globally PTPCs, while the other validation methods assign lower scores to these PTPCs. This outcome aligns closely with the expectation, that the baseline and RAG-based models are more sensitive to local coherence disruptions and are comparatively less effective at detecting global narrative inconsistencies. In contrast, the context-aware local validation method performs best under this setup, accurately identifying global incoherence as intended. These findings suggest that while the baseline and RAG-analysis methods are biased toward surface-level fluency and local consistency, the others are better suited for capturing coherence at the story level.

5. Evaluation

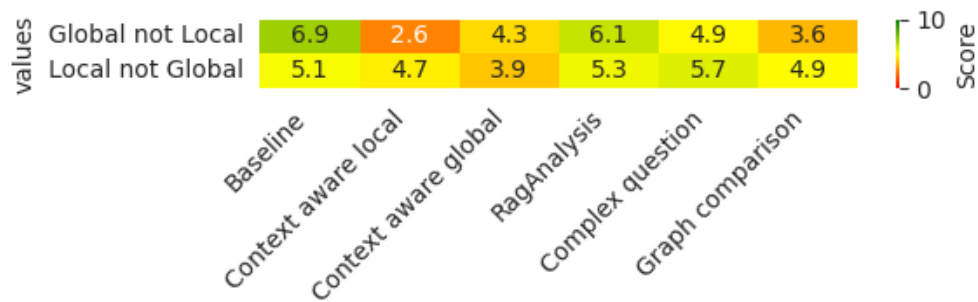


Figure 5.10.: Mean comparison of local and global perturbed samples with refined perturbations

Performance observations

To enable a brief performance analysis, each validation event was time-logged. The following section presents the differences in runtime across the various methods. A direct comparison is valid, as all methods involve negligible local computational overhead and primarily depend on LLM API calls. While API response times vary slightly due to server load or prompt size, these effects were found to be minimal. The resulting time-performance plot shown in Figure 5.11 aligns closely with the theoretical performance expectations of each validation method.

Figure 5.11 shows the time performance (in seconds) required to process different types of queries across 30 data samples (Books). The x-axis represents the index number of each book sample, while the y-axis denotes the processing time in seconds. Six query processing methods are compared: Baseline, Context Local, Context Global, RAG Analysis, Complex Question, and Graph Comparison. The Results are created by an experiment computed with the Deepseek V3 Model.

The Baseline (blue) method consistently exhibits the lowest processing time, remaining below 50 seconds for all samples, because this validation method uses total two LLM reequests. A ZS prompt for validation and the scoring prompt for scoring the prior LLM analysis.

The Context Local (orange) and Context Global (green) methods exhibit moderate processing times, typically ranging from 250–350 seconds and 60–90 seconds, respectively, with relatively low variance. These runtimes reflect the internal mechanics of the GRAG library’s search-based prompting, which both methods utilize for validation. It is important to note that the Context Global approach tends to be more computationally expensive due to its use of a map-reduce strategy. This strategy involves verifying numerous text chunks against the prompt, resulting in multiple individual API calls per validation. In contrast, the Context Local approach employs ranking and filtering techniques locally—without invoking external API calls—to retrieve relevant data. Only the final interpretation step involves the LLM, and it requires significantly fewer API requests. This explains why Context Local is more costly than the baseline but still considerably less expensive than

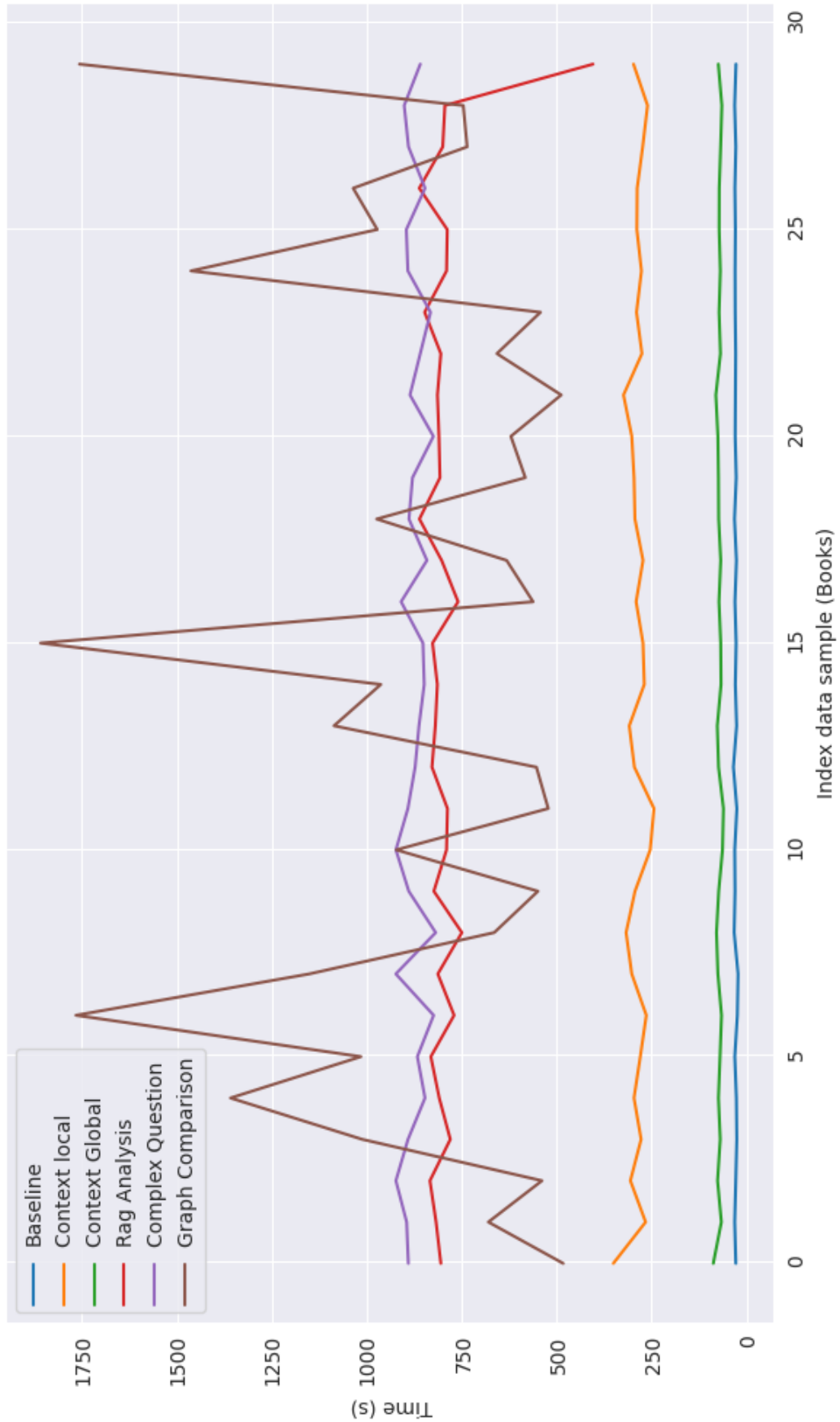


Figure 5.11.: Time measurement comparison of validation methods

5. Evaluation

Context Global.

The RAG Analysis (red) method demonstrates approximately three times longer runtime compared to the Context Local method. This increased processing time is primarily due to the structure of the RAG analysis approach, which issues three separate prompts per validation—each targeting a distinct coherence dimension: causal, cultural, and thematic. As a result, the number of required LLM API calls is higher, directly contributing to the elevated runtime.

The Complex Question (purple) method also incurs relatively high computational costs. At each validation step, N plot precondition questions are generated based on the validation text. This number N is fixed and directly determines the computational load of the method. In the conducted experiments, N was set to 9, resulting in nine local search queries per validation. Although the volume of queries increases, each query is comparatively simple—consisting of a single, focused question. Consequently, the average time per local search is significantly lower than that of a single query in the RAG Analysis method, which handles more complex, multi-part prompts.

The Graph Comparison (brown) method exhibits an alternating but consistently high runtime. This behavior arises from the method’s design, which involves generating a knowledge graph based on the content of the validated text. The number of relations within the graph varies depending on the complexity of the narrative—ranging from as few as three to significantly more. Each of these relations is individually validated against the SMB using a local search query. As a result, the number of LLM requests fluctuates across validation steps, explaining the observed behavior.

5.3. Parameters

Since the validation methods offer various parameters to fine-tune their behavior, this section serves as an entry point for initial observations on the effects of parameter changes. In particular, the impact of validation sample size and summarization size, which affect all validation methods, and the local and global query type in the context-aware validation method is examined. The goal is to identify patterns in how these parameters influence the results, providing a basis for potential optimization in future analyses.

Sample size

The following experiment demonstrates how the validation results differ between two samples of size 500 words and one sample of size 1000 words. Based on the assumption that the samples are summarized during processing, it is expected that information loss increases with larger sample sizes. Consequently, smaller samples should yield more detailed and accurate results. To test this, the experiment compares two chapters of 500 words each with one chapter of 1000 words. If the combined evaluation of the two 500-word chapters leads to the same conclusions as the single 1000-word chapter, then using larger sample sizes would be justified as a way to reduce computational costs.

The experiment was conducted using three validation methods, selected to reduce

computational costs based on earlier observation results. It was determined that the Baseline and RAG Simple validation methods exhibited similar behavior, as did the Graph Comparison and Complex Question Analysis methods. Additionally, to further manage costs, a sample of 10 books was used in the experiment.

The first set of results presents combined outcomes from two validation runs: one using a TTV sample size of 500 words, and another using an additional TTV of 500 words with an offset of 500 words—together representing a combined sample size of 1000 words. The second set of results is based on a single TTV validation run directly using a 1000-word sample.

Figure 5.12 shows this comparison of validation scores for two different TTVs sample sizes—500 words (left) and 1000 words (right)—across three validation methods: Baseline, context-aware local, and graph comparison. The scores are grouped by coherence type (causal, thematic, cultural) and polarity (positive, yellow/neutral, perturbed, negative). The results visible in Figure 5.12 indicate that, on the positive scale, the experiment using a direct 1000-word sample produced higher positive scores compared to the combined results. This discrepancy arises because the first 500-word segment of the combined run yields a relatively high score due to its direct continuation from the text, while the second 500-word segment scores lower, as the first 500 words are missing in the source material—reducing overall coherence and alignment, which significantly lowers the score. To mitigate this shortcoming, a more refined positive test would require a re-indexed source material specifically prepared for the source material + 500 words corpus from the first sample. However, due to high computational costs, this refinement was considered out of scope for the present study.

No significant differences were observed on the negative scale or for yellow text samples across the different validation runs.

In contrast, the perturbed samples showed a significant difference: the scores for the 1000-word sample were higher. This outcome is attributed to the fact that the quality or degree of perturbation decreases as the sample size increases—an expected result stemming from known limitations of zero-shot prompting in relation to large language models (LLMs). Unfortunately, this initial naive experiment did not reveal significant differences; however, it highlights important experiment configuration issues that must be addressed in future, more sophisticated iterations of the experiment.

Local and global

Another important parameter is the local or global query type, which influences how the graph is traversed in the context-aware validation process. As explained in the section on context-aware validation, the global search uses summaries of graph hotspots (i.e., communities) and retrieves responses based on these summaries. This approach is expected to improve the method’s ability to measure thematic coherence, which depends heavily on identifying and evaluating global themes across the text.

Figure 5.13 compares the performance of the local and global search parameter settings within the context-aware validation method. The heatmap illustrates the mean validation scores for different types of coherence—causal, thematic, and cultural—across multiple

5. Evaluation

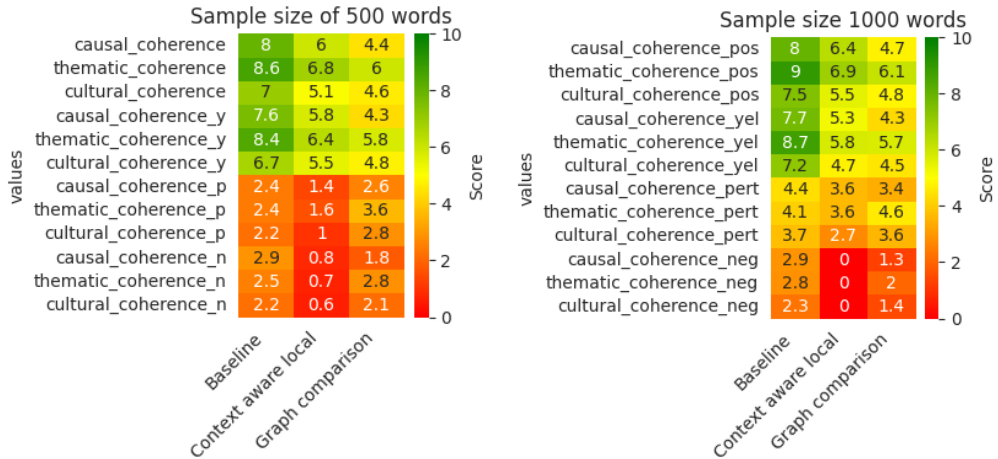


Figure 5.12.: Validation result comparison off two TTV sample sizes 500 words(left) and 1000 words (right)

sample categories: positive samples, yellow samples (`_y`), perturbed samples (`_p`), and negative samples (`_n`).

In the plot, it can be observed that the local context-aware validation method assigns relatively high thematic coherence scores to the negative samples. Based on the assumption that negative samples should exhibit no meaningful overlap with the SMB, scores of 0 are expected in this category. In contrast, the global context-aware validation detects the lack of coherence more reliably, assigning much lower scores to the negative samples as the local method.

This pattern is also evident in the thematic coherence scores of the perturbed samples. Here, the local method reports lower coherence, whereas the global method—consistent with theoretical expectations—assigns higher thematic coherence scores. Since the perturbed samples originate from the same story, some thematic relationship to the SMB is to be expected, which the global validation method successfully captures.

Although there are many additional parameters that could be adjusted and evaluated, the scope of this analysis is limited to the sample size. This restriction is due to the high computational cost of testing every possible parameter combination. As a result, parameter settings for the following experiments must be determined based on informed assumptions and educated guesses.

5.4. Case Study

Building on RQ3, the following section presents in-depth case studies of selected data samples. These case studies explore how effectively the different validation methods perform when applied to both coherent and intentionally non-coherent narratives. In addition to assessing the accuracy and reliability of the coherence scores, this section also

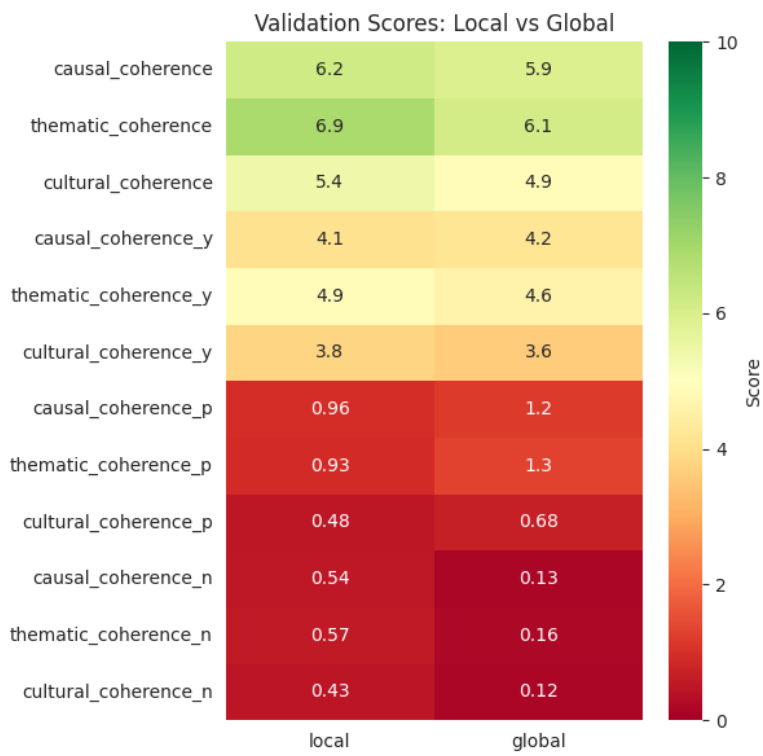


Figure 5.13.: Local and global search parameter comparison at the context-aware validation method

5. Evaluation

analyzes the qualitative output of each method. The goal is to evaluate how helpful the feedback is beyond the numeric score.

Validation against non-coherent data samples

The first analysis focuses on book data samples that do not have a coherent, continuous narrative structure. These include books such as *The Door in the Wall, and Other Stories*, which are collections of short stories rather than unified narratives. Such samples are categorized as *negative* samples based on the underlying assumption that those stories lacking global narrative coherence should yield low or zero coherence scores across all coherence metrics and evaluation samples.

A representative excerpt of raw coherence scores is shown in Figure 5.14. On the top side of the Figure, the raw results from the baseline validation method are presented for both a positive sample (a book with a continuous and coherent storyline) and a negative sample (a non-coherent collection of stories). Notably, the baseline method assigns relatively high coherence scores even to the negative sample. This outcome highlights a key limitation of the baseline approach—its inability to differentiate between coherent and non-coherent narratives.

Conversely, the bottom side of the Figure displays the evaluation results obtained using the context-aware validation method. In this case, the method demonstrates a clear ability to distinguish between the two types of samples. All coherence metrics for the non-coherent book yield values of zero, indicating a complete lack of continuity. These results suggest that the context-aware method is more effective in detecting anomalies and inconsistencies.

Figure 5.14 visualize coherence scores across different data samples derived from two books. The x-axis represents three distinct coherence metrics on four different sample types from each book. These sample types include: the positive sample (no suffix), the yellow sample (suffix `_y`), the perturbed sample (suffix `_p`), and the negative sample (suffix `_n`). The y-axis represents the books. The first book, *A Study in Scarlet*, serves as a representative of a coherent narrative, while the second book, *The Door in the Wall, and Other Stories*, exemplifies a non-coherent story collection.

While this pattern is evident in this example comparing the baseline and context-aware validation methods, it is necessary to evaluate whether this behavior generalizes across all non-coherent data samples. Additionally, the performance of alternative validation methods on non-coherent story structures is of particular interest.

For this extended analysis, all 120 book samples were manually annotated based on whether they exhibit obvious narrative coherence or not. In total, eight books were identified as non-coherent — i.e., lacking a global storyline — and were selected for further examination.

The corresponding plot 5.15 illustrates the mean coherence values across all non-coherent book samples, evaluated along different coherence metrics and sample types. The results reveal several insights:

- The baseline validation method fails to reflect the incoherent nature of the texts,

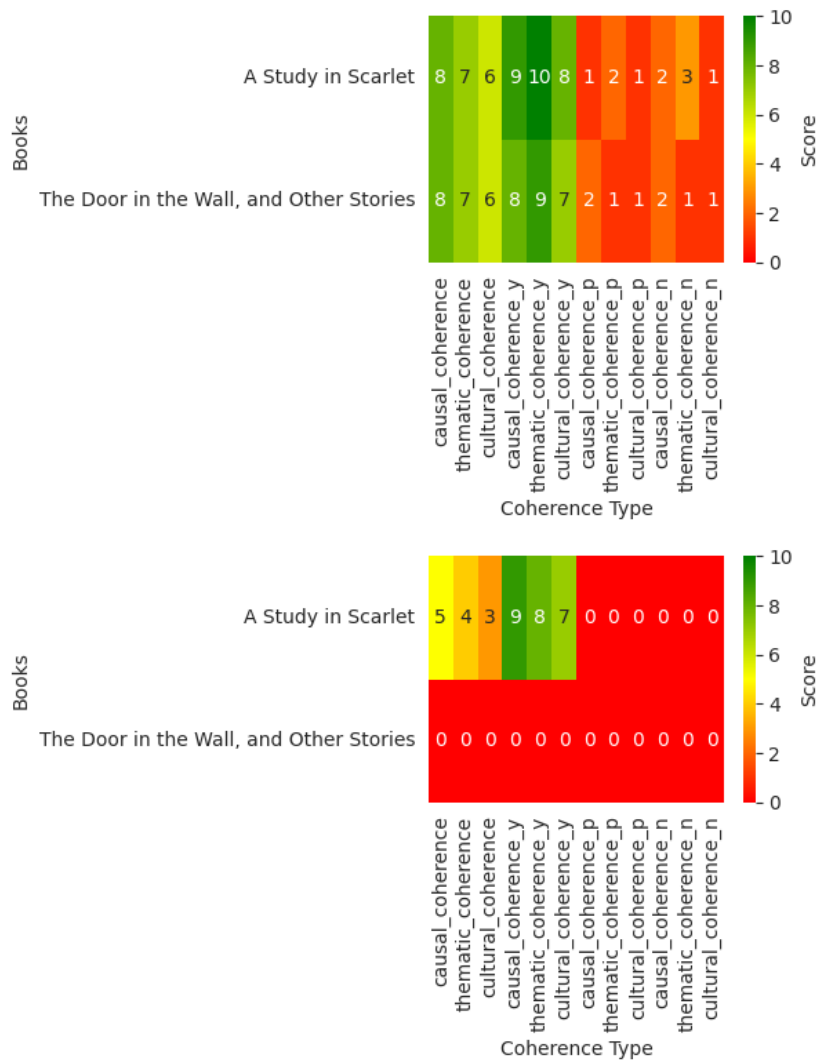


Figure 5.14.: Baseline validation (top) vs context-aware validation (bottom) results in focus of non-coherent story samples

5. Evaluation

producing mean coherence values that closely match those reported in the full validation dataset (cf. Section 5.2). This highlights a limitation in the method’s ability to detect global narrative incoherence.

- The RAG analysis validation method shows a slight negative deviation in coherence scores compared to the overall dataset. However, both the baseline and the RAG analysis scores remain unexpectedly high, due to the method’s emphasis on *local coherence*, which obscures global narrative discontinuity.
- The Complex Question Answering and Graph Comparison methods show a greater sensitivity to incoherence. While their average scores are still higher than the theoretical expectation of near-zero coherence, they are clearly lower than those of the baseline and RAG methods, indicating a more accurate assessment.
- The best results are achieved by the context-aware validation methods—both global and local—which consistently return the lowest coherence values for non-coherent samples. These methods most closely align with the assumption.

It is important to note that coherence scores across all methods are not strictly zero. This can be attributed to two main factors: (1) many story collections, while lacking global coherence, may share a consistent *narrative perspective* or stylistic unity that contributes to coherence; (2) the classification of samples as non-coherent was based on a naïve manual inspection, potentially introducing noise in the labeling process; (3) the cutoff point of a PTPC can overlap with the boundary between two chapters or stories within the book. This occurs because PTPC chunks are generated starting from exactly 50% of the book’s length, rather than aligning with the end of individual chapters or narrative units. As a result, a single sample simultaneously contains the conclusion of one chapter and the beginning of the next. Consequently, the validation method is expected to yield a positive coherence score at the start of the chunk—corresponding to the beginning of a new chapter—and a negative score toward the end. To mitigate this issue, a more precise selection process that defines cutoff points at actual chapter boundaries would yield more accurate results.

Validation feedback observation

Each validation result not only returns coherence scores but also provides a brief analysis explaining how these scores were derived. This analysis varies depending on the specific validation outcome. The primary objective of these analyses is to assist the user in identifying coherence issues and in interpreting the scores. This is particularly important, as some detected coherence inconsistencies may be misinterpreted or might stem from elements introduced within the same chapter. In the following, the outcome of a selected validation sample is examined. This sample includes numerous plot-related actions and is intended to challenge the robustness of the validation results. A summary of this sample is presented in the text presented in Figure 5.16. This summary is used by the validation methods to create the coherence scores.

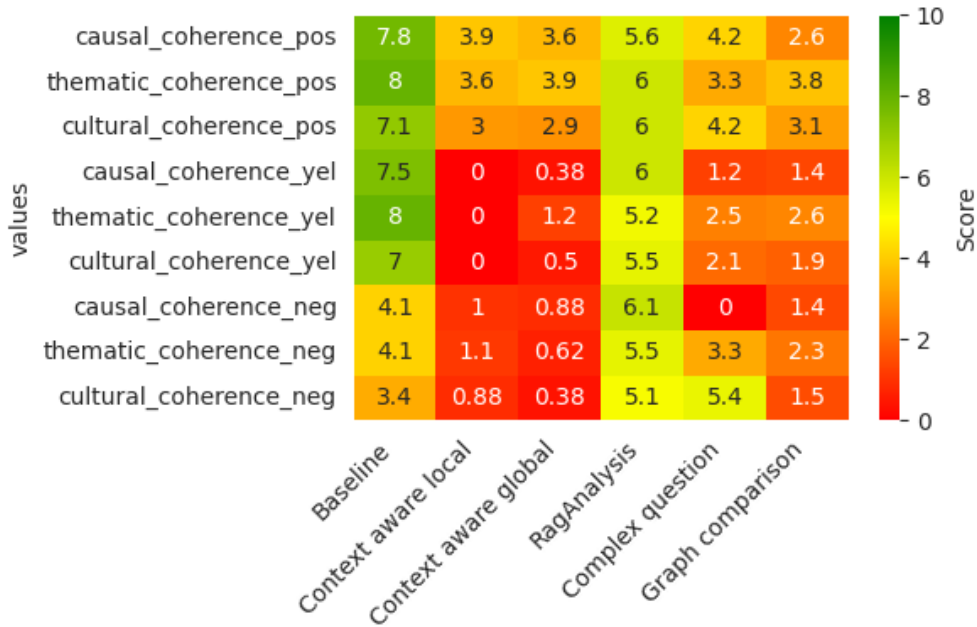


Figure 5.15.: Mean validation results on non-coherent data samples only

5.5. Analysis of Validation Results

The analysis of the output text produced during the validation process plays a critical role in this work. An examination of these texts not only helps guide users in interpreting validation outcomes, but also provides valuable feedback for ASG tools that leverage this information to improve and refine their future outputs generated with LLMs. In this section, the result text from one positive TTV validation run is analyzed for each validation method.

Baseline validation result An analysis of the validation results, as presented in Appendix A, reveals that the baseline evaluation offers limited insight into the narrative coherence or quality of the generated continuation. The results remain vague and lack specificity. They do not provide concrete explanations for the scores assigned, nor do they identify which textual features contribute to or detract from the SMB. The evaluation does not highlight strengths or weaknesses in terms of the analysed coherence metrics.

For instance, one example from the baseline results from Appendix 1 A states:

Causal Coherence (Score: 8): Story B maintains a strong causal coherence with Book A, particularly in its continuation of the plot involving Isabella's flight and the reactions of Manfred and Jerome. The events follow logically from the preceding chapters, with Manfred's suspicion and Jerome's concern for Theodore driving the narrative forward.

5. Evaluation

Result:

A monk, returns to the monastery distressed after failing to prevent Manfred pursuit of Isabella. He encounters another friar who mistakenly informs him that Princess Hippolita has died news Jerome disputes, having just left her in good health. The friar explains that the monastery believes the rumor and has gathered to pray for her soul. Jerome, alarmed, asks about Isabellas whereabouts but learns she vanished after being told of Hippolitas supposed death. Fearing Manfreds wrath, Jerome searches frantically for Isabella, suspecting she fled to avoid being forced into marriage with the Prince. Her disappearance deepens his anxiety, as it could provoke Manfred further, endangering Theodore, Jeromes son. To protect himself and Theodore, Jerome gathers fellow monks to return to the castle, intending to prove his innocence and plead for mercy. Meanwhile, Manfred, seething with suspicion, orders the castle gates opened, likely to escalate his pursuit. The core tensions revolve around mistaken death rumors, Isabellas flight, and Jeromes desperate efforts to shield Theodore from Manfreds vengeance. The plot thickens as misunderstandings and hidden motives drive the characters toward further conflict.

Figure 5.16.: The summary of the positive validation sample from the book: The Castle of Otranto

5.5. Analysis of Validation Results

In fact, a similar evaluation result can be observed when the continuation is based on a non-coherent story. This raises significant concerns regarding the validation framework. The results across both coherent and non-coherent samples suggests that the evaluation method is not sensitive to fundamental narrative flaws.

Context-Aware results: The context-aware validation framework provides two results of the generated story continuation. The first tier involves a context analysis. In this step, the RAG system identifies strengths and weaknesses of the submitted story by comparing it to the original SMB. This analysis produces feedback that highlights where the continuation aligns or diverges from the source, referencing concrete textual units. The second tier consists of a scoring mechanism derived from a LLM, which evaluates the story continuation in mind of the context-aware analysis. An example part from Appendix A of the context-aware validation reveals insights into the generated continuation, while also identifying several shortcomings:

“Thematic Simplification Although the theme of refuge is present, the continuation simplifies it to Isabella’s immediate escape, omitting broader societal implications—such as sanctuary-seeking in the context of supernatural or political threats. [Referenced data units: Reports (5, 15, 19)]”

“Narrative Implications The continuation isolates interpersonal conflicts from the broader narrative framework, thereby diminishing narrative tension. For example, Manfred’s rage is depicted as a personal reaction, rather than as a response rooted in the prophecy’s existential threat to his reign.”

This example demonstrates that context-aware validation offer a more helpful and grounded critique compared to baseline validation methods.

RAG Analysis results: While the graph-based RAG analysis provides more grounded results compared to the baseline model, it also reveals some conceptual issues with the overall approach. The first quote from the result example shows that the analysis can uncover meaningful insights that are important to the narrative:

“Isabella’s Flight Isabella’s disappearance aligns with her pattern of fleeing Manfred’s tyranny. Her fear that Manfred caused Hippolita’s death is a logical (though unstated) motive for escape. [Data: Reports (5, 8); Entities (18, 77); Relationships (13, 59)]”

However, the second quote illustrates a key conceptual limitation: the SMB does not inherently differentiate between events introduced in the current chapter and those introduced earlier. As a result, the validation process cannot determine whether an event is new or pre-existing. This leads to the generation of plot inconsistency alerts that are inaccurate:

5. Evaluation

“Hippolita’s Alleged Death She is alive in the source and central to the moral tension around divorce. Her death is fictional and unsupported. [Data: Entities (19); Relationships (12, 127)]”

This example highlights that the validation framework may misinterpret narrative developments, flagging them as inconsistencies when they are simply new plot points introduced in the chapter.

Complex question results: While the baseline validation methods, context-aware and RAG-based analysis, return full-text evaluations, the complex question analysis produces a concatenated output consisting of individually answered questions, each comprising the question itself and the corresponding response generated by the RAG system. In the following two questions-answer pairs from the validation results A give interesting insights:

“Does the monks’ reaction to the Princess’s rumored death align with their religious and social norms? No, the monks’ reaction does not align with their religious and social norms, as they show more concern for their own interests and political maneuvering than for the Princess’s well-being.”

“Does the extension respect the original’s cultural norms around nobility and the church? Yes, the narrative consistently portrays nobility and the church with reverence, aligning with medieval cultural norms where both institutions held significant power and moral authority. [Data: Reports (0, 9, 20); Entities (15, 77, 97); Relationships (106, 117, 132)]”

These examples demonstrate that complex question analysis can provide good insights into the coherence of the text. However, they also reveal the method’s limitations: when questions are vague or poorly formulated—as exemplified by the second pair—the answers tend to be very generic. As a result, the analysis quality is based on the system’s ability to generate good responses and on the method involved to craft good quality questions.

Graph comparison results: In the final validation method, each edge finding generated by the knowledge graph is individually validated by the Graph-RAG system, which provides an evaluative response for each statement. This process demonstrates that the statements can be effectively processed by the RAG system, and the individual responses are generally logically consistent.

However, the method also reveals problems—similar to those observed in the initial RAG analysis—where newly introduced plot points are mistakenly flagged as inconsistencies. This occurs because the SMB, by design, cannot yet contain knowledge of events introduced later in the narrative. As a result, some assessments produced by the validation system are misleading.

The first quote illustrates the system’s capacity to provide a high-quality explanation that captures narrative causality:

“The statement demonstrates strong causal coherence, clearly linking Isabella’s actions to subsequent events in the narrative. The cause-effect relationships are well-established, particularly in how her flight from Manfred’s pursuit leads to key plot developments. The references to specific data points (e.g., Reports, Relationships) further solidify this coherence.”

The second quote, while still offering a reasonable response, reveals the system’s difficulty in handling plot elements not yet introduced in the SMB:

“The statement partially aligns with the data, as Jerome’s distress over Hippolyta’s well-being is supported by his moral opposition to Manfred’s actions. However, the specific claim of ‘false news of her death’ lacks direct evidence, weakening the causal logic.”

5.6. Discussion

One of the recurring challenges encountered in this work was the handling of first-person narrative styles in the source material. Many of the books analyzed were written in first person, often interwoven with letters or dialogues where various characters also speak in first person. It became evident that both LLMs and RAG (or GRAG) systems struggled to correctly attribute first-person events to the appropriate entity. This issue affected the performance of all validation methods used in this work and raises important questions about how to mitigate these challenges—whether through data preparation or through adapting the validation techniques to better handle first-person narratives. Unfortunately, addressing this issue in depth was beyond the scope of this thesis; while the problem was acknowledged, it remains unresolved.

Another interesting aspect of story coherence that emerged from this work is the chaotic nature of narrative structures, particularly regarding the introduction of new characters or entities. This raises the question of whether newly introduced characters—those without a clear connection to the source material or an established story arc—should be treated as coherence errors. In practice, these entities are not necessarily errors, as they may later be integrated into the plot. This challenge influenced the design of the validation methods, ultimately leading to the development of the graph-based comparison approach. In this method, isolated entities are not automatically flagged as errors; instead, the absence of connections (edges) to other entities is used as an indicator, under the assumption that connections imply coherence with the source material. But, this design introduced another issue, which became apparent during the perturbation testing. A highly perturbed TTV can produce numerous isolated nodes—entities that have and will have no connection to the source material. These “negative” entities are not flagged as coherence errors, which lead into an higher coherence score.

Furthermore, the proposed validation methods assess coherence primarily based on contextual fit—that is, they check whether entities logically belong within the story’s context. However, they do not analyze the overall plot structure or narrative arc. For

5. Evaluation

example, a book that continuously introduces new characters without resolving any plotlines would typically be judged by human readers as incoherent or poorly structured, perhaps as an “open-ended” story. In contrast, a well-constructed narrative introduces only the necessary entities to advance the plot meaningfully. This dimension of narrative complexity—such as the branching and resolution of plotlines—is not explicitly captured by the current validation methods. While LLMs theoretically possess the ability to reason about what constitutes good plot design, it remains an open question whether, and how, this capability is interfering the validation results.

From another perspective, the runtime costs associated with executing the experiments led to small sample sizes and a selective experimental design. Given the breadth of the field—with dozens of varying parameters and implementations—many potentially valuable configurations that could have brought up more precise coherence scores or a more robust testing setup were unexplored or were discarded due to runtime constraints. Since the completion of this work, newer and more efficient GRAG frameworks such as LightRag [GXY⁺25] have been developed, offering considerably lower computational overhead compared to the Microsoft GraphRAG [Res24] framework used in this project. For instance, investigating the so-called First-Person Problem—specifically, validating the system’s performance with and without converting data samples from first to third person—could provide meaningful insights, but was excluded from the current study due to these limitations.

Finally results revealed interesting outcomes. In terms of overall performance, the context-aware methods consistently outperformed other, more complex approaches. Notably, while the graph-based analysis—characterized by its higher computational demands and a more “thoughtful” strategy—did achieve better results than the baseline, it still lagged behind the simpler context-aware methods in terms of effectiveness. Based on the findings of this study, one can observe that methods focusing on the retrieval of relevant contextual information demonstrate superior performance compared to question-answering approaches.

6. Conclusion

This thesis set out to explore how LLMs and RAG systems can be used to validate the coherence of long-form narratives. Early in the process, however, it became apparent that a key challenge lies in validating the validation methods themselves, especially in the absence of annotated benchmarks and datasets. The work therefore focused not only on developing techniques and methods for evaluating story coherence, but also on introducing new concepts for experimental design in this rapidly evolving field of ASG where best practices are still emerging. Throughout the research, as the foundations of validation methods were being established, advancements in RAG approaches were also integrated into the framework. The thesis presented several complementary validation methods, each addressing different challenges of story evaluation, highlighting their respective strengths and limitations. This process led to multiple optimization iterations and identified further areas for improvement, offering a foundation for future work.

In addition to introducing new ideas to the field, this thesis successfully demonstrated that the proposed validation methods are effective and can be reliably validated using the experimental setup presented. The results showed that these methods outperform the baseline in detecting incoherent or perturbed narrative samples, which the baseline method failed to identify. However, the experiments also revealed certain limitations—particularly in the interpretability of the results. It proved challenging to clearly demonstrate differences in coherence due to the interplay between dimensions of narrative coherence, such as local versus global coherence. In this work, this challenge was addressed through the use of a perturbation-based experimental setup designed to highlight both local and global coherence results.

Finally, this work provided insights into the validation methods through an analysis of the resulting texts, offering a understanding of the quality of the outputs that form the basis for coherence scoring. This analysis revealed several interesting features that could be valuable for future users or for tools seeking to implement and extend these methods, but also highlighted unique challenges—particularly those that arise in certain narrative styles and book-writing approaches.

There is future work to be done to improve the quality of the validation results—particularly in refining the methodology and experimental setup. Key areas for advancement include optimizing the datasets used, enhancing the validation methods themselves, and improving the overall experimental framework.

An important direction for future work is the further refinement and preparation of the dataset. Many of the uncertainties encountered during the analysis of the results in this study were ultimately traced back to issues within the dataset itself. Both the GRAG indexed source material and the s were selected without prior knowledge of the content, leading to the creation of data slices that did not account for the natural structure of

6. Conclusion

the material. In particular, the slicing process was unaware of chapter boundaries, which resulted in segments that could span the conclusion of one chapter and the beginning of another. This negatively impacted validation scenarios, especially in cases where books comprised multiple smaller stories. Future research should focus on defining data slices that respect chapter divisions or developing automated methods to achieve this. Although the book corpus was initially curated to exclude certain writing styles, such as rhymed texts and documentary works, selection like discarding first-person writing styles were not carried out due to time constraints. As a result, the corpus included a wide variety of book types — ranging from structured narrative texts with coherent storylines to collections of globally incoherent stories. A more systematic and selective curation of the corpus could provide clearer insights into the performance and limitations of the validation methods employed in this study. Additional methods could be employed to detect the writing styles and complexity of the books, for an automatic annotations within the dataset.

In addition to optimizing the dataset, the validation methods themselves present several opportunities for further exploration. One key challenge identified in this study was the presence of locally incoherent chapters, caused either by suboptimal slicing or by weaknesses in the writing. These incoherencies negatively impacted both the baseline validation and the RAG-based analysis. This issue became most evident in the perturbed experiments, where both local and global text perturbations were applied. Developing a validation setup that explicitly measures local coherence could help produce more reliable results. Similarly, quantifying the degree of perturbation, especially local perturbation, could offer valuable insights into the performance of various validation methods. Exploring how such coherence and perturbation measurements could lead to improved scoring and more informative experimental outcomes represents an interesting avenue for future work. Additional challenges were encountered in the complex question validation method, particularly during the stages of precondition extraction and question generation. While the extracted questions generally reflected the intended query, they often lacked sufficient contextual information from the validation samples. For example, names and specific references were generalized to terms such as “*the protagonist*” or other ambiguous descriptors. This lack of clarity prevented the RAG systems from retrieving the relevant context, leading to negative validation scores. Designing extraction prompts capable of generating good questions across a variety of writing styles proved to be a significant challenge. Improvements in prompt engineering could enhance the performance of this method. Furthermore, this approach proved to be particularly sensitive to complex writing styles. Another limitation identified in the graph comparison validation method concerns the handling of isolated nodes. This issue could be addressed by analyzing isolated nodes in relation to their corresponding text chunks, focusing on the entities themselves rather than solely on their relationships. Such an approach could also lead to improved validation outcomes. Moreover, this concept offers considerable potential for further development. For instance, incorporating graph community structures could provide additional dimensions for validation. While working on the method and reflecting on the isolated node problem, an alternative approach emerged: generating the knowledge graph directly through the GRAG system and subsequently analyzing the number of

isolated nodes or isolated clusters. Unfortunately, pursuing this line of investigation was beyond the scope of the available resources for this project.

Finally, the experimental setup itself could also be optimized. Some experiments did not yield clear results, or the outcomes were difficult to interpret, primarily because they were not conducted on a scale sufficient to validate the intended features. This issue was particularly evident in the parameter testing of sample sizes. Since this experiment was conducted late in the process, a comprehensive series of tests using varying sample sizes was beyond the project's scope and proved difficult to integrate with the perturbed experiments. Future work could focus on refining the experimental setup and developing an integration of a valuable test setup within the perturbed test. The primary reason for these limitations was the high cost associated with LLM API usage, as these methods require substantial amounts of both input and output tokens. Although the experiments were conducted entirely with DeepSeek v1 — one of the more affordable models available — the total API cost still reached approximately €74. Future work could focus on reducing validation costs, particularly the high expenses associated with the RAG framework, by exploring lighter alternatives such as Light-RAG [GXY⁺25]. Additionally, systematic evaluation of different models for validation purposes could offer further insights; such a comparative analysis was initially planned but ultimately discarded due to scope limitations.

Overall, this journey through a highly interdisciplinary field was both exciting and enriching. Valuable insights across many aspects of the topic were gained — from story writing and narrative structure to the automation of prompting techniques. Ultimately, this work highlights several areas of interest and opportunities for further optimization within the field, while achieving the goal of presenting methods capable of assessing the complex and chaotic nature of story coherence in book narratives.

Bibliography

- [ACBR20] Prithviraj Ammanabrolu, Wesley Cheung, William Broniec, and Mark O. Riedl. Automated storytelling via causal, commonsense plot ordering, 2020.
- [ACJR12] Apoorv Agarwal, Augusto Corvalan, Jacob Jensen, and Owen Rambow. Social network analysis of alice in wonderland. In David Elson, Anna Kazantseva, Rada Mihalcea, and Stan Szpakowicz, editors, *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 88–96, Montréal, Canada, June 2012. Association for Computational Linguistics.
- [ALP21] Amal Alabdulkarim, Siyan Li, and Xiangyu Peng. Automatic story generation: Challenges and attempts. In Nader Akoury, Faeze Brahman, Snigdha Chaturvedi, Elizabeth Clark, Mohit Iyyer, and Lara J. Martin, editors, *Proceedings of the Third Workshop on Narrative Understanding*, pages 72–83, Virtual, June 2021. Association for Computational Linguistics.
- [AWW⁺20] Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. Storium: A dataset and evaluation platform for machine-in-the-loop story generation, 2020.
- [BD24] Shaily Bhatt and Fernando Diaz. Extrinsic evaluation of cultural competence in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16055–16074, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
- [CCSC22] Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.
- [CFB⁺21] Louis Castricato, Spencer Frazier, Jonathan Balloch, Nitya Tarakad, and Mark Riedl. Automated story generation as question-answering, 2021.

Bibliography

- [CH11] Mark J. Weal Charlie Hargood, David E. Millard. Measuring narrative cohesion: A five variables approach. 2011.
- [CPR⁺25] John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max Kreminski. Modifying large language model post-training for diverse creative writing. *arXiv preprint arXiv:2503.17126*, 2025.
- [CSC24] Cyril Chhun, Fabian M. Suchanek, and Chloé Clavel. Do language models enjoy their own stories? prompting large language models for automatic story evaluation, 2024.
- [D⁺13] Julien Le Dem et al. Apache parquet: A columnar storage format for hadoop, 2013. Accessed: 2025-08-22.
- [DA25] Team DeepSeek-AI. Deepseek-v3 technical report, 2025.
- [ETC⁺25] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitanaky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2025.
- [FNJL23] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire, 2023.
- [FTD25] Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. From llm reasoning to autonomous ai agents: A comprehensive review, 2025.
- [GHZ⁺20] Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. A knowledge-enhanced pretraining model for commonsense story generation, 2020.
- [GLSM⁺21] Sarik Ghazarian, Zixi Liu, Akash S M, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. Plot-guided adversarial example construction for evaluating open-domain story generation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4334–4344, Online, June 2021. Association for Computational Linguistics.
- [GLT⁺20] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training, 2020.
- [GXG⁺24] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.

- [GXY⁺25] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation, 2025.
- [HB00] Tilmann Habermas and Susan Bluck. Getting a life: The emergence of the life story in adolescence. *Psychological Bulletin*, 126:748–769, 01 2000.
- [HCL⁺24] Senyu Han, Lu Chen, Li-Min Lin, Zhengshan Xu, and Kai Yu. IBSEN: Director-actor agent collaboration for controllable and interactive drama script generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1607–1619, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [Hen90] G. A. Henty. *Tales of Daring and Danger*. Blackie & Son, 1890. Accessed: 2025-06-05.
- [HSK⁺24] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What’s the real context size of your long-context language models?, 2024.
- [HWS⁺25] Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A. Rossi, Subhabrata Mukherjee, Xianfeng Tang, Qi He, Zhigang Hua, Bo Long, Tong Zhao, Neil Shah, Amin Javari, Yinglong Xia, and Jiliang Tang. Retrieval-augmented generation with graphs (graphrag), 2025.
- [IG21] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering, 2021.
- [ILL⁺22] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models, 2022.
- [JSH⁺25] Zhi Jing, Yongye Su, Yikun Han, Bo Yuan, Haiyun Xu, Chunjiang Liu, Kehai Chen, and Min Zhang. When large language models meet vector databases: A survey, 2025.
- [JYHA24] Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O. Arik. Long-context llms meet rag: Overcoming challenges for long inputs in rag, 2024.
- [Kan25] Sivasathivel Kandasamy. Control plane as a tool: A scalable design pattern for agentic ai systems, 2025.
- [Lin04] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Bibliography

- [LPP⁺21] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.
- [LW67] William Labov and Joshua Waletzky. Narrative analysis: Oral versions of personal experience. In *Oral Versions of Personal Experience*, pages 20–25. John Benjamins Publishing Company, 1967.
- [LZD⁺24] Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhui Chen. Long-context llms struggle with long in-context learning, 2024.
- [Mca06] Dan Mcadams. The problem of narrative coherence. *Journal of Constructivist Psychology - J CONSTR PSYCHOL*, 19:109–125, 04 2006.
- [MCH⁺16] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and evaluation framework for deeper understanding of commonsense stories, 2016.
- [Met24] Team Meta. The llama 3 herd of models, 2024.
- [MQL24] Yan Ma, Yu Qiao, and Pengfei Liu. Mops: Modular story premise synthesis for open-ended automatic story generation, 2024.
- [MS25] Carlo Merola and Jaspinder Singh. Reconstructing context: Evaluating advanced chunking strategies for retrieval-augmented generation, 2025.
- [MT23] Aviya Maimon and Reut Tsarfaty. COHESENTIA: A novel benchmark of incremental versus holistic assessment of coherence in generated texts. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5328–5343, Singapore, December 2023. Association for Computational Linguistics.
- [NGG20] Eric Nichols, Leo Gao, and Randy Gomez. Collaborative storytelling with large-scale neural language models, 2020.
- [now25] noworneverev. Graphrag visualizer: A web-based tool for visualizing and exploring artifacts from microsoft’s graphrag. <https://github.com/noworneverev/graphrag-visualizer>, 2025. Accessed: 2025-06-24.
- [Ope24] Team OpenAI. Gpt-4 technical report, 2024.
- [PA20] Wesley Cheung Zhaochen Luo William Ma Lara J. Martin Mark O. Riedl Prithviraj Ammanabrolu, Ethan Tien. Story realization: Expanding plot events into sentences. In *Proceedings of the AAAI Conference on Artificial Intelligence 7375-7382*, 2020.

- [PERPL24] Zeeshan Patel, Karim El-Refai, Jonathan Pei, and Tianle Li. Swag: Storytelling with action guidance, 2024.
- [PLS⁺23] Bhargavi Paranjape, Scott Lundberg, Sameer Singh, Hannaneh Hajishirzi, Luke Zettlemoyer, and Marco Tulio Ribeiro. Art: Automatic multi-step reasoning and tool-use for large language models, 2023.
- [PRWZ02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.
- [PSB21] Andrew Piper, Richard Jean So, and David Bamman. Narrative theory for computational narrative understanding. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [PyP24] Rafael Pérez y Pérez. Computer models as tools for the reflection on ideas: A model for evaluating the thematic consistency of a narrative. In *Proceedings of the 15th International Conference on Computational Creativity (ICCC)*, pages 75–82. Association for Computational Creativity, 2024.
- [PZL⁺24] Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. Graph retrieval-augmented generation: A survey, 2024.
- [QDH⁺24] Haoran Que, Feiyu Duan, Liqun He, Yutao Mou, Wangchunshu Zhou, Jiaheng Liu, Wenge Rong, Zekun Moore Wang, Jian Yang, Ge Zhang, Junran Peng, Zhaoxiang Zhang, Songyang Zhang, and Kai Chen. Hellobench: Evaluating long text generation capabilities of large language models, 2024.
- [Res24] Microsoft Research. Graphrag auto-tuning provides rapid adaptation to new domains. <https://www.microsoft.com/en-us/research/blog/graphrag-auto-tuning-provides-rapid-adaptation-to-new-domains/>, 2024. Accessed: 2025-07-12.
- [RPJL19] Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling, 2019.
- [SDS⁺21] Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. Perturbation checklists for evaluating nlg evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in*

Bibliography

- Natural Language Processing (EMNLP)*, pages 7219–7234. Association for Computational Linguistics, November 2021.
- [sed21] sedthh. `gutenberg_english`. https://huggingface.co/datasets/sedthh/gutenberg_english, 2021. Accessed: 2025-06-05.
- [SG79] Nancy Stein and Christine Glenn. An analysis of story comprehension in elementary school children. *New Directions in Discourse Processing*, 2, 01 1979.
- [SPS⁺19] Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. Do massively pretrained language models make better storytellers? In Mohit Bansal and Aline Villavicencio, editors, *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 843–861, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [SSS⁺25] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications, 2025.
- [TDM⁺19] Pradyumna Tambwekar, Murtaza Dhuliawala, Lara J. Martin, Animesh Mehta, Brent Harrison, and Mark O. Riedl. Controllable neural story plot generation via reward shaping. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-2019*, page 5982–5988. International Joint Conferences on Artificial Intelligence Organization, August 2019.
- [THL⁺24] Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. Are large language models capable of generating human-level narratives? *arXiv preprint arXiv:2407.13248*, 2024.
- [TNG⁺24] Pittawat Taveekitworachai, Chollakorn Nimpattanavong, Mustafa Can Gursesli, Antonio Lanata, Andrea Guazzini, and Ruck Thawonmas. Multi-verse of greatness: Generating story branches with llms, 2024.
- [TSD24] John Teller, Jane Smith, and Richard Doe. Evaluating thematic coherence in ai-generated stories using topic modeling. *Journal of Artificial Intelligence Research*, 70:123–145, 2024.
- [TSG⁺24] Chongyang Tao, Tao Shen, Shen Gao, Junshuo Zhang, Zhen Li, Zhengwei Tao, and Shuai Ma. Llms are also effective embedding models: An in-depth overview, 2024.
- [Tv85] Tom Trabasso and Paul van den Broek. Causal thinking and the representation of narrative events. *Journal of Memory and Language*, 24(5):612–630, 1985.

- [WFH⁺23] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt, 2023.
- [WHL⁺24] Qianyue Wang, Jinwu Hu, Zhengping Li, Yufeng Wang, daiyuan li, Yu Hu, and Mingkui Tan. Generating long-form story using dynamic hierarchical outlining with memory-enhancement, 2024.
- [WMY⁺25] Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. Writingbench: A comprehensive benchmark for generative writing, 2025.
- [WWS⁺23] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.
- [WYZ⁺24] Danqing Wang, Kevin Yang, Hanlin Zhu, Xiaomeng Yang, Andrew Cohen, Lei Li, and Yuandong Tian. Learning personalized alignment for evaluating open-ended text generation, 2024.
- [XCL23] Zhuohan Xie, Trevor Cohn, and Jey Han Lau. The next chapter: A study of large language models in storytelling. In C. Maria Keet, Hung-Yi Lee, and Sina Zarrieß, editors, *Proceedings of the 16th International Natural Language Generation Conference*, pages 323–351, Prague, Czechia, September 2023. Association for Computational Linguistics.
- [XLCL23] Zhuohan Xie, Miao Li, Trevor Cohn, and Jey Lau. Deltascore: Fine-grained story evaluation with perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5317–5331, Singapore, December 2023. Association for Computational Linguistics.
- [XPQ⁺25] Haotian Xia, Hao Peng, Yunjia Qi, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. Storywriter: A multi-agent framework for long story generation, 2025.
- [YCSR22] Anbang Ye, Christopher Cui, Taiwei Shi, and Mark O. Riedl. Neural story planning, 2022.
- [YHW⁺25] Qiang Yi, Yangfan He, Jianhui Wang, Xinyuan Song, Shiyao Qian, Xinhang Yuan, Li Sun, Yi Xin, Jingqun Tang, Keqin Li, Kuan Lu, Menghao Huo, Jiaqi Chen, and Tianyu Shi. Score: Story coherence and retrieval enhancement for ai narratives, 2025.
- [YJ24] Dingyi Yang and Qin Jin. What makes a good story and how can we measure it? a comprehensive survey of story evaluation, 2024.

Bibliography

- [YRB⁺22] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. Qa-gnn: Reasoning with language models and knowledge graphs for question answering, 2022.
- [YWL⁺23] Wang You, Wenshan Wu, Yaobo Liang, Shaoguang Mao, Chenfei Wu, Mao-song Cao, Yuzhe Cai, Yiduo Guo, Yan Xia, Furu Wei, and Nan Duan. Eipe-text: Evaluation-guided iterative plan extraction for long-form narrative text generation, 2023.
- [ZJC⁺23] Wangchunshu Zhou, Yuchen Eleanor Jiang, Peng Cui, Tiannan Wang, Zhenxin Xiao, Yifan Hou, Ryan Cotterell, and Mrinmaya Sachan. Recurrentgpt: Interactive generation of (arbitrarily) long text, 2023.
- [ZLCF24] Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets llm finetuning: The effect of data, model and finetuning method, 2024.

Acronyms

AI Artificial Intelligence. 2, 3, 10

ASE automatic story evaluation. 2–4, 6, 10

ASG automatic story generation. 1–4, 6, 10, 55, 61

FS few-shot. 18, 22

GRAG graph retrieval-augmented generation. vii, ix, 6–8, 15, 16, 18–33, 35, 46, 59–62

KG knowledge graph. 6, 15, 20, 22, 24, 28–30, 35

LLMs large language models. vii, 1–3, 7, 9–14, 16–18, 20, 25, 26, 28, 29, 39, 55, 59–61

NLP natural language processing. 3, 4

PTPC perturbed texts from positive text continuations. 42–45, 54

QA Question Answering. 4, 6

RAG retrieval-augmented generation. ix, 2, 4–7, 12, 14, 15, 18, 20, 21, 23, 25, 26, 29, 40, 45, 46, 48, 54, 57, 59, 61, 62

SMB source material book. 18, 23–26, 28–31, 36–38, 40, 42, 44, 48, 50, 55, 57–59

SNG social network graphs. 15, 20, 30

TTV text to validate. ix, 18, 22, 24, 28, 30, 31, 49, 50, 55, 59

ULNT ultra-long narrative texts. 2, 35

ULTG ultra-long text generation. 2, 24

ZS zero-shot. ix, 18, 19, 38, 42, 46

A. Appendix

Baseline results:

Causal Coherence (Score: 8) Story B maintains a strong causal coherence with Book A, particularly in its continuation of the plot involving Isabella's flight and the reactions of Manfred and Jerome. The events follow logically from the preceding chapters, with Manfred's suspicion and Jerome's concern for Theodore driving the narrative forward. The introduction of the false report of Hippolita's death adds a layer of intrigue that aligns with the Gothic tone of the original text.

Thematic Coherence (Score: 9) Thematic coherence is well-preserved in Story B, with themes of tyranny, fate, and supernatural intervention echoing those in Book A. The focus on Manfred's oppressive rule, Jerome's moral dilemmas, and Isabella's precarious situation reinforces the original themes. The false rumor of Hippolita's death also ties into the theme of deception and the fragility of life, which are central to the Gothic narrative of Book A.

Cultural Coherence (Score: 7) Cultural coherence is maintained through the continued emphasis on medieval settings, chivalric codes, and religious influences. However, Story B slightly diverges by not fully integrating the supernatural elements (like the enchanted helmet) that are prominent in Book A. The focus shifts more to human intrigue and moral conflict, which, while still fitting, lacks some of the Gothic cultural markers of the original.

Provided Context of the graph Rag system of global context aware validation of a text sample of the Book: The Castle of Otranto

Narrative Context and Divergences

****Established Context**** The broader narrative context involves a complex interplay of supernatural events, political challenges, and familial conflicts that shape the community's dynamics. Key elements include:

1. ****Supernatural Prophecies**** Manfred's actions are heavily influenced by a prophecy threatening his lineage, alongside other supernatural occurrences within the castle. [Data: Reports (9, 10, 20)]
2. ****Political Struggles**** The Herald's challenge to Manfred's rule and tensions over succession are central to the community's conflicts. [Data: Reports (2, 8, 13)]
3. ****Themes of Refuge and Escape**** Broader themes of seeking safety—such as Isabella's flight from forced marriage—are tied to these larger political and supernatural forces. [Data: Reports (5, 15, 19)]

—
****Where the Provided Story Fits**** The story aligns partially with the established context by:

- Focusing on ****Isabella's flight**** to avoid marriage to Manfred, a recurring motif. [Data: Reports (5, 15, 19)]
- Highlighting ****Jerome's protective role**** toward Theodore, though this relationship is less emphasized in the broader narrative.

—
****Divergences from the Established Context****

1. ****Lack of Supernatural Elements**** The story omits critical supernatural influences (e.g., the prophecy driving Manfred's actions) that define the community's tensions. [Data: Reports (9, 10, 20)]
2. ****Narrower Scope**** It centers on Jerome's personal anxieties and mistaken rumors, sidelining the ****political stakes**** (e.g., the Herald's challenge) and ****familial power struggles****. [Data: Reports (2, 8, 13)]
3. ****Thematic Simplification**** While refuge is a theme, the story reduces it to Isabella's immediate escape, missing the broader societal implications (e.g., sanctuary-seeking tied to supernatural or political threats). [Data: Reports (5, 15, 19)]

—
****Implications**** The provided story isolates interpersonal conflicts from the larger narrative framework, potentially undercutting the stakes. For instance, Manfred's rage is portrayed as personal rather than rooted in the prophecy's existential threat to his rule. This divergence may weaken the coherence of the broader plot.

—
****Key Missing Elements**** - The prophecy's role in Manfred's actions - The Herald's political challenge - Supernatural events in the castle
****Data References****: Reports (1, 2, 5, 8, 9, 10, 13, 15, 19, 20)

Result derived from the context:

Causal Coherence (Score: 6) The summary maintains basic cause-effect logic (e.g., Jerome's fear of Manfred's wrath driving his search for Isabella), but lacks deeper causal ties to the supernatural and political forces that shape the broader narrative.

Thematic Coherence (Score: 5) The summary captures themes of refuge and personal conflict (Isabella's flight, Jerome's protection of Theodore), but omits the overarching supernatural and political themes critical to the story's depth.

Cultural Coherence (Score: 4) The summary hints at cultural themes like protection and escape, but fails to engage with broader cultural narratives (e.g., prophecy, political legitimacy) that define the story's world.

Results from the graphrag analysis validation

Causal Coherence Analysis

Pros: Aligned with Source Material

Jerome's Role as Mediator The source material establishes Jerome as a moral authority and mediator between Manfred and Isabella. His distress over conflicting reports fits his tendency to intervene in crises, though the specific conflict about Hippolita's death is not directly supported. [Data: Reports (1); Entities (81); Relationships (123, 163)]

Isabella's Flight Isabella's disappearance aligns with her pattern of fleeing Manfred's tyranny. Her fear that Manfred caused Hippolita's death is a logical (though unstated) motive for escape. [Data: Reports (5, 8); Entities (18, 77); Relationships (13, 59)]

Manfred's Authoritarian Actions His command to open the castle gates reflects his volatile nature and is consistent with his established behavior. [Data: Sources (4, 22); Entities (15); Relationships (148)]

Theodore's Peril Jerome's concern for Theodore's safety matches the source, where Manfred does threaten Theodore. [Data: Relationships (148); Sources (22)]

Cons: Gaps in Causal Logic

Hippolita's Alleged Death She is alive in the source and central to the moral tension around divorce. Her death is fictional and unsupported. [Data: Entities (19); Relationships (12, 127)]

Jerome's Sudden Distress While he cares for Isabella and Theodore, there is no basis in the text for his frantic behavior over rumors of Hippolita's death. [Data: Reports (1); Relationships (163)]

Manfred's Suspected Role in Death He aims to divorce Hippolita, not kill her. Accusing him of murder has no source justification. [Data: Relationships (12, 127)]

Monks Mobilizing Against Manfred The story depicts Jerome acting largely alone. A collective uprising is a new invention. [Data: Reports (1)]

Conclusion

The story preserves core motivations but fabricates key causal events (e.g., Hippolita's death, monk mobilization) that disrupt narrative logic. For better coherence, new conflicts should be grounded in established text or foreshadowed with precedent.

Cultural Coherence Analysis Pros: Cultural Alignment with Source Religious Mediation Jerome's actions as a friar reflect his canon role as a moral guide. [Data: Entities (81); Relationships (118, 123); Sources (17, 22)]

Manfred's Tyranny His authoritarian traits and false accusations are well aligned. [Data: Entities (15); Relationships (109); Reports (3); Sources (4)]

Isabella's Flight Her escape reflects earlier incidents where she fled Manfred's advances. [Data: Entities (18, 77); Reports (8); Entities (15, 19)]

Cons: Cultural Deviations

Hippolita's Alleged Death Contradicts her core role as a virtuous, submissive figure in the divorce conflict. [Data: Entities (19); Relationships (123)]

78 **Jerome's Panic** Overstates his response. In the original, he acts alone through spiritual rather than militant means. [Data: Sources (17); Relationships (109, 118)]

...

Graph Comparison result:

Causal Coherence Score: 4.78

The analysis lacks direct causal support from the source material. Isabella's actions are driven by immediate threats rather than any influence from Princess Sanchia of Aragon, making the cause-effect logic weak.

The statement lacks clear cause-effect relationships as there is no evidence in the source material of Jerome meeting another monk in a porch. No direct causal links or narrative consequences are tied to this specific event.

The analysis lacks explicit cause-effect relationships linking Jerome to the convent or its brethren. His roles as a mediator and father are highlighted, but these do not directly tie to convent activities.

The statement lacks direct causal evidence in the source material. While Hippolita's piety and associations with religious institutions are noted, there is no explicit mention of convent brethren praying for her soul.

The statement lacks direct causal coherence as there is no explicit mention of Jerome spreading false news about Hippolita's death in the source material. The narrative focuses on other conflicts and themes.

The statement lacks clear causal links in the source material. There is no evidence of Martelli spreading false news about Hippolita's death, and the documented events do not support this claim.

The analysis demonstrates strong causal coherence, clearly linking Manfred's tyrannical rule and obsession to his actions and Isabella's defiance. The cause-effect relationships are well-articulated and supported by narrative evidence.

The analysis demonstrates strong causal coherence, clearly linking the Friar's role as a mediator and spiritual guide to his advice to Isabella. The cause-effect chain is well-supported by the narrative context, showing how his intervention influences her actions, such as seeking sanctuary.

The analysis demonstrates strong causal coherence, clearly linking Jerome's concern for Theodore to Manfred's actions and threats. The cause-effect relationships are well-supported by narrative evidence, though a minor deduction is made for not explicitly addressing potential alternative interpretations or counterarguments.

The statement lacks a clear cause-effect chain, particularly around the "false news" element. While Jerome's actions are documented in relation to Isabella's sanctuary and Manfred's tyranny, there's no evidence linking his concern to misinformation.

The statement demonstrates strong causal coherence, clearly linking Isabella's actions to subsequent events in the narrative. The cause-effect relationships are well-established, particularly in how her flight from Manfred's pursuit leads to key plot developments. The references to specific data points (e.g., Reports, Relationships) further solidify this coherence.

The statement partially aligns with the data, as Jerome's distress over Hippolita's well-being is supported by his moral opposition to Manfred's actions. However, the specific claim of "false news of her death" lacks direct evidence, weakening the causal logic.

The statement "Princess Hippolita resided in the castle" is a static description and lacks explicit cause-effect logic. However, it indirectly connects to broader causal dynamics, such as her submission to Manfred's tyranny and her moral conflicts.

Complex question result:

Thematic Coherence Score: 10.0

Questions & Answers Q1: Does the extension maintain the theme of religious influence and monastic life present in the original story? A: Yes, the extension maintains the theme of religious influence and monastic life, as Isabella seeks refuge in the Church of St. Nicholas and considers the convent as a permanent sanctuary [Data: Reports (5, 15); Entities (46, 55, 82); Relationships (51, 55)].

Q2: Does the extension align with the original's portrayal of courtly intrigue and familial tension? A: Yes, the extension maintains the original's themes of courtly intrigue and familial tension, focusing on Manfred's tyrannical rule and the conflicts surrounding Isabella's flight [Data: Reports (8, 1, 0); Entities (15, 18, 77); Relationships (13, 59, 117, 12, 10)].

No, the extension lacks the depth of political maneuvering and supernatural elements present in the original, simplifying the narrative [Data: Reports (8, 1); Entities (36, 49); Relationships (31, 71)].

Commentary: The extension partially aligns with the original's portrayal but diverges in key areas. While it captures Manfred's authoritarianism and Isabella's plight, it omits or reduces elements like the prophecy and supernatural occurrences that enriched the original's tension [Data: Reports (0); Entities (24, 50); Relationships (15, 48)].

Q3: Does the extension respect the original's cultural norms around nobility and the church? A: Yes, the narrative consistently portrays nobility and the church with reverence, aligning with medieval cultural norms where both institutions held significant power and moral authority [Data: Reports (0, 9, 20); Entities (15, 77, 97); Relationships (106, 117, 132)].

No, Manfred's tyrannical actions and disregard for clerical intervention challenge traditional norms, reflecting tension between secular and religious authority [Data: Reports (0, 9, 20); Entities (15, 86); Relationships (111, 132, 181)].

Commentary: The text upholds medieval hierarchies (e.g., Alfonso's veneration, Isabella's sanctuary) while subverting them through Manfred's defiance, mirroring the era's complex power dynamics [Data: Entities (35, 77); Relationships (106, 119)].

Supernatural elements further blur these norms, as prophecies and specters influence noble actions [Data: Reports (0, 9); Entities (24); Relationships (15)].

Cultural Coherence Score: 3.33

Questions & Answers Q1: Does the monks' reaction to the Princess's rumored death align with their religious and social norms? A: No, the monks' reaction does not align with their religious and social norms, as they show more concern for their own interests and political maneuvering than for the Princess's well-being.

Commentary:

Jerome is more focused on negotiating with Manfred and securing his own position than on mourning or addressing the Princess's rumored death [Data: Sources (17); Relationships (124)].

Religious figures prioritize political influence over spiritual duty, contradicting expected monastic behavior [Data: Reports (1, 19); Entities (10, 97)].