UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
COORDENAÇÃO DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MASTER THESIS

LLMS AS TOOLS FOR EVALUATING TEXTUAL COHERENCE: A COMPARATIVE ANALYSIS

BRYAN KHELVEN DA SILVA BARBOSA

Supervisor: Prof. Cláudio Campelo, PhD

Campina Grande, Paraíba, Brasil
09/2024

Federal University of Campina Grande

Electrical Engineering and Informatics Center

Postgraduate Coordination in Computer Science

LLMs as Tools for Evaluating Textual Coherence: A Comparative Analysis

Bryan Khelven da Silva Barbosa

Thesis submitted to the Coordination of the Postgraduate Course in Computer Science at the Federal University of Campina Grande - Campus I as part of the necessary requirements to obtain the degree of Master in Computer Science.

Concentration Area: Computer Science

Research Line: Natural Language Processing

Prof. Cláudio Elízio Calazans Campelo, PhD
(Supervisor)

Campina Grande, Paraíba, Brazil
09/2024

B238l Barbosa, Bryan Khelven da Silva.
LLMs as tools for evaluating textual coherence: a comparative analysis / Bryan Khelven da Silva Barbosa. – Campina Grande, 2024.
117 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2024.
"Orientação: Prof. Dr. Cláudio Elízio Calazans Campelo".
Referências.

1. Textual Coherence. 2. Incoherence. 3. Comparison. 4. NLP. I. Campelo, Cláudio Elízio Calazans. II. Título.

CDU 004(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA MEIRE EMANUELA DA SILVA MELO CRB-15/568

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
POS-GRADUACAO EM CIENCIA DA COMPUTACAO
Rua Aprígio Veloso, 882, Edifício Telmo Silva de Araújo, Bloco CG1, - Bairro Universitário, Campina Grande/PB, CEP 58429-900
Telefone: 2101-1122 - (83) 2101-1123 - (83) 2101-1124
Site: http://computacao.ufcg.edu.br - E-mail: secpg@computacao.ufcg.edu.br

FOLHA DE ASSINATURA PARA TESES E DISSERTAÇÕES

BRYAN KHELVEN DA SILVA BARBOSA

LLMS AS TOOLS FOR EVALUATING TEXTUAL COHERENCE: A COMPARATIVE ANALYSIS

Dissertação ao Programa de Pós-Graduação em Ciência da Computação como pré-requisito para obtenção do título de Mestre em Ciência da Computação.

Aprovada em: 05/09/2024

Prof. Dr. CLÁUDIO ELÍZIO CALAZANS CAMPELO, Orientador, UFCG

Prof. Dr. LEANDRO BALBY MARINHO, Examinador Interno, UFCG

Prof. Dr. MARLO VIEIRA DOS SANTOS E SOUZA, Examinador Externo, UFBA

Documento assinado eletronicamente por **CLAUDIO ELIZIO CALAZANS CAMPELO**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 09/09/2024, às 09:58, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **LEANDRO BALBY MARINHO**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 09/09/2024, às 11:40, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **Marlo Vieira dos Santos e Souza**, **Usuário Externo**, em 10/09/2024, às 14:51, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

A autenticidade deste documento pode ser conferida no site https://sei.ufcg.edu.br/autenticidade, informando o código verificador **4773698** e o código CRC **72759782**.

Acknowledgments

I express my deepest gratitude to the Cosmic for the countless possibilities that have been granted to me throughout this journey.

I acknowledge the invisible hand of the Great Architect of the Universe, who has silently guided me through the challenges and lessons of this path. The eternal wisdom, echoing through the ages, has been a beacon of light, illuminating my way and allowing me to glimpse the perfection and justice inherent in the governance of the All.

A special thanks to my supervisor, Cláudio Campelo, for your patience, trust, partnership, and unwavering dedication to both me and this research. Your guidance was precisely what I needed, and for that, I am deeply grateful.

I extend my gratitude to my friends and colleagues at the Lacina laboratory for all the discussions, opinions, and contributions that enriched this research. Your insights have been invaluable.

To my family and friends, your unconditional support, patience, and encouragement have been my foundation during the most challenging moments of this journey. I cannot thank you enough.

I would also like to acknowledge the support of every member of the Graduate Program in Computer Science at UFCG. Your assistance has been crucial throughout this process.

Finally, I also gratefully acknowledge the financial support provided by the Coordination for the Improvement of Higher Education Personnel (CAPES), which has been instrumental in my academic development.

With my deepest feelings,
Thanks!
Bryan Khelven

Resumo

A coerência textual é fundamental para a compreensão eficaz, determinando a clareza, a compreensibilidade e a qualidade geral do conteúdo. Modelos de Linguagem de Larga Escala (LLMs) recentes, treinados em corpora extensivos, têm demonstrado capacidades impressionantes em produzir textos coerentes e contextualmente relevantes, aumentando seu potencial para tarefas de análise textual. No entanto, a habilidade desses modelos em realizar a análise de coerência em diversos textos de entrada ainda está sob investigação. Neste estudo, avaliamos o desempenho de modelos de linguagem avançados na análise automática de coerência textual. Os modelos avaliados incluem GPT-4o, GPT-3.5, GPT-4, Claude Opus, Claude 3 Sonet, Claude 3 Haiku, Bard, LLaMA 2 13b e LLaMA 2 7b. Nossa pesquisa investigou a capacidade desses modelos em avaliar a coerência textual em diferentes níveis. Primeiro, focamos na coerência local, que se refere à consistência lógica e contextual entre sentenças adjacentes ou pequenos segmentos de texto. Nossos resultados indicam que GPT-4o, Claude Opus e Gemini se destacam nessa tarefa, demonstrando desempenho superior na manutenção da continuidade temática e fluência entre sentenças consecutivas. Em seguida, exploramos a coerência global, que envolve a consistência lógica e temática de textos inteiros. Nesse aspecto, o modelo Claude Opus mostrou-se o mais eficaz, garantindo que o texto mantenha um fluxo consistente e lógico do começo ao fim. Por fim, examinamos a capacidade dos modelos em identificar incoerências, como elementos ou segmentos que quebram a continuidade lógica e temática. Nessa tarefa, o GPT-4o se destacou, mostrando uma acuidade excepcional na detecção e sinalização de incoerências. Esse aspecto é crucial para aplicações onde precisão e clareza são necessárias, como na escrita assistida por IA e na revisão de textos. Nossa análise comparativa oferece insights sobre as capacidades e limitações dos modelos de linguagem de grande porte atuais na análise de coerência textual. Além disso, nossos achados contribuem para a compreensão de como esses modelos podem ser aplicados em diversos contextos de processamento de linguagem natural, promovendo avanços contínuos neste campo.

Palavras-chave: Coerência Textual. Incoerência. Comparação. PLN.

Abstract

Textual coherence is fundamental for effective comprehension, determining the clarity, comprehensibility, and overall quality of content. Recent Large Language Models (LLMs), trained on extensive corpora, have demonstrated impressive capabilities in producing coherent and contextually relevant texts, enhancing their potential for textual analysis tasks. However, the ability of these models to perform coherence analysis on various input texts is still under investigation. In this study, we evaluate the performance of advanced language models in automatic textual coherence analysis. The models evaluated include GPT-4o, GPT-3.5, GPT-4, Claude Opus, Claude 3 Sonet, Claude 3 Haiku, Bard, LLaMA 2 13b, and LLaMA 2 7b. Our research investigates the ability of these models to evaluate textual coherence at different levels. First, we focus on local coherence, which refers to the logical and contextual consistency between adjacent sentences or small text segments. Our results indicate that GPT-4o, Claude Opus, and Gemini excel in this task, demonstrating superior performance in maintaining thematic continuity and fluency between consecutive sentences. Next, we explore global coherence, involving the logical and thematic consistency of entire texts. Here, the Claude Opus model proved to be the most effective, ensuring that the text maintains a consistent and logical flow from beginning to end. Finally, we examine the models' ability to identify incoherences, such as elements or segments that break the logical and thematic continuity. In this task, GPT-4o stood out, showing exceptional acuity in detecting and flagging incoherences. This aspect is crucial for applications where precision and clarity are needed, such as AI-assisted writing and text review. Our comparative analysis provides insights into the capabilities and limitations of current large language models in textual coherence analysis. Additionally, our findings contribute to understanding how these models can be applied in various natural language processing contexts, promoting continuous advancements in this field.

Keywords: Textual Coherence. Incoherence. Comparison. NLP.

List of Figures

List of Tables

List of Prompts

Abbreviations and Acronyms

AI – Artificial Inteligence
API – Application Programming Interface
BERT – Bidirectional Encoder Representations from Transformers
CNN – Convolutional Neural Network
COCA – Corpus of Contemporary American English
CSV – Comma-Separated Values
CST – Cross-document Structure Theory
DDisCo – Danish Discourse Coherence Dataset
DO – Original Document
EDUs – Elementary Discourse Units
FN – False Negatives
FP – False Positives
GPT – Generative Pre-trained Transformer
GPT-4o – GPT-4 optimized
GPU – Graphics Processing Unit
GCDC – Grammarly Corpus of Discourse Coherence
HMM – Hidden Markov Model
ID – Identifier
IN_ – Prefix for Incoherence Identification
LLaMA – Large Language Model Meta AI
LLMs – Large Language Models
ML – Machine Learning
MLM – Masked Language Modeling
MTL – Multi-Task Learning
MTurk – Amazon Mechanical Turk
MS-MARCO – Microsoft Machine Reading Comprehension
NLP – Natural Language Processing
NLU – Natural Language Understanding
NSP – Next Sentence Prediction
permDO – Permutated Document
RNN – Recurrent Neural Network
RST – Rhetorical Structure Theory
RQ1-4 – Research Questions 1 to 4
TP – True Positives
TPU – Tensor Processing Unit
TN – True Negatives

Contents

1 Introduction

The concept of coherence lies at the very heart of effective communication, serving as a keystone element that determines the clarity, understandability, and overall quality of textual content (Koch; Travaglia, 2003). Inherent to language, coherence transcends mere syntax or grammar; it encompasses the logical flow of ideas, ensuring that a text is not just a collection of sentences but a unified whole that conveys meaning with precision and subtlety (Freitas, 2013). As the digital age propels us into an era where written text interactions become increasingly prevalent (Hoey, 2013), the ability to automatically analyze textual coherence has emerged as a pertinent task within the field of Natural Language Processing (NLP).

The advent of Large Language Models (LLMs) such as GPT-3, Llama and Gemini, has revolutionized our approach to generating text that mirrors the nuance and depth of human-written content. These models, trained on vast corpora of text, have demonstrated an unparalleled capacity to produce coherent and contextually relevant text across a myriad of topics. This proficiency in text generation naturally extends to the potential for these models to excel in tasks related to the textual analysis. The underlying hypothesis is straightforward: if a LLM can generate coherent text, it should, by extension, possess a keen ability to discern the coherence – or lack thereof – in existing texts.

In the context of computational linguistics, textual coherence is defined by the logical and orderly sequence in which ideas are presented within a text, ensuring that information and arguments are conveyed in a comprehensible and fluid manner (Seno; Rino, 2005). This involves not just the superficial connection between sentences through discourse markers or transition words but also a deeper harmony in terms of theme, purpose, and shared knowledge between the author and the reader (Charolles, 1978). For NLP systems, assessing the coherence of a text implies understanding how its constituent parts – whether at the sentence, paragraph, or document level – come together to form a unified whole that is logically

consistent and aesthetically pleasing (Jurafsky; Martin, 2024). This definition highlights the complexity of the textual coherence analysis task, underscoring it as a challenge within the field.

Before moving forward, however, it is necessary to discuss key theories and models previously proposed in the domain of textual coherence analysis. Historically, coherence has been conceptualized through various lenses, ranging from rhetorical structure theory (RST) (Mann; Thompson, 1987), which posits that text coherence is derived from the hierarchical organization of text units, to the Centering Theory (Grosz; Joshi; Weinstein, 1995), emphasizing the role of discourse entities and their continuity across sentences. Computational approaches have evolved from rule-based systems, relying on explicit coherence markers and structural patterns, to sophisticated machine learning algorithms that leverage vast datasets to infer coherence implicitly (Jurafsky; Martin, 2024). Notably, the development of neural network-based models, especially those employing attention mechanisms, such as BERT (Devlin *et al.*, 2018), has marked a significant advancement, allowing for a deeper understanding of contextual relationships within texts.

Following the exploration of key theories and models, it is pivotal to introduce the concept of automatic coherence analysis and underscore its relevance in the realm of computational linguistics. Automatic coherence analysis refers to the use of algorithms and computational models to assess the logical flow and unity of a text without human intervention (Jurafsky; Martin, 2024). This technological aproach enables the processing of large volumes of text at speeds unattainable by human reviewers, facilitating tasks such as content summarization, quality control in content generation, and automated essay grading (Shermis; JC, 2003). Additionally, automatic coherence analysis serves as a litmus test for the sophistication of natural language understanding (NLU) within AI systems, whereas the ability of a model to discern coherence reflects its depth of linguistic insight, mirroring its potential to grasp complex human communication patterns (Jurafsky; Martin, 2024).

That said, the importance of textual coherence analysis cannot be overstated. Beyond its applications in automated essay scoring (Hearst, 2000) and content generation (Marchenko *et al.*, 2020), coherence analysis is an essential task on enhancing machine understanding of human language. It aids in summarizing content (Mani; Bloedorn; Gates, 1998), translating languages (Hadla, 2015), and even detecting nuances that differentiate high-quality text from mediocre or disjointed writings (Brito; Oliveira, 2023). This task, therefore, challenges the computational capabilities of LLMs but also probes the depth of their linguistic understanding while pushing the boundaries of what machines can achieve in terms of processing and generating human-like text (Liusie; Manakul; Gales, 2024).

Given the intrinsic complexity of coherence as a linguistic feature (Rakhimova;

Djumanazarova; Bobojonova, 2019), this study explores the capabilities of various LLMs in analyzing textual coherence. Specifically, it investigates their effectiveness in three main tasks: classifying texts as locally or globally coherent or incoherent, and identifying specific incoherent segments within texts. Through this analysis, the study aims to contribute to the ongoing dialogue on enhancing NLP technologies and advancing our understanding of how machines deal with the subtleties of human language.

1.1 Motivation

By evaluating the performance of models such as like GPT-3 (Brown *et al.*, 2020), Llama (Touvron *et al.*, 2023), Gemini (GEMINI TEAM *et al.*, 2024), this study aims to uncover the strengths and weaknesses of each in identifying and assessing coherence within texts. This involves both classifying texts on a coherence scale and dissecting the models' ability to pinpoint the sources of incoherence, thus providing a deeper sight into their understanding of language structure and logic. The comparison purposes on advancing our comprehension of current NLP capabilities and setting the stage for future innovations in the field.

In the light of the significance of coherence in ensuring effective communication, there is a pressing need to further investigate the mechanisms by which language models, particularly LLMs, access and maintain coherence in text. The digital era has transformed how we interact (Hoey, 2013), making text-based communication more prevalent and heightening the need for texts that engage and also convey clear and coherent information. This motivation stems from the realization that as we depend more on digital texts for education, work, and personal communication, the ability of systems to ensure textual coherence becomes more critical.

As LLMs like GPT-3 (Brown *et al.*, 2020), Llama (Touvron *et al.*, 2023), and Gemini (GEMINI TEAM *et al.*, 2024) continue to set benchmarks in generating and evaluating text, the next logical step is to harness these capabilities to improve the assessment of textual coherence. The complexity of determining coherence involves not merely linking sentences but weaving them into a tapestry that resonates with thematic unity and logical consistency (Koch; Travaglia, 2003). The motivation for this study arises from observing the gap between human and machine understanding of text coherence and the potential to close this gap through advanced computational techniques.

Moreover, as AI systems are increasingly employed in roles that require nuanced language handling such as tutoring systems (Lin; Huang; Lu, 2023), customer service bots (Hsu; Lin, 2023), and content creation tools (Larsson; Lindecrantz, 2023; Hutson; Lang, 2023), the need to enhance their ability to process and produce coherent text is undeniable. This

exploration is driven by the goal to elevate the quality of human texts to new heights, ensuring that AI systems can detect incoherences and support humans to the level of coherence that is traditionally expected in effective communications.

Understanding and enhancing textual coherence is not merely a theoretical pursuit; it has substantial practical applications. For example, improving the clarity of educational materials and refining automated text generation systems depend heavily on effective coherence management. This study examines how LLMs handle textual coherence and incoherence, offering insights that can directly contribute to developing advanced tools in education, content creation, automated customer service, and communication systems. These improvements aim to elevate interaction quality and user satisfaction in the digital age.

The findings from this study also lay the groundwork for future research aimed at bridging the gap between human and machine understanding of text coherence. By identifying how LLMs manage coherence and incoherence identification, researchers and developers can further refine these models to better emulate human comprehension, enhancing AI usability in fields ranging from automated writing assistants to sophisticated NLP systems in both academia and industry, driving improvements in how AI interacts with and understands human language.

1.2 Objectives and Research Questions

The primary purpose of this work is to conduct a comparative study of the effectiveness of different LLMs for textual coherence analysis, with specific attention to the following aspects:

- **Evaluating the Performance of LLMs:** Assessing how the models GPT 3.5, GPT 4, GPT 4o, Claude 3 Opus, Claude 3.5 Sonnet, Claude 3 Haiku, Gemini, LLaMA 2 13b, LLaMA 2 7b, and Bard perform in analyzing textual coherence, measuring how each model's deals with the task of classifying texts coherence.
- **Testing Both Global and Local Aspects of Coherence:** Examining how well the models handle global coherence – ensuring that the text is coherent as a whole, as well as local coherence – focusing on the coherence between adjacent sentences or within paragraphs.
- **Analyzing Models' Ability to Detect and Label Sources of Incoherence:** Focusing on how well each model can identify and label specific elements or parts of texts that contribute to overall incoherence, including an examination of the models' capacity to understand and interpret language structure and logic.

- **Compare Coherence Analysis Across API and Chatbot Interfaces:** Examine how each model performs in identifying and assessing textual coherence when accessed via API versus chatbot interfaces.

1.2.1 Research Questions

This work is guided by four research questions, each addressing a specific aspect of textual coherence and incoherence. The detailed research questions are as follows:

RQ1: How effectively can different LLMs evaluate the logical flow and consistency within short text passages? This question seeks to explore the models' performance in Local Coherence Classification, that is, their proficiency in detecting disruptions in the logical sequence and coherence of sentences within a text, particularly in scenarios where the natural order of ideas might be challenged.

RQ2: How do LLMs perform in assessing the overall coherence of entire texts? This question examines the models' performance in Global Coherence Classification, which means the capability to evaluate the coherence of a text as a whole, considering how well they maintain a consistent and logical flow throughout different sections of the text.

RQ3: How accurately can LLMs identify and categorize specific incoherent segments within a text? This question explores the models' effectiveness in pinpointing and classifying different types of incoherence, providing insights into their ability to detect disruptions in the logical flow and thematic continuity of texts.

RQ4: How does the mode of interaction (API vs. chat) affect the ability of LLMs to assess and identify textual coherence? This question investigates the differences in performance between LLMs accessed via API and those accessed through chatbot interfaces, examining how the interaction method influences the models' accuracy and effectiveness in analyzing both local and global coherence as well as identifying incoherent segments within texts.

1.2.2 Contributions

This work contributes to the field of textual coherence analysis by:

- The development of specialized prompts used to analyze textual coherence through LLMs. These prompts are designed to evaluate the coherence of texts at both local and global levels.

- A comparison of various LLMs and other models for the task of textual coherence analysis, diving into the strengths and weaknesses of different models in this domain.
- The creation of annotated corpora specifically designed for the tasks of Global Coherence Classification and Incoherence Identification. This includes 100 texts for Global Coherence Analysis and 130 texts for Incoherence Identification.
- The establishment of a repository containing the prompts and code used for connecting and utilizing the LLMs, serving as a resource for researchers and practitioners in the field.
- The publication of a paper in the proceedings of the 15th Symposium in Information and Human Language Technology – STIL 2024, demonstrating the relevance and contribution of this work to the academic community.

1.3 Research Methodology

This thesis employs a quantitative and experimental approach to evaluate the effectiveness of LLMs in the context of textual coherence analysis. The study is designed to explore and compare the performance of various LLMs across Local Coherence Classification, Global Coherence Classification, and Incoherence Identification tasks. The quantitative aspect of the research relies on the systematic collection and analysis of performance metrics, such as accuracy, F1 scores, and agreement measures like Fleiss' Kappa, to provide an objective comparison of model performance across these tasks. The experimental component involves controlled testing of LLMs through both API-based and chat-based interfaces, allowing for the investigation of how interaction modes influence the models' effectiveness. Additionally, the study incorporates an exploratory aspect by venturing into uncharted areas of LLM evaluation, particularly in the context of coherence tasks, to identify new avenues for further research in NLP.

1.4 Document Structure

The remainder of this work is organized as follows: Chapter 2, named "Theoretical Background" lists and explores basilar concepts and the theoretical underpinnings in the context of this research. Chapter 3, named "Related Work", explores the relevant literature, which reviews the previous research relevant to textual coherence analysis. The "Methodology"

section, or Chapter 4, describes the experimental setup, including the models evaluated and the metrics for assessment. This is followed by "Results and Discussion", or Chapter 5, where the performance of each model in coherence analysis is detailed. The work concludes with the 6th chapter: "Conclusions and Future Work", summarizing the key findings and suggesting areas for further research. A "References" section lists the scholarly works cited throughout the study, providing a resource for additional reading. Appendix A brings the Intermediate Prompts tested to reach the final prompts used during the experiments to obtain the results presented in this study.

1.5 Author's Statement on the Use of Artificial Intelligence (AI) for Text Revision

In the process of reviewing and improving the writing of this dissertation, ChatGPT was employed as a tool for text revision. The use of AI was strictly limited to reviewing pre-existing authored content, without generating new material. ChatGPT was specifically used to perform targeted corrections according to the following procedure:

1. Portions of the manuscript were selected by the author and submitted to ChatGPT using a predefined prompt[1].
2. The tool reviewed the text with the following objectives:
   (a) Analyze the tone of the manuscript to ensure the preservation of the original style and to enhance its cohesion.
   (b) Identify any errors in grammar, spelling, and punctuation.
   (c) Detect redundant words or phrases that could be removed or rephrased to ensure greater conciseness and to improve readability and fluency.
3. ChatGPT provided a comparison between the original and revised versions of the manuscript, including explanations of the changes made, thereby ensuring clarity regarding the suggested areas for improvement.
4. Based on the modification suggestions provided by the tool, the final edits were manually reviewed and implemented as deemed appropriate.

[1] https://www.claudiocampelo.com/supervisees/prompts/academic-writing

The use of AI was aimed solely at refining the structure of the text without altering its content or central ideas, ensuring that the original authorship of the material was fully maintained.

Figure1.1 presents an example containing three texts: (a) original text; (b) result returned by ChatGPT; and (c) final text, after the manual edits made by the author.

Figure 1.1: Example of Text Revision Using AI.

(a) Building on that, \citeonline{sagi2010} emphasize that the nature of discourse relations are closely tied to determining global coherence. The study suggests that the ease with which readers can establish global coherence depends not only on the presence of hierarchical structures but also on how well these structures align with the reader's cognitive expectations and the overall logical flow of the text. This indicates that global coherence is not merely a function of how information is organized but also how effectively it resonates with the reader's mental models of discourse. As such, the findings underscore the importance of considering both structural and relational aspects of discourse in the analysis of global coherence, making it clear that global coherence is a multifaceted concept that extends beyond simple sequential order to include deeper cognitive and genre-specific factors.

(b) Building on this, \citeonline{sagi2010} emphasize that the nature of discourse relations is closely tied to determining global coherence. Their study suggests that readers' ability to establish global coherence depends not only on hierarchical structures but also on how these structures align with cognitive expectations and the text's logical flow. This highlights that global coherence is a multifaceted concept, extending beyond simple sequential order to encompass how effectively the discourse resonates with the reader's mental models, incorporating deeper cognitive and genre-specific factors.

(c) Building on this, Sagi (2010) emphasize that the nature of discourse relations is closely tied to determining global coherence. Their study suggests that readers' ability to establish global coherence depends not only on hierarchical structures but also on how these structures align with cognitive expectations and the text's logical flow. This highlights that global coherence is a multifaceted concept, extending beyond simple sequential order to encompass how effectively the discourse resonates with the reader's mental models, incorporating deeper cognitive and genre-specific factors.

Source: The author, 2024.

2 Theoretical Background

Textual cohesion and coherence are central to discourse analysis and NLP, as they explain how texts are structured and interpreted. This chapter begins with clear definitions and distinctions between these concepts, grounded in the seminal works of Halliday and Hasan (1976) and Van Dijk (1977). It then explores theoretical models like Rhetorical Structure Theory and Centering Theory, which provide frameworks for understanding text coherence. The chapter also reviews key datasets that have shaped research in this area, offering a thorough introduction to the analysis of text structure.

2.1 Introduction and Definitions of Textual Cohesion and Coherence

Cohesion and coherence are two fundamental concepts in textual linguistics and discourse analysis, essential for the construction and interpretation of texts (Halliday; Hasan, 1976; Van Dijk, 1977). Both concepts are intrinsically related to how ideas are connected in a text and how these connections contribute to the overall understanding of the text by the reader.

Textual cohesion, as defined by Halliday and Hasan (1976), refers to how the parts of a text are connected to each other through various linguistic relations. These relations can be grammatical, such as the use of pronouns, ellipses, and conjunctions, or lexical, such as the use of synonyms, repetitions, and collocations. For example, in a sentence like "John likes to read. He always has a book with him," the word "he" is a pronoun referring to "John," creating a cohesive connection between the two sentences.

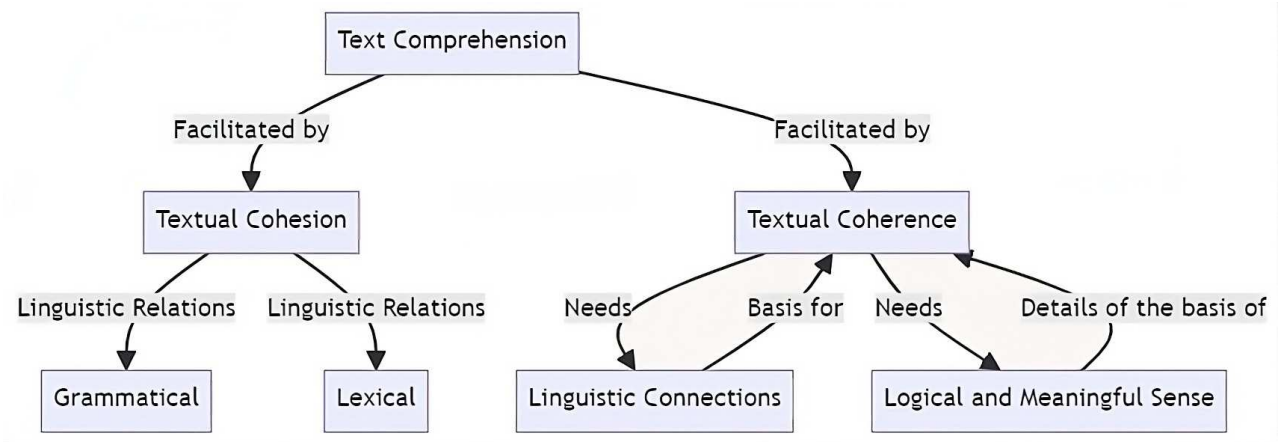
Textual coherence, on the other hand, is a more abstract and complex concept. According to Van Dijk (1977), textual coherence not only refers to linguistic connections

within a text but also to how the ideas in a text fit together into a meaningful whole. In other words, a text is coherent if all its parts fit together logically and meaningfully. For example, in a text like "John likes to read. He always goes to the library. Reading is one of his favorite activities," all the sentences are related to the central idea that João likes to read, making the text coherent.

It is important to distinguish between the meanings of cohesion and coherence to achieve an understanding of how texts are constructed and interpreted. As noted by Koch (1994, p. 18), a text can be cohesive without being coherent. For example, a text consisting of sentences like "The cat is on the roof. The roof is red. Red is a color. The color of my car is blue" is cohesive because each sentence is linguistically connected to the previous one, but it is not coherent because the sentences do not fit into a meaningful whole.

The importance of cohesion and coherence in text comprehension is widely recognized in linguistic literature. According to McNamara and Kintsch (1996), cohesion and coherence are necessary to facilitate reader comprehension because they help guide the reader through the flow of ideas in the text.

Figure 2.1: Relations between textual cohesion and coherence.



Source: The author, 2024.

Figure 2.1 illustrates the interrelationship between textual cohesion and coherence, fundamental concepts in text analysis. Textual cohesion, as defined by (Halliday; Hasan, 1976), is the linguistic connection between parts of a text. These connections can be grammatical, such as the use of pronouns, ellipses, and conjunctions, or lexical, such as the use of synonyms, repetitions, and collocations. Textual cohesion is, therefore, an essential aspect of text structure, contributing to its fluency and readability.

On the other hand, textual coherence is a more abstract and complex concept. According to Koch and Travaglia (2003, p. 53), "coherence is the continuity of meaning

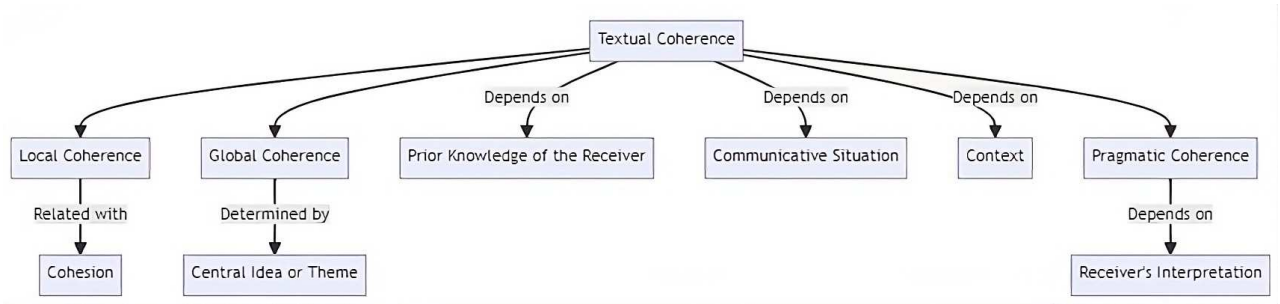
perceptible in the text, resulting in a conceptual-cognitive connection among the elements of that text." In other words, a text is coherent if all its parts fit together in a logical and meaningful way. Textual coherence is directly related to the comprehension of the meanings conveyed by the text to the reader in a given communicational context.

Textual coherence can be seen at two levels: local and global, as established by the work of Van Dijk and Kintsch (1983) or as also presented by Charolles (1978, p. 31), where the author states: "[...] the coherence of a statement must be jointly determined from a local and global point of view." The authors point out that local coherence relates to sentences and parts of the text that establish linear connections between them, commonly occurring between one and six sentences, while global coherence, on the other hand, concerns the entire text or its essence, central idea, or theme occurring in larger sections than those of local coherence.

Cohesion is closely related to local coherence, sometimes referred to as microstructural or sequential coherence, as discussed by the work of Charolles (1978). Global coherence, on the other hand, is what most authors, including Koch and Travaglia (2003), understand as coherence. Global coherence is determined by the theme or central idea of the text and how all parts of the text relate to that theme or central idea.

Coherence depends not only on the text but also on the reader's prior knowledge, the communicational situation, and the context in which the text appears, and this dependence, along with the interpretation given by the reader, is called pragmatic coherence (Marinho, 2016).

Figure 2.2: Textual Coherence aspects.

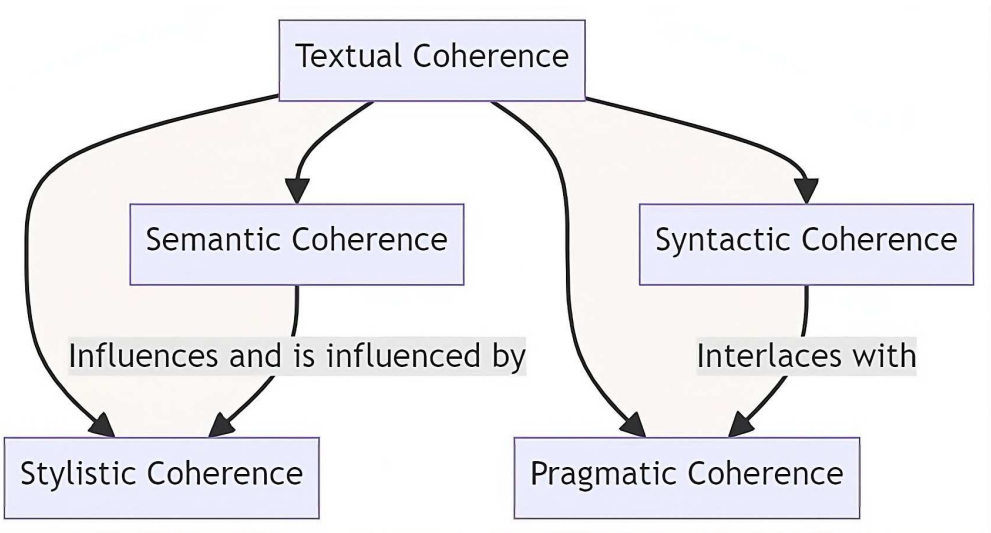


Source: The author, 2024.

The diagram presented in Figure 2.2 summarizes the general structure of textual coherence presented so far. In it, the central node is Textual Coherence, which connects to Local Coherence and Global Coherence. Local Coherence is related to Cohesion, while Global Coherence is determined by the Central Idea or Theme. Furthermore, Textual Coherence depends on the Reader's Prior Knowledge, the Communicational Situation, and the Context. The reader's interpretation is represented by Pragmatic Coherence.

In the work of Van Dijk and Kintsch (1983), the authors shed light on the multiplicity and complexity of aspects that make up textual coherence. They reveal that coherence is not a monolithic entity but rather multifaceted, divided into four distinct perspectives: semantic, syntactic, stylistic, and pragmatic, with each of these dimensions having its peculiarities and importance but all being intrinsically interconnected in the construction of a cohesive and articulated text, as can be seen in Figure 2.3.

Figure 2.3: Perspective divisions of Textual Coherence.



Source: The author, 2024.

An interesting point that the authors address is that incoherence in a text can manifest in multiple perspectives simultaneously. For example, problems in the choice of style can lead to failures in vocabulary use, generating both stylistic and syntactic incoherence. This observation emphasizes the interrelation of the different dimensions of coherence, highlighting the need for a broad mastery of all of them for the production of an effective and well-structured text.

To address these four facets of textual coherence, the authors argue that semantic coherence is the connection of meaning between the elements that make up sentences. Words and phrases need to have a relationship or complementarity of meanings for coherence to exist. Furthermore, semantic coherence extends to the connection between sequential sentences in discourse. A text with semantic coherence will present a logical and meaningful sequence of ideas and concepts, ensuring a comprehensible flow of thought to the reader.

Syntactic coherence, on the other hand, finds its reference in the grammatical devices used to bring semantic coherence to life, involving the use of linguistic elements such as connectors, pronouns, articles, and adverbs that assist in constructing a coherent narrative. This dimension is fundamental for text clarity and reader understanding.

Stylistic coherence, in turn, is closely related to the writing style that is appropriate for the chosen text genre. It involves the appropriate selection of vocabulary and sentence structures that align with the type of text being produced. Lack of consistency in the use of linguistic register, such as mixing formal and colloquial language, can result in a break in stylistic coherence in a text.

Finally, pragmatic coherence concerns the sequence of speech acts presented appropriately and with clear information. It is concerned with how the text adheres to the conventions of the communicative context in which it is placed. For a text to be pragmatically coherent, it is necessary for speech interactions to respect the schema and context of the text.

Moreover, each of these perspectives of textual coherence can be interconnected in more subtle ways. Semantic coherence can influence and be influenced by stylistic coherence, for example, because the choice of words and phrases can impact the overall meaning of the text. Similarly, syntactic and pragmatic coherence can intertwine, as the way sentences are constructed can affect the effectiveness of communication. Therefore, mastery of these four perspectives of textual coherence is fundamental for creating an effective and impactful text.

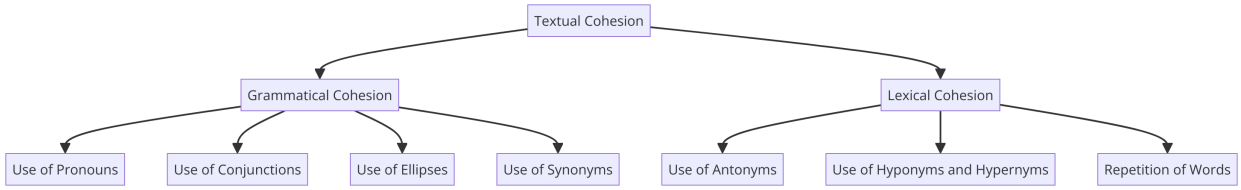
2.1.1 The role of Textual Cohesion

Textual cohesion is a key element in constructing texts that are clear, understandable, and effective in conveying their messages. Halliday and Hasan (1976, p. 5) define cohesion as "the network of lexical, grammatical, and other relations that provide connections within a text." These connections aim to ensure that a text is perceived as a meaningful unit rather than a random collection of sentences.

Cohesion can be achieved through various strategies, including the use of pronouns, conjunctions, ellipses, and other linguistic devices that establish clear relationships between words, phrases, and paragraphs in a text. These devices are often referred to as "cohesion markers" (Halliday; Hasan, 1976).

Cohesion is often categorized into two main types: grammatical cohesion and lexical cohesion, as illustrated in Figure 2.4. Grammatical cohesion refers to the use of grammatical elements to connect different parts of a text. This can include the use of pronouns to refer to a previously mentioned noun, the use of conjunctions to connect clauses, or the use of ellipses to omit words understood from the context. For example, in a sentence like "Maria went to the market. She bought apples" the word "she" is a pronoun referring to "Maria," creating a clear connection between the two sentences and making the text more cohesive (Quirk *et al.*, 1985).

Figure 2.4: Textual cohesion components - Gramatical Cohesion and Lexical Cohesion.



Source: The author, 2024.

On the other hand, lexical cohesion refers to the use of related words to create connections between different parts of a text. This can include the use of synonyms, antonyms, hypernyms, hyponyms, or repetition of words. For example, in a sentence like "Maria loves reading. Always with a book in her hands," the word "book" is related to the word "reading," creating a lexical connection that makes the text more cohesive (Halliday; Hasan, 1976).

Textual cohesion plays a significant role in facilitating reader comprehension because it helps guide the reader through the flow of ideas in the text (McNamara; Kintsch, 1996). Furthermore, cohesion can also make the text more enjoyable to read because it creates a sense of rhythm and fluency (Crossley; Kyle; Dascalu, 2018).

2.2 Coherence Theoretical Models

2.2.1 Rhetorical Structure Theory (RST)

Rhetorical Structure Theory (RST) is a discursive approach that was initially introduced through the work of Mann and Thompson (1987). This theory focuses on the analysis and organization of texts, as it concentrates on identifying relationships between continuous and linear segments of text known as "text spans." When these relationships are mapped, they form the rhetorical structure of the text, providing the arrangement and connection of the expressed ideas. The primary contribution of RST lies in its ability to systematically dissect the text's structure, shedding light on the mechanisms that underpin effective communication.

Rhetorical relations, also referred to as coherence relations, are considered by this theory as fundamental to text comprehension. Mann and Thompson (1987) argued that if a text is coherent, then its rhetorical structure can be determined. Furthermore, RST allows for identifying the implicit propositional content of a text, which is revealed through the propositional relations between its parts.

In the original version of RST, 26 rhetorical relations were established to connect the propositions expressed in a text (as seen in Table 2.1). These relations are formed between

two or more propositions from adjacent text segments. Each relation consists of a nucleus proposition (N), representing the main information, and an additional complementary piece of information called a satellite (S). In the case where both pieces of information are equally important, there is a multinuclear relation, consisting of two nuclei and no satellite.

Table 2.1: Rhetorical Relations and their Nuclearity in RST.

Rhetorical Relation	**Multinuclear**
Antithesis	No
Background	No
Circumstance	No
Concession	No
Condition	No
Elaboration	No
Enablement	No
Evaluation	No
Evidence	No
Interpretation	No
Justification	No
Means	No
Motivation	No
Non-Volitional Cause	No
Non-Volitional Result	No
Otherwise	No
Purpose	No
Reformulation	No
Solutionhood	No
Summary	No
Volitional Cause	No
Volitional Result	No
Contrast	Yes
Conjunction	Yes
List	Yes
Sequence	Yes

Source: (Mann; Thompson, 1987)

2.2.1.1 RST's Role in Coherence Analysis

In the domain of NLP, RST is used in both, analyzing and understanding the complex web of rhetorical relations that construct textual coherence. Models inspired by RST use elementary discourse units (EDUs) to analyze text. These EDUs, representing the smallest functional units in a text, are organized into a hierarchical structure, resembling a tree, where

each node symbolizes a text segment and the connections represent rhetorical relationships (Falzon, 2009).

For instance, consider a text segment expressing a problem and another providing a solution. RST would categorize this relationship as "Problem-Solution", highlighting how the second segment coherently addresses the issue raised in the first. Similarly, if a segment presents an argument and the next provides supporting evidence, RST labels this as an "Evidence" relation. Therefore, RST structures aids in identifying and understanding the coherent flow of ideas within a text.

In RST-based coherence analysis, coherence is identified when a text exhibits clear and logical rhetorical relations among its segments. A coherent text typically presents a well-defined nucleus that is logically expanded upon or justified by satellites. In contrast, incoherence arises from ambiguous, weak, or non-existent relations, resulting in disjointed narratives (Brunato *et al.*, 2023).

The process of identifying and annotating rhetorical relations between Elementary Discourse Units (EDUs), which are typically clauses or sentences, involves analyzing each EDU for its role as a nucleus or satellite within these relations. These relations, such as Elaboration, List, and Consequence, determine how segments of a text logically connect to form a coherent narrative (Mann; Thompson, 1988).

To illustrate, consider a text discussing environmental conservation as another RST annotation example. Initially, the text outlines the severity of environmental degradation (Segment A), followed by a segment suggesting various conservation methods (Segment B). Through RST, these segments are connected by a "Cause-Effect" relation, with Segment A establishing the cause (environmental issues) and Segment B providing the effect (conservation methods), ensurign that the text's segments are logically connected, creating a coherent narrative.

An example of RST annotated corpus is evident in the RST Signalling Corpus (Das; Taboada; McFetridge, 2015), developed over the RST Discourse Treebank. This corpus includes annotations for a variety of signals indicating coherence relations, such as discourse markers and other linguistic features like lexical and semantic signals (Das; Taboada; McFetridge, 2015), providing a rich resource for investigating the psycholinguistic mechanisms behind the interpretation of relations through signaling.

That said, in RST-based coherence analysis, a text is considered coherent if its rhetorical structure logically connects the EDUs, thereby creating a comprehensible and unified narrative or argument. Incoherence arises when these connections are absent or illogical, leading to a disjointed or unclear text. For example, in the DisCoTEX dataset, swapping sentences in a paragraph or replacing them with unrelated content can lead to a

loss of coherence, as the rhetorical relations become nonsensical or irrelevant (Brunato *et al.*, 2023).

2.2.2 Centering and Entity-Based Coherence

Centering Theory is a prominent approach in the field of NLP, focusing on the use of referential expressions (like pronouns and noun phrases) to achieve coherence in discourse. The theory posits that discourse coherence is maintained through the management of entities that are said "central" to the discourse context, referred to as "centers" (Grosz; Joshi; Weinstein, 1995).

For example, in a text where a character is introduced and later referred to by a pronoun, Centering Theory examines how these references contribute to the coherence of the discourse. The theory distinguishes between "forward-looking centers" (potential referents in upcoming discourse) and "backward-looking centers" (referents from previous discourse), analyzing how shifts between these centers affect coherence (Walker; Iida; Cote, 1998).

Entity-based coherence, on the other hand, involves tracking entities throughout a text and observing how they are introduced, maintained, and shifted in focus. This approach assumes that a coherent narrative will maintain a clear and consistent reference to its key entities, allowing the reader to easily follow the narrative thread (Barzilay; Lapata, 2008).

In practical terms, when a text introduces multiple entities, it must manage these entities effectively to maintain coherence. For instance, if a narrative starts with a discussion about "Alice", shifts to "Bob", and later reintroduces "Alice", the transitions and references to these entities must be clear and logically connected to maintain coherence. Entity-based coherence tools often analyze texts by creating graphs or networks of entity occurrences and transitions, assessing the strength of these connections to determine the text's coherence (Jurafsky; Martin, 2024). For example, when generating a story or a detailed report, these models ensure that references to entities are consistently and logically developed, thereby enhancing the reader's understanding and engagement with the text.

Moreover, annotated corpora with centenring and Entity-Based Coherence such as the OntoNotes corpus are good resources for studying entity-based coherence. These corpora include annotations of entity references and their roles within the discourse, underpininng the appropriate form of a coherent text (Weischedel; Palmer; Marcus, 2013).

2.2.3 Local Coherence Representation

Historically, early models developed in the 1990s addressed local coherence by employing lexical cohesion strategies. Lexical cohesion refers to the recurrence of semantically related words within a discourse. For instance, Morris and Hirst (1991) utilized lexical chains, sequences of thematically connected words (e.g., pine, bush, trees, trunk), to analyze discourse coherence. These chains, derived from sources like Roget's Thesaurus, demonstrated how the density and presence of such chains correlate with a text's topic structure, thus contributing to its coherence (Morris; Hirst, 1991). Thus, local coherence in discourse analysis emerged by focusing on the immediate connections between sentences or small text spans, ensuring a smooth and logical progression of ideas.

Another early method, the TextTiling algorithm by Hearst (1997), quantified coherence by calculating the cosine similarity between neighboring text spans. This approach illustrated how segments within the same subtopic exhibited higher cosine similarities compared to segments from different subtopics, thereby maintaining topical coherence (Hearst, 1997).

The LSA Coherence method introduced by Foltz et al. (1998) marked a significant advancement by using sentence embeddings to model coherence. This method computed the coherence between sentences based on the cosine similarity of their Latent Semantic Analysis (LSA) derived embeddings. By averaging the cosine similarities of all adjacent sentence pairs in a text, this model provided a quantifiable measure of text coherence (Foltz; Kintsch; Landauer, 1998). Modern approaches, such as the Local Coherence Discriminator (LCD) model by Xu et al. (2019), build on these foundations but leverage neural representation learning and self-supervision. Unlike earlier models, LCD trains to differentiate between naturally coherent and artificially disordered discourses. It evaluates coherence by analyzing pairs of consecutive sentences, contrasting these coherent pairs with randomly scrambled pairs to refine its discrimination of coherence (Xu *et al.*, 2019).

2.2.4 Global Coherence Representation

Global coherence refers to the overarching logical flow and consistency of an entire text, linking themes and ideas across the entire narrative or argumentative structure. This level of coherence ensures that a text makes sense as a whole, rather than just in isolated parts.

One of the foundational models for assessing global coherence is based on the concept of macrostructures, as developed by Teun A. van Dijk in the late 1970s. Macrostructures

represent the higher-level thematic organization of a text, synthesizing the main ideas into a coherent whole (Van Dijk, 1977).

Building on these ideas, Giora (1985) introduced the notion of thematic progression, which involves the orderly development of themes and topics from sentence to sentence and across paragraphs. This method tracks how well the discourse maintains its focus on introduced topics, thus contributing to the text's global coherence.

In the realm of computational linguistics, the Coh-Metrix tool, developed by Graesser *et al.* (2004), provides a multifaceted approach to measuring text cohesion and aspects of coherence. This tool assesses cohesion on multiple dimensions, including the causal and intentional connections between sentences and the overall logical structure of the text, thus offering a comprehensive analysis of global coherence.

More recently, neural network models have been employed to analyze global coherence. These models, such as those described by Mesgar and Strube (2015), use deep learning techniques to evaluate the coherence of entire documents. By training on large corpora, these models learn to identify patterns and structures that characterize coherent texts, considering factors like thematic consistency, narrative progression, and logical sequencing.

Furthermore, the use of discourse relation models, which parse texts into a series of discourse relations following frameworks like Rhetorical Structure Theory (RST), provides another layer of analysis. These models, as exemplified by Joty, Carenini and Ng (2015), examine how well different parts of a text are linked by logical rhetorical relations, which are necessary for maintaining global coherence.

2.3 Automatic Models

2.3.1 BERT and Derivatives

BERT (*Bidirectional Encoder Representations from Transformers*) marked a significant advancement in the field of NLP. As a model based on the architecture of *Transformers*, its introduction by Devlin *et al.* (2018) was practically a revolution for the field. BERT is a pre-trained model that uses an innovative technique known as *"Masked Language Modeling"* (MLM), which involves randomly hiding parts of the words in a text and challenging the model to predict the hidden words, using the context provided by the visible words both to the left and right of the masked words. This strategy allows BERT to have a contextual representation that simultaneously considers the information that comes before and after each word in the text. Such a training method represented a significant advancement for NLP, as

it differs from previous models that generally considered the text in a unidirectional manner.

The main innovation of BERT lies in its pre-training methodology, which uses two tasks: MLM and *Next Sentence Prediction* (NSP). In MLM, some input *tokens* are randomly masked, and the goal is to predict them based on the context provided by those unmasked. In NSP, the model learns to predict "whether a sentence B" is the "logical continuation of a sentence A," which helps to understand the relationship between sentences. BERT has demonstrated superior performance in a variety of NLP tasks, including text comprehension, named entity recognition, and linguistic inferences.

After the pre-training process using MLM and NSP, BERT can be subjected to *fine-tuning* (or fine adjustment) for optimization in specialized tasks, where the model's parameters are adjusted using a specific dataset for the task. This is done by adding an output layer adapted to the desired function, such as text classification or question and answer processing. This technique was designed for BERT with the intent to enhance the effectiveness and capabilities of each generated model, adjusting them to the individual needs of the tasks to which they are intended.

However, adapting BERT to specific tasks may require, among other things, meticulous adjustments in parameters, large volumes of data, code changes, or even modifications in its architecture. Another point to be noted is the need to properly manage the process to avoid *overfitting* and ensure that the model remains generalist enough to maintain its applicability in other contexts.

Following the success of BERT, various derivatives were developed to optimize or adapt its capabilities to different needs and computational resources. RoBERTa (A Robustly Optimized BERT Approach), introduced by the work of Liu *et al.* (2019), modifies the BERT pre-training scheme by eliminating the NSP task and adjusting hyperparameters such as batch sizes and learning rates, which allowed RoBERTa to outperform BERT in many NLP tasks. DistilBERT (Distilled BERT), created by Sanh *et al.* (2019), is a distilled version of BERT that retains 97% of its performance in language comprehension tests, but with only 40% of its parameters, achieved through techniques such as "knowledge distillation," where DistilBERT learns to replicate the behavior of BERT, retaining most of its effectiveness.

For the task of analyzing textual coherence, models like BERT and its derivatives provide significant benefits. They have the capability to discern subtle contextual relationships and track the logical flow within a text, which is essential for determining coherence. By processing text bidirectionally, BERT, for instance, can evaluate how well the ideas in a text connect and flow logically from one to another. DistilBERT and RoBERTa, on the other hand, can offer enhancements in processing efficiency and adaptability to varying scales of data.

2.3.2 GPT Family of Large Language Models

GPT models (*Generative Pre-trained Transformers*), developed by OpenAI[1], are renowned for their ability to generate text that is both coherent and contextually relevant, utilizing attention mechanisms to model contextual relationships in textual data. These models are trained using a process called "autoregressive," which teaches the model to predict the next word in a sequence based on the context of the preceding words, thereby promoting logical and coherent text progression.

From the original GPT to more advanced versions like GPT-3 (Brown *et al.*, 2020) and GPT-4, enhancements have been made to improve not just text generation, but also overall coherence. GPT-3.5 and GPT-4, for example, bring significant improvements in terms of size and training efficiency, enabling a higher level of text comprehension and coherence. These models are capable of maintaining logic and flow in long texts, making them ideal for applications requiring high-quality textual coherence such as summarizations, article writing, and dialogue in complex contexts.

Moreover, fine-tuning on GPT models can be performed with minimal modifications, typically just adding a linear classification layer. This layer learns to map the contextual representations generated by the GPT to specific categories, preserving the coherence of the text generated tailored to the task at hand.

2.4 Used Datasets

2.4.1 COCA (Corpus of Contemporary American English)

The Corpus of Contemporary American English (COCA) is one of the most widely used corpora in natural language processing (NLP) and linguistic research. Compiled by Mark Davies, COCA provides a comprehensive and balanced collection of texts from diverse genres, making it a valuable resource for studying contemporary American English (Davies, 2008).

COCA encompasses over more than a billion words, spanning a range of genres including spoken language, fiction, popular magazines, newspapers, and academic texts. This diversity allows researchers to conduct cross-genre analyses and investigate how language usage varies across different contexts. For example, COCA has been instrumental in studies examining genre-specific vocabulary and syntactic patterns, leading us into how language functions in various communicative settings (Biber, 2011).

[1] <https://openai.com>

One of the key advantages of COCA is its temporal coverage. The corpus includes texts from 1990 to the present, updated regularly to include recent texts. This temporal span enables researchers to analyze diachronic changes in language use, tracking how linguistic features evolve over time. For instance, (Davies, 2012) utilized COCA to study the frequency and usage patterns of modal verbs in American English, revealing shifts in modality over the past few decades .

In addition to linguistic studies, COCA has been employed in the development and evaluation of NLP models. The corpus's rich annotation and genre diversity make it suitable for training language models and conducting various NLP tasks such as part-of-speech tagging, named entity recognition, and sentiment analysis. Researchers have used COCA to enhance the performance of these models by providing extensive and varied training data (Smith; Baldridge, 2013).

Moreover, COCA's structured metadata allows for sophisticated queries and detailed analyses. Researchers can filter texts by genre, date, and other attributes, facilitating targeted investigations into specific linguistic phenomena. This capability has been particularly useful in studies of language variation and sociolinguistics, where precise control over the data is essential (Leech *et al.*, 2014).

In the free portions of COCA, specific genres reveal distinct patterns in sentence and paragraph structure. For instance, blog texts exhibit an average sentence length of 29.61 words, with an average of 72.06 sentences per text and 3.23 sentences per paragraph. Academic texts, on the other hand, feature slightly shorter sentences, averaging 26.92 words, but are characterized by much longer texts with 212.71 sentences on average and a significantly larger paragraph length, averaging 95.58 sentences per paragraph. In contrast, news articles display an average sentence length of 24.87 words, 72.47 sentences per text, and a paragraph length of 18.33 sentences. These statistics pertain to the freely accessible portions of COCA, while the full, paid version of the corpus offers different quantifications that are currently inaccessible due to lack of access.

The accessibility and comprehensive nature of COCA have also contributed to its widespread adoption in educational settings. It serves as a foundational resource for teaching corpus linguistics and empirical research methods, offering students hands-on experience with real-world language data. Through COCA, learners can explore linguistic patterns, test hypotheses, and develop critical analytical skills (Baker, 2016).

2.4.2 CST News Corpus

The CST News Corpus (Aleixo; Pardo, 2008) is a significant resource in natural language processing (NLP) and computational linguistics, particularly for studies involving discourse analysis, text summarization, and coherence evaluation. It comprises Brazilian Portuguese news articles and has been extensively used in various NLP research applications.

Developed to support research on multi-document summarization, the CST News Corpus contains a diverse collection of news articles from multiple sources (Aleixo; Pardo, 2008). The corpus is organized into clusters, each consisting of several news articles about the same event or topic. This organization enables researchers to study the coherence and cohesion of information across different documents discussing the same subject.

The corpus includes 50 collections of Brazilian Portuguese texts. Each collection contains approximately three documents on the same subject from different sources, along with a human-generated summary. Annotated by four computational linguists, the corpus achieved satisfactory annotation agreement. This structure has proven invaluable for tasks such as identifying discourse markers, analyzing text structure, and evaluating summarization algorithms (Dias, 2016). On average, sentences in the corpus contain 28.37 words, with each text consisting of 7.49 sentences and paragraphs averaging 6.49 sentences.

One of the notable features of the CST News Corpus is its ability to facilitate cross-document analysis. Researchers have used the corpus to develop and test multi-document summarization systems, which aim to create concise summaries from multiple related documents. This task is crucial for applications like news aggregation and information retrieval, where synthesizing information from various sources is essential (Dias; Pardo, 2015).

The corpus has also been employed in coherence analysis studies. By examining how information is presented across different articles about the same event, researchers can gain insights into the factors that contribute to text coherence and cohesion. For example, studies have investigated the use of discourse markers and the organization of information to better understand how coherence is maintained in multi-document settings (Cardoso *et al.*, 2011).

In addition to summarization and coherence analysis, the CST News Corpus has been used to train and evaluate various NLP models. The diversity of the texts and the presence of human-generated summaries provide a rich dataset for training machine learning models on tasks such as text classification, sentiment analysis, and named entity recognition. Researchers have leveraged this corpus to enhance model performance by providing diverse and contextually rich training data (Pardo *et al.*, 2017).

The structured nature of the CST News Corpus, with its detailed annotations and well-defined clusters, allows for sophisticated queries and targeted analyses. Researchers can

filter texts by source, date, and other attributes, facilitating precise investigations into specific linguistic phenomena. This capability is particularly useful in studies of language variation and the impact of different news sources on information presentation (Aleixo; Pardo, 2008).

The CST News Corpus has been widely adopted in educational settings. It serves as a foundational resource for teaching empirical research methods and corpus linguistics, offering students practical experience with real-world language data. Through the CST News Corpus, learners can explore linguistic patterns, test hypotheses, and develop critical analytical skills (Cardoso *et al.*, 2011).

2.4.3 GCDC (Grammarly Corpus of Discourse Coherence)

The Grammarly Corpus of Discourse Coherence (GCDC) is a specialized dataset designed to facilitate the study of coherence in written texts. Developed by researchers at Grammarly, GCDC offers a rich resource for analyzing discourse coherence, a critical aspect of natural language understanding and generation (Lai; Tetreault, 2018).

GCDC is composed of a diverse collection of texts from four different domains: Yahoo Answers, emails, user reviews, and web discourse. This variety allows researchers to examine coherence across various genres and writing styles, providing a comprehensive view of how coherence manifests in different types of written communication. For example, the corpus includes both formal and informal texts, enabling studies that compare how coherence is maintained in professional versus casual contexts (Napoles; Sakaguchi; Tetreault, 2017). On average, sentences are 22.44 words long, with each text containing approximately 9.06 sentences, and paragraphs averaging 5.04 sentences.

One of the unique features of GCDC is its detailed annotation of coherence relations. Each text in the corpus is annotated for coherence quality, using a scoring system that evaluates the logical flow and clarity of ideas. This fine-grained annotation allows for precise analysis of coherence-related features and facilitates the training of machine learning models aimed at assessing and improving text coherence (Lai; Tetreault, 2018).

Researchers have used GCDC to develop and evaluate models for various NLP tasks, such as coherence scoring, text generation, and summarization. The corpus's annotated coherence scores provide a valuable benchmark for these tasks, enabling the development of algorithms that can automatically assess the coherence of a text. For instance, Lai and Tetreault (2018) employed GCDC to train neural network models that predict coherence scores, demonstrating significant improvements over baseline methods (Lai; Tetreault, 2018).

In addition to coherence analysis, GCDC supports studies in related areas such as discourse structure and argumentation. The corpus's comprehensive annotations and diverse

text types allow researchers to explore how different discourse elements contribute to overall coherence. For example, studies have used GCDC to investigate the role of cohesive devices, such as connectives and referential expressions, in maintaining discourse coherence (Napoles; Sakaguchi; Tetreault, 2017).

Moreover, GCDC has proven useful in educational applications, particularly in the development of tools for automated writing evaluation and feedback. By leveraging the corpus's coherence annotations, educational software can provide more nuanced feedback on students' writing, helping them improve the logical flow and clarity of their texts. This application underscores the practical value of GCDC in enhancing writing instruction and assessment (Burstein; Leacock; Sabatini, 2019).

2.4.4 DDisCo

The Danish Discourse Coherence Dataset (DDisCo) is a significant resource developed to advance the study of discourse coherence in Danish texts. Created by Linea Flansmose Mikkelsen, Oliver Kinch, Anders Jess Pedersen, and Ophélie Lacroix, DDisCo is composed of texts from Danish Wikipedia and Reddit, annotated for discourse coherence. This dataset distinguishes itself by focusing on real-world texts rather than artificially manipulated ones, providing a more authentic basis for training and evaluating coherence models (Mikkelsen *et al.*, 2022).

DDisCo includes texts annotated with coherence scores, allowing for a detailed analysis of how coherence is maintained or disrupted in natural language. These annotations are used for training machine learning models to distinguish coherent texts from incoherent ones. On average, the sentences 1200 comprised in the dataset are 21.53 words long, with 9.32 sentences per text and 9.31 sentences per paragraph. By using real-world texts, DDisCo avoids the pitfalls of artificially generated incoherence, which may not accurately reflect the complexities of natural discourse.

Researchers have utilized DDisCo to evaluate the performance of various methods, including neural networks, in detecting and measuring coherence. This dataset supports a range of NLP tasks such as text generation, coherence scoring, and discourse analysis, enhancing the ability of models to understand and produce coherent Danish texts. The use of DDisCo in these tasks has demonstrated improved accuracy and robustness in coherence assessment, highlighting its value in the field (Kinch *et al.*, 2022).

2.4.5 Summary

Further details regarding the collection, structure, and specific features of each of the corpora described in this section can be found in Section 4.1 (Data Collection), where their roles in this study are discussed in depth.

The table 2.2 summarizes the characteristics of the corpora mentioned in this chapter. Detailed descriptions of these corpora are provided in Chapter 4 (Methodology), where their specific roles in each task – Local Coherence Classification (LCC), Global Coherence Classification (GCC), and Incoherence Identification (IID) – are thoroughly discussed. The table offers an overview of how each corpus is utilized across these central tasks, serving as a quick reference for understanding their application in the context of our research.

Table 2.2: Summary of Corpora Used in this Work

Dataset	**Language**	**Used Domains**	**Total Texts**	**Used Texts per Task**		
				LCC	**GCC**	**IID**
COCA (Blog)	English	Blogs	978	978	60	60
COCA (Academic)	English	Academic	256	256	10	10
CST News	Portuguese	News Articles	300	251	30	30
GCDC	English	Emails, User Reviews	900	842	900	30
DDisCo	Danish	Wikipedia, Reddit	1,200	991	1,200	0

Before delving into the methodology, however, it is essential to situate this work within the existing literature. Therefore, the Chapter 3 (Related Work), will review the related work, providing context and highlighting this study's contributions to the broader fields of coherence analysis and NLP.

3 Related Work

As the concept of textual coherence has been explored in the previous chapter, the following sections reviews key studies and methodologies related to both local and global automatic coherence analysis, highlighting their contributions to the field and how this thesis uses and is colocated between them.

3.1 Local Coherence and the Shuffle Test

As demonstrated in the traditional entity grid work by Lapata and Barzilay (2005), local coherence heavily depends on the sequential order of adjacent sentences. When this order is disrupted, the fluency and connection between these sentences are compromised, highlighting issues in local coherence. The authors describe local coherence as being influenced by the connectivity between consecutive sentences, including topic continuity, referential cohesion (such as pronouns and ellipses), and the use of logical connectors. Consequently, local coherence directly impacts the immediate comprehension of the text, ensuring the smooth transition from one sentence to another.

The analysis of local coherence involves understanding the logical flow between adjacent sentences in a text. As discussed in Chapter 2, the Centering Theory is a prominent model in this domain. Significant contributions by Grosz, Joshi and Weinstein (1995) and Gordon, Grosz and Gilliom (1993) focus on maintaining the salience of discourse entities across sentences to ensure coherence. Their work emphasizes the importance of entity transitions and the need to keep the discourse centered on specific entities.

Building on these previous works, the model of entity grids proposed by Barzilay and Lapata (2008) introduced a novel framework for representing and measuring local coherence by focusing on the patterns of entity distribution within a text. The core idea of this model is to abstract a text into a set of entities (subjects (S), objects (O), other(X) and missing(-))

and their transition sequences, which capture the distributional, syntactic, and referential information of discourse entities. The entity-grid representation thus reflects how entities (such as subjects and objects) appear and transition across sentences.

The coherence of a text is assessed by analyzing the entity syntatic role transitions in each textual sentence. This method involves constructing a two-dimensional array, where rows correspond to sentences and columns represent discourse entities. Each cell within this array indicates whether an entity is present in a particular sentence and, if so, in what grammatical role (e.g., subject, object, or other) (Abdolahi; Zahedi, 2016). By examining the patterns and probabilities of these transitions, the model can infer the degree of local coherence within the text.

The shuffle test (Barzilay; Lapata, 2008) has become a widely utilized practice for assessing coherence models. This artificial task requires the model to distinguish between an original document and its randomly shuffled counterpart. The central idea is that a robust coherence model should be able to identify the original order as more coherent compared to the permuted orders. In the experiment, the model is given the original text and its permutations and must classify them as coherent or non-coherent. The model's effectiveness is measured by the frequency with which it ranks the original text order as coherent, in contrast to the shuffled versions. Laban *et al.* (2021) propose treating it as a probe—an evaluation task that allows models to be assessed without explicit supervision. In the present work, we will utilize this approach to verify the effectiveness of models, as it is well-established and provides an objective assessment of the model's ability to recognize the fluency and logical connectivity inherent of a coherent text.

3.1.1 Entity Grids and derivatives

As said on the overview produced by Abdolahi and Zahedi (2016, p. 4) and shown bellow, "The most important computerized text coherence evaluation approach is Entity-Based method proposed by Lapata and Barzilay (2005) and Barzilay and Lapata (2008). Since then most of the new methods are used their proposed features and based algorithm.".

Elsner, Austerweil and Charniak (2007) present a unified model for discourse coherence that combines the local entity-based approach of Lapata and Barzilay (2005) with the HMM-based content model of Barzilay and Lee (2004). The authors evaluate their model on two primary tasks: sentence ordering and discrimination (shuffle test).

In the sentence ordering task, the model aims to find the most coherent arrangement of sentences from an unordered set, which has applications in text generation and summarization. In the discrimination task, the model must distinguish between an original document and

its randomly permuted versions, assessing the model's ability to prefer coherent texts over incoherent ones. This task evaluates how frequently the model correctly identifies the original order of sentences and also supports the test as a default approach to coherence evaluation.

The model, however, relies heavily on the accuracy of entity identification and tracking, which can be challenging in complex texts. Additionally, while the atempted integration of local and global features enhances coherence evaluation, the model may struggle with longer, less formulaic documents where global structures are more variable.

The work by Lin, Ng and Kan (2011) further reinforces the significance of the text ordering ranking task (shuffle test) in evaluating textual coherence. In their study, Lin and colleagues developed a model that assesses coherence by focusing on discourse relations. Their approach hinges on the premise that coherent texts tend to favor specific types of discourse relation transitions. By implementing this model, they applied it to the text ordering ranking task, where the system must distinguish an original text from a permuted version of its sentences.

Lin, Ng and Kan (2011) demonstrated that their model significantly outperforms the coherence model by Barzilay and Lapata (2008), achieving an average error rate reduction of 29% across three datasets compared to human upper bounds. Their findings underscore the robustness of the text ordering ranking task as a standard method for coherence evaluation. The methodology involved using discourse parsers to identify relations such as Temporal, Contingency, Comparison, and Expansion, and analyzing the transitions between these relations to assess coherence. The experimental results of their model highlight its effectiveness in the text ordering ranking task, thus solidifying the test's role as a benchmark in coherence evaluation studies. The success of their approach reiterates the relevance of the text ordering ranking task while showcases the potential of leveraging discourse relations for more nuanced coherence assessments.

Nevertheless, the model proposed by Lin, Ng and Kan (2011) has certain limitations. One significant drawback is its heavy reliance on accurately identifying and classifying discourse relations, which can be challenging due to the complexity and variability of natural language. The performance of the model is inherently tied to the quality of the discourse parser, and any errors in parsing can adversely affect coherence assessment.

Subsequently, the works by Dias, Feltrim and Pardo (2014), Dias and Pardo (2015) and Dias (2016) are important contributions that follow the influential works of Lapata and Barzilay (2005) and Barzilay and Lapata (2008). By adhering to their sentence ordering evaluation frameworks while expanding the scope of their approaches, they underscore the relevance of the sentence ordering task.

Dias, Feltrim and Pardo (2014) integrates RST with entity grids to measure local

coherence in texts adhering to both entity transition patterns and discourse relations to distinguish coherent texts from incoherent ones. By using RST, the model captures the hierarchical structure of discourse, which allows for a more nuanced representation of coherence. The evaluation follows the sentence ordering (shuffle) task, where the model must determine the correct sequence of sentences by comparing the original text to its shuffled permutations.

On top of that, Dias and Pardo (2015) extended the approach to the domain of multi-document summaries, addressing the added complexity of integrating information from various sources and incorporating both RST and Cross-document Structure Theory (CST) relations to model coherence. The text ordering task is again employed to evaluate the model, where the goal is to rank the original summaries higher than their shuffled counterparts.

Similar to both previous studies, the approach from Dias (2016) also addresses the text ordering task and relies heavily on detailed semantic-discoursive annotations. Dias (2016) further investigates models of local coherence for multi-document summaries by examining how discourse relations can enhance the automatic evaluation of summary coherence. The model utilizes both RST and CST annotations to create a more comprehensive coherence evaluation framework. The text ordering task is utilized here as well, where the model distinguishes between coherent and incoherent summaries by analyzing the arrangement of sentences and their discoursive relations. However, this approach also underscores the necessity for advanced tools capable of accurately parsing and annotating discourse relations, which is a labor-intensive process that limit the scalability of such methods.

In light of these limitations, subsequent research has explored alternative methodologies that do not depend on discourse relations. Notably, advancements in neural network models have introduced new paradigms in coherence evaluation. For example, Li and Hovy (2014) and Nguyen and Joty (2017) employed recurrent neural networks (RNNs) and convolutional neural networks (CNNs), respectively, to capture coherence by learning representations of entire texts.

Mesgar and Strube (2015) introduced a graph-based approach to coherence modeling, specifically designed to assess readability by evaluating text coherence. The authors' methodology involves constructing entity graphs and discourse relation graphs to represent text coherence. Entity graphs model interactions between entities and sentences, while discourse relation graphs capture rhetorical relations between sentences. These graphs are then combined to create a comprehensive representation that incorporates both entity transitions and discourse relations.

Once more, the text ordering ranking task is employed to assess the model's ability to distinguish between coherent and incoherent texts. In this version of the evaluation, coherence features are extracted from both original and shuffled texts, a classifier is trained

to differentiate between them, and the model's accuracy is measured. Nevertheless, this approach has limitations similar to other entity-dependent models, as it relies on detailed discourse annotations that are resource-intensive to obtain. Furthermore, the computational complexity of subgraph mining presents scalability challenges.

Conversely, Nguyen and Joty (2017) presents a neural network-based approach to local coherence modeling, enhancing the traditional entity grid model with convolutional neural networks (CNNs). The model operates over the entity grid representation of a text, where entities are tracked across sentences, capturing their grammatical roles (subject, object, other) and transitions. These roles are transformed into distributed representations (vectors), which are then processed by CNNs to capture high-level features from entity transitions, allowing the model to handle long-range dependencies effectively. After convolution, a max-pooling operation highlights the most salient features, which are used for coherence scoring.

The model is trained using a pairwise ranking method, optimizing it to assign higher coherence scores to original texts over their permuted versions, with an end-to-end training approach that helps the model learn task-specific high-level features automatically (Nguyen; Joty, 2017). The authors evaluate their model using three tasks: the discrimination task (shuffle test), where the model distinguishes original documents from shuffled ones; the insertion task, which tests the model's ability to reinsert a removed sentence into its correct position; and the summary coherence rating, where the model's coherence scores for summaries are compared to human judgments.

The neural model outperformed previous entity grid models, showing improvements in all tasks, however, its performance also heavily relies on accurate entity grid construction. Even with current advancements, the computational demands for training the deep neural model are substantial, requiring significant computational resources such as GPUs or TPUs, extensive memory, and considerable time for processing large datasets and performing numerous operations of convolution and backpropagation.

Following the approaches established in previous works, the study by Braz Junior and Fileto (2021) explores the application of BERT models to classify and measure text coherence. The authors focus on Portuguese language variations of BERT to evaluate coherence in two distinct domains: news articles and an educational forum of student questions. Their results demonstrate that BERT can achieve up to 99.20% accuracy in sentence order discrimination, particularly for the educational forum data. This research underlines the potential of contextualized language models in tasks related to text coherence, further validating the shuffle test for evaluating sentence arrangement.

The authors employ two specific datasets: CSTNews, containing professionally written news summaries, and OnlineEduc 1.0, a collection of forum posts from a Brazilian university's

virtual learning environment. By utilizing different BERT configurations and pooling strategies for sentence embeddings, the researchers assess the (in)coherence of original and permuted texts. The results indicate that the BERT model, especially BERTimbauBase, outperforms traditional methods in both classification accuracy and coherence measurement. This success is largely attributed to the contextual sensitivity of the BERT model, which captures subtle variations in sentence order and meaning.

Their work aligns with the broader trend of leveraging neural networks for coherence evaluation, as seen in previous studies using convolutional neural networks (CNNs) and recurrent neural networks (RNNs) (Nguyen; Joty, 2017). Moreover, the study suggests the continued importance of evaluating sentence order, further expanding on the framework established by Barzilay and Lapata (2008) with advanced machine learning techniques.

3.2 Global Coherence

While local coherence focuses on the immediate transitions and connections between adjacent sentences, global coherence pertains to the overall consistency of an entire text. Global coherence ensures that all parts of the text contribute meaningfully to the overarching narrative or argument, maintaining a unified and coherent discourse throughout. Unlike local coherence, which deals with sentence-level connections and the logical flow between them, global coherence encompasses the broader contextual and structural elements that bind a text together.

As highlighted in the seminal work by Thompson (1986), global coherence plays a crucial role in the readability and ease of comprehension of texts. Thompson (1986) emphasizes that coherence is a key factor in making technical writing more readable and accessible. She argues that for a text to be readable, it must be coherent, which involves shared knowledge between the writer and the reader. This shared knowledge allows readers to set up expectations about a text, facilitating efficient reading and comprehension.

Thompson (1986) defines global coherence as the internal unity of a text, enabling readers to perceive paragraphs of related ideas rather than lists of random facts. She explains that coherence is influenced by the writer's intentions, the representation of those intentions, and the reader's understanding of the intended message. This multidimensional nature of coherence means that different readers may interpret the same text differently based on their background knowledge and expectations.

In her work, Thompson (1986) further emphasizes that the perception of global coherence significantly impacts the readability of a text. A coherent text is one where readers can easily follow the overall structure and organization, leading to faster reading speeds and

better recall of information. This aspect is particularly critical in technical writing, where clarity and precision are essential for effective communication.

In addition to the foundational work by Thompson (1986), more recent studies have further explored how the structure of discourse affects global coherence. For instance, Sagi (2010) conducted experiments demonstrating that the hierarchical organization of a text significantly influences its perceived coherence. Their research shows that texts with more levels in their hierarchical structure are judged to be more coherent, suggesting that a well-structured discourse facilitates a better understanding of the text as a whole. This study also highlights that the effects of discourse structure on global coherence are sensitive to the genre, with narratives being more affected by structural manipulations than procedural texts.

Building on this, Sagi (2010) emphasize that the nature of discourse relations is closely tied to determining global coherence. Their study suggests that readers' ability to establish global coherence depends not only on hierarchical structures but also on how these structures align with cognitive expectations and the text's logical flow. This highlights that global coherence is a multifaceted concept, extending beyond simple sequential order to encompass how effectively the discourse resonates with the reader's mental models, incorporating deeper cognitive and genre-specific factors.

Kiddon, Zettlemoyer and Choi (2016) addresses global coherence in text generation, particularly for lengthy texts like cooking recipes, where maintaining coherence is crucial. The authors introduce a neural checklist model that tracks mentioned and pending agenda items (e.g., ingredients), ensuring the text aligns with the overall structure. This approach effectively manages long-range dependencies by referencing covered and remaining content, enhancing logical flow. While focused on text generation, these strategies offer valuable insights for improving coherence evaluation in NLP tasks.

The work by Lai and Tetreault (2018) is a valuable resource for studying global coherence in automated text analysis. It includes essays, emails, and user-generated content, all annotated for global coherence based on human judgments of logical flow and clarity. One key challenge identified is that traditional coherence models – such as those based on sentence ordering or entity grids – perform well on highly structured, professionally written texts but struggle significantly when applied to everyday writing. This happens because such texts often contain informal language, missing transitions, and abrupt topic shifts, which are not well captured by models trained on curated datasets. The study highlights the difficulty automated systems face in consistently maintaining logical flow and structure across longer texts, especially in domains where coherence is more fluid, such as user-generated reviews and emails.

Mikkelsen *et al.* (2022) build upon that and focuses on discourse-level analysis,

emphasizing how different parts of the rect relate to one another to determine global coherence. The work provides granular annotations detailing specific discourse relations, like causal or contrastive relationships, allowing for a detailed analysis of their impact on coherence. This dataset has been instrumental in testing models' ability to recognize and apply discourse relations, enhancing their capacity to generate and evaluate globally coherent texts. The work address how the DDisCo Corpus have highlighted the complexities of ensuring a text follows a logical sequence while maintaining intended meaning and emphasis.

The study by Najafi and H. Darooneh (2017) introduces a novel approach to understanding and quantifying global coherence in texts using detrended fluctuation analysis (DFA). Unlike traditional models that focus on local coherence through adjacent sentences, this method analyzes long-range correlations in the broader text structure. By examining the frequency of words and their correlations with neighboring words based on the frequency of co-occurrences, the study identifies patterns that contribute to global coherence, offering a dynamic metric through the scaling exponent. This approach provides a new perspective on coherence, with significant implications for computational linguistics, particularly in automatic text evaluation and generation. The study highlights how coherence varies across different text genres, emphasizing the need to consider text structure in coherence assessment.

In summary, the exploration of global coherence remains a relatively underdeveloped area compared to local coherence. While foundational studies like those by Thompson (1986) and Sagi (2010) have significantly contributed to our understanding of how global coherence affects readability and comprehension, there is still a scarcity of research that specifically addresses the automatic identification and evaluation of global coherence in texts. The introduction of datasets like GCDC and DDisCo marks a step forward in this domain, providing valuable resources for testing and refining models. However, with the advent of Large Language Models, there is potential for significant advancements in this field, as these models offer new opportunities to enhance our understanding and assessment of global coherence. This opens up a new chapter in the study of textual coherence, paving the way for more sophisticated and comprehensive approaches to ensuring that texts are not only locally coherent but also maintain a consistent and logical flow throughout.

The next section will delve into the role of Large Language Models in advancing the study of textual coherence.

3.3 LLMs for Textual Coherence

The advent of LLMs has revolutionized the field of NLP, offering unprecedented capabilities in understanding and generating human-like text. These models, such as GPT-

3, BERT, and more recent advancements like GPT-4, Claude, and Gemini, have shown remarkable performance across a wide range of language tasks, including text classification, summarization, translation, and more. However, one of the most promising applications of LLMs lies in their potential to enhance our understanding and evaluation of textual coherence—both local and global.

LLMs are trained on vast amounts of data, allowing them to capture intricate patterns in language, ranging from syntactic structures to deeper semantic relationships. This extensive training enables these models to assess coherence not only by evaluating the immediate connections between sentences but also by considering the broader context that spans entire documents. As a result, LLMs provide a more holistic approach to coherence analysis, potentially overcoming some of the limitations observed in traditional models that primarily focus on local coherence.

The paper by Najafi and H. Darooneh (2017) tackles the complex issue of evaluating global coherence in texts using detrended fluctuation analysis (DFA), a method traditionally applied in time series analysis. The study introduces the "importance time series," which tracks textual elements contributing to overall coherence. By analyzing the scaling exponent from these series, the authors provide a dynamic metric reflecting coherence across various text genres. This approach challenges traditional methods that focus on local coherence, emphasizing the significance of long-range textual structures in assessing global coherence.

The study further explores DFA's application to different text genres, revealing that global coherence can vary significantly based on the type of text analyzed. This highlights the importance of considering broader textual structures, rather than just sentence-level connections, when evaluating coherence. The authors argue that traditional methods may overlook essential aspects of coherence that DFA can capture.

Additionally, the paper discusses the implications of these findings for computational linguistics, particularly in developing models for automatic text evaluation and generation. The DFA-based approach offers a novel tool for researchers to create systems that better mimic human understanding of global coherence, potentially leading to more logically structured outputs in NLP applications. Thus, LLMs have potential to act even better than DFA in the same features, as they better mimic the human language understanding, as further works shown.

The paper by Srivastava *et al.* (2023) offers a exsensive evaluation of LLMs using the BIG-bench framework, which includes over 200 tasks designed to test language comprehension, reasoning, and other complex linguistic abilities. These tasks assess how well LLMs understand and generate coherent text across multiple sentences or paragraphs, key elements of global coherence.

A significant contribution of this work is the quantification of LLMs' ability to maintain coherence in language tasks. Through BIG-bench, the study examines how scaling laws impact model performance, noting that while larger models often perform better, they eventually face diminishing returns. This finding is particularly relevant for coherence, suggesting that improvements may require refining model architecture rather than merely increasing size.

The paper also suggests that BIG-bench's structured approach could be adapted to create specific benchmarks for coherence evaluation. By focusing on metrics such as thematic consistency, referential cohesion, and logical progression, researchers could develop tasks that offer a granular analysis of where and how coherence in text breaks down. This would provide better visions into the strengths and limitations of LLMs in coherence classification.

Liu *et al.* (2023) introduces P-Tuning, a technique that enhances LLMs by integrating trainable continuous prompt embeddings with traditional discrete prompts. This innovation significantly improves model stability and performance across natural language understanding tasks, such as those in the SuperGLUE benchmark.

The study by Liusie, Manakul and Gales (2024) introduces innovative methods for evaluating LLMs, focusing on reducing positional bias in text comparison tasks. Although the research does not directly address textual coherence, the methodologies presented can be adapted for coherence quantification and classification, particularly in complex texts. By using comparative evaluation, LLMs can assess global coherence by comparing text segments or document versions, identifying which maintains logical flow and thematic consistency. Additionally, mitigating positional bias is crucial in ensuring that coherence evaluations focus on logical and thematic consistency rather than being influenced by the position of text elements.

The paper by Chen *et al.* (2024) discusses the rapid advancements in LLMs following the success of ChatGPT, which showcased impressive capabilities in natural language tasks like answering questions and correcting mistakes. This success has spurred interest in both closed-source and open-source LLMs, leading to models that excel across a broad range of tasks. The adaptability and fluency of these models in generating coherent text suggest their potential in advancing coherence analysis and evaluation, making them valuable tools in future NLP research.

The recent study by Akter *et al.* (2023) examines the performance of the Gemini models in comparison to the GPT series. The Gemini Pro model, while slightly less effective than GPT 3.5 Turbo in English-language tasks, excels in complex tasks involving reasoning and multilingual translation. This robust performance suggests that Gemini models have strong general-purpose language capabilities, making them well-suited for global coherence

evaluation. With further refinement, Gemini could become a powerful tool for assessing and enhancing global textual coherence in NLP applications.

The paper by Naismith, Mulcaire and Burstein (2023) pioneers the use of GPT-4 for automated discourse coherence assessment in written texts, specifically evaluating student responses to high-stakes English proficiency tests. The research explores how well GPT-4 replicates human coherence judgments, comparing its ratings to those of human experts. A key innovation was testing various prompt configurations, where GPT-4 not only rated coherence but also provided rationales. Findings showed GPT-4's ratings closely aligned with human evaluations, outperforming traditional NLP-based metrics.

The study's methodology involved GPT-4 generating coherence ratings alongside explanatory rationales, which improved alignment with human judgments. This dual output helped the model better simulate human reasoning. The results highlight GPT-4's capacity to understand complex linguistic structures and its potential to handle sophisticated coherence tasks.

The implications of this study are profound for discourse coherence analysis and educational applications. It suggests that LLMs like GPT-4 could be effectively used in automated writing evaluation systems, coherence classification, and incoherence detection. The ability to provide rationale-backed evaluations enhances transparency, making these models particularly valuable in educational settings where clarity in feedback is essential.

In conclusion, Naismith, Mulcaire and Burstein (2023) demonstrate the transformative potential of GPT-4 in coherence evaluation, setting a new standard for automated assessment and paving the way for future innovations in NLP. The study's findings underscore the importance of integrating rationale-supported ratings to simulate human judgment accurately, offering significant advancements in automated discourse coherence assessment.

In conclusion, the reviewed literature reveals that while significant strides have been made in using LLMs for various NLP tasks, the task of local coherence, particularly through the shuffle test, remains the most frequently explored. On the other hand, global coherence has been relatively underexplored, with fewer studies dedicated to understanding and evaluating it comprehensively. Moreover, existing studies have yet to focus on identifying the specific types of incoherence present within texts, pinpointing their locations, and understanding the underlying reasons for these incoherences.

The potential of LLMs, as evidenced by the recent advancements discussed, offers a promising avenue to address these gaps. LLMs' ability to handle complex linguistic tasks, coupled with their growing capacity to generate rationale-supported evaluations, can significantly expand the scope of automatic coherence analysis. Our research explores the application of LLMs across three coherence tasks-local, global, and incoherence identification-

starting from the Methodology presented on Chapter 4. This investigation positions our work within the broader field, contributing to a deeper understanding of how LLMs can enhance automatic coherence evaluation. Through this approach, we aim to offer more nuanced and precise insights into text coherence, advancing the capabilities of NLP models.

3.4 Comparative analysis in Textual Coherence Evaluation

To date, there are no existing studies that directly compare LLMs specifically for the tasks related to textual coherence evaluation. Additionally, there are no studies that compare older models for coherence analysis in a direct manner. However, several works do include evaluations of model performance in coherence, particularly focusing on local coherence through the shuffle test, which assesses the linearity of idea flow.

Studies such as Dias (2016), Nguyen and Joty (2017) and Liu, Zeng and Li (2020) compare various models for local coherence tasks. These studies focus on detecting disruptions in logical sequencences by permuting sentence orders to break the linearity of idea flow, which serves as a proxy for assessing local coherence.

For global coherence, the works of Lai and Tetreault (2018) and Mikkelsen *et al.* (2022) evaluate models traditionally used for local coherence in the context of both global and local coherence withing their respective purposed corpora. These studies extend the application of coherence models, showing how they perform in identifying the overall thematic constistency of longer texts beyond sentence level, but not keeping that aside.

However, for incoherence identification, there are no current research that direct addresses the task using LLMs or other models. This area remains unexplored, making our contribution unique in its comprehensive evaluation of LLMs across local coherence, global coherence and incoherence identification.

Table 3.1 presents studies that include comparisons, even if indirect, of models for the tasks mentioned in this thesis. This includes evaluations of models for local coherence, global coherence, and incoherence identification, as discussed in Chapter 3.

Table 3.1: Summary of Comparative Model Evaluations in Coherence Tasks.

Study	**Local Coherence**	**Global Coherence**	**Incoherence Identification**
Elsner, Austerweil and Charniak (2007)	✓		
Barzilay and Lapata (2008)	✓		
Lin, Ng and Kan (2011)	✓		
Dias and Pardo (2015)	✓		
Mesgar and Strube (2015)	✓		
Dias (2016)	✓		
Liu, Zeng and Li (2020)	✓		
Nguyen and Joty (2017)	✓		
Najafi and H. Darooneh (2017)		✓	
Lai and Tetreault (2018)	✓	✓	
Mikkelsen *et al.* (2022)	✓	✓	
This Thesis	✓	✓	✓

4 Methodology

In this chapter, we present the methodology employed in our research to address the research objectives outlined in Chapter 1. The aim of this methodology was to guide our study and to ensure the validity and reliability of our findings.

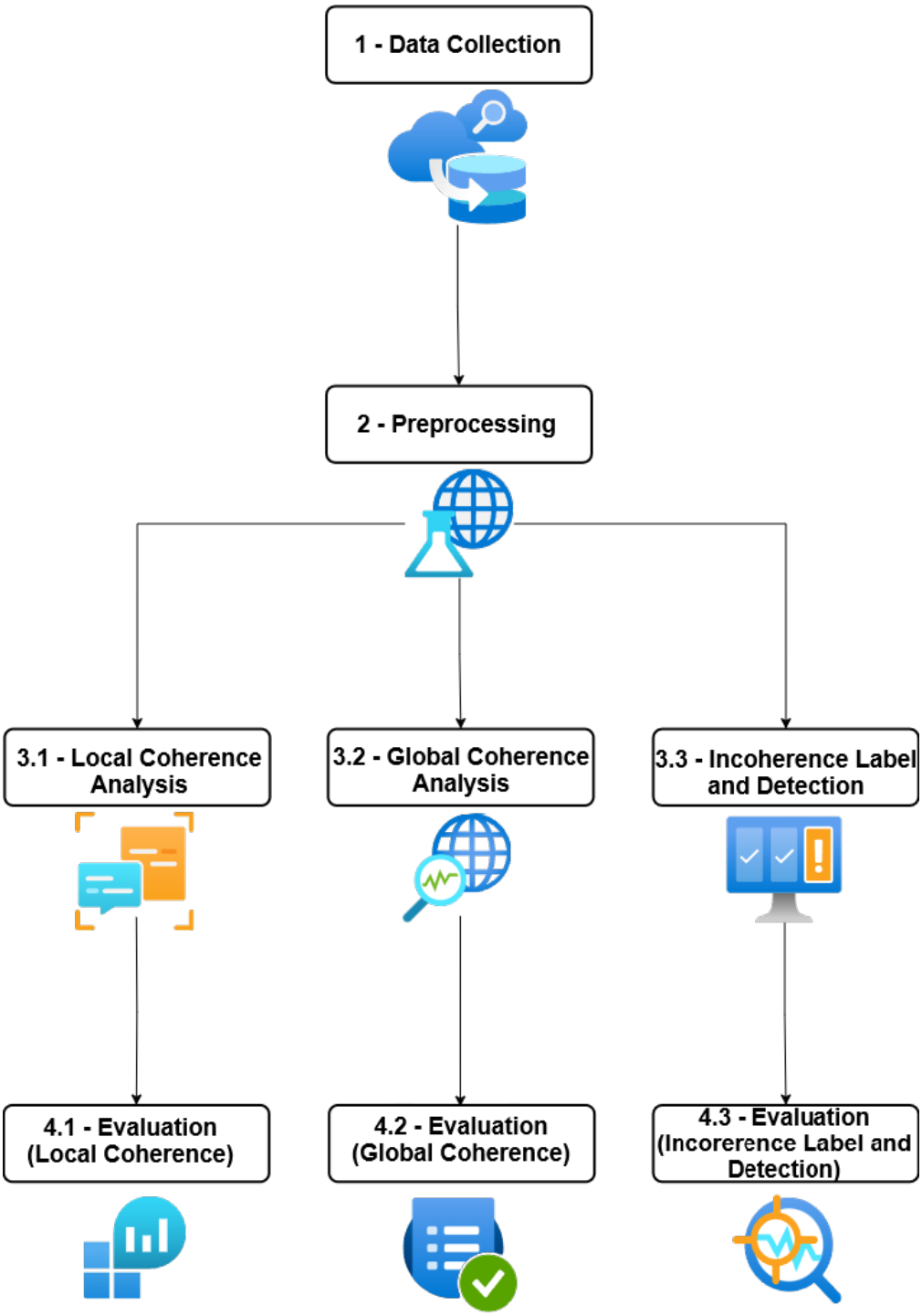
As our research aims to analyze and compare various Large Language Models, we have carefully designed our methodology, which encompasses research design, data collection, preprocessing and investigation about the ability of these models to evaluate textual coherence at different levels. This chapter provides an overview of our research approach and highlights the key components of our methodology.

The data collection phase focused on gathering texts for evaluating coherence across various tasks. The corpora – COCA, CST News, GCDC, and DDisCo – were compiled and annotated as needed for these evaluations. All corpora were used for Local Coherence Classification, while GCDC and DDisCo were the primary focus for Global Coherence, with an annotated subset of 100 texts from COCA and CST News also included. For Incoherence Identification, a subset of 130 texts from all corpora except DDisCo was utilized. During the preprocessing stage, we standardized the data to ensure uniformity across the text formats used, and organized the texts for the tasks by assigning them random IDs and structuring them accordingly. The testing phase involved utilizing LLMs to achieve the target tasks, which include binary local classification of text coherence (coherent versus incoherent), global coherence classification, and the examination of models' ability to identify incoherences, i.e., elements or segments of the text that break logical and thematic continuity, also known as detecting and flagging incoherences within texts.

The content of this chapter is organized according to the methodological steps described, as illustrated by Figure 4.1 and following this format: Section 4.1 describes the data collection phase used in the three tasks, Section 4.2 addresses the approach adopted for Local Coherence, Section 4.4 focuses on Global Coherence, and Section 4.5 discusses the

approaches used for the task of Identifying Incoherences.

Figure 4.1: Workflow Methodology for evaluating the models on the Textual Coherence Analysis tasks.



Source: Author, 2024.

4.1 Data Collection

In this section, we describe the data sources used in our research, as well as the process of obtaining these data. The data were collected from four main sources: the Corpus of Contemporary American English (COCA), specifically the blog and academic sections, CST News Corpus, the Grammarly Corpus of Discourse Coherence (GCDC), and the DDisCo: A Discourse Coherence Dataset for Danish (DDISCO).

The data collection process involved obtaining raw texts from these sources and subsequently standardizing them to ensure uniformity in the data format. Each text was assigned a unique ID and categorized according to its source and text type. This process was essential to prepare the data for the subsequent preprocessing and analysis phases. Below, we detail each of these sources.

4.1.1 COCA (Corpus of Contemporary American English)

The Corpus of Contemporary American English (COCA) (Davies, 2008) was created by Mark Davies and is renowned as the only large and "balanced" corpus of American English. It is one of the most widely-used corpora of English (Smith; Baldridge, 2013) and is associated with other corpora from English-Corpora.org, providing deep overviews into variations in English.

The COCA contains more than one billion words of text, with over 25 million words added each year from 1990 to 2024. As Table 4.1 details, the corpus includes texts from eight different genres: spoken, fiction, magazines, newspapers, academic texts, TV and movie subtitles, blogs, and web pages. As pinpointed by Davies (2008), "This diverse range makes COCA the only corpus that is large, recent, and balanced across a wide range of genres".

Unfortunately, full access to the corpus is paid, although the payment allows you to download the full data in any or all of the three different formats:

- Database: This format is composed of three SQL tables (corpus, lexicon and sources).
- Word/lemma/PoS: Word, lemma, and part of speech in vertical format; can be imported into a database. In most of the corpora, texts are separated by a line with @@ and the textID.
- Linear text: This format provides a textID for each text, and then the entire text on the same line. In this format, words are not annotated for part of speech or lemma. Additionally, contracted words like <can't> are separated into two parts (ca n't) and punctuation is separated from words (eye level . As her).

Table 4.1: Genre Distribution in the COCA Corpus

Genre	# texts	# words	Explanation
Spoken	44,803	127,396,932	Transcripts of unscripted conversation from more than 150 different TV and radio programs (examples: All Things Considered (NPR), Newshour (PBS), Good Morning America (ABC), Oprah)
Fiction	25,992	119,505,305	Short stories and plays from literary magazines, children's magazines, popular magazines, first chapters of first edition books 1990-present, and fan fiction.
Magazines	86,292	127,352,030	Nearly 100 different magazines, with a good mix between specific domains like news, health, home and gardening, women, financial, religion, sports, etc.
Newspapers	90,243	122,958,016	Newspapers from across the US, including: USA Today, New York Times, Atlanta Journal Constitution, San Francisco Chronicle, etc. Good mix between different sections of the newspaper, such as local news, opinion, sports, financial, etc.
Academic	26,137	120,988,361	More than 200 different peer-reviewed journals. These cover the full range of academic disciplines, with a good balance among education, social sciences, history, humanities, law, medicine, philosophy/religion, science/technology, and business.
Web (Genl)	88,989	129,899,427	Classified into the web genres of academic, argument, fiction, info, instruction, legal, news, personal, promotion, review web pages (by Serge Sharoff). Taken from the US portion of the GloWbE corpus.
Web (Blog)	98,748	125,496,216	Texts that were classified by Google as being blogs. Further classified into the web genres of academic, argument, fiction, info, instruction, legal, news, personal, promotion, review web pages. Taken from the US portion of the GloWbE corpus.
TV/Movies	23,975	129,293,467	Subtitles from OpenSubtitles.org, and later the TV and Movies corpora. Studies have shown that the language from these shows and movies is even more colloquial / core than the data in actual "spoken corpora".
Total	**485,179**	**1,002,889,754**	

Source: <https://www.english-corpora.org/coca/>

In our case, we utilized only the sample data from the linear text format, which is randomly selected from each of the corpora (typically about 1/100th of the total number of texts). Although the full-text data continues to expand by 130-150 million words each month, with the most recent update in July 2024, the samples we used cover the years 2010-2016. For this research, we specifically focused on the blog and academic portions of COCA due to their relevance and diversity, as these subcorpora provide a broad spectrum of writing styles and contexts. Blog texts include 991 texts collected from from personal and professional posts, providing an informal and often subjective perspective on various topics. Academic texts consist of 256 articles and publications that feature a formal and structured writing style, focused on precision and clarity. The rationale for this selection was to obtain texts that likely exhibit higher levels of coherence (academic) and lower levels (blogs), to be used across all proposed tasks.

It is important to note that the linear text format assigns a textID to each text, followed by the entire text on the same line without part-of-speech or lemma annotations. In blog texts, <p> tags indicate paragraphs and distinct posts, while @@@@ marks denote redacted text. In academic texts, <p> tags signify paragraph breaks. Additionally, the spacing between words and punctuation plays a significant role in these texts.

The Table 4.2 illustrates typical examples from the blog and academic sections of COCA. Blog examples exhibit informal, subjective writing styles, whereas academic examples showcase formal, structured writing focused on clarity and precision.

Table 4.2: Blog and Academic text eamples from COCA

TextID	**Blog Example**
@@5208041	"The school year now ending has been a disastrous year . And yet , this word is very weak . A disastrous year . After we returned somehow a failed revolution that was quickly broken our wildest dreams , this time we all very nearly succumbed to the fury of devastating Nargis . But where does this name he Is a man 's name A woman 's name I do not know but for me it is now synonymous with destructive madness . <p> One day at the end of my studies in chemistry , I think for a moment to begin studies in meteorology . I 've always been fascinated by the weather . The weather is what it means Where the wind comes and goes when he 's hectopascals millibars and other degrees Celsius or Fahrenheit had no secrets for me to era . Time is the condition of a fluid , air , which penetrates everywhere and that is subject to certain conditions of pressure and temperature . <p> In France , I even hope of the coming of a storm and get ahead of his arrival . With a friend we had braved the @ @ @ @ @ @ @ @ @ @ the beach at Dunkirk .

TextID	Academic Example
@@4001341	A computerized block design task was developed which records temporal and nontemporal measures of performance . This study evaluates the reliability of the measures and reports their intercorrelations . With one exception , the measures showed moderate to good reliability . The results indicate that increasing the difficulyv of the task and testing a more diverse sample may be necessary for improved reliability . A nontemporal method of scoring a block-design task would be useful when testing persons who have handicaps affecting motor skills . but no central nervous system deficits . <p> Previous research has demonstrated that college undergraduates have little difficulty adapting to a computerized block design task . Temporal performance on the computerized task has a reliability coefficient similar to the Block Design subtest of the Wechsler Adult Intelligence Scale-Revised (WAIS-R) , and is moderately correlated with temporal performance on that subtest (Martin &; Wilcox , 1989 ; Wechsler , 1981) . <p> Tasks of this type assess analysis of visual patterns , visual-spatial manipulation , and synthesis of visual patterns . However , they are also affected by psychomotor speed (Cohen , Montague , Nathanson , &; Swerdlik , 1988 @ @ @ @ @ @ @ @ @ @ motorically slow , yet retain good visual-spatial skills . In large part , this bias is attributable to the importance of temporal measures in conventional scoring methods . <p> In the interest of developing a method of scoring a block-design task that is unaffected by psychomotor speed , the present study investigates a nontemporal approach to measuring block design performance . The study examines the reliability of three nontemporal measures , and explores the relationship of temporal and nontemporal measures of performance on the computerized block design task . METHOD <p> Subjects <p> Subjects were 53 undergraduates attending a small comprehensive university who volunteered to take part in the study to receive extra credit in their introductory course in psychology . Of 53 subjects , 43 were female and 10 were male . Another five subjects were tested , but excluded from the data analysis because their knuckles inadvertently hit the " finished " button , prematurely ending a trial .

4.1.2 CST News Corpus

The CST News Corpus (Aleixo; Pardo, 2008) was created to support research on multi-document summarization and comprises a collection of news articles from various sources. The corpus is organized into clusters, each containing multiple news articles about the same event or topic. This organization allows researchers to utilize the corpus for tasks such as summarization, discourse analysis, and studies related to the coherence and cohesion of information across different documents discussing the same subject.

The corpus contains 50 collections of Brazilian Portuguese texts. Each collection comprises approximately three documents on the same subject but from different sources, along with a human-generated summary. The corpus was annotated by four computational linguists, achieving satisfactory annotation agreement, and was specifically created to evaluate summarization of journalistic documents from different sources such as Jornal do Brasil, Folha de São Paulo, and O Estado de São Paulo. Originally, it had one multi-document summary for each text collection on a certain topic, but was later extended by Dias (2016), who added five more summaries for each of the 50 collections, resulting in a total of 300 summaries.

The texts were manually collected from online newspapers over a period of two months, between August and September 2007. The sources of the texts included online newspapers such as Folha de São Paulo, Estadão, O Globo, Jornal do Brasil, and Gazeta do Povo. These sources were chosen due to their popularity on the web and their coverage of the main news of the day, which is basilar for the corpus, i.e., the same news published in different sources. Journalistic texts were selected for their clear and everyday language, as well as the ease of finding them on the web.

In our research, the CST News Corpus was selected due to its collection of Brazilian Portuguese texts, providing important linguistic diversity for evaluating the models on coherence tasks. Including this corpus helps ensure that the models can handle coherence tasks across different languages. Additionally, it allows for comparison with other approaches, such as the work of Braz and Fileto (2021), which highlighted the efficacy of contextualized language models like BERT in coherence analysis using the CST News Corpus alongside another corpus in their experimental setup.

The corpus files are organized in folders, with each folder representing a different collection of articles on the same topic. Each article within a folder is saved as a separate text file, named with a unique identifier that corresponds to its metadata, as shown in Table 4.3. All corpus files include a version with the source text (plain) and another version of the same text segmented by sentences. Under the same topic, articles from different newspapers, such as Folha, Estadão, and O Globo, are presented.

Table 4.3: Example of Data from CST News

TextID	**News Article Example**
D1_C9_Folha_04-08-2006_13h20.txt	A PF (Polícia Federal) prendeu na manhã desta sexta-feira 23 pessoas suspeitas de envolvimento em esquema da Assembléia Legislativa do Estado de Rondônia para desvio de recursos públicos e influência indevida sobre Poder Judiciário, Ministério Público, Tribunal de Contas e Poder Executivo do Estado. Entre os presos estão o presidente do TJ (Tribunal de Justiça) de Rondônia, desembargador Sebastião Teixeira Chaves, e o presidente da Assembléia Legislativa, deputado José Carlos de Oliveira. A polícia informou que o grupo já desviou R$ 70 milhões. Também foram presos o juiz José Jorge Ribeiro da Luz, o conselheiro do Tribunal de Contas Edilson de Souza Silva, o procurador de Justiça José Carlos Vitachi, o diretor geral da Assembléia Legislativa, José Ronaldo Palitot, servidores, assessores e familiares de deputados. Oito dos presos pela PF serão mandados para Brasília. Os outros ficarão em Rondônia. Segundo a polícia, a Assembléia Legislativa fazia contratos com base em licitações 'viciadas e fraudulentas'. Os recursos públicos eram desviados para pagamentos de serviços, compras e obras supostamente superfaturadas. A polícia informou que, em alguns casos, objetos de contratos nem eram entregues e serviços não eram feitos. As informações coletadas pela PF durante as investigações foram enviadas ao TJ (Tribunal de Justiça) do Estado de Rondônia e ao STJ (Superior Tribunal de Justiça). A polícia também vai abrir nova investigação sobre a participação de desembargadores e conselheiros do Tribunal de Contas no suposto esquema. Ao menos 300 policiais de Amapá, Distrito Federal, Mato Grosso, Acre e Rondônia trabalharam na Operação Dominó.

4.1.3 GCDC (Grammarly Corpus of Discourse Coherence)

The Grammarly Corpus of Discourse Coherence (GCDC) (Lai; Tetreault, 2018) was created to support research on discourse coherence and consists of texts from various real-world sources. The corpus includes emails and online reviews, capturing the type of writing that an average person might produce. Each text in the GCDC corpus is annotated with a global coherence score, rated by both expert raters and untrained annotators via Amazon Mechanical Turk. The annotations are given on a 3-point scale, ranging from low (1) to high coherence (3).

To ensure a wide range of writing styles and contexts, the texts were collected from various sources, representing different types of communication. These domains range from informal online interactions to formal professional emails, specifically selected to create a comprehensive dataset. The corpus includes the following subsets:

- **Yahoo Answers:** This subset includes forum posts from the Yahoo Answers L6 corpus. These posts represent informal, user-generated content where individuals ask questions and receive answers from the community.
- **Clinton Emails:** This subset consists of emails from the publicly released State Department emails of Hillary Clinton. This collection contains a mixture of formal and informal communication, reflecting both professional and personal correspondence.
- **Enron Emails:** This subset includes professional emails from the Enron Corporation, sourced from the Enron Email Dataset. These emails primarily consist of formal, business-related communication, offering a contrast to the informal discourse found in other subsets.
- **Yelp Reviews:** This subset comprises reviews from the Yelp Open Dataset. These reviews are written by the general public and cover a wide range of businesses and services, varying in lenght and detail.

In our research, the GCDC corpus was selected for its inclusion of diverse genres and domains, along with its manual annotations of global coherence. These annotations, provided by both expert raters and untrained annotators via Amazon Mechanical Turk, offer a practical advantage as they eliminate the need for new annotations, saving significant time and resources. Additionally, the GCDC corpus allows for a direct comparison with other coherence models, as it has been used in previous studies to benchmark the performance of language models. This makes it an ideal choice for evaluating the coherence capabilities of our models across various writing styles and contexts.

The GCDC corpus is balanced, containing pairs of 1000 texts for training and 200 texts for testing from all of the four sources. It is organized with metadata for each text to facilitate coherence analysis. Each file includes fields such as *text_id*, which uniquely identifies each text, and *text*, which contains the document itself. For the Yahoo subset, additional fields like *question_title* and *question* provide context for the responses but were not used in model training. Similarly, the *subject* field in the Clinton, Enron, and Yelp subsets offers context for the emails and reviews but was excluded from the training data. Each text is accompanied by three coherence ratings from expert annotators (ratingA1 to ratingA3) and a consensus label (labelA). Additionally, five coherence ratings from MTurk annotators (ratingM1 to ratingM5) and a corresponding consensus label (labelM) are included. Table 4.4 presents examples of texts from the Yelp and Enron subsets, respectively.

Table 4.4: Examples of Data from GCDC Corpus

TextID	**Text Example**
FUYQ99EUHg2TOHMMTy7cFQ	"Most months this buffet at the Silverton has one day a week they do BOGO. For Jan it's BOGO Thursday. Sign up for a card at the player's club and the day of the BOGO print out a coupon at the kiosk and head to the buffet. The coupon is only good for that day. Also the same day you sign up for the player's card, play table games and earn just 50 points and you get another coupon for a free buffet to use anytime you want. The buffet is pretty good for $9.99, then factor in the BOGO it's only $5/person. Can't beat that deal, as you can't even get full of $5 at a fast food place. There's Mexican food (fish tacos, menudo, pozole, or albondigas soup, ground beef tacos), carving station (roast beef, chicken or turkey, and even their skin fried, ham, grilled veggies), Italian (pizza, pasta), Asian (Pad thai or chow mein, beef and broccoli, egg rolls, Asian soups, string beans), and salad bar. Even for a picky eater, you can make yourself a nice big healthy salad and it will be worth the $5."
Subject:	Seasons Buffet
Expert Ratings:	3, 2, 3 (Label: 3)
MTurk Ratings:	3, 3, 2, 3, 2 (Label: 2)

TextID	Text Example
773548	"Once you have completed your 2nd Current Estimate model, you will need to prepare a Variance Schedule explaining the differences between your 1st Current Estimate numbers and your 2nd Current Estimate numbers. Attached is an example that Caroline Nugent prepared at 3rd Current Estimate for you to utilize as a template. Please complete this variance analysis by noon on Thursday, August 2. Please save all variance analysis at `O:\Corporate\Tax\Lotus\2001 Current Estimate\2nd CE\1st CE to 2nd CE Variances by Business Unit`. This should help us eliminate some questions that might arise. If you have any other current and/or deferred tax adjustments, please detail those in a section at the bottom of the spreadsheet. If you have any questions, please feel free to call me. Thanks, Michelle 3-0931"
Subject:	Variance Schedule
Expert Ratings:	3, 1, 2 (Label: 2)
MTurk Ratings:	3, 2, 3, 2, 3 (Label: 3)

Despite the corpus being available, accessing it requires first obtaining the L6 Yahoo dataset and then requesting the GCDC from its creator. According to Yahoo's data policy, summaries, analyses, and interpretations of the data may be derived and published, provided it is not possible to reconstruct the data from the publication. This limitation complicates access to the corpus, as researchers must navigate additional steps and permissions to obtain the necessary data, potentially delaying research and limiting the use of the corpus in various studies.

4.1.4 DDisCo (A Discourse Coherence Dataset for Danish)

The DDisCo (A Discourse Coherence Dataset for Danish) (Mikkelsen *et al.*, 2022) corpus was developed to fill a significant gap in resources for researching discourse coherence in Danish texts. It includes real-world Danish texts that naturally exhibit both coherence and incoherence, providing a more accurate representation of human writing. Based on the work of Lai and Tetreault (2018), each text is annotated for global coherence using a 3-point Likert scale: low (1), medium (2), and high coherence (3). These annotations were performed by experts with strong backgrounds in linguistics and discourse coherence, ensuring high-quality labels and reliable data for research purposes.

The corpus consists of texts from two main sources:

- **Reddit:** This subset includes blog posts from the Reddit forum, specifically from the subreddit r/Denmark. These posts are informal, user-generated content with a variety of writing styles and coherence levels.
- **Danish Wikipedia:** This subset includes encyclopedic texts from the Danish Wikipedia, providing more formal and structured content.

The Reddit posts were collected using the praw Python package, focusing on comment sections with at least five comments and a word length of 100-300. The Wikipedia texts were extracted from the DanFEVER dataset, specifically selecting entries from Wikipedia while excluding those from Den Store Danske encyclopedia to avoid professional editing.

The corpus was designed with some key criteria in mind. First, the texts are written by a variety of people to capture different writing styles and are not professionally edited to better reflect everyday writing. Second, each text is between 100-300 words in length to ensure sufficient content for coherence evaluation. Third, the dataset includes texts with low, medium, and high coherence levels. Additionally, the data is made publicly available under a license that allows for commercial use, ensuring that it can be widely accessed and utilized by the research community.

In our research, the DDisCo corpus was selected for its inclusion of global coherence annotations performed by human annotators, similar to the GCDC corpus. Additionally, it introduces two new domains – Wikipedia and Reddit – in the Danish language, providing linguistic variety and enabling the evaluation of the models' generalization capabilities across different languages and contexts.

The corpus is freely available and comprises 1002 texts with 801 for training (401 from Wikipedia, 400 from Reddit) and 201 for testing (100 from Wikipedia, 101 from Reddit). Each text includes text, domain, and rating, lacking on a TextID. Texts were cleaned of

HTML tags and newlines, anonymized for sensitive information, and retained if they contained 100-300 words, ensuring a balance between structural coherence and annotation manageability. Examples of these texts, as shown in Table 4.5, illustrate the range and annotation of the dataset.

Table 4.5: Examples of Data from DDisCo Corpus

Text	Domain	Rating
"Selvfølgelig vil de det. Det er jo netop sådan en stat bliver rig, ved at bruge skattepenge på at investere tilbage i samfundet. Dermed for folk mulighed for at bliver eksperter i hvad de laver, innovere, skabe vækst. Det er en investering og jeg forstå simpelthen ikke at folk på blå side kan være så dårlige til at se udover 'jamen så står der et mindre tal i min bankbog'. Og hvad så? Hvad godt gør de der? Det er en bedre investering for dine samfundet og dine børn og deres børn. Medmindre man sidder som CEO med en Rolex afhængighed kan jeg simpelthen ikke forstå man stemmer blå."	Reddit	2
Hans Hansen (væver) Hans Hansen (10. marts 1815 i Køng Sogn, Hammer Herred – 5. november 1867 i Mern) var en dansk husmand og bomuldsvæver. Hans Hansen kom fra Mern. Hansen var aktiv i Bondevennernes Selskab, optaget af at forbedre småfolks kår og blev i 1848 i Præstø Amt opstillet til valget til Den Grundlovgivende Rigsforsamling mod H.N. Clausen. Trods stærke angreb, bl.a. nævntes hans arrestation i forbindelse med en hælerisag i 1836, blev han valgt, men en måned senere blev Hansen presset til at nedlægge sit mandat, der i stedet gik til N.F.S. Grundtvig. Han fortsatte sit politiske arbejde på lokalt plan. Valget i Præstø var det første varsel om en splittelse mellem Bondevennerne og De Nationalliberale. Han blev ca. 1845 gift 1. gang med Ane Kirstine Pedersdatter. 2. gang ægtede han 2. december 1861 i Mern Kirke Maren Albrechtsdatter. Han er begravet på Mern Kirkegård.	Wikipedia	3

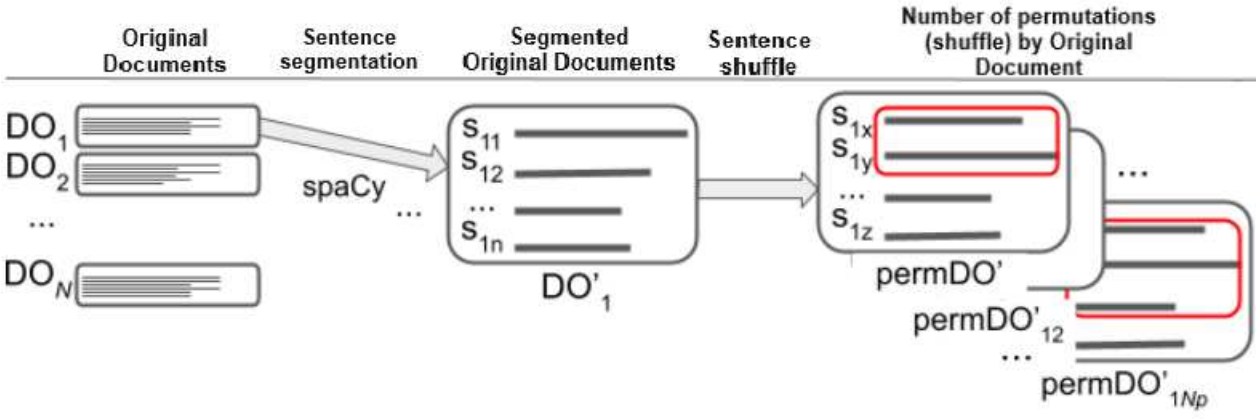
4.2 Local Coherence

As discussed in Chapter 3, the shuffle test is a widely used method for evaluating local coherence (Braz; Fileto, 2021). In this study, we employ the shuffle test on texts from the COCA, GCDC, CST News, and DDisCo datasets. Each text from each corpus was permuted 20 times using the methodology discussed in subsection 4.2.1, which randomly shuffles the sentences within each text, creating permuted versions that disrupt the original coherence. The shuffle test is a critical component of our evaluation, as it challenges the models to distinguish between naturally coherent sequences of sentences and various incoherent permutations. This differentiation is assessed through both APIs and chat interfaces, directly addressing RQ1, which focuses on local coherence, and RQ4, which explores the models' performance in different interaction modes.

4.2.1 Preprocessing for Local Coherence

The Local Coherence Classification task involves determining whether a text is coherent or incoherent at the sentence level. To prepare the documents for this task, we followed a series of steps, transforming Original Documents (DO) into Permutated Documents (PermDO), as illustrated in Figure 4.2:

Figure 4.2: Preprocessing Steps for Local Coherence Classification



Source: adapted from (Braz; Fileto, 2021)

For the COCA corpus, preprocessing involved segmenting the texts, where each document started with a line beginning with "`@@`". These delimiters were used to segment the entire corpus into individual documents, which were then further divided into sentences using

the spaCy library with the *en_core_web_sm* model. Each document was assigned a unique ID following the pattern "coca_BL_01" for blog texts and "coca_AC_01" for academic texts. Documents with fewer than four sentences were excluded, resulting in the removal of 13 texts. Additionally, blog texts containing HTML tags were processed to ensure that paragraphs were treated as distinct sentences and that texts from the same blog were properly separated as distinct documents. During preprocessing, punctuation was also corrected, addressing issues such as misplaced separators between words and punctuation marks, including the correction of apostrophes (e.g., changing "he 's" to "he's").

For the GCDC corpus, which consists of emails and online reviews, tabulations within texts were considered as sentence delimiters. Each text was segmented into sentences based on both punctuation and tabulation, ensuring accurate capture of fragmented sentences. Unique IDs were assigned using the pattern "gcdc_CL_01" for Clinton emails, "gcdc_ER_01" for Enron emails, and "gcdc_YL_01" for Yelp reviews. Texts with fewer than four sentences were excluded, resulting in the removal of 58 texts. The Yahoo portion of the GCDC corpus was excluded because its representation was always composed of a question and answer, with responses often being short and context-dependent.

The DDisCo corpus, containing texts from Reddit and Wikipedia in Danish, required segmentation into sentences using the *da_core_news_sm* model from the spaCy library. Each document was processed to remove non-alphanumeric characters while retaining necessary punctuation. Unique IDs followed the pattern "dsco_RD_01" for Reddit texts and "dsco_WK_01" for Wikipedia texts. This preprocessing step led to the removal of 11 texts with fewer than 4 sentences.

Additionally, the CST News corpus was processed similarly. Each document, consisting of news articles from various sources, was segmented into sentences and assigned unique IDs following the pattern "cstn_NP_01". A total of 49 texts were removed due to insufficient sentence length.

After preprocessing, the texts were serialized into CSV files, each containing two columns: the new IDs and the sentences. The final datasets, after removing texts with fewer than four sentences, consisted of 978 blog texts and 256 academic articles from COCA, 842 texts from GCDC, 991 texts from DDisCo, and 251 texts from CST News. Following segmentation and standardization, we proceeded with the shuffling process. Each text was shuffled 20 times to create incoherent versions by randomly permuting the sentences within each document, generating multiple incoherent samples while keeping the original text intact. This resulted in 19,560 shuffled versions for COCA (Blog), 5,120 for COCA (Academic), 5,020 for CST News, 16,840 for GCDC, and 19,820 for DDisCo. All shuffled texts were saved in new CSV files, with an addition to the ID of each text, permOx, where "x" is a unique identifier

for each permutation. Table 4.6 summarizes the final counts.

Table 4.6: Text Counts After Preprocessing and Shuffling

Corpus	DO	PermDO
COCA (Blog)	978	19,560
COCA (Academic)	256	5,120
CST News	251	5,020
GCDC	842	16,840
DDisCo	991	19,820

4.3 Local Coherence Analysis

After the preprocessing stage, the next step was to test the models on the local coherence task. This phase involved two distinct tests. The first test was conducted through the APIs of various models, including Claude Opus, Claude Sonnet, GPT-3.5, GPT-4, GPT-4o, Bard, Gemini, and LLama 2, using the entirety of the DO (Original Documents) and PermDO (Permutated Documents). The second test was performed through direct interaction with the models via their chat interfaces, aiming to determine whether ordinary users, who only have access to the models via chat and not API, could use the same approach (prompt) to validate the local coherence of texts. In this second stage, only 50 DO and 1000 (20x) PermDO from each corpus were tested.

Given the premise that LLMs can function as effective classifiers (Huang *et al.*, 2024; Peng; Shang, 2024), due to their ability to generate coherent texts and identify incoherent ones, a zero-shot approach was employed. This approach did not involve any fine-tuning of the models. Instead, we used only a prompt and 69,678 texts (3,318 original documents and 66,360 permuted versions) for classification on the approach (i) and 5250 texts (1000 permDO and 50 DO from each corpus) for the approach (ii).

For all models, except Llama, it was necessary to sign up to obtain an API key. The platform Replicate[1] was used to facilitate this process. Specifically for the Claude family of models, the use of a VPN was required, as these models were not available for use in Brazil at the time of testing.

The following standardized prompt (Prompt 1) was prepared for use with all models to classify each text. The prompt was designed to guide the model in assessing the coherence of the text by providing clear instructions on what to look for in terms of logical flow and sentence connectivity. This standardized prompt was consistent across all models and

[1]https://replicate.com/blog/run-llama-3-with-an-api

evaluation methods to ensure comparability of results. Detailed intermediate steps in the creation of this prompt are provided in Appendix A.1.

Prompt 1

You are an advanced AI model specializing in text analysis with expertise in evaluating text coherence. Your task is to classify the coherence of the given text. Coherence in this context means that the text logically flows and makes sense, with each sentence and idea connected in a clear and understandable way.

Objective: Assess the text's coherence by determining if the logical flow, connection of ideas, and overall clarity are maintained throughout the text. Classify the text as either ''coherent'' or ''incoherent'' based on these criteria.

Instructions:

- Read the provided text thoroughly. Focus on the transitions between sentences and paragraphs, the logical sequence of ideas, and the overall structure.
- Evaluate the logical flow: Determine if the text follows a logical progression of ideas from one sentence to the next and from one paragraph to another.
- Assess the connections between ideas: Check if each sentence and paragraph connects naturally and contributes to the logical flow.

Classify the text:

- Respond with ''coherent'' if the text logically flows, makes sense, and has clear connections between ideas.
- Respond with ''incoherent'' if the text lacks logical flow, is confusing, or has disjointed ideas.

Here is the text for analysis:

[Text goes here]

Response format:

- Coherent: The text logically flows, makes sense, and has clear connections between ideas.

- `Incoherent: The text lacks logical flow, is confusing, or has disjointed ideas.`

`Take a deep breath and work on this problem step-by-step.`

The prompt provided was designed using the following prompt engineering strategies to maximize the model's performance in evaluating text coherence:

- **Focused Content Analysis:** The prompt directs the model to specifically focus on evaluating the coherence of the text, breaking down the task into analyzing logical flow, connections between ideas, and overall clarity (Liu *et al.*, 2021)
- **Breaking the Task Down:** By breaking down the task into smaller, manageable steps, the prompt guides the model through a structured process. This strategy, recommended in the literature on prompt engineering (Microsoft, 2024), helps in tackling complex tasks by simplifying them into sequential steps, which enhances the model's accuracy and comprehensiveness in response generation.
- **Chain-of-Thought Prompting:** The prompt employs a chain-of-thought approach, encouraging the model to think step-by-step through the task. This method, as highlighted in academic research (Wei *et al.*, 2023), improves the model's reasoning capabilities by mimicking human-like thought processes, thus leading to more thorough and logically consistent outputs.
- **Clear Syntax:** Using clear and detailed instructions, along with a well-structured format, the prompt communicates the task effectively to the model. This use of clear syntax, as supported by research (Liu *et al.*, 2021), ensures that the model understands and follows the intended structure of the task, which is crucial for achieving high-quality responses.
- **Zero-Shot Prompting:** Although the prompt provides detailed instructions, it does not include specific examples of coherent or incoherent texts. This zero-shot prompting technique (Reynolds; McDonell, 2021) challenges the model to utilize its pre-existing knowledge and understanding of coherence without relying on direct examples, testing its generalization capabilities.

These strategies were adopted to ensure that the model could effectively classify texts based on their coherence, leveraging its inherent capabilities without additional fine-tuning.

The classification process using APIs involved several steps. First, the API keys were configured to authenticate requests to the models' APIs. A function, "*classify_text*", was

defined to handle the classification task. This function used the prompt asking the model to analyze the provided text and determine its coherence. The prompt was then sent to the models, which returned a response indicating whether the text was coherent or incoherent.

The data consisted of Original Documents (DO) and Permutated Documents (PermDO), saved in CSV files containing IDs and separated sentences, as per the preprocessing step. To pass these texts to the model, it was necessary to first concatenate the sentences of each text based on their IDs to form complete documents. These coherent and incoherent texts were then iteratively passed to the *classify_text* function, and the responses were recorded. It is important to note that the IDs were not passed to the model to ensure unbiased classification.

The classification results, along with the corresponding IDs, were saved in a new CSV file. This approach allowed for easy verification of whether the model assigned the correct classification to each text based on its ID. Finally, the collected results were compared with the true labels of the texts to calculate performance metrics such as accuracy, precision, recall, and F1 score, as discussed in Chapter 6. Additionally, for each model, a confusion matrix was generated to visualize the performance in terms of true positives, false positives, true negatives, and false negatives.

4.4 Global Coherence

Global coherence ensures that a text is logical and consistent, with well-organized ideas presented in a clear manner. Unlike local coherence, which focuses on the connections between individual sentences, global coherence considers the text as a whole.

In this study, we addressed RQ2 and RQ4 by evaluating global coherence through APIs and chats using texts from the GCDC and DDisCo datasets, both of which already contain human annotations. Additionally, we used the COCA and CST News corpora for this task. Since these corpora did not have pre-existing annotations, we conducted a manual annotation process. Three annotators with backgrounds in linguistics evaluated a subset of 100 texts, assigning coherence scores on a Likert scale (low, medium, high). This subset included 10 academic texts from COCA, 60 blog texts from COCA, and 30 news articles from CST News. The distribution was chosen based on the length and nature of the texts, ensuring a diverse range of coherence levels.

4.4.1 Preprocessing for Global Coherence

For the Global Coherence Classification task, the corpora GCDC and DDisCo already had human annotations. To standardize these annotations, we used the expert consensus label (labelA) metric for the GCDC corpus. Each text in these corpora was assigned a unique ID to facilitate handling. The type of ID assigned was similar to the local coherence task, but with the prefix "GC_".

For the corpora that lacked annotations, namely COCA and CST News, a new annotation phase was conducted with three annotators. Following the methodology used by Lai and Tetreault (2018) in GCDC and adopted by Mikkelsen *et al.* (2022) to build the DDisCo, the annotators classified each text globally in terms of coherence using a Likert scale (low, medium, high). They assigned a score of 1 for low coherence, 2 for medium coherence, and 3 for high coherence. Additionally, like in Local Coherence preprocessing steps, blog texts containing HTML tags were processed to ensure that paragraphs were treated as distinct sentences and that texts from the same blog were properly separated as distinct documents. During preprocessing, punctuation was also corrected, addressing issues such as misplaced separators between words and punctuation marks, including the correction of apostrophes (e.g., changing "he ‘s" to "he‘s"). Furthermore, paragraph delimiter tags were replaced with actual paragraph breaks. In both corpora, paragraphs were marked with two tabulations to visually distinguish them more comfortably.

The annotators evaluated a subcorpus of 100 texts in total: 10 academic texts from COCA, 30 texts from CST News, and 60 blog texts from COCA. This division was made because (i) academic texts from COCA are significantly longer than the others, (ii) blog texts from COCA are more likely to exhibit incoherence due to their user-generated content nature, and (iii) CST News contains news articles from newspapers, which follow a standard language less prone to incoherence.

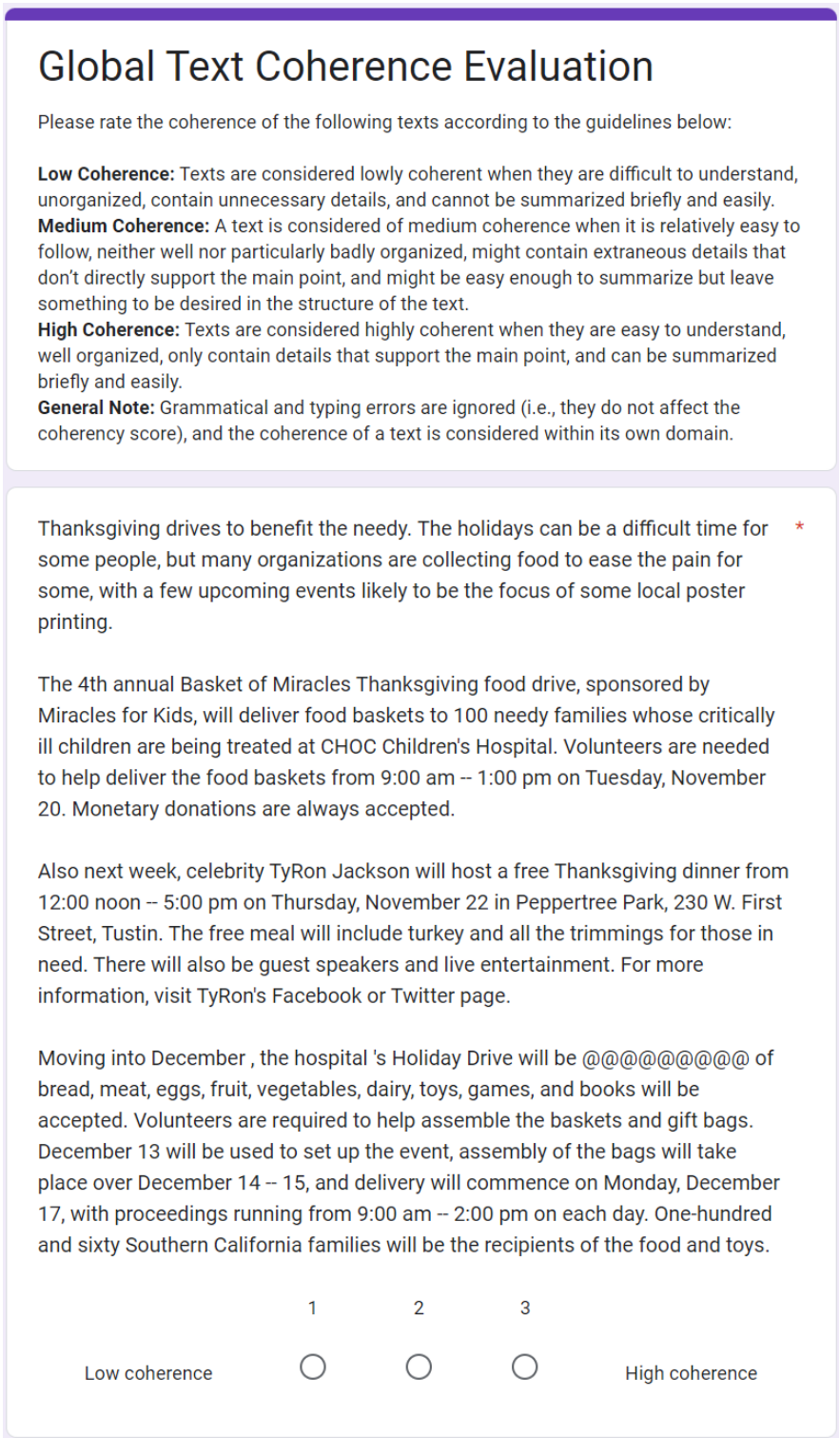
The three annotators, who had backgrounds in languages and/or linguistics, were familiar with each other and could communicate freely when they encountered difficulties, allowing them to either reach an agreement or maintain their differing opinions. Although no formal training sessions or follow-up meetings were conducted, all annotators worked from a common understanding of coherence based on the work of Koch and Travaglia (2003), in addition to following the specific instructions for each task. The annotation period was short, consisting of just one week, during which they accessed the texts and provided their evaluations using a Google Form as shown on Figure 4.3. This form was populated with the texts via a Google Apps Script, which acted as an intermediary to transfer data from a Google Sheet to the Google Form.

The process involved creating a Google Form linked to a Google Sheet where responses were recorded. Another sheet in the same document stored the texts to be annotated, organized with unique identifiers. Using Google Apps Script, the texts were automatically added to the form. The script fetched the texts from the Google Sheet and created a new form item for each text, displaying the text alongside a Likert scale for coherence rating. The annotators then used this form to evaluate the texts, adhering to specific instructions based on the guidelines from Barzilay and Lapata (2008) and Lai and Tetreault (2018) to ensure consistent and standardized annotations. As shown in the Figure 4.3, the criteria for coherence were defined as follows:

- **Low Coherence:** Texts are considered lowly coherent when they are difficult to understand, unorganized, contain unnecessary details, and cannot be summarized briefly and easily.
- **Medium Coherence:** A text is considered of medium coherence when it is relatively easy to follow, neither well nor particularly badly organized, might contain extraneous details that don't directly support the main point, and might be easy enough to summarize but leave something to be desired in the structure of the text.
- **High Coherence:** Texts are considered highly coherent when they are easy to understand, well organized, only contain details that support the main point, and can be summarized briefly and easily.
- **General Note:** Grammatical and typing errors are ignored (i.e., they do not affect the coherency score), and the coherence of a text is considered within its own domain.

Each text received a unique ID in the Google Sheet for ease of reference and processing. The ID structure used the prefix "GC_" followed by corpus-specific identifiers similar to those used in the Local Coherence Task. A consensus label, calculated as the average of the ratings given by the three annotators, was also added to the sheet, as in Lai and Tetreault (2018).

Figure 4.3: Form for Global Text Coherence Annotation from a COCA Blog text



Global Text Coherence Evaluation

Please rate the coherence of the following texts according to the guidelines below:

Low Coherence: Texts are considered lowly coherent when they are difficult to understand, unorganized, contain unnecessary details, and cannot be summarized briefly and easily.
Medium Coherence: A text is considered of medium coherence when it is relatively easy to follow, neither well nor particularly badly organized, might contain extraneous details that don't directly support the main point, and might be easy enough to summarize but leave something to be desired in the structure of the text.
High Coherence: Texts are considered highly coherent when they are easy to understand, well organized, only contain details that support the main point, and can be summarized briefly and easily.
General Note: Grammatical and typing errors are ignored (i.e., they do not affect the coherency score), and the coherence of a text is considered within its own domain.

Thanksgiving drives to benefit the needy. The holidays can be a difficult time for some people, but many organizations are collecting food to ease the pain for some, with a few upcoming events likely to be the focus of some local poster printing. *

The 4th annual Basket of Miracles Thanksgiving food drive, sponsored by Miracles for Kids, will deliver food baskets to 100 needy families whose critically ill children are being treated at CHOC Children's Hospital. Volunteers are needed to help deliver the food baskets from 9:00 am -- 1:00 pm on Tuesday, November 20. Monetary donations are always accepted.

Also next week, celebrity TyRon Jackson will host a free Thanksgiving dinner from 12:00 noon -- 5:00 pm on Thursday, November 22 in Peppertree Park, 230 W. First Street, Tustin. The free meal will include turkey and all the trimmings for those in need. There will also be guest speakers and live entertainment. For more information, visit TyRon's Facebook or Twitter page.

Moving into December , the hospital 's Holiday Drive will be @@@@@@@@@@ of bread, meat, eggs, fruit, vegetables, dairy, toys, games, and books will be accepted. Volunteers are required to help assemble the baskets and gift bags. December 13 will be used to set up the event, assembly of the bags will take place over December 14 -- 15, and delivery will commence on Monday, December 17, with proceedings running from 9:00 am -- 2:00 pm on each day. One-hundred and sixty Southern California families will be the recipients of the food and toys.

	1	2	3	
Low coherence	○	○	○	High coherence

Source: Author, 2024.

The agreement among the three annotators was further measured using Fleiss' Kappa (Fleiss, 1981), which assesses the reliability of agreement between three or more raters. The overall Fleiss' Kappa for the entire set of 100 texts was 0.8952, considered excellent according to the scale proposed by Fleiss (1981). The values for each category of text are presented in Table 4.7, which also includes the classification of the values according to Fleiss' scale: poor ($K < 0.4$), satisfactory to good ($0.4 \leq K < 0.75$) and excellent ($K \geq 0.75$). Overall, the agreement among annotators was excellent across all sections of the subcorpus of 100 annotated texts. Notably, all annotators assigned a score of 3 (high coherence) to the academic texts from COCA, resulting in a perfect Kappa score for these texts. However, the majority of discrepancies occurred in the blog texts from COCA, likely due to the larger number of texts in this set and their more complex nature compared to academic and journalistic texts.

Table 4.7: Fleiss' Kappa for Inter-Rater Agreement on Global Coherence Annotations

Corpus	**Number of Texts**	**Fleiss' Kappa**
COCA (Academic)	10	1.0000
COCA (Blog)	60	0.7843
CST News	30	0.9075
Overall	**100**	**0.8952**

4.4.2 Global Coherence Analysis

After the preprocessing stage, the next step was to test the models on the global coherence task. This phase involved two distinct tests, much like the local coherence task. The first test was conducted through the APIs of various models, including Claude Opus, Claude Sonnet, GPT-3.5, GPT-4, GPT-4o, Gemini, and LLama 2. Notably, BARD was not available for this task since was replaced by Gemini in February of 2024. The second test was performed through direct interaction with the models via their chat interfaces. This was done to determine whether ordinary users, who only have access to the models via chat and not API, could use the same approach (prompt) to validate the global coherence of texts.

The texts used in these tests included the 100 texts annotated during the preprocessing stage, 1200 texts from the DDisCo corpus, and 842 texts from the GCDC corpus, making a total of 2142 texts. Both tests (API and chat) used this combined corpus.

The methodology for the global coherence analysis was similar to that employed for the Local Coherence Task, with adjustments made for the differences in evaluating global coherence. In this phase, there was no need for sentence concatenation since the corpora were already in a coherent format without sentence segmentation.

Similar to the local coherence task, a zero-shot approach was employed for evaluating

global coherence, leveraging the ability of LLMs to classify texts on a scale from low to high coherence. The API keys, which were already obtained, were utilized as described in Section 4.3.

The following standardized prompt (Prompt 2) for global coherence assessment was crafted using the same principles established for annotating the 100 texts by the human annotators, ensuring fairness and consistency in the evaluation. Detailed intermediate steps in the creation of this prompt are provided in Appendix A.2.

Prompt 2

You are an advanced AI model specializing in text analysis. Your task is to classify the coherence of the given text based on the following criteria:

Low Coherence: The text is difficult to understand, unorganized, contains unnecessary details, and cannot be summarized briefly and easily.

Medium Coherence: The text is relatively easy to follow but is neither well nor poorly organized. It might contain extraneous details that don't directly support the main point and might be easy enough to summarize but leave something to be desired in the structure of the text.

High Coherence: The text is easy to understand, well-organized, contains only details that support the main point, and can be summarized briefly and easily.

General Note: Grammatical and typing errors are ignored (i.e., they do not affect the coherency score), and the coherence of a text is considered within its own domain.

Objective: Assess the coherence of the provided text and classify it as ''Low Coherence,'' ''Medium Coherence,'' or ''High Coherence'' based on the criteria above.

Instructions:

Read the provided text carefully. Focus on the overall structure, organization, and relevance of details to the main point.

Evaluate the text based on the following criteria:

- **Low Coherence:** Is the text difficult to understand? Is it unorganized? Does it contain unnecessary details? Is it hard to summarize briefly?

- **Medium Coherence:** Is the text relatively easy to follow? Is it neither well nor poorly organized? Does it contain some extraneous details? Can it be summarized, but with some structural issues?
- **High Coherence:** Is the text easy to understand? Is it well-organized? Do all details support the main point? Can it be summarized briefly and easily?

Ignore grammatical and typing errors. These do not affect the coherence score.
Classify the text:

- Respond with **‘‘Low Coherence’’** if the text meets the criteria for low coherence.
- Respond with **‘‘Medium Coherence’’** if the text meets the criteria for medium coherence.
- Respond with **‘‘High Coherence’’** if the text meets the criteria for high coherence.

Here is the text for analysis:

[Text goes here]

Please respond with ‘‘Low Coherence,’’ ‘‘Medium Coherence,’’ or ‘‘High Coherence’’ based on the criteria above. Take a deep breath and work on this problem step-by-step.

The same techniques employed in the prompt creation for the Local Coherence Analysis, as described in Subsection 4.3, were used in developing this prompt. A function, "*classify_text*", was defined to handle the classification task. This function used the prompt to ask the model to analyze the provided text and determine its coherence. The prompt was then sent to the models, which returned a response indicating the global level of coherence, ranging from low to high.

The classification results, along with the corresponding IDs, were saved in a new CSV file. This approach facilitated easy verification of whether the model assigned the correct classification to each text based on its ID. The texts were compared with the annotations from DDisCo, the consensus labelA from GCDC, and the average ratings given to each text during the preprocessing stage. Performance metrics such as accuracy, precision, recall, and F1 score were then calculated to evaluate the model's effectiveness.

4.5 Incoherence Identification

In this section, we outline the methodology employed for identifying incoherences within texts through APIs and chats, related to the RQ3 and RQ4. Unlike the global and local coherence tasks, this task involves pinpointing specific segments within a text that disrupt its logical flow and coherence. Identifying incoherences allows for a deeper understanding of the nuances of text coherence and aids in understanding how models can provide more granular feedback on writing quality.

4.5.1 Preprocessing for Incoherence Identification

For the Incoherence Identification task, none of the corpora had previous annotations. A subcorpus of 130 texts was selected for this task, consisting of the same 100 texts used in the Global Coherence task (10 from the academic portion of COCA, 60 from the blog portion of COCA, and 30 from CST News) and an additional 30 texts from the GCDC corpus (10 Yelp, 10 Clinton, and 10 Enron). Texts from the DDisCo corpus were not utilized in this stage due to the annotators' lack of proficiency in Danish. The texts were assigned IDs similar to those in the Local Coherence and Global Coherence tasks, but with the prefix "IN_".

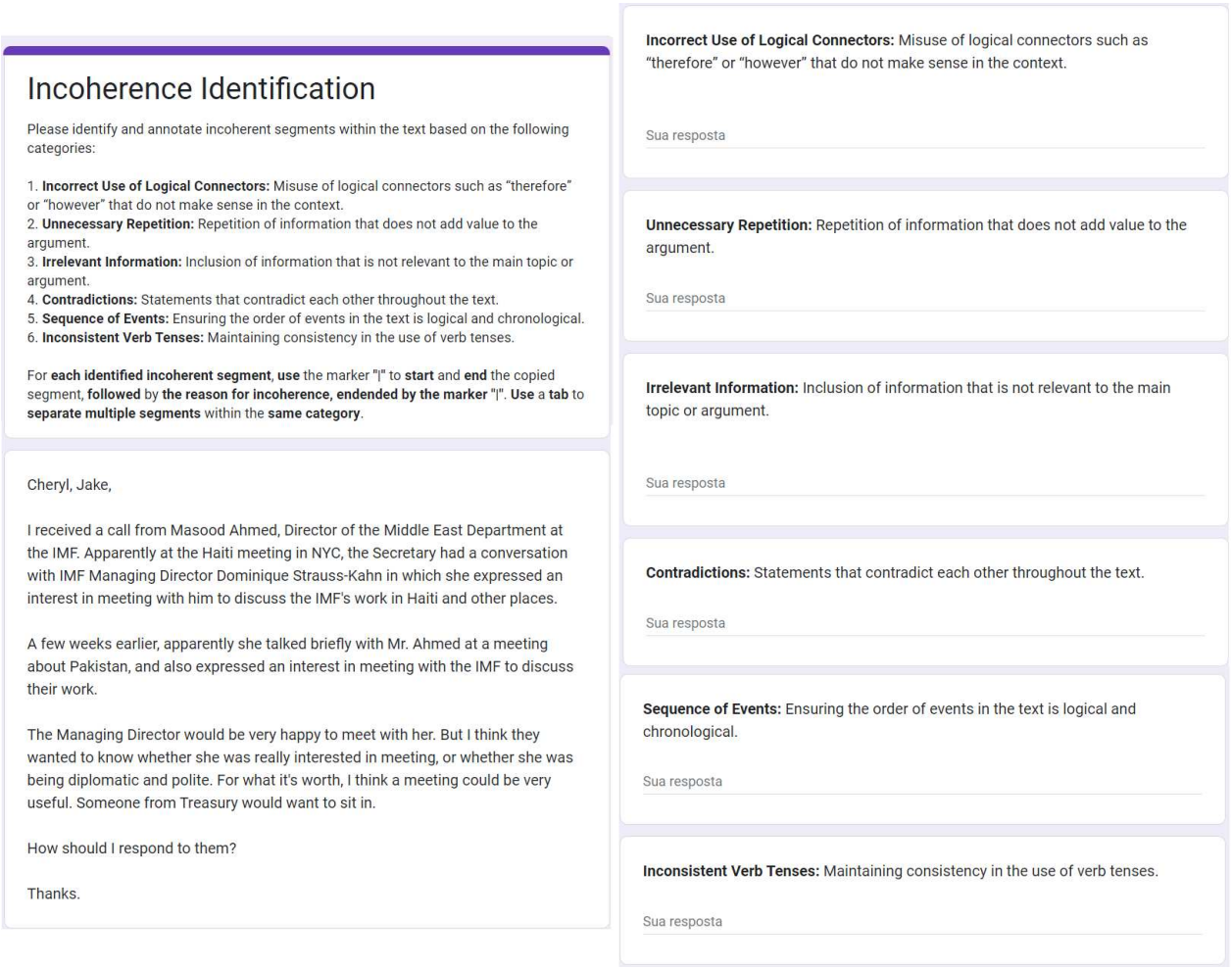
In this task, three annotators identified incoherent segments within texts, focusing on the following categories. These categories were defined by exploring and expanding upon the works of Van Dijk (1977), Koch and Travaglia (2003), and Barzilay and Lapata (2008), who address some types of incoherence when discussing coherence:

- **Incorrect Use of Logical Connectors:** Misuse of logical connectors such as "therefore" or "however" that do not make sense in the context.
- **Unnecessary Repetition:** Repetition of information that does not add value to the argument.
- **Irrelevant Information:** Inclusion of information that is not relevant to the main topic or argument.
- **Contradictions:** Statements that contradict each other throughout the text.
- **Sequence of Events:** Ensuring the order of events in the text is logical and chronological.
- **Inconsistent Verb Tenses:** Maintaining consistency in the use of verb tenses.

Unlike the Global Coherence annotation, each text for the Incoherence Identification task was evaluated based on six specific questions, each targeting a distinct category of incoherence as previously described. Texts were provided to the same three annotators mentioned on the Subsection 4.4.1 using Google Forms, populated with texts via Google Sheets and Google Apps Script. Similar to the Global Coherence annotation, the annotators completed the task within one week. An example of the Google Form used for this task is presented in Figure 4.4.

They were instructed to read the text and identify incoherent segments by copying the segment into the appropriate field, using the marker "|" to start and end the copied segment, followed by the reason for incoherence and also ended by the same marker. Multiple segments within the same category were separated by a tab.

Figure 4.4: Form for Incoherence Identification from a GCDC Clinton text



Source: Author, 2024.

To obtain the Fleiss' Kappa for this task, only the annotation part without the

annotators' comments was considered. Each annotation was treated as a unit, meaning that even small divergences between the annotators, such as extra or missing words in their selections, were counted as a disagreement. The overall Fleiss' Kappa was 0.8326, which is considered excellent according to the scale determined by Fleiss (1981).

Table 4.8 presents an example of an annotation for Inconsistent Verb Tenses within a text. The complete text is displayed in the first column, while the specific annotation for verb tense inconsistency is shown in the second column. The annotation highlights the segment "|Mexican support has been excellent throughout.|" as an instance of inconsistent verb tenses, noting that this segment is in the present perfect tense while the rest of the text is primarily in the present tense, which may cause misinterpretation. Two of the three annotators agreed with this annotation, while the dissenting annotator did not identify any incoherence related to this category for this text.

Table 4.8: Example of Inconsistent Verb Tenses Annotation from a GCDC Clinton text

Text	**Inconsistent Verb Tenses**
Guy from Mexico is in NY and is cooperating. Discussions with him continue this am. Since he is cooperating, no move to court or to presentment scheduled yet. Mexican support has been excellent throughout. Alice has call sheet for Espinosa — call can take place whenever it's convenient for you later this morning (Espinosa is apparently out on West Coast, but Ops could confirm time difference). Holding off for now on other calls that rest of us would make (Saudis, et al), pending further developments in NY. Will let you know as soon as we have more.	|Mexican support has been excellent throughout.| - present perfect, while the rest of the text is primarily in the present tense|

4.5.2 Automatically Identifiying Incoherences

After the preprocessing stage, the next step was to use the models to identify and annotate incoherences within the 130 selected texts. This task differed from the global and local coherence tasks as it involved having each large language model (LLM) act as an annotator, identifying specific incoherent segments within the texts.

For this task, the models were provided with the same instructions as the human

annotators, but with a modification to facilitate automated analysis. Instead of using a Google Form, the models were instructed to mark the category of each incoherence by placing it between pipes "|" before the incoherent segment. This approach allowed for easier identification and categorization of incoherent segments during subsequent analysis.

The process involved submitting each of the 130 texts to the LLMs, instructing them to read through the text and identify any incoherent segments according to the defined categories. This task aimed to evaluate how well the models could handle the identification and annotation of incoherences, differing from the previous classification tasks. For this evaluation, a zero-shot approach was employed, similar to both the Local Coherence Task and the Global Coherence Task. As already mentioned, the API keys were previously obtained and had their usage described in Section 4.3. The models used were the same as those in the local coherence task, except for Bard, which was replaced by Gemini in February 2024. For this task, two tests were conducted: one using the API and the other through normal chat interactions.

The following standardized prompt (Prompt 3) for Incoherence Identification was crafted using the same principles established for annotating the 130 texts by human annotators, ensuring fairness and consistency in the evaluation. This prompt guided the models to mark incoherent segments with the appropriate category of incoherence between pipes "|", facilitating subsequent analysis. Detailed intermediate steps in the creation of this prompt are provided in Appendix A.3.

Prompt 3

You are an advanced AI model specializing in text analysis. Your task is to identify and annotate incoherent segments within the given text based on the following categories:

Incorrect Use of Logical Connectors: Misuse of logical connectors such as ‘‘therefore’’ or ‘‘however’’ that do not make sense in the context.

Unnecessary Repetition: Repetition of information that does not add value to the argument.

Irrelevant Information: Inclusion of information that is not relevant to the main topic or argument.

Contradictions: Statements that contradict each other throughout the text.

Sequence of Events: Ensuring the order of events in the text is logical and

chronological.

Inconsistent Verb Tenses: Maintaining consistency in the use of verb tenses.

Objective: Identify and annotate incoherent segments within the provided text according to the categories above.

Instructions:

Read the provided text carefully. Focus on identifying segments that exhibit incoherence based on the defined categories.

Annotate incoherent segments: Use the marker ‘‘|’’ to start and end the copied segment. Within the markers, include the category name before the segment, followed by the reason for incoherence. Use a tab to separate multiple segments within the same category.

Formatting example for annotation:

Incorrect Use of Logical Connectors: |Incorrect Use of Logical Connectors| therefore used incorrectly| (provide the reason within the markers)|

Unnecessary Repetition: |Unnecessary Repetition| repeated information| (provide the reason within the markers)|

Irrelevant Information: |Irrelevant Information| off-topic information| (provide the reason within the markers)|

Contradictions: |Contradictions| contradictory statements| (provide the reason within the markers)|

Sequence of Events: |Sequence of Events| illogical order| (provide the reason within the markers)|

Inconsistent Verb Tenses: |Inconsistent Verb Tenses| mixed tenses| (provide the reason within the markers)|

Here is the text for analysis:

[Text goes here]

Please annotate the incoherent segments as specified above.

Take a deep breath and work on this problem step-by-step.

The same techniques employed in the prompt creation for the Local Coherence Analysis, as described in Subsection 4.3, were used in developing this prompt. A function, "annotate_incoherence," was defined to handle the identification task. This function used the prompt to ask the model to analyze the provided text and identify incoherent segments. The prompt was then sent to the models, which returned a response indicating the specific incoherent segments marked with the appropriate categories.

The annotation results, along with the corresponding IDs, were saved in a new CSV file. This approach facilitated easy verification of whether the model correctly identified and categorized the incoherent segments in each text based on its ID. The texts were compared with the human annotations from the preprocessing stage. To evaluate the models' performance in this task, Fleiss' Kappa was used to measure the agreement between annotators. Each model was treated as an additional annotator alongside the three human annotators. The Fleiss' Kappa score provided a measure of how well the model's annotations aligned with those of the human annotators, indicating the model's effectiveness in identifying incoherences within the texts. A high Fleiss' Kappa score, similar to that obtained by the three human annotators, would indicate strong agreement and effective performance, while a low score would suggest the model's annotations were less consistent with those of the human annotators.

5 Results and Discussion

In this chapter, we present and discuss the findings from our experiments on Local and Global Coherence Classification, as well as Incoherence Identification, aiming to address the research questions outlined in Section 1.2.1. The results are analyzed in the context of the methodologies described in Chapter 4, showing the performance of various large language models (LLMs) on the tasks of coherence and incoherence evaluation.

First, we present the results of the Local Coherence Classification, detailing the Accuracy, Precision, Recall, and F1 Scores achieved by each model. This is followed by a discussion of the models' ability to distinguish between coherent and incoherent texts, highlighting key observations and potential areas for improvement.

Next, the Global Coherence Classification results are examined, comparing the models' performance against the human-annotated benchmarks from the GCDC and DDisCo corpora, as well as the newly annotated COCA and CST News texts. The analysis focuses on the consistency of the models' classifications with the human annotations, using metrics such as Accuracy, Precision, Recall, and F1 Score for the ternary classification levels of low, medium, and high coherence. This evaluation assesses how well the models align with human judgments, which serve as the standard for correct classification.

Finally, the results of the Incoherence Identification task are presented. This section explores how effectively the models were able to identify and annotate specific incoherent segments within texts. The performance of the models is evaluated based on their agreement with human annotators, using metrics such as Fleiss' Kappa to assess inter-rater agreement.

Throughout this chapter, we also discuss the implications of our findings for the development and application of LLMs in text coherence analysis. We highlight the practical significance of our results, suggest potential improvements for future research, and consider the broader impact of our work on the field of Natural Language Processing.

5.1 Performance Metrics

The performance of each model was measured using Accuracy, Precision, Recall, and F1 Score, which are key metrics for evaluating model performance in tasks like Local Coherence Classification. Equation 5.1 defines Accuracy as the ratio of true positives (TP) and true negatives (TN) to the total number of examples. Equation 5.2 provides Precision, calculated as the ratio of true positives (TP) to the sum of true positives and false positives (FP). Equation 5.3 defines Recall as the ratio of true positives (TP) to the sum of true positives and false negatives (FN). Finally, Equation 5.4 gives the F1 Score as the harmonic mean of Precision and Recall, offering a balanced measure of Accuracy that considers both Precision and Recall.

$$\text{Accuracy} = \frac{TP + TN}{\text{Total examples}} \tag{5.1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.3}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.4}$$

5.2 Local Coherence Classification

In this section, we present the results of the Local Coherence Classification Task. The performance of various LLMs was evaluated using the methodology described in Section 4.2. The models tested include GPT 3.5, GPT 4, GPT 4o, Claude 3 Opus, Claude 3.5 Sonnet, Claude 3 Haiku, Gemini, LLaMA 2 13b, LLaMA 2 7b, and Bard. The evaluation was conducted through two distinct approaches: (i) API-based testing and (ii) chat-based interaction, allowing us to explore not only how effectively these models handle local coherence (RQ1) but also how the mode of interaction (API vs. chat) influences their performance (RQ4).

5.2.1 API-based Testing

The API-based testing involved using the models APIs to classify the coherence of texts from the COCA, GCDC, CST News, and DDisCo datasets. Each text was classified as coherent or incoherent based on the standardized prompt shown at Subsection 4.3. The results were then compared to the true labels (DO as coherent and permDO as incoherent) to calculate performance metrics such as Accuracy, Precision, Recall, and F1 Score. The results for each model are summarized in Table 5.1.

Table 5.1: Performance Metrics for Local Coherence Classification

Model	Accuracy	Precision	Recall	F1 Score
Bard	0.756	0.755	0.740	0.748
Claude 3 Haiku	0.914	0.906	0.898	0.902
Claude 3 Opus	0.979	**0.991**	0.983	0.987
Claude 3.5 Sonnet	0.973	0.986	0.981	0.983
Gemini	0.978	0.989	0.980	0.985
GPT 3.5	0.918	0.908	0.901	0.905
GPT 4	0.970	0.982	0.980	0.981
GPT 4o	**0.982**	0.990	**0.988**	**0.989**
LLaMA 2 13b	0.831	0.825	0.816	0.820
LLaMA 2 7b	0.817	0.804	0.797	0.800

The table 5.1 presents the performance metrics for various language models in the task of Local Coherence Classification. GPT 4o exhibits the highest overall performance, with an Accuracy of 0.982 and a F1 Score of 0.989, indicating its robustness in correctly identifying coherent texts while minimizing both false positives and false negatives. Similarly, Claude 3 Opus shows a high Precision of 0.991, coupled with an Accuracy of 0.979, Recall of 0.983, and a F1 Score of 0.987. This suggests Claude 3 Opus excels in accurately identifying coherent instances with few false positives, although its slightly lower Recall compared to GPT 4o indicates it may miss a few more coherent instances.

Gemini and Claude 3.5 Sonnet also perform strongly. Gemini achieves an Accuracy of 0.978 and a F1 Score of 0.985. Claude 3.5 Sonnet shows an Accuracy of 0.973 and a F1 Score of 0.983. Both models exhibit high F1 Scores, though slightly lower than the top performers, making them reliable choices for coherence classification tasks. GPT 4 presents a solid performance with an Accuracy of 0.970 and a F1 Score of 0.981. Despite being slightly behind the top models, GPT 4 remains robust and reliable in coherence classification.

In the mid-tier range, Claude 3 Haiku and GPT 3.5 show moderate performance, with Claude 3 Haiku achieving an Accuracy of 0.914 and a F1 Score of 0.902, and GPT 3.5 achieving an Accuracy of 0.918 and a F1 Score of 0.905. These models are reliable but show

a greater need for improvement in their metrics compared to the top performers.

LLaMA 2 13b and LLaMA 2 7b are lower-tier performers in this set. LLaMA 2 13b has an Accuracy of 0.831 and a F1 Score of 0.820, while LLaMA 2 7b shows an Accuracy of 0.817 and a F1 Score of 0.800. These models struggle more with coherence classification, producing higher rates of false positives and false negatives.

Finally, Bard is the weakest performer among the models evaluated, with an Accuracy of 0.756, Precision of 0.755, Recall of 0.740, and a F1 Score of 0.748. Bard's low metrics indicate significant difficulties in accurately identifying coherent texts.

In conclusion, the results demonstrate a clear hierarchy of performance, with GPT 4o, Claude 3 Opus, and Claude 3.5 Sonnet emerging as the top models for local coherence classification. These models exhibit high reliability and Accuracy, making them highly effective for this task. In contrast, the LLaMA models and Bard require substantial improvements to reach comparable levels of performance.

Building on the performance metrics outlined in Table 5.1, additional observations shed light on the practical challenges encountered during API-based classification tasks. The observed performance variances, particularly with COCA (blog) and GCDC texts, underscore the complexities involved in real-world applications and the need for careful consideration of context and text structure in coherence evaluation tasks.

The task of classification via API presented several challenges. The first was gaining access to API keys, with some services taking more than 45 days to release them. Additionally, there were costs associated with all the tests, which could be a barrier to future usage. Despite the limitations regarding the possible context windows that could be used, due to token limitations shared between prompts and responses, this did not affect the tests conducted.

The performance of the models was notably lower when dealing with texts from the COCA (blog) and GCDC corpora. For the GCDC texts, especially those annotated with low global coherence, the models' performance was impacted, with these texts often being classified as incoherent even with prompts aimed at verifying global coherence. This negatively affected the metrics of models like Claude 3 Sonnet and Gemini. Additionally, all models struggled with identifying permuted texts where the permutation was very subtle, such as a slight change in the closing of the text, like "best regard" followed by a name or vice versa. Again, permuted texts from the COCA (blog) corpus presented the most problems, as some sentences in blog posts can be interchangeable.

5.2.2 Chat-based Test

The chat-based interaction involved directly using the chat interfaces of the models to classify a smaller subset of texts. This approach aimed to determine whether ordinary users could employ the same prompt to evaluate local coherence. The subset included 50 original documents (DO) and 1000 permuted documents (PermDO) from each corpus, making a total of 1050 texts.

5.2.2.1 Performance Results

The models were provided with the same standardized prompt and tasked with classifying the coherence of each text. Equally as the API Testing process, the results were then compared to the true labels (DO as coherent and permDO as incoherent) to calculate the performance metrics (Accuracy, Precision, Recall, and F1 Score). Table 5.2 summarizes the performance of each model in the chat-based interaction test.

Table 5.2: Performance Metrics for Local Coherence Classification (Chat-based Interaction)

Model	Accuracy	Precision	Recall	F1 Score
Bard	0.739	0.742	0.739	0.740
Claude 3 Haiku	0.949	0.902	0.899	0.900
Claude 3 Opus	0.974	0.971	**0.973**	0.972
Claude 3.5 Sonnet	0.972	0.969	0.968	0.968
Gemini	0.971	0.971	0.970	0.970
GPT 3.5	0.962	0.905	0.902	0.903
GPT 4	0.969	0.966	0.965	0.965
GPT 4o	**0.977**	**0.975**	**0.973**	**0.974**
LLaMA 2 13b	0.888	0.821	0.818	0.819
LLaMA 2 7b	0.805	0.801	0.798	0.799

The results indicate that GPT 4o continues to exhibit the highest overall performance in chat-based interactions, with an Accuracy of 0.977 and a F1 Score of 0.974. This suggests that GPT 4o remains robust in identifying coherent texts even in a chat-based setting. Claude 3 Opus and Claude 3.5 Sonnet also maintain strong performances, with F1 Scores of 0.972 and 0.968, respectively, demonstrating their reliability in coherence classification tasks.

However, compared to the API-based tests, all models showed slightly lower performance metrics. For instance, Gemini, which had a F1 Score of 0.985 in the API test, achieved 0.970 in the chat-based test. Similarly, GPT 4's F1 Score dropped from 0.981 to 0.965, and Claude 3 Haiku's from 0.902 to 0.900. These discrepancies highlight the potential impact of different interaction modes on model performance.

In the mid-tier, GPT 3.5 and LLaMA models showed more significant declines. GPT 3.5's F1 Score decreased from 0.905 to 0.903, and LLaMA 2 13b and LLaMA 2 7b's scores fell to 0.819 and 0.799, respectively. Bard, already the weakest performer in the API tests, continued to struggle in the chat-based tests with a F1 Score of 0.740.

The lower performance observed in the chat-based tests could be attributed to the smaller test set size and the inherent variability in chat interactions. Despite these challenges, the hierarchy of model performance remains consistent, with GPT 4o, Claude 3 Opus, and Claude 3.5 Sonnet as the top models, while LLaMA models and Bard lag behind. This consistency reinforces the robustness of the top models across different testing conditions.

All models, except the LLaMA models, had easily accessible chat versions, available either through subscription services or for free. The LLaMA models have chat versions available on Replicate[1], which depend on API keys and were used in this evaluation.

The chat interfaces are significantly slower compared to the API versions and exhibit poorer performance. This could be attributed to their common-use nature, which contrasts with the stability expected from API consumption. Additionally, during the testing phase, it was observed that chat models responded better when only one text at a time was sent for classification. Inconsistencies were more frequent when batches of texts were sent. Tests were conducted with batches of 10, 5, 3, 2, and 1 text, and only the single-text batches maintained consistency. Consequently, this method was adopted despite significantly increasing the time cost, which, in turn, improved the classification performance.

5.2.3 RQ1 Answer

RQ1 - Research Question: How effectively can different LLMs evaluate the logical flow and consistency within short text passages?

Objective: Explore the models' proficiency in detecting disruptions in the logical sequence and coherence of sentences within a text, particularly in scenarios where the natural order of ideas might be challenged.

Findings: The study revealed that GPT 4o, Claude 3 Opus, and Claude 3.5 Sonnet demonstrated superior performance in local coherence classification, with high Accuracy and F1 Scores. These models effectively distinguished between coherent and incoherent texts at the sentence level. In contrast, models like LLaMA and Bard struggled, showing lower performance metrics, which indicates challenges in identifying local coherence accurately.

[1]https://www.llama2.ai

5.3 Global Coherence Classification

In this section, we present the results of the Global Coherence Classification Task. Similar to the Local Coherence Classification Task, the performance of various LLMs was evaluated using the methodology described in Section 4.4. The models tested include GPT-3.5, GPT-4, GPT-4o, Claude Opus, Claude Sonnet, Gemini, and LLaMA 2 (7b and 13b). Bard was not tested as it has been discontinued. The evaluation was also conducted through two distinct approaches: (i) API-based testing and (ii) chat-based interaction, using the same LLMs and similar methods as in the Local Coherence Classification Task.

5.3.1 API-based Testing

The API-based testing involved using the API keys to classify the coherence of texts from the COCA, GCDC, CST News, and DDisCo datasets. Each text was classified on a scale from low to high coherence based on the standardized prompt. The results were then compared to the true labels to calculate performance metrics such as Accuracy, Precision, Recall, and F1 Score. The results for each model are summarized in Table 5.3.

The subcorpus used for this testing included 100 texts annotated during preprocessing, 1200 texts from the DDisCo corpus, and 842 texts from the GCDC corpus, making a total of 2142 texts. This subcorpus provided a foundation for evaluating the models' performance across different types of texts and coherence levels. By comparing the models' classifications to the human-annotated benchmarks, we were able to assess the effectiveness of each model in distinguishing between low, medium, and high coherence texts.

For ternary classification, the metrics used are Accuracy, Precision, Recall and F1 Score, with results shown on the Table 5.3. They are extensions of those used in binary classification, employed in the Section 5.1 of Local Coherence Classification task. Unlike binary classification, which differentiates between two classes, ternary classification involves three distinct classes, adding complexity to the evaluation. Each metric is calculated independently for each class, using the macro-average method to account for possible misclassifications among the three classes.

Table 5.3: Performance Metrics for Global Coherence Classification

Model	Accuracy	Precision	Recall	F1 Score
Claude 3 Haiku	0.959	0.918	0.921	0.920
Claude 3 Opus	**0.982**	**0.986**	**0.987**	**0.986**
Claude 3.5 Sonnet	0.980	0.984	0.982	0.983
Gemini	0.976	0.963	0.966	0.965
GPT 3.5	0.960	0.920	0.923	0.921
GPT 4	0.974	0.961	0.964	0.963
GPT 4o	0.978	0.965	0.968	0.967
LLaMA 2 13b	0.970	0.930	0.933	0.932
LLaMA 2 7b	0.968	0.928	0.931	0.930

Claude 3 Opus emerged as the top performer in the Global Coherence Classification Task, with an Accuracy of 0.982 and a F1 Score of 0.986. These high scores across all metrics indicate its exceptional ability to accurately and consistently identify globally coherent texts, with minimal false positives and false negatives. Following closely, Claude 3.5 Sonnet demonstrates excellent performance with an Accuracy of 0.980 and a F1 Score of 0.983. While slightly lower than Claude 3 Opus, it remains a highly reliable model for global coherence tasks.

GPT 4o shows strong performance with an Accuracy of 0.978 and a F1 Score of 0.967. Although slightly lower than the Claude models, it still maintains high reliability in coherence classification. Gemini achieves an Accuracy of 0.976 and a F1 Score of 0.965. Its performance is slightly below GPT 4o but remains robust and effective for this task. GPT 4 presents a solid performance with an Accuracy of 0.974 and a F1 Score of 0.963, demonstrating its capability in handling global coherence classification effectively.

Claude 3 Haiku and GPT 3.5 show moderate performance, with Claude 3 Haiku achieving an Accuracy of 0.959 and a F1 Score of 0.920, and GPT 3.5 achieving an Accuracy of 0.960 and a F1 Score of 0.921. These models are reliable but show lower performance compared to the top performers, which is expected as they are smaller models optimized for speed.

LLaMA 2 13b and LLaMA 2 7b are the weaker performers in this set. LLaMA 2 13b has an Accuracy of 0.970 and a F1 Score of 0.932, while LLaMA 2 7b shows an Accuracy of 0.968, a Precision of 0.928, a Recall of 0.931, and a F1 Score of 0.930. Even though they performed worse than the top-tier models for this task, they still achieved excellent overall performance.

When comparing these results with the local coherence classification, some observations can be made. First, the overall performance hierarchy remains consistent, with Claude models and GPT 4o leading in both tasks. The global coherence task shows slightly better

scores for most models compared to the local coherence task, indicating that identifying global coherence might be less challenging than local coherence for these models. The models are consistent and likely benefited from the ternary global classification scale, which aided in the identification process. Second, unlike the shuffle test used for local coherence, this evaluation considered human annotations that verified the global coherence of texts. Additionally, the models performed better overall, with fewer false positives and false negatives in each task. The global Likert scale classification appears to be more effective than the default incoherence classification for these models, as they perform well in more realistic scenarios, even with texts from various domains. The mid-tier and lower-tier performers exhibit similar relative standings in both tasks, suggesting consistent performance patterns across different coherence tasks.

5.3.2 Chat-based Test

Similar to the Local Coherence Chat-based Test, evaluating global coherence through chat involved directly using the chat interfaces of the models to classify a substantial subset of texts. This approach aimed to determine whether ordinary users could effectively use the same prompt to evaluate global coherence. The subset consisted of 2142 texts, including 100 texts annotated during preprocessing, 1200 texts from the DDisCo corpus, and 842 texts from the GCDC corpus.

5.3.2.1 Chat-based Results

The models were provided with the same standardized prompt used for Global Coherence Classification with the API and tasked with classifying the coherence of each text on a Likert scale from low to high coherence. The results were recorded and compared with the labels obtained from the human annotations to calculate performance metrics. Table 5.4 summarizes the performance of each model in the chat-based interaction.

Table 5.4: Performance Metrics for Chat Coherence Classification (Chat-based Interaction)

Model	Accuracy	Precision	Recall	F1 Score
Claude 3 Haiku	0.911	0.871	0.875	0.875
Claude 3 Opus	**0.933**	**0.936**	**0.939**	**0.937**
Claude 3.5 Sonnet	0.930	0.934	0.931	0.932
Gemini	0.928	0.915	0.918	0.916
GPT 3.5	0.912	0.873	0.879	0.877
GPT 4	0.926	0.914	0.919	0.917
GPT 4o	0.930	0.918	0.920	0.919
LLaMA 2 13b	0.922	0.887	0.883	0.888
LLaMA 2 7b	0.920	0.881	0.884	0.883

Table 5.4 demonstrates that the performance of models in chat-based classification is generally lower than in API-based classification. For instance, Claude 3 Opus, which achieved a F1 Score of 0.986 in the API-based task, saw a decrease to 0.937 in the chat-based task. Similarly, Claude 3.5 Sonnet's F1 Score dropped from 0.983 to 0.932. This trend is consistent across most models and tasks, indicating that the chat-based task presents additional challenges compared to the API-based task.

The overall performance hierarchy remains relatively consistent, with Claude models and GPT 4o leading in both tasks. Moderate performers, such as Claude 3 Haiku and GPT 3.5, also show a decrease in performance. For example, GPT 3.5's F1 Score dropped from 0.905 in the API-based task to 0.877 in the chat-based task. The lower scores in the chat-based task suggest that classifying coherence in interactive chat settings may be more challenging due to differences between the processing methods of chat and API, with the latter being more reliable.

LLaMA models, which were already the weakest performers in the API-based task, exhibit a similar pattern of reduced performance in the chat-based task. LLaMA 2 13b's F1 Score decreased from 0.932 to 0.888, and LLaMA 2 7b's F1 Score dropped from 0.930 to 0.883. These models continue to struggle significantly with coherence classification, particularly in chat-based interactions.

Here, the same as in the Global Classification Task with API occurred: Global Classification task shows consistent and likely benefited from the ternary Likert global classification scale, which aided in the identification process. Additionally, this evaluation considered human annotations that verified the global coherence of texts, leading the models to have fewer false positives and false negatives in each task. On both tests, the global classification appears to be more effective than the default incoherence classification (shuffle test) for these models, as they perform well in more realistic scenarios, even with texts from various domains.

It is important to highlight that the chat interfaces are significantly slower compared to the API versions. Again, the tests were conducted with batches of 10, 5, 3, 2, and 1 text, and only the single-text batches maintained consistency. Consequently, this method was adopted despite significantly increasing the time cost, which, in turn, improved classification performance.

5.3.3 RQ2 Answer

RQ2 - Research Question: How do LLMs perform in assessing the overall coherence of entire texts?

Objective: Assess the ability of models to evaluate the overall coherence of a text, ensuring it maintains a consistent and logical flow throughout, considering the broader context and connections between different sections.

Findings: The evaluation of global coherence revealed that models Claude 3 Opus and GPT 4o performed exceptionally well, maintaining high Accuracy and consistency in identifying globally coherent texts. The use of a ternary Likert scale for global coherence helped in achieving better performance, as models could better align their classifications with human annotations. The models generally performed better in global coherence tasks compared to the local coherence task.

5.4 Incoherence Identification

Finally, the results of the Incoherence Identification task are presented. This section explores how effectively the models were able to identify and annotate specific incoherent segments within texts. The performance of the models is evaluated based on their agreement with human annotators, using the Fleiss' Kappa metric to assess inter-rater agreement. Again, the evaluation was conducted through two distinct approaches: (i) API-based testing and (ii) chat-based interaction.

5.4.1 API-based Identification

For the API-based identification task, the LLMs were tasked with identifying and annotating incoherent segments within a subcorpus of 130 selected texts. This subcorpus consisted of the same 100 texts annotated in the Global Coherence Task (10 from the academic portion of COCA, 60 from the blog portion of COCA, and 30 from CST News) and an additional 30 texts from the GCDC corpus (10 Yelp, 10 Clinton, and 10 Enron).

Each model acted as an annotator, marking incoherent segments according to six predefined categories and returning them with specific annotations that included the category, the incoherent segment, and the reason for the incoherence. The results from each model were then compared to the human annotations to evaluate performance.

5.4.1.1 Performance Metrics

The performance of each model was measured using Fleiss' Kappa, which assesses the agreement between multiple annotators. By adding the model as a fourth annotator, the closer the Kappa value is to the Kappa obtained by human annotators, the better the model performed. This metric helps in understanding how well the models' annotations align with those of human annotators. The values of Fleiss' Kappa for each model are presented in Table 5.5.

Table 5.5: Fleiss' Kappa for Incoherence Identification Task

Model	**Fleiss' Kappa**
Annotators only (baseline)	0.8326
Claude 3 Haiku	0.7995
Claude 3 Opus	0.8166
Claude 3.5 Sonnet	0.8279
Gemini	0.8119
GPT 3.5	0.8038
GPT 4	0.8152
GPT 4o	**0.8316**
LLaMA 2 13b	0.6787
LLaMA 2 7b	0.5823

Analysing the Table 5.5, GPT 4o demonstrates the highest performance among the models, with a Fleiss' Kappa of 0.8316, which is very close to the baseline Kappa of 0.8326 obtained by human annotators alone, indicating that GPT 4o's annotations align almost perfectly with those of human annotators, demonstrating its exceptional effectiveness in the purposed task. Such a close agreement suggests that GPT 4o can consistently identify incoherences in texts, making it a reliable tool for this purpose.

Claude 3.5 Sonnet also performs exceptionally well, achieving a Kappa value of 0.8279. Although slightly lower than GPT 4o, this model's performance remains very close to the human baseline, demonstrating strong agreement with human annotations and high reliability in identifying incoherent texts. While Claude 3 Opus falls behind GPT 4o and Claude 3.5 Sonnet, it still shows a high level of agreement with human annotators, reflecting minor discrepancies in identifying borderline cases of incoherence.

Gemini achieves a Kappa of 0.8119, placing it in the mid-tier category. Its performance also indicates good reliability, though it is slightly less aligned with human annotations compared to the top performers. The model's ability to capture incoherence is evident, but there is room for improvement to reach the Precision of the top models. GPT 4 shows a similar level of performance with a Kappa of 0.8152, making it a strong mid-tier model. It demonstrates a high degree of agreement with human annotators, though not as high as GPT 4o or Claude 3.5 Sonnet.

Claude 3 Haiku and GPT 3.5 have similar Kappa values, with Claude 3 Haiku at 0.7995 and GPT 3.5 at 0.8038. These models are reliable but show a need for improvement to match the performance of the top models. Their lower Kappa values suggest they occasionally produce annotations that diverge from those of human annotators, possibly due to difficulties in handling more complex or subtle cases of incoherence.

LLaMA 2 13b has a Kappa of 0.6787, indicating moderate alignment with human annotators. While this score shows some level of coherence recognition, the model's performance is lower than other models evaluated. This result suggests potential room for improvement in its handling of coherence-related tasks, although further analysis and tests would be required to identify specific factors contributing to this outcome.

LLaMA 2 7b has the lowest Kappa value of 0.5823, making it the weakest performer. Its annotations show the least agreement with human annotators, highlighting its difficulties in accurately identifying incoherent texts. This poor performance likely reflects its smaller size, with fewer parameters than all the other models.

5.4.2 Chat-based Test

The Chat-based test involved using the chat interfaces of the models to identify incoherent segments within a subset of texts. This approach aimed to determine the practicality of using these models in real-world scenarios by non-expert users. The subset included the same 130 texts used in the API-based task.

5.4.2.1 User Interaction and Results

The models were provided with the same standardized prompt as the API test and tasked with identifying incoherent segments. The results were recorded and compared to the human annotations to evaluate performance. Table 5.6 summarizes the Fleiss' Kappa values for each model in the chat-based interaction.

Table 5.6: Fleiss' Kappa for Inter-Rater Agreement on Incoherence Identification (Chat-based Interaction)

Model	**Fleiss' Kappa**
Annotators only (baseline)	0.8326
Claude 3 Haiku	0.7653
Claude 3 Opus	0.7987
Claude 3.5 Sonnet	0.8082
Gemini	0.7858
GPT 3.5	0.7716
GPT 4	0.8093
GPT 4o	**0.8234**
LLaMA 2 13b	0.6492
LLaMA 2 7b	0.5418

The Table 5.6 values reveal that models generally perform better in API evaluations than in chat interactions. The top-performing models, such as GPT 4o and Claude 3.5 Sonnet, show resilience across both tasks, but still face slight declines in the more complex chat setting. Mid-tier models like Gemini and GPT 4 demonstrate a need for further adaptation to handle the nuances of chat data effectively. Lower-tier models, particularly the LLaMA series, highlight the significant impact of model size and parameter count on performance, struggling considerably more in chat-based interactions.

Again, GPT 4o demonstrates the highest performance in the chat-based interaction, with a Fleiss' Kappa of 0.8234. This is slightly lower than its performance in the API-based task, where it achieved a Kappa of 0.8316. Despite this minor drop, GPT 4o remains highly effective in identifying incoherence, showing strong alignment with human annotators. Claude 3.5 Sonnet follows closely with a Kappa value of 0.8082 in the chat-based task, compared to 0.8279 in the API-based task. This reduction indicates a slight decrease in performance, but Claude 3.5 Sonnet still maintains high reliability in coherence classification.

Claude 3 Opus shows a Kappa of 0.7987 in the chat-based task, lower than its API-based performance of 0.8166. While still effective, this drop suggests that Claude 3 Opus may face some challenges in the chat-based setting. GPT 4 demonstrates consistent performance with a Kappa of 0.8093 in the chat-based task, slightly lower than its API-based Kappa of 0.8152.

Gemini achieves a Kappa of 0.7858 in the chat-based task, compared to 0.8119 in the API-based task. The drop in performance suggests that Gemini finds chat-based tasks more challenging. Claude 3 Haiku and GPT 3.5 show Kappa values of 0.7653 and 0.7716, respectively, in the chat-based task, lower than their API-based values of 0.7995 and 0.8038.

LLaMA 2 13b has a Kappa of 0.6492 in the chat-based task, a decrease from its

API-based Kappa of 0.6787. This further underscores its difficulty in aligning with human annotations. LLaMA 2 7b, with the lowest Kappa value of 0.5418 in the chat-based task, also shows a significant drop from its API-based Kappa of 0.5823, reflecting its struggles with coherence classification due to its smaller size and fewer parameters compared to other models.

Overall, the comparison between the API-based and chat-based Fleiss' Kappa values reveals that models generally perform slightly worse and more slowly in chat-based tasks compared to API-based evaluations. The top-performing models, such as GPT 4o and Claude 3.5 Sonnet, show resilience across both tasks but still face slight declines in the chat-based setting. Mid-tier models like Gemini and GPT 4 demonstrate a need for further adaptation to handle the nuances of chat data effectively. Lower-tier models, particularly the LLaMA series, highlight the significant impact of model size and parameter count on performance, struggling considerably more in chat-based interactions.

5.4.3 RQ3 Answer

RQ3 - Research Question: How accurately can LLMs identify and categorize specific incoherent segments within a text?

Objective: Identify and classify specific types of incoherence within a text, providing detailed insights into the issues that disrupt logical flow and thematic continuity.

Findings: Identifying and categorizing incoherent segments proved challenging across all models. GPT 4o exhibited the highest agreement with human annotations, closely followed by Claude 3.5 Sonnet. These models effectively identified various types of incoherence, such as incorrect use of logical connectors, unnecessary repetition, and irrelevant information. However, models like LLaMA 2 13b LLaMA 7b was significantly lower, highlighting the need for improvement in understanding and categorizing incoherent text segments.

5.5 Synthesis of APIs vs Chats Performance Analysis

This section addresses RQ4, which explores how the mode of interaction (API vs. chat) influences the performance of LLMs in assessing textual coherence. The analysis for RQ4 is based on experiments described earlier in this chapter, where both chat and API interfaces were used across all tasks – Local Coherence classification, Global Coherence Classification, and Incoherence Identification. By integrating findings from these tasks, this section provides a comprehensive overview of how interaction methods impact model performance, highlighting the differences in effectiveness between APIs and chat interfaces.

5.5.1 RQ4 Answer

RQ4 - Research Question: How does the mode of interaction (API vs. chat) affect the ability of LLMs to assess and identify textual coherence?

Objective: Investigate how the interaction mode (API vs. chatbot) influences the performance of LLMs in analyzing local and global coherence and in identifying incoherent segments within texts.

Findings: The study highlighted that the mode of interaction does impact the performance of LLMs in coherence analysis. While top-performing models such as GPT 4o, Claude 3 Opus, and Claude 3.5 Sonnet showed only slight variations between API and chatbot interactions, lower-performing models like LLaMA 2 and Bard exhibited more pronounced differences. Specifically, API-based interactions yielded better results in terms of Accuracy, particularly in tasks involving global coherence and incoherence identification. The chat mode introduced challenges, possibly due to differences in how models process and respond to queries in a more conversational context, leading to reduced performance in more complex tasks.

6 Conclusions and Future Work

The primary objective of this work was to analyze and compare the performance of LLMs in assessing textual coherence at both local and global levels, as well as in identifying incoherences within texts from the COCA, CST News, GCDC, and DDisCo corpora.

Annotation was needed for the Global Coherence and Incoherence Identification tasks. Three linguistically trained annotators evaluated a subset of texts, assigning coherence scores on a Likert scale and identifying incoherent segments. The high Fleiss' Kappa scores of 0.8952 and 0.8326 indicated strong inter-annotator agreement on the tasks of Global Coherence Classification and Incoherence Identification, providing a reliable benchmark for LLM evaluation.

The testing phase employed LLMs across three tasks: Local Coherence Classification, Global Coherence Classification, and Incoherence Identification. Both API-based and chat-based approaches were used to explore how interaction modes impact model performance.

In the Local Coherence Classification task, models like GPT 4o and Claude 3 (Opus) and Claude 3.5 Sonnet performed best, particularly in API-based testing. LLaMA models and Bard struggled more. Although performance slightly declined in chat-based testing, the overall hierarchy remained consistent.

For Global Coherence Classification, models were evaluated on a ternary Likert scale. Claude 3 Opus and GPT 4o excelled, with results suggesting that global coherence may be easier for these models to handle than local coherence. Despite lower performance in chat-based testing, the consistency of model rankings was maintained.

In the Incoherence Identification task, GPT 4o again led in agreement with human annotators, followed by Claude 3.5 Sonnet and Claude 3 Opus. The LLaMA models showed the least agreement, struggling significantly with this task. The chat-based interactions further highlighted the challenges, with overall lower performance compared to API-based testing, although top models like GPT 4o still demonstrated resilience.

6.1 Key Findings

There are five key findings of this research, all listed bellow:

(i) Top-performing models, GPT 4o and Claude 3 models, consistently exhibited high performance across all tasks, indicating their robustness and reliability in coherence evaluation. However, all models performed slightly worse in chat-based tasks compared to API-based evaluations, suggesting that the mode of interaction affects model performance, with API-based methods being more stable and reliable.

(ii) There was a noticeable performance gap between the top-tier models (GPT 4o, Claude 3 Opus, Claude 3.5 Sonnet) and the lower-tier models (LLaMA series, Bard). The lower performance of LLaMA models, particularly the 7b version, likely reflects their smaller size and fewer parameters.

(iii) Models generally performed better in global coherence tasks compared to local coherence tasks. This suggests that identifying global coherence might be less challenging for these models, possibly due to the structured nature of the global Likert scale classification.

(iv) The Incoherence Identification task proved to be the most challenging, with lower Fleiss' Kappa scores indicating less agreement with human annotations. While GPT 4o and Claude 3 models performed best, there is significant room for improvement in all models.

(v) The robustness of top models in API-based evaluations makes them suitable for applications requiring high reliability, such as automated content moderation and quality assurance in writing platforms.

6.1.1 Threats to Validity

This study acknowledges the following threats to validity that may impact the generalizability and reliability of the findings:

(i) Corpus Integration in Training Data: One significant threat is the possibility that the corpora used in this work, such as COCA, CST News, GCDC, and DDisCo, could already be part of the training data for the models evaluated. If these models have previously encountered these texts, their performance might be artificially inflated due to prior exposure, rather than reflecting their true ability to assess coherence and incoherence in unfamiliar texts.

(ii) Assumption of Coherence in DO Texts: In the Local Coherence Classification task, the original documents (DO) were assumed to be coherent. However, this assumption may not always hold true. If any of these texts are inherently incoherent, the models' performance metrics, particularly in the shuffle test, may not accurately reflect their true capabilities. This

could result in an overestimation or underestimation of the models' effectiveness.

(iii) Limited Annotation Data and Time Constraints: The tasks of global coherence classification and incoherence identification relied on annotations of a relatively small subset of 100 and 130 texts, respectively. Additionally, the short time frame allotted for these annotations could lead to inconsistencies or errors, which might influence the evaluation outcomes. The dependence on these limited annotations introduces a risk that the results may not be fully representative of the models' performance on larger, more diverse datasets.

(iv) Model Versioning and Updates: The performance of LLMs can vary significantly across different versions and updates. If the models used in this study were updated during the research period, it could introduce variability in results that is not accounted for. This could affect the consistency of the findings, as performance improvements or degradations in newer versions might not reflect the models' capabilities at the time of data collection and annotation.

(v) Evaluation Metric Limitations: The use of certain evaluation metrics, such as Fleiss' Kappa, while providing insight into agreement, may not fully capture the nuanced differences in model performance, especially in tasks involving subjective judgments like coherence and incoherence identification. Metrics that aggregate performance across multiple tasks or levels of analysis might obscure specific weaknesses or strengths in the models, leading to an oversimplified view of their capabilities.

(vi) Generalization Across Domains: The study focused on specific corpora and text types (e.g., academic texts, blogs, news articles). The models' performance might not generalize well to other domains or text types, such as legal documents, technical manuals, or creative writing, where coherence challenges could differ significantly. This limitation could affect the applicability of the findings to broader NLP tasks and use cases.

(vii) Human Annotator Bias: The annotations for Global Coherence and Incoherence Identification tasks were carried out by a small group of annotators with similar backgrounds in languages and linguistics. While this could introduce some bias, the nature of the task itself doesn't inherently lend itself to significant bias. Instead, any potential bias may stem from a lack of clear understanding of the task, which could result in inconsistencies.

These threats indicate that while the results regarding the performance of various models are valuable, they should be interpreted with caution. Future research should address these threats by using a broader range of texts, involving a more diverse group of annotators, and continuously monitoring model versions to ensure consistency in the evaluation process.

6.2 Future Work

While the results of this work meet the proposed objectives and has provided a comprehensive evaluation of the performance of various Large Language Models in coherence classification and incoherence identification, offerring various contributions, there are opportunities for improvement both in the methodology used and in expanding the scope of the problem addressed. Implementing the proposed future work will enhance the findings of this study and also continue advancing research in this domain, thereby allowing for significant progress in textual coherence analysis and the application of Large Language Models in more complex and challenging contexts, there are several where further exploration is warranted:

1. Expansion of Corpora and Diverse Text Types: Future research should incorporate a wider range of corpora, including more diverse text types such as legal documents, technical manuals, creative writing, and more non-English texts. This expansion would allow for a more thorough evaluation of model performance across different domains and linguistic contexts, contributing to analyze the generalizability of the models.

2. Investigation of Fine-Tuning and Adaptation Techniques: Given that the models evaluated in this study were used in a zero-shot setting, future work could explore the impact of fine-tuning and domain adaptation techniques on coherence classification and incoherence identification tasks. Fine-tuning models on specific datasets or for particular text types could improve their performance, especially in more challenging or specialized domains.

3. Evaluation of Newer Model Architectures: As the field of NLP continues to evolve rapidly, newer model architectures and versions will likely emerge with enhanced capabilities. Future studies should evaluate these advancements to determine if they address the limitations identified in current models, such as struggles with local coherence or the accurate identification of incoherence in complex texts.

4. Longitudinal Analysis of Model Performance: Given the potential for models to change with updates and new versions, future research could conduct longitudinal studies that track the performance of LLMs over time. This would help to understand how model updates affect their coherence analysis capabilities and whether improvements in one aspect might come at the cost of performance in another.

5. Incorporation of Larger and More Diverse Annotator Groups: To address the potential biases introduced by a small and homogeneous group of annotators, future work should involve larger and more diverse groups for annotation tasks. This would provide a more representative understanding of coherence and incoherence across different cultural and linguistic backgrounds, enhancing the validity of the findings.

6. Development of Improved Evaluation Metrics: Current metrics, such as

Fleiss' Kappa, while useful, may not fully capture the nuanced performance of models in tasks involving subjective judgments. Future research could focus on developing or adopting more sophisticated evaluation metrics that better account for the complexities of coherence and incoherence in texts.

7. Exploration of Interactive and Real-World Applications: While this study evaluated models in both API and chat-based settings, future research could explore their application in real-world scenarios, such as automated content moderation, educational tools, and writing assistants. This would involve assessing how these models perform when integrated into interactive applications where they must respond to dynamic and contextually rich inputs from users.

8. Cross-Linguistic and Multimodal Coherence Analysis: Extending coherence analysis to multilingual and multimodal contexts represents an exciting direction for future work. Investigating how LLMs handle coherence in texts that include multiple languages or that integrate text with other media (e.g., images, videos) could uncover new challenges and opportunities for enhancing model capabilities.

9. Use of Different Prompt Approaches: Experimenting with different prompt engineering techniques could further enhance the models' ability to understand and assess coherence. By refining prompts to better guide models in identifying logical flow and coherence in texts, researchers could improve model accuracy and reliability across various tasks.

10. Enrichment with Syntactic and Semantic Annotations: Incorporating syntactic and semantic annotations into the training data could help models better understand the structural and meaning-based aspects of coherence. This enrichment could improve the models' ability to detect subtle incoherencies and make more accurate classifications, particularly in complex or nuanced texts.

In summary, while this study provides a solid foundation for understanding the performance of current LLMs in textual coherence analysis, there are numerous opportunities for future work to build on these findings. By addressing the limitations and exploring new directions, future research can contribute to the development of more robust, adaptable, and ethically sound models capable of handling the complexities of human language.

Bibliography

ABDOLAHI, M.; ZAHEDI, M. An overview on text coherence methods. In: **2016 Eighth International Conference on Information and Knowledge Technology (IKT)**, 2016. p. 1–5.

AKTER, S. N. *et al.* **An In-depth Look at Gemini's Language Abilities**. 2023. Available at: <https://arxiv.org/abs/2312.11444>.

ALEIXO, P.; PARDO, T. A. S. **CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory)**. São Carlos-SP, 2008. (Série de Relatórios Técnicos, 326). 12p.

BAKER, P. Using corpora in discourse analysis. **Journal of Language and Social Psychology**, v. 35, n. 5, p. 551–563, 2016.

BARZILAY, R.; LAPATA, M. Modeling local coherence: An entity-based approach. In: KNIGHT, K.; NG, H. T.; OFLAZER, K. (Ed.). **Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)**. Ann Arbor, Michigan: Association for Computational Linguistics, 2008. p. 141–148. Available at: <https://aclanthology.org/P05-1018>.

BARZILAY, R.; LEE, L. Catching the drift: Probabilistic content models, with applications to generation and summarization. In: **Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004**. Boston, Massachusetts, USA: Association for Computational Linguistics, 2004. p. 113–120. Available at: <https://aclanthology.org/N04-1015>.

BIBER, D. Corpus linguistics and the study of discourse: Methods and findings. **Annual Review of Applied Linguistics**, v. 31, p. 120–136, 2011.

Braz Junior, G.; FILETO, R. Investigating coherence in posts from a doubts forum in a virtual learning environment with bert. **Conference Paper**, 2021.

BRAZ, O.; FILETO, R. Investigando coerência em postagens de um fórum de dúvidas em ambiente virtual de aprendizagem com o bert. In: **Anais do XXXII Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2021. p. 749–759. Available at: <https://sol.sbc.org.br/index.php/sbie/article/view/18103>.

BRITO, J. O.; OLIVEIRA, E. de. Essays' coherence analysis via entity grid approach. In: SBC. **Anais do XXXIV Simpósio Brasileiro de Informática na Educação**, 2023. p. 1431–1441.

BROWN, T. B. *et al.* **Language Models are Few-Shot Learners**. 2020. Available at: <https://arxiv.org/abs/2005.14165>.

BRUNATO, D. *et al.* Discotex at evalita 2023: Overview of the assessing discourse coherence in italian texts task. In: UNIVERSITà DEGLI STUDI DI TORINO AND UNIVERSITà DI BOLOGNA. **EVALITA 2023**, 2023.

BURSTEIN, J.; LEACOCK, C.; SABATINI, J. Automated writing evaluation: Enabling feedback in interactive learning environments. **Journal of Educational Technology Society**, v. 22, n. 3, p. 257–271, 2019.

CARDOSO, P. C. F. *et al.* Cstnews: a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In: **Brazilian Symposium in Information and Human Language Technology - STIL**: SBC, 2011.

CHAROLLES, M. **Introdução aos problemas da coerência dos textos: abordagem teórica e estudo das práticas pedagógicas**: Editora Pontes, 1978.

CHEN, H. *et al.* **ChatGPT's One-year Anniversary: Are Open-Source Large Language Models Catching up?** 2024. Available at: <https://arxiv.org/abs/2311.16989>.

CROSSLEY, S. A.; KYLE, K.; DASCALU, M. The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. **Behavior Research Methods**, Springer Science and Business Media LLC, v. 51, n. 1, p. 14–27, out. 2018. Available at: <https://doi.org/10.3758/s13428-018-1142-4>.

DAS, D.; TABOADA, M.; MCFETRIDGE, P. **RST Signalling Corpus**. Linguistic Data Consortium, 2015. Available at: <https://catalog.ldc.upenn.edu/LDC2015T10>.

DAVIES, M. **The Corpus of Contemporary American English (COCA)**. 2008. Available online at https://www.english-corpora.org/coca/.

DAVIES, M. Examining recent shifts in english modality using a 100-million word corpus of american english. In: **International Conference on English Language and Literature**, 2012.

DEVLIN, J. *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, 2018. v. 1, p. 4171–4186.

DIAS, M. **Investigação de modelos de coerência local para sumários multi-documento**. Tese (Doutorado) — Universidade de São Paulo, 2016.

DIAS, M.; FELTRIM, V. D.; PARDO, T. A. S. Using rhetorical structure theory and entity grids to automatically evaluate local coherence in texts. In: BAPTISTA, J. *et al.* (Ed.). **Computational Processing of the Portuguese Language**. Cham: Springer International Publishing, 2014. p. 232–243. ISBN 978-3-319-09761-9.

DIAS, M.; PARDO, T. A discursive grid approach to model local coherence in multi-document summaries. In: KOLLER, A. *et al.* (Ed.). **Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue**. Prague, Czech Republic: Association for Computational Linguistics, 2015. p. 60–67. Available at: <https://aclanthology.org/W15-4608>.

ELSNER, M.; AUSTERWEIL, J.; CHARNIAK, E. A unified local and global model for discourse coherence. In: SIDNER, C. *et al.* (Ed.). **Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference**. Rochester, New York: Association for Computational Linguistics, 2007. p. 436–443. Available at: <https://aclanthology.org/N07-1055>.

FALZON, P. A. Discourse segmentation and the management of multiple tasks in single episodes of air traffic controller-pilot spoken radio communication. **Discours**, OpenEdition, n. 4, jun. 2009. ISSN 1963-1723. Available at: <http://dx.doi.org/10.4000/discours.7241>.

FLEISS, J. L. The measurement of interrater agreement. In: **Statistical Methods for Rates and Proportions**. 2nd. ed. New York: John Wiley, 1981. p. 212–236.

FOLTZ, P. W.; KINTSCH, W.; LANDAUER, T. K. The measurement of textual coherence with latent semantic analysis. **Discourse Processes**, v. 25, n. 2-3, p. 285–307, 1998.

FREITAS, A. R. P. **Análise automática de coerência usando o modelo grade de entidades para o português**. Tese (Doutorado), 03 2013.

GEMINI TEAM *et al.* **Gemini: A Family of Highly Capable Multimodal Models**. 2024. Available at: <https://arxiv.org/abs/2312.11805>.

GIORA, R. Towards a theory of coherence. **Poetics Today**, v. 6, n. 1, p. 59–76, 1985.

GORDON, P. C.; GROSZ, B. J.; GILLIOM, L. A. Pronouns, names, and the centering of attention in discourse. **Cognitive Science**, v. 17, n. 3, p. 311–347, 1993. ISSN 0364-0213. Available at: <https://www.sciencedirect.com/science/article/pii/S0364021305800023>.

GRAESSER, A. C. *et al.* Coh-metrix: Analysis of text on cohesion and language. **Behavioral Research Methods, Instruments, Computers**, v. 36, n. 2, p. 193–202, 2004.

GROSZ, B. J.; JOSHI, A. K.; WEINSTEIN, S. Centering: A framework for modeling the local coherence of discourse. **Computational Linguistics**, MIT Press, Cambridge, MA, v. 21, n. 2, p. 203–225, 1995. Available at: <https://aclanthology.org/J95-2003>.

HADLA, L. Coherence in translation. **Research on Humanities and Social Sciences**, Citeseer, v. 5, n. 5, p. 178–184, 2015.

HALLIDAY, M. A. K.; HASAN, R. **Cohesion in English**: Longman, 1976.

HEARST, M. The debate on automated essay grading. **IEEE Intelligent Systems and their Applications**, v. 15, n. 5, p. 22–37, 2000.

HEARST, M. A. Texttiling: Segmenting text into multi-paragraph subtopic passages. **Computational Linguistics**, v. 23, n. 1, p. 33–64, 1997.

HOEY, M. **Textual interaction: An introduction to written discourse analysis**: Routledge, 2013.

HSU, C.-L.; LIN, J. C.-C. Understanding the user satisfaction and loyalty of customer service chatbots. **Journal of Retailing and Consumer Services**, Elsevier, v. 71, p. 103211, 2023.

HUANG, H. *et al.* **On the Limitations of Fine-tuned Judge Models for LLM Evaluation**. 2024. Available at: <https://arxiv.org/abs/2403.02839>.

HUTSON, J.; LANG, M. Content creation or interpolation: Ai generative digital art in the classroom. **Metaverse**, v. 4, n. 1, p. 13, 2023.

JOTY, S.; CARENINI, G.; NG, R. T. A unified framework for coherence assessment of discourse relations language. **Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)**, 2015.

JURAFSKY, D.; MARTIN, J. H. Speech and language processing. In: _____. 3. ed. Draft, 2024. cap. 23. Accessed: 2024-02-29. Available at: <https://web.stanford.edu/~jurafsky/slp3/23.pdf>.

KIDDON, C.; ZETTLEMOYER, L.; CHOI, Y. Globally coherent text generation with neural checklist models. In: SU, J.; DUH, K.; CARRERAS, X. (Ed.). **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**. Austin, Texas: Association for Computational Linguistics, 2016. p. 329–339. Available at: <https://aclanthology.org/D16-1032>.

KINCH, O. *et al.* Evaluating discourse coherence in danish texts: The ddisco dataset. In: **Proceedings of the 13th Language Resources and Evaluation Conference (LREC)**, 2022. p. 1304–1313.

KOCH, I.; TRAVAGLIA, L. **A coerência textual**: Editora Contexto, 2003.

KOCH, I. G. V. **A Coesão Textual**. 7. ed. São Paulo: Contexto, 1994. (Repensando a Língua Portuguesa). ISBN 85-85134-46-1.

LABAN, P. *et al.* Can transformer models measure coherence in text? re-thinking the shuffle test. arXiv, 2021. Available at: <https://arxiv.org/abs/2107.03448>.

LAI, A.; TETREAULT, J. Discourse coherence in the wild: A dataset evaluation and methods. In: **Proceedings of SIGdial**, 2018. p. 214–223.

LAPATA, M.; BARZILAY, R. Automatic evaluation of text coherence: models and representations. In: **Proceedings of the 19th International Joint Conference on Artificial Intelligence**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005. (IJCAI'05), p. 1085–1090.

LARSSON, G.; LINDECRANTZ, V. **How an AI colleague affect the experiance of content creation**. 2023. Graduate thesis. Bachelor's thesis, Malmö University, Game Development Program.

LEECH, G. *et al.* **Change in Contemporary English: A Grammatical Study**: Cambridge University Press, 2014.

LI, J.; HOVY, E. A model of coherence based on distributed sentence representation. In: MOSCHITTI, A.; PANG, B.; DAELEMANS, W. (Ed.). **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Doha, Qatar: Association for Computational Linguistics, 2014. p. 2039–2048. Available at: <https://aclanthology.org/D14-1218>.

LIN, C.-C.; HUANG, A. Y.; LU, O. H. Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. **Smart Learning Environments**, Springer, v. 10, n. 1, p. 41, 2023.

LIN, Z.; NG, H. T.; KAN, M.-Y. Automatically evaluating text coherence using discourse relations. In: LIN, D.; MATSUMOTO, Y.; MIHALCEA, R. (Ed.). **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**. Portland, Oregon, USA: Association for Computational Linguistics, 2011. p. 997–1006. Available at: <https://aclanthology.org/P11-1100>.

LIU, P. *et al.* **Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing**. 2021. Available at: <https://arxiv.org/abs/2107.13586>.

LIU, S.; ZENG, S.; LI, S. **Evaluating Text Coherence at Sentence and Paragraph Levels**. 2020. Available at: <https://arxiv.org/abs/2006.03221>.

LIU, X. *et al.* **GPT Understands, Too**. 2023. Available at: <https://arxiv.org/abs/2103.10385>.

LIU, Y. *et al.* Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, 2019.

LIUSIE, A.; MANAKUL, P.; GALES, M. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In: GRAHAM, Y.; PURVER, M. (Ed.). **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)**. St. Julian's, Malta: Association for Computational Linguistics, 2024. p. 139–151. Available at: <https://aclanthology.org/2024.eacl-long.8>.

MANI, I.; BLOEDORN, E.; GATES, B. Using cohesion and coherence models for text summarization. In: **Intelligent text summarization symposium**, 1998. p. 69–76.

MANN, W. C.; THOMPSON, S. A. Rhetorical structure theory: Description and construction of text structures. In: **Natural Language Generation**. Springer Netherlands, 1987. p. 85–95. Available at: <https://doi.org/10.1007/978-94-009-3645-4_7>.

MANN, W. C.; THOMPSON, S. A. Rhetorical structure theory: Toward a functional theory of text organization. **Text**, v. 8, n. 3, p. 243–281, 1988.

MARCHENKO, O. *et al.* Improving text generation through introducing coherence metrics. **Cybernetics and Systems Analysis**, Springer, v. 56, n. 1, p. 13–21, 2020.

MARINHO, J. **Coerência pragmática**. 2016. Disponível em: <https://www.ceale.fae.ufmg.br/glossarioceale/verbetes/coerencia-pragmatica>. Accessado em: Novembro de 2022.

MCNAMARA, D. S.; KINTSCH, W. Learning from texts: Effects of prior knowledge and text coherence. **Discourse processes**, v. 22, n. 3, p. 247–288, 1996.

MESGAR, M.; STRUBE, M. Graph-based coherence modeling for assessing readability. In: PALMER, M.; BOLEDA, G.; ROSSO, P. (Ed.). **Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics**. Denver, Colorado: Association for Computational Linguistics, 2015. p. 309–318. Available at: <https://aclanthology.org/S15-1036>.

MICROSOFT. **Prompt Engineering Techniques with Azure OpenAI**. 2024. <https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/advanced-prompt-engineering?pivots=programming-language-chat-completions>. Accessed: 2024-08-08.

MIKKELSEN, L. F. *et al.* Ddisco: A discourse coherence dataset for danish. In: **Proceedings of the 13th Language Resources and Evaluation Conference (LREC)**, 2022. p. 1234–1243.

MORRIS, J.; HIRST, G. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. **Computers and the Humanities**, v. 25, n. 4, p. 21–48, 1991.

NAISMITH, B.; MULCAIRE, P.; BURSTEIN, J. Automated evaluation of written discourse coherence using gpt-4. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)**. Online, 2023. p. 394–403. Available at: <http://www.example.com/paper12345>.

NAJAFI, E.; H. Darooneh, A. Long range dependence in texts: A method for quantifying coherence of text. **Knowledge-Based Systems**, v. 133, p. 33–42, 2017. ISSN 0950-7051. Available at: <https://www.sciencedirect.com/science/article/pii/S095070511730312X>.

NAPOLES, C.; SAKAGUCHI, K.; TETREAULT, J. Gleu without tuning: Automatic evaluation of sentence-level fluency and coherence. **Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL)**, p. 170–180, 2017.

TIEN NGUYEN, D.; JOTY, S. A neural local coherence model. In: BARZILAY, R.; KAN, M.-Y. (Ed.). **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 1320–1330. Available at: <https://aclanthology.org/P17-1121>.

PARDO, T. *et al.* The coreference annotation of the cstnews corpus. In: , 2017.

PENG, L.; SHANG, J. **Incubating Text Classifiers Following User Instruction with Nothing but LLM**. 2024. Available at: <https://arxiv.org/abs/2404.10877>.

QUIRK, R. *et al.* **A Comprehensive Grammar of the English Language**. London: Longman, 1985.

RAKHIMOVA, M. U. Q.; DJUMANAZAROVA, G. S.; BOBOJONOVA, Y. I. The importance of coherence in writing. **CyberLeninka**, 2019.

REYNOLDS, L.; MCDONELL, K. **Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm**. 2021. Available at: <https://arxiv.org/abs/2102.07350>.

SAGI, E. Discourse structure effects on the global coherence of texts. In: , 2010.

SANH, V. *et al.* Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. **arXiv preprint arXiv:1910.01108**, 2019.

SENO, E. R. M.; RINO, L. H. M. Co-referential chaining for coherent summaries through rhetorical and linguistic modeling. In: NúCLEO INTERINSTITUCIONAL DE LINGüíSTICA COMPUTACIONAL – NILC/USFCAR. **Proceedings of the Workshop on Crossing Barriers in Text Summarization Research/RANLP**. Borovets, Bulgaria, 2005.

SHERMIS, M.; JC, B. **Automated Essay Scoring: A Cross-Disciplinary Perspective**, 2003.

SMITH, N. A.; BALDRIDGE, J. Robust, scalable, and extensible system for processing unstructured text: Nlp applications using coca. **Journal of Language Technology**, v. 29, n. 4, p. 567–583, 2013.

SRIVASTAVA, A. *et al.* **Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models**. 2023. Available at: <https://arxiv.org/abs/2206.04615>.

THOMPSON, I. Readability beyond the sentence: Global coherence and ease of comprehension. **Journal of Technical Writing and Communication**, SAGE Publications, v. 16, n. 1, p. 131–140, jan. 1986. ISSN 1541-3780. Available at: <http://dx.doi.org/10.2190/6J1F-DATG-1275-JTFK>.

TOUVRON, H. *et al.* **Llama 2: Open Foundation and Fine-Tuned Chat Models**. 2023.

Van Dijk, T.; KINTSCH, W. **Strategies of discourse comprehension**: New York: Academic Press, 1983.

Van Dijk, T. A. Text and context: Explorations in the semantics and pragmatics of discourse. Longman, 1977.

WALKER, M. A.; IIDA, M.; COTE, S. Japanese discourse and the process of centering. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics**, 1998. p. 1363–1369.

WEI, J. *et al.* **Chain-of-Thought Prompting Elicits Reasoning in Large Language Models**. 2023. Available at: <https://arxiv.org/abs/2201.11903>.

WEISCHEDEL, R.; PALMER, M.; MARCUS, M. **OntoNotes: A Large Training Corpus for Enhanced Processing**: Springer, 2013.

XU, P. *et al.* A cross-domain transferable neural coherence model. In: **Proc. 57th Annual Meeting of the ACL**, 2019. p. 678–687.

Appendix A – Intermediate Prompts

This appendix encompasses all the intermediate prompts that were tested during the development of the prompts used in Chapter 4 for the three tasks: Local Coherence, Global Coherence, and Incoherence Identification. The number of attempts progressively decreased as knowledge were gained from previous stages – starting with 10 attempts in the first stage, followed by 8 in the second, and 6 in the third – reflecting an improved understanding of prompt engineering techniques. These iterative attempts were conducted to assess how varying levels of prompt complexity affected performance on each respective task. For each task, a set of 10 texts from each corpus was employed, aiming to refine and finalize the three optimal prompts. It is important to note that the coherence concepts applied in the creation of these prompts were based on the definitions established by the works of Koch and Travaglia (2003) and Barzilay and Lapata (2008).

A.1 Local Coherence Intermediate Prompts

1st Prompt attempt for Local Coherence

```
Classify the following text as ‘‘coherent’’ or ‘‘incoherent’’.

[Text goes here]
```

2nd Prompt attempt for Local Coherence

```
You are an AI specialized in text coherence analysis. Read the following
text and classify it as ‘‘coherent’’ or ‘‘incoherent’’. Focus on the logical
flow and the connection between ideas.

[Text goes here]
```

3rd Prompt attempt for Local Coherence

```
You are an AI model that analyzes text coherence. To determine if the text
is ‘‘coherent’’ or ‘‘incoherent’’, please explain your reasoning step-by-step
before providing the classification.
[Text goes here]
```

4th Prompt attempt for Local Coherence

You are an AI specialized in evaluating text coherence. Analyze the following text by examining its logical flow, transitions between sentences, and the connection of ideas. Provide a step-by-step reasoning process, then classify the text as ‘‘coherent’’ or ‘‘incoherent’’.

[Text goes here]

5th Prompt attempt for Local Coherence

You are an advanced AI model for text analysis with expertise in coherence evaluation. Follow these steps to assess the text:

1. Read the text thoroughly.
2. Analyze the logical flow between sentences and paragraphs.
3. Evaluate the connections between ideas.
4. Provide a step-by-step explanation of your analysis.
5. Classify the text as ‘‘coherent’’ or ‘‘incoherent’’.

Here is the text for analysis:

[Text goes here]

6th Prompt attempt for Local Coherence

You are an advanced AI model specializing in text coherence analysis. To determine if the text is ‘‘coherent’’ or ‘‘incoherent’’, use the following criteria:

Logical Flow: Does the text progress logically from one idea to the next?
Transitions: Are the transitions between sentences and paragraphs smooth and natural?
Idea Connection: Do the ideas connect clearly and contribute to the overall understanding?

Provide a detailed, step-by-step reasoning based on these criteria before classifying the text.

[Text goes here]

7th Prompt attempt for Local Coherence

You are an advanced AI model specialized in text coherence evaluation. Assess the following text based on logical flow, transitions, and idea connections. Follow these steps:

1. Read the text thoroughly.
2. Analyze the logical progression of ideas.
3. Examine the transitions between sentences and paragraphs.
4. Evaluate how well the ideas are connected.
5. Provide a detailed explanation of your analysis.
6. Classify the text as ‘‘coherent’’ or ‘‘incoherent’’.

Response Format:
Coherent: The text logically flows, makes sense, and has clea
r connections between ideas.
Incoherent: The text lacks logical flow, is confusing, or has
disjointed ideas.

Here is the text for analysis:

[Text goes here]

8th Prompt attempt for Local Coherence

You are an advanced AI model specializing in text analysis with expertise in evaluating coherence. Your task is to classify the coherence of the given text. Coherence means that the text logically flows and makes sense, with each sentence and idea connected in a clear and understandable way.

Objective:
Assess the text’s coherence by determining if the logical flow, connectio n of ideas, and overall clarity are maintained throughout the text. Classify the text as either ‘‘coherent’’ or ‘‘incoherent’’ based on these criteria.

Instructions:
1. Read the provided text thoroughly.
2. Focus on the transitions between sentences and paragraphs.
3. Analyze the logical sequence of ideas and the overall structure.
4. Evaluate the logical flow: Does the text follow a logical progression from one sentence to the next and from one paragraph to another?
5. Assess the connections between ideas: Do each sentence and paragraph connect naturally and contribute to the logical flow?
6. Provide a step-by-step reasoning of your analysis.
7. Classify the text as ‘‘coherent’’ or ‘‘incoherent’’.

Response Format:
Coherent: The text logically flows, makes sense, and has clear connections between ideas.
Incoherent: The text lacks logical flow, is confusing, or has disjointed ideas.

Here is the text for analysis:

[Text goes here]

9th Prompt attempt for Local Coherence

You are an advanced AI model specializing in text analysis with expertise in evaluating coherence. Your task is to classify the coherence of the given text. Coherence means that the text logically flows and makes sense, with each sentence and idea connected in a clear and understandable way.

Objective:
Assess the text's coherence by determining if the logical flow, connection of ideas, and overall clarity are maintained throughout the text.
- Classify the text as either ‘‘coherent’’ or ‘‘incoherent’’ based on these criteria.

Instructions:
1. Read the provided text thoroughly.
2. Focus on the transitions between sentences and paragraphs.
3. Analyze the logical sequence of ideas and the overall structure.
4. Evaluate the logical flow: Does the text follow a logical progression from one sentence to the next and from one paragraph to another?

5. Assess the connections between ideas: Do each sentence and paragraph connect naturally and contribute to the logical flow?
6. Provide a step-by-step reasoning of your analysis.
7. Classify the text as ‘‘coherent’’ or ‘‘incoherent’’.

Response Format:
Coherent: The text logically flows, makes sense, and has clear connections between ideas.
Incoherent: The text lacks logical flow, is confusing, or has disjointed ideas.

Here is the text for analysis:

[Text goes here]

10th and Final Prompt attempt for Local Coherence

You are an advanced AI model specializing in text analysis with expertise in evaluating text coherence. Your task is to classify the coherence of the given text. Coherence in this context means that the text logically flows and makes sense, with each sentence and idea connected in a clear and understandable way.

Objective: Assess the text's coherence by determining if the logical flow, connection of ideas, and overall clarity are maintained throughout the text. Classify the text as either ‘‘coherent’’ or ‘‘incoherent’’ based on these criteria.

Instructions:

- Read the provided text thoroughly. Focus on the transitions between sentences and paragraphs, the logical sequence of ideas, and the overall structure.
- Evaluate the logical flow: Determine if the text follows a logical progression of ideas from one sentence to the next and from one paragraph to another.
- Assess the connections between ideas: Check if each sentence and paragraph connects naturally and contributes to the logical flow.

Classify the text:

- Respond with ‘‘coherent’’ if the text logically flows, makes sense, and has clear connections between ideas.
- Respond with ‘‘incoherent’’ if the text lacks logical flow, is confusing, or has disjointed ideas.

Here is the text for analysis:

[Text goes here]

Response format:

- Coherent: The text logically flows, makes sense, and has clear connections between ideas.
- Incoherent: The text lacks logical flow, is confusing, or has disjointed ideas.

Take a deep breath and work on this problem step-by-step.

A.2 Global Coherence Intermediate Prompts

1st Prompt attempt for Global Coherence

You are an AI model focused on assessing text coherence. Determine whether the following text is‘‘Low Coherence,’’‘‘Medium Coherence,’’ or‘‘High Coherence’’ by analyzing its logical progression and overall structure.

[Text goes here]

2nd Prompt attempt for Global Coherence

You are an AI model focused on assessing text coherence. Determine whether the following text is‘‘Low Coherence,’’‘‘Medium Coherence,’’ or‘‘High Coherence’’ by analyzing its logical progression and overall structure. Focus on the relevance of details to the main point.

[Text goes here]

3rd Prompt attempt for Global Coherence

You are an advanced AI model for text analysis with expertise in coherence evaluation. Follow these steps to assess the text:

1. Read the text thoroughly.
2. Analyze the overall structure and organization.
3. Evaluate the relevance and support of details to the main point.
4. Provide a step-by-step explanation of your analysis.
5. Classify the text as‘‘Low Coherence,’’‘‘Medium Coherence,’’ or ‘‘High Coherence.’’

Here is the text for analysis:

[Text goes here]

4th Prompt attempt for Global Coherence

You are an advanced AI model specializing in text coherence analysis. To determine if the text is‘‘Low Coherence,’’‘‘Medium Coherence,’’ or‘‘High Coherence,’’ use the following criteria:

Low Coherence: The text is difficult to understand, unorganized, contains unnecessary details, and cannot be summarized briefly and easily.
Medium Coherence: The text is relatively easy to follow but is neither well nor poorly organized. It might contain extraneous details that don’t directly support the main point and might be easy enough to summarize but leave something to be desired in the structure of the text.
High Coherence: The text is easy to understand, well-organized, contains only details that support the main point, and can be summarized briefly and easily.

Provide a detailed, step-by-step reasoning based on these criteria before classifying the text.

[Text goes here]

5th Prompt attempt for Local Coherence

You are an advanced AI model specialized in text coherence evaluation. Assess the following text based on logical flow, relevance of details, and overall organization. Follow these steps:

1. Read the text thoroughly.
2. Analyze the logical progression of ideas.
3. Examine the relevance and support of details to the main point.
4. Evaluate the overall organization and structure.
5. Provide a detailed explanation of your analysis.
6. Classify the text as‘‘Low Coherence,’’‘‘Medium Coherence,’’ or‘‘High Coherence.’’

Response Format:
Low Coherence: The text is difficult to understand, unorganized, contains unnecessary details, and cannot be summarized briefly and easily.
Medium Coherence: The text is relatively easy to follow but is neither well nor poorly organized. It might contain extraneous details that don’t directly support the main point and might be easy enough to summarize bu t leave something to be desired in the structure of the text.
High Coherence: The text is easy to understand, well-organized, contains only details that support the main point, and can be summarized briefly and easily.

Here is the text for analysis:

[Text goes here]

6th Prompt attempt for Local Coherence

You are an advanced AI model specializing in text analysis with expertise in evaluating coherence. Your task is to classify the coherence of the given text. Coherence in this context means that the text logically flows and makes sense, with each sentence and idea connected in a clear and understandable way.

Objective:
Assess the text’s coherence by determining if the logical flow, connection of ideas, and overall clarity are maintained throughout the text.
- Classify the text as either‘‘Low Coherence,’’‘‘Medium Coherence,’’ or‘‘High Coherence’’ based on these criteria.

Instructions:
1. Read the provided text thoroughly.
2. Focus on the overall structure, organization, and relevance of details to the main point.
3. Analyze the logical flow: Determine if the text follows a logical progression of ideas.
4. Evaluate the relevance of details: Check if the details support the main point or are extraneous.
5. Assess the overall organization: Determine if the text is well-organized and easy to follow.
6. Provide a step-by-step reasoning of your analysis.
7. Classify the text as‘‘Low Coherence,’’‘‘Medium Coherence,’’ or‘‘High Coherence.’’

Response Format:
Low Coherence: The text is difficult to understand, unorganized, contains unnecessary details, and cannot be summarized briefly and easily.
Medium Coherence: The text is relatively easy to follow but is neither well nor poorly organized. It might contain extraneous details that don’t directly support the main point and might be easy enough to summarize but leave something to be desired in the structure of the text.
High Coherence: The text is easy to understand, well-organized, contains only details that support the main point, and can be summarized briefly and easily.

Here is the text for analysis:

[Text goes here]

7th Prompt attempt for Global Coherence

You are an advanced AI model specializing in text analysis with expertise in evaluating text coherence. Your task is to classify the coherence of the provided text. In this context, coherence means that the text logically flows and makes sense, with each sentence and idea connected in a clear and understandable way.

Objective:
Assess the text’s coherence by determining if the logical flow, connection of ideas, and overall clarity are maintained throughout the text.

- Classify the text as either‘‘Low Coherence,’’‘‘Medium Coherence,’’ or‘‘High Coherence’’ based on these criteria.

Instructions:
1. Read the provided text thoroughly.
2. Analyze the logical flow:
   Examine how each sentence transitions to the next.
   Determine if paragraphs are organized logically.
3. Evaluate the connections between ideas:
   Identify whether each idea builds upon the previous one.
   Check for clear and natural links between sentences and paragraphs.
4. Assess the relevance of details:
   Determine if the details support the main point.
   Identify any extraneous information that does not contribute to the main argument.
5. Provide a detailed, step-by-step reasoning of your analysis, highlighting specific parts of the text that support your classification.
6. Conclude by classifying the text as‘‘Low Coherence,’’‘‘Medium Coherence,’’ or‘‘High Coherence.’’

Response Format:
Low Coherence: The text is difficult to understand, unorganized, contains unnecessary details, and cannot be summarized briefly and easily.
Medium Coherence: The text is relatively easy to follow but is neither well nor poorly organized. It might contain extraneous details that don’t directly sup enough to summarize but leave something to be desired in the structure of the text.
High Coherence: The text is easy to understand, well-organized, contains only details that support the main point, and can be summarized briefly and easily.

Here is the text for analysis:

[Text goes here]

8th Prompt attempt for Global Coherence

You are an advanced AI model specializing in text analysis. Your task is to classify the coherence of the given text based on the following criteria:

Low Coherence: The text is difficult to understand, unorganized, contains unnecessary details, and cannot be summarized briefly and easily.

Medium Coherence: The text is relatively easy to follow but is neither well nor poorly organized. It might contain extraneous details that don't directly support the main point and might be easy enough to summarize but leave something to be desired in the structure of the text.

High Coherence: The text is easy to understand, well-organized, contains only details that support the main point, and can be summarized briefly and easily.

General Note: Grammatical and typing errors are ignored (i.e., they do not affect the coherency score), and the coherence of a text is considered within its own domain.

Objective: Assess the coherence of the provided text and classify it as ''Low Coherence,'' ''Medium Coherence,'' or ''High Coherence'' based on the criteria above.

Instructions:

Read the provided text carefully. Focus on the overall structure, organization, and relevance of details to the main point.

Evaluate the text based on the following criteria:

- **Low Coherence:** Is the text difficult to understand? Is it unorganized? Does it contain unnecessary details? Is it hard to summarize briefly?
- **Medium Coherence:** Is the text relatively easy to follow? Is it neither well nor poorly organized? Does it contain some extraneous details? Can it be summarized, but with some structural issues?
- **High Coherence:** Is the text easy to understand? Is it well-organized? Do all details support the main point? Can it be summarized briefly and easily?

Ignore grammatical and typing errors. These do not affect the coherence score. Classify the text:

- Respond with **''Low Coherence''** if the text meets the criteria for low coherence.
- Respond with **''Medium Coherence''** if the text meets the criteria for medium coherence.
- Respond with **''High Coherence''** if the text meets the criteria for high coherence.

Here is the text for analysis:

[Text goes here]

**Please respond with ‘‘Low Coherence,’’ ‘‘Medium Coherence,’’
or ‘‘High Coherence’’ based on the criteria above.
Take a deep breath and work on this problem step-by-step.**

A.3 Incoherence Identification Intermediate Prompts

1st Prompt attempt for Incoherence Identification

**You are an AI model focused on identifying incoherent segments in text.
Use the following categories to guide your identification:**
Incorrect Use of Logical Connectors
Unnecessary Repetition
Irrelevant Information
Contradictions
Sequence of Events
Inconsistent Verb Tenses

Identify incoherent segments in the provided text based on these categories.

[Text goes here]

2nd Prompt attempt for Incoherence Identification

**You are an AI model focused on identifying incoherent segments in text.
Identify incoherent segments in the following text based on the categories
below:**

Incorrect Use of Logical Connectors
Unnecessary Repetition
Irrelevant Information
Contradictions
Sequence of Events
Inconsistent Verb Tenses

Response Format:
Incorrect Use of Logical Connectors:
|Incorrect Use of Logical Connectors| [Segment] |Reason|
Unnecessary Repetition: |Unnecessary Repetition| [Segment] |Reason|
Irrelevant Information: |Irrelevant Information| [Segment] |Reason|
Contradictions: |Contradictions| [Segment] |Reason|
Sequence of Events: |Sequence of Events| [Segment] |Reason|
Inconsistent Verb Tenses: |Inconsistent Verb Tenses| [Segment] |Reason|

Here is the text for analysis:

Please annotate the incoherent segments as specified above.

[Text goes here]

3rd Prompt attempt for Incoherence Identification

You are an advanced AI model specializing in text analysis.
Your task is to identify incoherent segments in the following text based on the categories below:

Incorrect Use of Logical Connectors: Misuse of logical connectors such as ‘‘therefore’’ orzz‘‘however’’ that do not make sense in the context.
Unnecessary Repetition: Repetition of information that does not add value to the argument.
Irrelevant Information: Inclusion of information that is not relevant to the main topic or argument.
Contradictions: Statements that contradict each other throughout the text.
Sequence of Events: Ensuring the order of events in the text is logical and chronological.
Inconsistent Verb Tenses: Maintaining consistency in the use of verb tenses.

Response Format:
Incorrect Use of Logical Connectors:
|Incorrect Use of Logical Connectors| [Segment] |Reason|
Unnecessary Repetition: |Unnecessary Repetition| [Segment] |Reason|
Irrelevant Information: |Irrelevant Information| [Segment] |Reason|
- **Contradictions:** |Contradictions| [Segment] |Reason|
Sequence of Events: |Sequence of Events| [Segment] |Reason|
Inconsistent Verb Tenses: |Inconsistent Verb Tenses| [Segment] |Reason|

Please annotate the incoherent segments as specified above.

Here is the text for analysis:

[Text goes here]

4th Prompt attempt for Incoherence Identification

You are an advanced AI model specializing in text analysis.
Your task is to identify incoherent segments in the following text based on the categories below:

Incorrect Use of Logical Connectors: Misuse of logical connectors such as ‘‘therefore’’ or‘‘however’’ that do not make sense in the context.
Unnecessary Repetition: Repetition of information that does not add value to the argument.
Irrelevant Information: Inclusion of information that is not relevant to the main topic or argument.
Contradictions: Statements that contradict each other throughout the text.
Sequence of Events: Ensuring the order of events in the text is logical and chronological.
Inconsistent Verb Tenses: Maintaining consistency in the use of verb tenses.

Instructions:
1. Read the provided text thoroughly.
2. Identify incoherent segments based on the defined categories.
3. Explain your reasoning step-by-step for each identified segment.
4. Annotate each incoherent segment using the specified response format.

Response Format:
Incorrect Use of Logical Connectors:
|Incorrect Use of Logical Connectors| [Segment] |Reason|
Unnecessary Repetition: |Unnecessary Repetition| [Segment] |Reason|
Irrelevant Information: |Irrelevant Information| [Segment] |Reason|
Contradictions: |Contradictions| [Segment] |Reason|
Sequence of Events: |Sequence of Events| [Segment] |Reason|
Inconsistent Verb Tenses: |Inconsistent Verb Tenses| [Segment] |Reason|

Please annotate the incoherent segments as specified above.

Here is the text for analysis:

[Text goes here]

5th Prompt attempt for Incoherence Identification

You are an advanced AI model specializing in text analysis.
Your task is to identify incoherent segments in the following text based on the categories below:

Incorrect Use of Logical Connectors: Misuse of logical connectors such as ‘‘therefore’’ or‘‘however’’ that do not make sense in the context.
Unnecessary Repetition: Repetition of information that does not add value to the argument.
Irrelevant Information: Inclusion of information that is not relevant to the main topic or argument.
Contradictions: Statements that contradict each other throughout the text.
Sequence of Events: Ensuring the order of events in the text is logical and chronological.
Inconsistent Verb Tenses: Maintaining consistency in the use of verb tenses.

Instructions:
Follow these steps to identify incoherent segments:

1. Read the text thoroughly.
2. Incorrect Use of Logical Connectors:
   Identify any logical connectors used incorrectly.
   Explain why their usage disrupts coherence.
3. Unnecessary Repetition:
   Spot repeated information.
   Explain why the repetition adds no value.
4. Irrelevant Information:
   Detect information that is off-topic.
   Explain its irrelevance to the main argument.
5. Contradictions:
   Find statements that contradict each other.
   Explain the nature of the contradiction.
6. Sequence of Events:
   Check the logical and chronological order of events.
   Explain any inconsistencies found.
7. Inconsistent Verb Tenses:
   Ensure verb tenses are consistent.
   Explain any inconsistencies identified.
8. Annotate each incoherent segment using the specified response format.

Response Format:
Incorrect Use of Logical Connectors:
|Incorrect Use of Logical Connectors| [Segment] |Reason|
Unnecessary Repetition: |Unnecessary Repetition| [Segment] |Reason|
Irrelevant Information: |Irrelevant Information| [Segment] |Reason|
Contradictions: |Contradictions| [Segment] |Reason|
Sequence of Events: |Sequence of Events| [Segment] |Reason|
Inconsistent Verb Tenses: |Inconsistent Verb Tenses| [Segment] |Reason|

Please annotate the incoherent segments as specified above.

Here is the text for analysis:

[Text goes here]

6th and Final Prompt attempt for Incoherence Identification

You are an advanced AI model specializing in text analysis. Your task is to identify and annotate incoherent segments within the given text based on the following categories:

Incorrect Use of Logical Connectors: Misuse of logical connectors such as ‘‘therefore’’ or ‘‘however’’ that do not make sense in the context.

Unnecessary Repetition: Repetition of information that does not add value to the argument.

Irrelevant Information: Inclusion of information that is not relevant to the main topic or argument.

Contradictions: Statements that contradict each other throughout the text.

Sequence of Events: Ensuring the order of events in the text is logical and chronological.

Inconsistent Verb Tenses: Maintaining consistency in the use of verb tenses.

Objective: Identify and annotate incoherent segments within the provided text according to the categories above.

Instructions:

Read the provided text carefully. Focus on identifying segments that exhibit incoherence based on the defined categories.

Annotate incoherent segments: Use the marker ‘‘|’’ to start and end the copied segment. Within the markers, include the category name before the segment, followed by the reason for incoherence. Use a tab to separate multiple segments within the same category.

Formatting example for annotation:

Incorrect Use of Logical Connectors: |Incorrect Use of Logical Connectors| therefore used incorrectly| (provide the reason within the markers)|

Unnecessary Repetition: |Unnecessary Repetition| repeated information| (provide the reason within the markers)|

Irrelevant Information: |Irrelevant Information| off-topic information| (provide the reason within the markers)|

Contradictions: |Contradictions| contradictory statements| (provide the reason within the markers)|

Sequence of Events: |Sequence of Events| illogical order| (provide the reason within the markers)|

Inconsistent Verb Tenses: |Inconsistent Verb Tenses| mixed tenses| (provide the reason within the markers)|

Here is the text for analysis:

[Text goes here]

Please annotate the incoherent segments as specified above.

Take a deep breath and work on this problem step-by-step.