



MSc. IT & Cognition

Leveraging Large Language Model Capabilities for Big Five Personality Trait Classification

Exploring the Impact of Input Features and Prompting
Techniques Using the Essay Dataset

Sophia De Spiegeleire

Supervised by Alberto Parola

January 2024

Sophia De Spiegeleire

Leveraging Large Language Model Capabilities for Big Five Personality Trait Classification

MSc. IT & Cognition, January 2024

Supervisor: Alberto Parola

Character count: 122,778

Standard page count: 51

University of Copenhagen

Faculty of Humanities

Department of Nordic Studies and Linguistics

Emil Holms Kanal 2

2300 Copenhagen S

Executive Summary

This thesis explores how Large Language Models (LLMs) can be leveraged, through prompt engineering, to predict Big Five personality traits from unstructured stream-of-consciousness text. It begins by framing personality as a foundational construct in psychology, highlighting the widespread use of Big Five assessments in diverse domains ranging from clinical settings to organizational behavior. However, conventional measures, often reliant on self-reports, are costly, time-consuming, and prone to bias.

To address these challenges, a novel methodology was designed, starting with semantically chunking the well-known Essay Dataset into coherent segments. Each chunk was then enriched with either programmatic features or LLM-based semantic features. Then, the study systematically compared prompting techniques, baseline (zero-shot) versus Chain-of-Thought (CoT), to evaluate their impact on classification accuracy and interpretability. Results reveal that including additional programmatic or semantic features, coupled with well-crafted prompts, substantially improves prediction performance compared to text-only inputs. While CoT prompting boosts classification rates, it occasionally leads to overconfident judgements, highlighting the trade-offs between model decisiveness and interpretability. Overall, these findings underscore the promise of LLMs as flexible, scalable tools for personality analysis, laying the groundwork for broader interdisciplinary applications in personality psychology and computational social science.

Reader's Guide

This thesis is structured to present the research in a clear, chronological progression. However, the following reading path is suggested to help efficiently grasp the core experiments and key concepts:

1. **Introduction (Chapter 1):** Start here for an overview of the research questions, objectives, and the significance of using LLMs for personality classification.
2. **Methodology (Chapter 4):** Focus on the Big Five Personality Trait Classification Experiment (Section 4.4) first, followed by Semantic Feature Extraction (Section 4.3) and Programmatic Feature Extraction (Section 4.2).
3. **Results and Discussion (Chapters 5 and 6):** Review the experiment outcomes and key insights from the findings.
4. **Related Work (Chapter 2):** Explore this chapter to understand the broader context and existing approaches in personality prediction using LLMs.
5. **Conclusion (Chapter 7):** The conclusion summarizes the thesis contributions and proposes future research directions.

Contents

1	Introduction	1
2	Related Work	4
2.1	Foundations of Personality and Language	4
2.1.1	Psychological Foundations of Personality	4
2.1.2	Linguistic Correlates of Personality	6
2.2	Computational Personality Prediction	7
2.2.1	Machine Learning	7
2.2.2	State-of-the-Art	9
2.3	Large Language Models	10
2.3.1	Chunking Strategies	11
2.3.2	Applications in Feature Extraction and Personality	11
2.3.3	Prompt Engineering	13
2.3.4	Output Evaluation	14
3	Dataset	18
3.1	Introduction & Background	18
3.2	Original Dataset EDA	19
3.3	Semantic Chunks Dataset EDA	22
4	Methodology	25
4.1	Semantic Chunking	26
4.1.1	Rationale	26
4.1.2	Method	27
4.2	Programmatic Feature Extraction	29
4.2.1	Feature Set Rationale	29

4.2.2	Normalization	31
4.3	LLM-Based Semantic Feature Extraction	32
4.3.1	Feature Set Rationale	32
4.3.2	GPT-4o : CoT Prompt	34
4.3.3	Confidence Scores Evaluation	36
4.4	Big Five Personality Trait Classification Experiment	37
4.4.1	Overview	38
4.4.2	Task Description	38
4.4.3	Experimental Setup	39
4.4.4	Prompt Design	40
4.4.5	Evaluation	44
4.4.6	Summary	47
5	Results	49
5.1	Programmatic Feature Extraction	49
5.2	LLM-Based Semantic Feature Extraction	50
5.2.1	Evaluation of Confidence Scores	51
5.3	Big 5 Personality Trait Classification	52
5.3.1	Chunk-level Analysis	53
5.3.2	Author-level Analysis	56
5.3.3	LLM as a Judge Quality Assessment	57
6	Discussion	59
6.1	The Role of Semantic Chunking	59
6.2	The Impact of Programmatic and Semantic Features	60
6.3	The Impact of Prompting Techniques	61
6.4	Comparing Chunk and Author-Level Performance	62
6.5	Trait-Specific Nuances	63
6.6	About Ground Truth	64
6.7	Evaluation Methods	65
6.8	Addressing Over- and Underestimation	66
6.9	Practical Transferability of the Method	67
6.10	Ethical Considerations	68

6.11 Further Limitations	69
6.12 Discussion Summary	70
7 Conclusion	71
7.1 Main Findings and Contributions	71
7.2 Broader Applications	72
7.3 Future Directions	73
7.4 Closing Remarks	73
Appendices	75
I Big 5 Trait Classification Prompts	75
II GPT-o1 Evaluation Prompt	78
III GPT-o1 Evaluation Transcript	79
IV GitHub Repository	84
Bibliography	85

Introduction

Personality, defined as the consistent patterns of thoughts, emotions, and behaviors that characterize individuals, plays a crucial role in understanding human behavior, decision-making, and interaction [2, 21, 39]. Its influence spans numerous domains, including mental health, education, organizational behavior, and personal relationships [4, 69]. While traditional personality assessments, particularly those based on the Big Five Model [49], have proven valuable in contexts ranging from therapy to marketing [48], they are often costly, time-consuming, and susceptible to biases such as social desirability and limited self-awareness [37]. These limitations highlight the need for more efficient and objective methods of personality assessment.

This thesis addresses this need by exploring the transformative potential of Large Language Models (LLMs) a branch of Artificial Intelligence (AI) with advanced generative capabilities. Unlike previous research that primarily employed LLMs for generating text embeddings to be used as input into separate classifiers [84], this study utilizes LLMs as independent, generative decision-makers. In this innovative approach, the LLM acts as the sole agent for personality trait predictions, directly interpreting text and providing assessments in response to carefully designed prompts, similar to how conversational AI systems like ChatGPT operate. The rapid advancement of LLMs, particularly since the introduction of the Transformer architecture in 2017 [89] and the subsequent 'LLM Boom' of 2022, presents a unique opportunity to revolutionize personality assessment. Central to this advancement is the development of prompt engineering [42], the art and science of crafting precise input prompts to elicit desired outputs. This has been critical in maximizing the effectiveness

of these models, enabling them to generate more accurate and relevant results [8].

Building on these advancements, this thesis investigates the efficacy of LLMs in predicting Big Five personality traits directly from text. It focuses on a comparative analysis to identify optimal conditions for accurate and interpretable predictions. The core methodology involves semantic chunking of the Essay Dataset [65], programmatic and semantic feature extraction using LLMs, and experimentation with personality trait prediction using various input configurations and prompting techniques. This study explores the impact of both Zero-shot (Baseline) and Chain-of-Thought (CoT) reasoning [92] on model performance.

The contributions of this research are both methodological and empirical. Methodologically, it introduces a novel hybrid approach, integrating traditional linguistic metrics with cutting-edge LLM techniques, pushing the boundaries of current methods in computational personality assessment. Empirically, it provides a systematic evaluation of multiple LLM configurations for personality prediction, offering insights into the strengths and limitations of different approaches. The findings have practical implications for fields such as mental health, education, and human-computer interaction, offering the potential for scalable and accessible personality assessments. These assessments could enhance AI-based therapy, personalize learning experiences, or inform the development of more human-like AI agents.

This study is guided by the following research questions: 1) How effectively can LLMs predict personality traits from short, semantically coherent text chunks? 2) What is the impact of incorporating programmatic features and LLM-based semantic features on classification accuracy? 3) How do different prompting techniques, specifically CoT, influence model performance and the interpretability of results? By exploring these questions, this research aims to advance the understanding of personality through language, leveraging the potential

of LLMs to create more accessible, scalable, and nuanced approaches to personality assessment.

The remainder of this thesis is structured as follows: Chapter 2 reviews the relevant literature on personality psychology, computational methods for personality assessment, and the evolution and capabilities of LLMs. Chapter 3 introduces the Essay Dataset and details its preparation for use in this study. Chapter 4 outlines the methodology, including the semantic chunking process, feature extraction techniques, and the experimental design. Chapter 5 presents the experimental results, providing a quantitative analysis of model performance across different conditions. Chapter 6 discusses these findings in detail, exploring their theoretical and practical implications, as well as acknowledging the limitations of the study. Chapter 7 concludes the thesis by summarizing the key contributions and outlining promising directions for future research.

Related Work

The foundation and objectives established above require careful examination of existing research across multiple domains. The following chapter reviews key developments across multiple domains: psychological foundations of personality theory (Section 2.1), evolution of computational personality analysis (Section 2.2), recent advances in LLMs (Section 2.3) and prompt engineering (Section 2.3.3), providing the theoretical groundwork for understanding how LLMs can be leveraged for personality trait classification.

2.1 Foundations of Personality and Language

Building on the introductions overview of personalitys importance across diverse contexts, this section lays out the theoretical underpinnings of personality. It highlights the Big Five personality model and examines how language use can reflect deeper psychological constructs.

2.1.1 Psychological Foundations of Personality

Definitions. Personality, broadly defined, is the unique configuration of traits, characteristics, and dispositions that shape an individuals consistent patterns of thought, emotion, and behavior [2, 21, 24]. It is shaped by biological, environmental, and social factors [10, 19] and plays a significant role in predicting

behaviors and outcomes across clinical, educational, and organizational contexts [2, 4, 10]

The Big Five. Among the various theoretical approaches to personality, the Big Five model (also known as the Five-Factor Model) has achieved substantial scholarly recognition [16, 18, 29, 31, 41]. This model emerged from factor analytic and lexical studies of personality descriptors found in everyday language [17, 18, 21, 24]. It identifies five broad dimensions: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (often abbreviated as OCEAN) [49]. These factors have shown consistent cross-cultural generalizability [15, 38, 50] and provide a comprehensive yet flexible framework for understanding individual differences [18, 29]. The model's broad applicability has influenced research in areas such as academic achievement, work performance, mental health, and interpersonal relationships (e.g., [32, 57, 69, 96]).

Criticism. The Big Five model is not without its critics. Some researchers argue that it oversimplifies the complexity of human personality by reducing it to five broad dimensions [6]. Others have questioned the comprehensiveness of the model, suggesting that it may not capture all relevant personality traits, particularly those related to morality, values, or specific skills [63]. Additionally, the reliance on factor analysis as a primary method for identifying personality dimensions has been debated, with some arguing that it may lead to the identification of statistical artifacts rather than true underlying constructs [7]. Despite these criticisms, the Big Five remains a dominant framework in personality research due to its robust empirical support and practicality.

Measuring Personality. Traditional methods of assessing personality often rely on self-report inventories (e.g., the Big Five Inventory) or clinical interviews [16, 31]. While these approaches have proven useful for assessments, they can be limited by biases such as social desirability, distorted self-views, or reference-group effects [37, 62]. Researchers have thus increasingly turned to alternative

assessments based on naturalistic data, especially textual data, to overcome these obstacles.

2.1.2 Linguistic Correlates of Personality

The Lexical Hypothesis. Rooted in the idea that language reflects fundamental aspects of human experience [3], the lexical hypothesis provides a theoretical basis for computational personality analysis. It suggests that personality attributes are encoded in the words people choose to describe themselves and others [24, 46, 60]. This encoding manifests not only in the use of specific trait adjectives (e.g., 'outgoing,' 'anxious') but also in more subtle linguistic patterns to infer underlying personality traits.

Pioneering Psycholinguistic Methods and Tools. Early studies established links between specific linguistic features and personality traits. Notably, the Linguistic Inquiry and Word Count (LIWC) tool [65–67] systematically categorized words into psychologically meaningful groups, demonstrating that even function words (e.g., pronouns, articles) can reflect stable aspects of personality [46, 64, 85]. For instance, Pennebaker and King (1999) found that individuals scoring high on Extraversion used more upbeat, social words, while those higher in Neuroticism employed more anxiety-related terms. Additionally, individuals high in Openness to Experience tended to use more articles and longer words, potentially reflecting their intellectual curiosity and engagement with complex ideas. Conversely, those high in Conscientiousness used fewer discrepancy words (e.g., "would," "should"), suggesting a more organized and less uncertain cognitive style [65].

These findings highlighted the potential of using relatively simple linguistic features to gain insights into personality. The Essay Dataset [65], comprising stream-of-consciousness texts paired with personality labels, provided a valuable resource for this work. However, while these early psycholinguistic approaches were influential in linking linguistic style to underlying personality

traits [95], they also revealed challenges in fully capturing complex constructs. Traits such as Openness may depend on more nuanced indicators like rhetorical structure, irony, or contextual shifts, which simple lexical analysis struggles to detect [66].

2.2 Computational Personality Prediction

The desire to move beyond manual coding and simple feature counts, coupled with the need to address the limitations of early approaches, has driven the field towards more advanced computational methods capable of handling larger and more diverse datasets [83, 94]. In this section, the shift to expanded computational tools and the use of social media data is explored, which provides extensive textual materials reflecting authentic, spontaneous communication [11, 23].

2.2.1 Machine Learning

Early Machine Learning and Feature Engineering. Traditional machine learning approaches leveraged handcrafted linguistic features derived from established psycholinguistic insights. Mairesse et al. (2007) were among the first to build a personality prediction system that used a combination of lexical and syntactic features. This work demonstrated that lexical diversity, function word use, and sentence length were highly predictive of certain personality traits. For example, extroverts tend to use more positive words and have a higher proportion of frequently used n-grams from the spoken language, news, and magazine registers. On the other hand, individuals with neurotic personality traits tend to use more words related to anxiety, negative tones, and mental health [46]. However, this reliance on pre-defined, handcrafted features presented several limitations such as significant required domain expertise and manual effort to develop and validate features. Features often captured only surface-

level linguistic patterns, failing to account for the nuances of context that can significantly influence the meaning of text [95].

Transition to Deep Learning. As computational capacities expanded, researchers shifted to deep neural networks, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to better model linguistic complexity. RNN-based architectures, including LSTMs and GRUs, were particularly effective in analyzing sequential text data, enabling models to identify temporal dependencies and patterns in language [47, 83]. For example, Majumder et al. (2017) proposed a deep learning framework using recurrent neural networks (RNNs) to extract sequential patterns in text for personality prediction. They found that neural models significantly improved accuracy over traditional methods, especially in capturing traits like neuroticism and conscientiousness [47]. Although these models automatically learned features from raw text and reduced reliance on handcrafted indicators, deep learning models still faced challenges. They often required large amounts of labeled data for training, and their 'black box' nature made it difficult to interpret the specific linguistic features driving their predictions. Furthermore, they still struggled with highly context-dependent phenomena, such as sarcasm or abrupt shifts in emotional tone [83].

Transformer-Based Advances and Context-Aware Embeddings. A significant leap forward came with the emergence of transformer-based architectures, including BERT, RoBERTa, and GPT variants [20, 89]. These models introduced a robust self-attention mechanism that simultaneously processes all positions in the input sequence and thus brings in more context, thereby overcoming the limitations of deep learning methodologies. In the context of personality prediction, this capacity for long-range contextual awareness is essential for detecting subtle cues in user-generated texts such as rapid emotional transitions, ironic undertones, or thematically shifting content [66, 83]. By producing context-aware embeddings, transformers enable a richer representation of linguistic features than simpler lexical or even simpler semantic features can offer... Researchers

have profited from this by fine-tuning large pre-trained models for personality classification tasks, often reporting superior performance compared to earlier approaches [75]. This shift toward transformer-based models thus addressed many of the context-related limitations faced by traditional neural architectures and further advanced the field of computational personality classification.

2.2.2 State-of-the-Art

While traditional NLP approaches, such as context-aware embeddings, are effective in capturing sequential patterns within text, recent advances in natural language understanding (NLU) and LLMs have considerably deepened models ability to interpret meaning and intent [13]. This shift toward more sophisticated NLU has driven the development of advanced and unexpectedly popular chat interfaces and contributed to the widespread adoption of LLMs. Moreover, the increasing accessibility of powerful LLMs through APIs and open-source initiatives has democratized their use, enabling researchers with varying levels of computational resources and expertise to explore their potential in diverse domains, including personality psychology.

Leveraging vast training corpora, LLMs can uncover subtle linguistic cues that signal underlying personality traits [58]. Consequently, further exploration of prompt engineering strategies and LLM-based feature extraction techniques holds the potential to significantly enhance the accuracy, scalability, and robustness of personality classification systems.

Building on these advancements, this thesis focuses on leveraging the capabilities of LLMs, accessed through a conversational interface, for personality trait classification from stream-of-consciousness text. By exploring different prompt engineering strategies and input configurations, this research aims to contribute to a deeper understanding of whether and - if so - how LLMs can be effectively employed for nuanced and context-aware personality analysis

2.3 Large Language Models

Large language models (LLMs) have evolved rapidly, from early models like BERT, which pioneered bidirectional attention mechanisms for capturing contextual relationships in text [20], to advanced architectures like the Generative Pre-trained Transformer (GPT) series developed by OpenAI [8, 55]. While models like RoBERTa [43] and T5 [73] refined specific aspects of these approaches, the GPT series, starting with in GPT-2 [53, 72], pushed the boundaries of scale in both model parameters and training data. This evolution has led to significant improvements in language understanding, generation, and the ability to perform complex reasoning tasks. While each architecture has its own nuances, they all share the core principle of using the Transformer architecture and self-attention mechanisms to process and generate text [89].

The Transformer utilizes 'self-attention' to focus on relevant parts of the text, allowing it to select the most suitable words for the context. Essentially, it transforms an input sequence of words into a new representation that captures the relationships between them, enabling coherent and contextually relevant text generation [75]. LLMs have been applied in various fields, such as healthcare, finance, customer service, and education, and have shown remarkable state-of-the-art performance in NLP applications [22].

Importantly, LLMs promise deeper semantic understanding, which can be crucial for nuanced tasks like personality analysis. Rather than solely counting word frequencies or using shallow syntactic cues, LLMs can capture broader thematic elements, interpret subtle language patterns, and generate structured responses. As discussed earlier, personality research often requires detecting cues that are buried within broader discourse structures making advanced models like LLMs especially relevant to this thesis, which relies on chunked stream-of-consciousness essays to classify Big Five traits.

2.3.1 Chunking Strategies

While LLMs have demonstrated impressive capabilities in handling lengthy texts, researchers have recognized that strategically "chunking" or segmenting text can significantly impact LLM performance in various retrieval and generation tasks [71, 77, 80, 90]. Semantic chunking methods aim to identify boundaries between text segments based on shifts in meaning or topic, often using techniques like measuring cosine similarity between sentence embeddings or identifying points of low lexical cohesion [25]. By grouping together sentences that are semantically related, these methods aim to create chunks that are more coherent and meaningful units of information for the LLM to process. This can be particularly beneficial for tasks like personality assessment, where understanding the context and flow of ideas within a text is crucial for accurate trait inference.

Chunking addresses several LLM limitations by improving computational efficiency for long texts, overcoming memory constraints, helping maintain focus, and improving coherence and contextual understanding. Different chunking strategies offer varying benefits. Fixed-size chunking, while simple, can disrupt semantic units. Semantic chunking, conversely, preserves natural text structures, potentially leading to more coherent LLM processing [1, 45, 80]. However, optimal chunking depends heavily on the task and text. In personality analysis, where inputs vary in length and coherence, the use of chunking methods remains under explored.

2.3.2 Applications in Feature Extraction and Personality

LLMs and Feature Extraction. Advanced LLMs like GPT-4o offer significant potential for extracting nuanced linguistic and semantic features from text. While earlier models like BERT demonstrated strong capabilities in capturing context-

sensitive semantic representations, advancing tasks such as sentiment analysis and emotion detection [20], the scale and architectural refinements of newer LLMs provide even greater capacity for interpreting complex textual data and generating fluent, human-like text [8, 55]. This enhanced ability to understand and process nuanced language makes them promising tools for a wide range of applications involving subtle feature extraction from text, including but not limited to personality assessment. However, research specifically exploring the application of these newer LLMs, with their enhanced capabilities, for fine-grained feature extraction across various domains remains limited.

LLMs and Personality. Research coupling LLMs with personality inference remains relatively new but is rapidly expanding. This work can be categorized into two main lines of inquiry: 1) investigating whether LLMs themselves exhibit consistent 'personalities', and 2) assessing the capacity of LLMs to infer human personality traits from text.

In the first line of research, studies have found that manipulating prompts can yield variable apparent traits in LLMs. For instance, [41] tested ChatGPT against standard personality inventories and found partial alignment with known trait patterns, but also inconsistencies in the model's responses. Similarly, [81] demonstrated that changes in prompts or conversational context could significantly shift an LLM's "apparent" Big Five trait scores. Furthermore, one study showed that assigning a 'customer service' persona caused a model to adopt more polite and agreeable language [28], while another study found that an LLM could maintain a specific personality 'role' across extended dialogues [59]. These findings suggest that while LLMs may simulate personality traits, this behavior is highly sensitive to the specific input they receive.

The second line of research focuses on using LLMs to assess human personality, an area that has seen limited but growing investigation. Rao et al. (2023) introduced a framework to assess human personality using ChatGPT and GPT-4 based on the Myers-Briggs Type Indicator (MBTI) test. Their method included

unbiased prompts to reduce sensitivity to prompt biases, subject-replaced queries to allow flexible assessments, and correctness-evaluated instructions to clarify responses. While their findings showed that LLMs could provide consistent and fairer assessments compared to InstructGPT, they also revealed ongoing challenges with prompt sensitivity and variability in outputs, underscoring the need for more robust evaluation techniques [74].

Ganesan et al. (2023) evaluated the zero-shot capabilities of GPT-3 for classifying Big Five personality traits from social media posts. They examined the impact of different knowledge infusion techniques, such as using textbook definitions, word lists, or survey items. They found that prompts containing survey items yielded the highest classification accuracy, particularly for Extraversion and Conscientiousness. However, GPT-3 struggled with more nuanced classifications and often defaulted to predicting high Openness and low Neuroticism. The study highlighted that while LLMs encode relevant linguistic patterns, they face limitations in handling fine-grained personality distinctions [87].

Taken together, these studies highlight both the promise and the limitations of using LLMs for personality analysis. While recent advancements demonstrate that LLMs can be leveraged for personality assessment tasks, challenges remain, particularly regarding prompt sensitivity, bias, and the models' capacity to handle complex, multi-class distinctions. Notably, the second line of research remains sparse, indicating a significant gap in the literature. This underscores the importance of the present work, which aims to further develop methods for LLM-based personality predictions by addressing these gaps through improved prompt engineering and evaluation strategies.

2.3.3 Prompt Engineering

Prompt engineering is a critical aspect of working with LLMs. It involves crafting effective prompts or instructions to guide the model towards generating desired outputs that are accurate, relevant, and useful [33, 82]. By carefully

adjusting the wording, context, and examples within a prompt, we can mitigate issues like hallucinations (generating incorrect or nonsensical information) and better align the model’s responses with user goals [12].

Strategies for Prompt Design. Various prompting techniques exist, including zero-shot prompting, which leverages the LLM’s inherent knowledge with minimal instructions [70], and few-shot prompting, which incorporates a limited number of examples to provide context and improve performance [8, 70]. Chain-of-Thought (CoT) prompting encourages the LLM to articulate its reasoning process, enhancing performance in complex or multi-step tasks [34, 92]. Additionally, knowledge-infused prompting integrates domain-specific knowledge into the prompt to improve accuracy in specialized tasks [14, 87, 91].

Challenges of Prompt Engineering. However, prompt engineering presents challenges, such as finding the optimal balance between clarity and specificity [76], managing the sensitivity of LLMs to minor wording changes, and handling ambiguous input to prevent irrelevant or hallucinated content [74, 102]. These challenges highlight the need for a nuanced understanding of LLM behavior and careful consideration in prompt design to effectively harness the capabilities of these powerful language models. This thesis explores how different prompt techniques affect classification outcomes and model confidence.

Across the studies introduced in the previous sections, key open issues include interpretability (understanding how LLMs derive specific trait inferences), calibration (ensuring consistent and well-calibrated confidence in outputs under varied prompts), and validation (comparing LLM assessments against established psychometric standards).

2.3.4 Output Evaluation

Evaluating the outputs generated by LLMs for personality assessment presents unique challenges. The subjective nature of personality traits, combined with

the potential for biases in LLM-generated interpretations, necessitates careful consideration of appropriate evaluation metrics.

LLM-Based Personality Classification Studies. This becomes apparent when considering the few studies that have directly explored LLM-based personality assessment. Firstly, Rao et al. (2023) evaluated LLMs on personality assessment through three key metrics: consistency, robustness, and fairness. Consistency measured the stability of assessments across multiple tests, robustness evaluated the model’s sensitivity to prompt perturbations, and fairness assessed potential biases across different demographic groups. While their framework provides a valuable starting point, it primarily relies on quantitative measures that may not fully capture the qualitative nuances of personality. Furthermore, the MBTI itself has been subject to criticism regarding its psychometric properties, potentially impacting the validity of the evaluation [74]. Secondly, Ganesan et al. (2023), primarily employed Macro F1 scores, benchmarking the model’s performance against a lexical regression baseline (WT-LEX) in both binary and multi-class settings. However, the applicability of F1 score evaluation is questionable, as this metric is typically used for models trained on labeled data, unlike GPT-3’s zero-shot application in the study. This aligns with the study’s own observation that GPT-3 appears unsuited for regression personality classification. This, combined with their reliance on a lexical baseline for comparison, suggest that their evaluation approach, while informative, may not fully capture the complexities of assessing LLM performance on this nuanced task [87].

These studies underscore the limitations of relying solely on traditional quantitative metrics like accuracy or F1 scores for evaluating LLM-based personality assessments. Such metrics fail to capture crucial aspects of model behavior, including the quality of reasoning, the appropriate use of textual evidence, and potential subtle biases. This gap highlights the need for more comprehensive approaches that incorporate qualitative analysis and domain-specific considerations in the complex field of personality assessment. Existing LLM evaluation

frameworks, such as self-criticism methods that rely on factuality [100], self-contrast [99], and probability discrepancy analysis [93], are often designed for tasks with more objective ground truths, making them less suitable for the nuanced and interpretive nature of personality assessment.

LLM Self-Evaluation Confidence Scores. This thesis addresses this gap by employing a multifaceted evaluation strategy. It leverages LLM-generated confidence scores and accompanying explanations as internal measures of the model's certainty, offering insights into its decision-making process. While LLMs can exhibit a confidence-competence gap [79], where confidence does not always align with accuracy, these scores, particularly when combined with detailed justifications, can still provide valuable information about the model's assessment of its own output. Models like GPT-4 have shown potential in gauging their competence accurately [78], and exploring confidence scores in this context may reveal patterns related to task difficulty or trait ambiguity, similar to the Dunning-Kruger effect observed in humans [36].

LLM as a Judge. The thesis also incorporates a qualitative "LLM as a judge" assessment. This approach, gaining traction in recent research, leverages the capabilities of LLMs to evaluate the quality of text generated by other models or systems [40]. In this study, a separate LLM (GPT-1o) is employed to assess the primary model's (GPT-4o) reasoning and textual grounding, effectively acting as a judge of its personality classifications. This method allows for fine-grained evaluations, assessing not only the final classification but also specific attributes such as helpfulness, harmlessness, reliability, relevance, and overall quality [40]. While "LLM as a judge" offers significant potential, it also presents challenges, including potential biases and vulnerabilities [86]. For example, studies have shown that LLMs may exhibit a tendency to judge positively when uncertain, particularly smaller models [86]. To mitigate these risks, careful calibration and the use of techniques such as pairwise contrastive training can be employed [101]. In this study, the "LLM as a judge" framework is supplemented by a human "smell check" for alignment with human intuition.

Conclusion. This study bridges the gap in existing research by moving beyond traditional quantitative metrics and incorporating novel approaches like LLM-generated confidence scores and a qualitative "LLM as a judge" assessment, validated by human judgment. This multifaceted evaluation strategy provides a more nuanced understanding of LLM performance in the complex domain of personality assessment, paving the way for more reliable and interpretable models.

This chapter reviewed the evolution of personality analysis, from its psychological foundations to the latest advancements in computational methods and LLMs. While traditional approaches have provided valuable insights, the emergence of LLMs offers new opportunities to analyze personality through language with greater depth and nuance. As this field is evolving, significant gaps in literature remain. First, there is minimal research on adapting chunking strategies especially for freeform, stream-of-consciousness texts to ensure that [non-chaotic] context is preserved for trait classification. Second, prompt engineering has proven effective in various NLP tasks, yet its direct impact on personality classification remains underexplored, particularly with well performing approaches such as CoT. Third, methods for reliably evaluating LLM outputs in subjective domains are still evolving, raising questions about how to interpret model confidence, calibrate predictions, and ensure logical consistency.

These gaps underpin this thesis methodology, which combines semantic chunking, programmatic and LLM-based feature extraction, and systematic prompt engineering to classify Big Five personality traits. By addressing these open issues, the study aims to provide a more robust, interpretable framework for LLM-driven personality assessment that could ultimately extend to other complex, subjective domains beyond computational personality psychology.

Dataset

The Essay Dataset serves as a benchmark in personality research due to its unique combination of stream-of-consciousness writing and validated personality measures. This chapter examines the dataset's background and collection methodology (Section 3.1), followed by exploratory data analysis of both the original essays (Section 3.2) and semantically chunked segments (Section 3.2).

3.1 Introduction & Background

The dataset used in this study comprises stream-of-consciousness essays collected by Pennebaker & King (1999) [65]. Often referred to as the "Stream of Consciousness Essay Dataset" or simply the "Essays Dataset," this collection was developed as part of a series of studies exploring the relationship between language use and personality traits. The essays were written by undergraduate psychology students at the University of Texas at Austin and Southern Methodist University between 1993 and 1996. Participants were asked to write continuously for 20 minutes about their thoughts, feelings, and reflections on but not limited to starting and attending college. This assignment was designed to elicit natural, unfiltered language that could provide insights into the writers' personality traits.

Personality Assessment: In addition to their essays, participants completed the Big Five Inventory (BFI) [30], a well-established measure of the five major dimensions of personality: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. The BFI consists of 44 items rated on a 5-point

Likert scale. These scores provide standardized measures of personality traits, enabling researchers to examine the connections between linguistic patterns and personality characteristics.

Significance and Relevance: The Essays Dataset has become a cornerstone in the field of text-based personality prediction. Its significance lies in its combination of naturalistic language samples with validated personality measures. Some notable studies that have utilized this dataset include: Pennebaker and King's (1999) work, which established the initial connections between linguistic features and personality traits [65]; Mairesse et al.'s (2007) study, which compared various machine learning approaches for personality recognition from text [46]; Yarkoni's (2010) large-scale analysis of personality and word use, which identified specific linguistic markers associated with each of the Big Five traits [95]; Park et al.'s (2015) research on automatic personality assessment using open-vocabulary approaches [60].

These studies, among others, have demonstrated the dataset's utility in advancing our understanding of how personality is reflected in language use. It has been instrumental in developing and validating computational models of personality prediction from text, making it a valuable asset for interdisciplinary research at the intersection of psychology, linguistics, and artificial intelligence.

3.2 Original Dataset EDA

The dataset originally comprises 1,480 stream-of-consciousness essays authored by unique participants, ensuring that each essay represents an independent data point. To investigate the structure and variability of the dataset...

Structure. The dataset utilized in this study is organized as a dataframe with the following columns:

- #AUTHID: A unique identifier for each essay, including a year and a numerical ID (e.g., "1997-504851.txt").
- TEXT: The full stream-of-consciousness essay.
- cEXT: Binary indicator for Extraversion (y/n).
- cNEU: Binary indicator for Neuroticism (y/n).
- cAGR: Binary classification for Agreeableness (y/n).
- cCON: Binary classification for Conscientiousness (y/n).
- cOPN: Binary classification for Openness to Experience (y/n).

Word and Character Count Statistics. The word and character count statistics in Table 3.1 are visualized in Figure 3.1, revealing a right-skewed distribution for both metrics. Most essays fall within a reasonable range, with the majority clustering around the mean and fewer essays at the extremes.

Table 3.1: Word and Character Count Statistics for Essays Dataset

Statistic	Word Count	Character Count
Mean	651.1	3,286.8
Variance	65,879.83	1,689,340.38
Standard Deviation	256.67	1,299.75
Maximum	2,500	12,846
Minimum	33	158

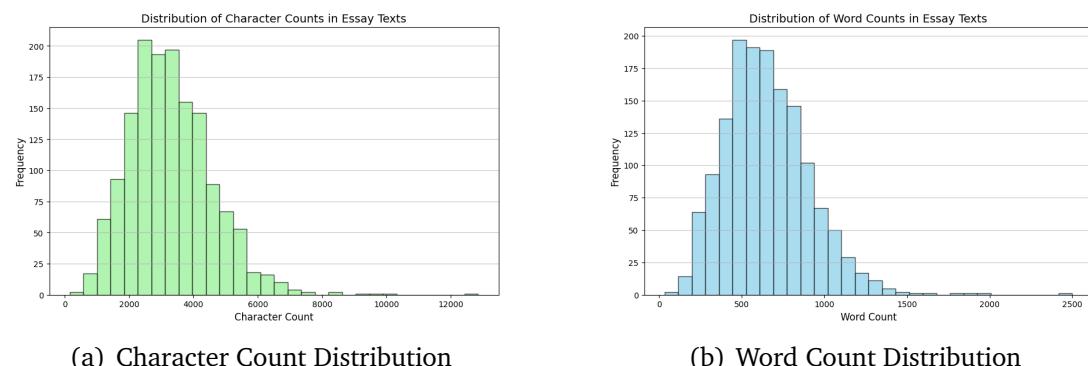


Figure 3.1: Distribution of Word and Character Counts in the Dataset.

Lexical Diversity. The cleaned text (after stopword removal, punctuation removal, and stemming) was analyzed for lexical diversity using both Type-Token Ratio (TTR) and the Measure of Textual Lexical Diversity (MTLD). TTR measures the ratio of unique words to total words, while MTLD provides a more robust measure that is less sensitive to text length. As shown in Table 3.2 and Figure 3.2, the mean TTR of 0.66 indicates a moderate level of lexical diversity across the essays, with relatively low variance (0.0061) and standard deviation (0.078). The MTLD scores, however, reveal greater variability in vocabulary use, with a mean of 139.65 and a wide range from 20 to 500, highlighting significant differences in lexical richness among the essays.

Table 3.2: Lexical Diversity Statistics for Essays Dataset

Statistic	TTR	MTLD
Mean	0.66	139.65
Variance	0.0061	3524.65
Standard Deviation	0.078	59.37
Maximum	0.92	500
Minimum	0.30	20

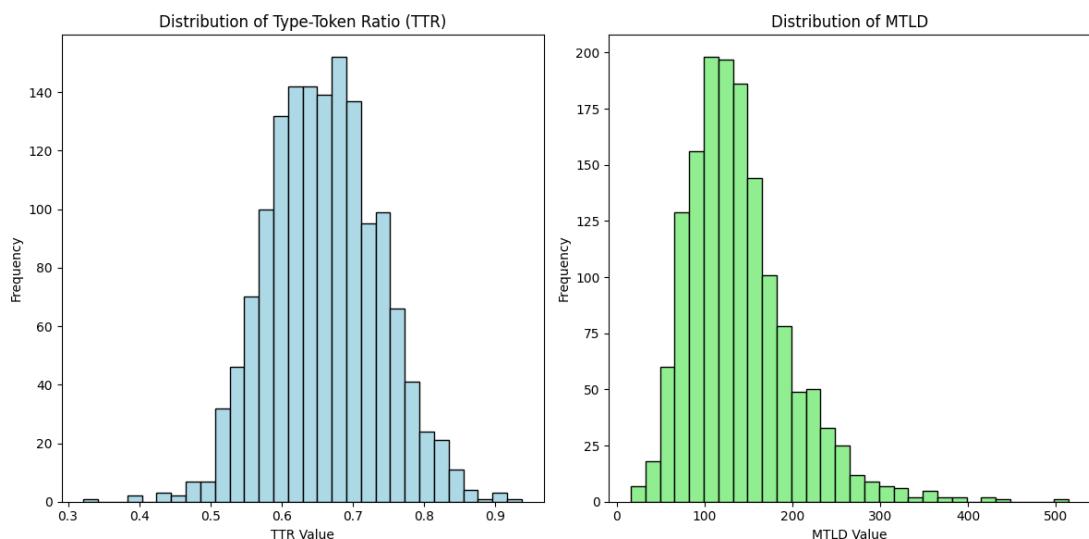


Figure 3.2: TTR and MTLD Distribution in Essay Dataset

3.3 Semantic Chunks Dataset EDA

After semantic chunking (detailed in Section 4.1), the original essays were divided into 15,000 semantically coherent chunks. Due to resource limitations and an error in the experimental setup that resulted in a non-randomized selection process, the first 2,000 chunks were used for the experiments. This subset includes contributions from 191 unique authors.

To understand the characteristics of these chunks, an exploratory data analysis was conducted, both on the full set of 2,000 chunks and on a subset excluding extreme outliers. Outliers were identified as chunks with word counts exceeding three times the interquartile range (IQR) above the third quartile. This resulted in the identification of 62 outlier chunks. While these outliers were excluded for visualization purposes to better illustrate central tendencies, they were included in all subsequent experiments and analyses. Figures 3.3 and 3.4 present these distributions with and without outliers, respectively.

Table 3.3: Dataset Statistics Overview

Metric	Full Dataset	Without Outliers*
Total Chunks	2,000	1,938
Unique Authors	191	189
Mean Words/Chunk	63.41	53.89
Median Words/Chunk	46.0	45.0
Word Count SD	73.74	37.75

*Outliers removed using $3 \times \text{IQR}$ method for visualization purposes only

The word count distribution of the 2,000 chunks (Figure 3.3) exhibits significant right-skewness, with chunk lengths ranging from 2 to 1,371 words. The mean chunk length is 63 words, but the median (46 words) provides a better representation of the typical chunk size, as it is less influenced by the long right tail. The large standard deviation (73.74) reflects the considerable variation in chunk lengths (Table 3.3).

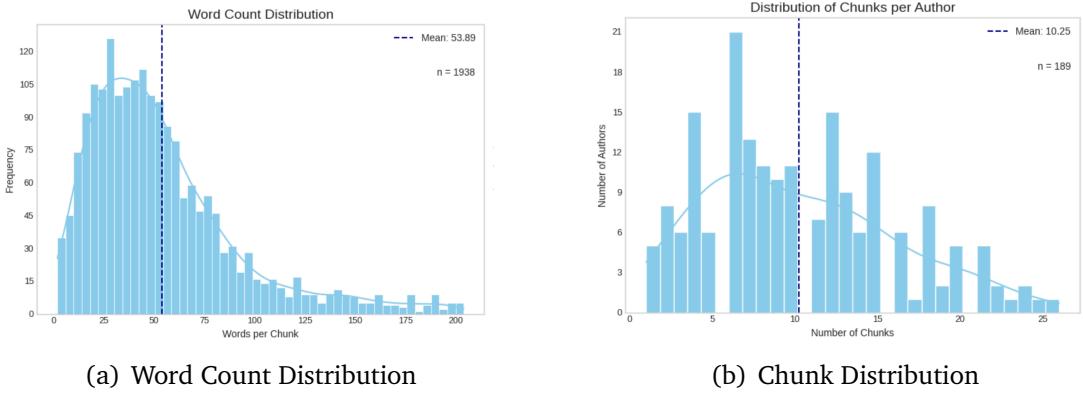


Figure 3.3: Distribution of Word and Chunk Counts with Outliers Excluded

Excluding the 63 outliers (Figure 3.4) provides a clearer picture of the chunk distribution across authors. In this filtered set, the mean number of chunks per author is 10.53, with a median of 10. Most authors contribute between 5 and 15 chunks, with no single author contributing more than 26 chunks.

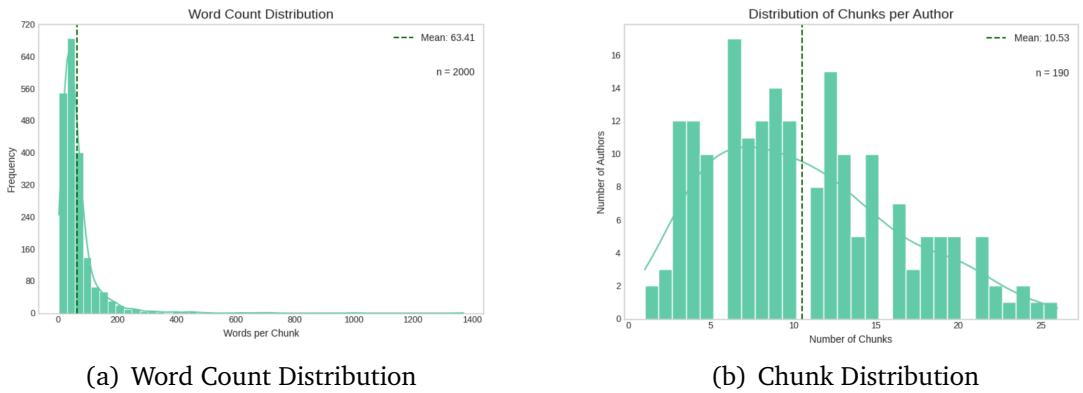


Figure 3.4: Distribution of Word and Chunk Counts with Outliers Included

Personality Combination and Trait Distribution. Each chunk inherits the personality trait labels (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) from its corresponding author. These traits are represented as binary labels ("y" or "n"), resulting in 32 possible trait combinations. Figure 3.5(a)

shows the frequency of each combination. The most frequent combination appears in 164 chunks (8.2% of the dataset), while the least frequent appears in only 17 chunks (0.85%). This indicates that while certain personality profiles are more common, no single profile dominates the dataset.

The distribution of individual trait labels (Figure 3.5(b)) is nearly balanced across all five traits, with each trait exhibiting close to a 50/50 split between "y" and "n" labels. This balance is beneficial for downstream classification tasks, mitigating potential biases arising from imbalanced classes. The variety in trait combinations and the near-balanced distribution of individual traits underscore the dataset's richness for studying personality.

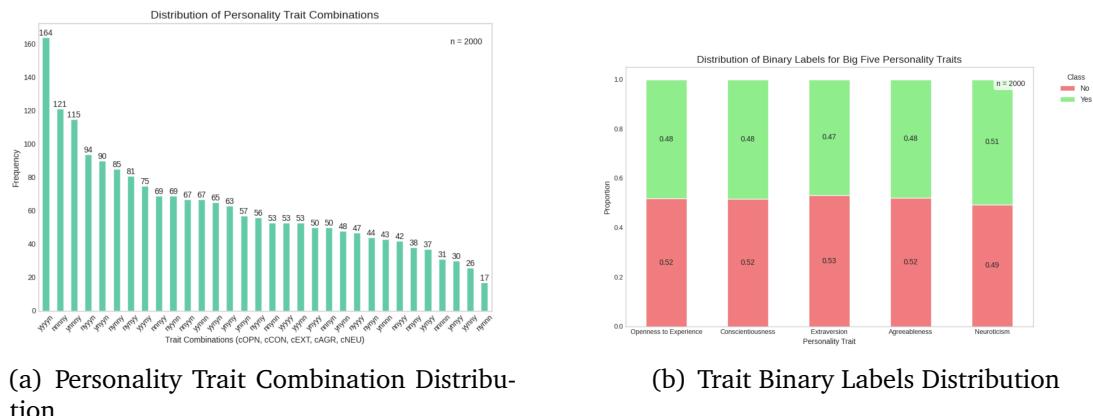


Figure 3.5: Distribution of Personality Combinations and Trait Labels Across the Dataset

Methodology

To effectively analyze the Essay Dataset and explore the capabilities of LLMs for personality prediction, Chapter 4 presents a four-part methodological framework: semantic chunking for text preparation (Section 4.1), feature extraction through both programmatic (Section 4.2) and semantic LLM-based approaches (Section 4.3), and a Big Five personality trait classification experiment investigating combinations of input types and prompting techniques on classification performance (Section 4.4). Each component builds upon established techniques while introducing novel adaptations for personality classification. See Figure 4.1 for an overview of the methodological framework.

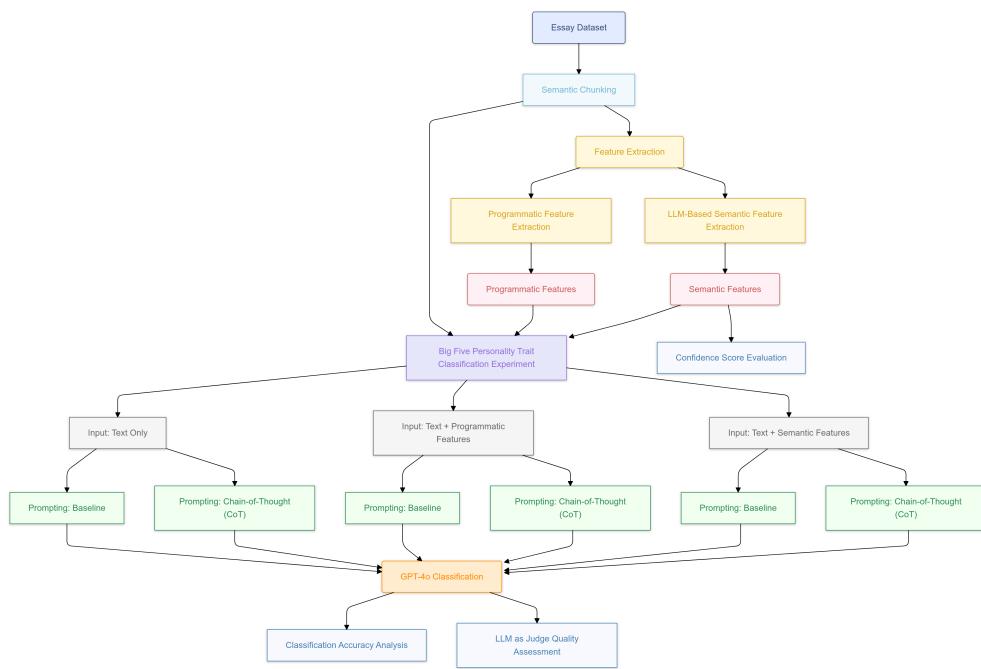


Figure 4.1: Higher-Level Methodology Overview

4.1 Semantic Chunking

This section details the semantic chunking methodology employed to prepare the stream-of-consciousness essays for semantic feature extraction and personality classification. The rationale behind this approach (Section 4.1.1), followed by a step-by-step description of the method, is presented below (Section 4.1.2).

4.1.1 Rationale

The stream-of-consciousness essays in the dataset presented a unique challenge for the task of personality classification using LLMs. Unlike well-structured text with clear paragraphs and topic sentences, these essays often lack traditional organizational cues, instead mimicking the natural flow of thought through a continuous, and at times, fragmented narrative [65]. While LLMs have demonstrated remarkable capabilities in natural language processing, they are often sensitive to the coherence and context of the input text. This is particularly crucial for tasks like personality analysis, where subtle linguistic cues, thematic shifts, and the overall flow of ideas can be highly indicative of underlying personality traits.

In this study, the inherent fragmented and associative nature of the stream-of-consciousness essays posed a significant obstacle to accurate personality classification. Fixed-size chunking, a common technique that divides text into segments of a predetermined length, would likely disrupt the natural flow of these essays, severing connections between related thoughts and potentially obscuring subtle personality cues. For instance, a recurring theme or a gradual shift in tone, both potentially indicative of a specific trait, might be split across multiple fixed-size chunks, diminishing their interpretability for the LLM. To ensure that the LLM could effectively process and interpret the text for accurate personality classification, a semantic chunking strategy was employed. Semantic chunking

aims to identify boundaries based on shifts in meaning or topic, thus preserving the natural flow of ideas within the text [44, 52, 61]. It is analogous to dividing a story into meaningful scenes rather than arbitrary paragraphs, which helps in understanding the text’s underlying structure and themes [68, 80]. By grouping together semantically related sentences, this method creates chunks that are more likely to represent coherent units of meaning, providing the LLM with the necessary context to identify subtle linguistic patterns and thematic elements crucial for inferring personality traits. In essence, semantic chunking aims to present the LLM with text segments that more closely resemble the coherent thought processes that underlie personality expression, thereby enhancing the accuracy and reliability of the classification task.

4.1.2 Method

The semantic chunking method used in this thesis employed a two-pass approach to transform the original stream-of-consciousness essays into thematically coherent segments.

Pass 1: Identifying Fine-Grained Thematic Boundaries

The first pass aimed to identify fine-grained thematic boundaries within each essay, adapting to the natural flow of the text without imposing arbitrary length constraints. This was achieved through the following steps:

- (1) **Sentence Embeddings:** Each essay was segmented into sentences and each sentence was converted into a 768-dimensional numerical representation (embedding) using the Stella 1.5B model [97], a 1.5 billion parameter English language model. Stella 1.5B was selected for its demonstrated ability to generate high-quality semantic representations, capturing nuanced meanings and contextual relationships within text, making it well-suited for this task.

- (2) **Cosine Similarity Calculation:** The cosine similarity between consecutive sentence embeddings was calculated. Cosine similarity measures the angle between two vectors, providing a score between -1 and 1 that reflects the degree of semantic similarity between the corresponding sentences. Higher scores indicate greater similarity.
- (3) **Local Minima Detection:** "Local minima" in the sequence of cosine similarity scores were identified using a custom function. These points represent significant dips in similarity between consecutive sentences, indicating shifts in topic or tone and thus serving as natural thematic boundaries [61]. A sentence was considered a boundary if its similarity score was lower than both the preceding and succeeding similarity scores.

Pass 2: Merging Small Chunks

The initial granular segmentation, while preserving subtle thematic transitions, often resulted in numerous small chunks. The average length of chunks after the first pass was approximately 44 words, and each author's text was divided into an average of 15 chunks. To create slightly larger, more contextually rich segments better suited for LLM-based personality analysis, a second pass was implemented:

- (1) **Chunk Embeddings:** Each chunk generated in the first pass was converted into a new embedding. This was accomplished by averaging the sentence embeddings (generated in Pass 1) of all sentences within the chunk, again using the Stella 1.5B model [97]. This provided a single vector representation summarizing the overall meaning of each chunk.
- (2) **Similarity-Based Merging:** Adjacent chunks were iteratively merged if their cosine similarity exceeded a predetermined threshold [44, 52].
- (3) **Threshold Optimization:** To determine the optimal threshold for merging, different values (0.6, 0.5, and 0.4) were tested on a subset of the data. A threshold of 0.4 was found to strike the best balance between merging semantically related chunks and preserving distinct thematic units. This

resulted in an average chunk size of approximately 63 words, providing more contextually rich segments for subsequent analysis

This two-pass approach generates text chunks that capture coherent themes or ideas, offering a more suitable input for LLMs compared to fixed-length segmentation. Importantly, this is the first known application of semantic chunking to prepare the Essay Dataset specifically for LLM-based personality classification. The code used for this process is available at [\[link to code repository\]](#), and the exploratory data analysis in Section 3.3 provides further details on the resulting chunk characteristics.

4.2 Programmatic Feature Extraction

Following the semantic chunking of the Essay Dataset, a programmatic feature extraction process was undertaken to quantify specific linguistic characteristics within each text chunk. These features were then used in conjunction with the semantic chunks as input for the LLM in the Text + Programmatic Features experimental conditions for the personality classification experiment. This section details the rationale for selecting these features (Section 4.2.1), the methods employed for their extraction, and their subsequent normalization for use as input to the personality classification task (Section 4.2.2).

4.2.1 Feature Set Rationale

To prepare input for the Text + Programmatic conditions, a range of programmatic features were systematically extracted from each text chunk. This involved quantifying specific, pre-defined linguistic elements using a spaCy-based NLP pipeline (the en_core_web_sm model) [51] for tokenization, part-of-speech tagging, and sentence segmentation, and NLTK’s VADER sentiment analyzer [26] for sentiment scoring. These code-driven methods are efficient, reproducible, and provide stable, low-level linguistic indicators. While these features

may not capture the same level of nuanced meaning as semantic feature extraction (see Section 4.3), they represent a set of empirically validated variables that have been shown to correlate with personality traits in previous research [64, 65, 67, 95]. The extracted programmatic features span three broad categories:

(1) Lexical and Structural Metrics

- **Word Count and Character Count:** Basic text length metrics that can provide a general indication of verbosity or conciseness, potentially related to traits like Extraversion or Conscientiousness.
- **Type-Token Ratio (TTR):** Calculated as the number of unique words divided by the total number of words, TTR serves as a measure of lexical diversity. Higher lexical diversity has been associated with Openness to Experience [95], reflecting a broader vocabulary and potentially greater intellectual curiosity.

(2) Psycholinguistic and Sentiment Cues

- **Personal Pronoun Percentage:** The proportion of personal pronouns (e.g., "I," "me," "we," "us") relative to the total number of words. The use of first-person singular pronouns has been linked to traits like Neuroticism [64], while first-person plural pronouns can be indicative of social orientation and Extraversion [46].
- **Sentiment Polarity (VADER):** A compound sentiment score ranging from -1 (most extreme negative) to +1 (most extreme positive), reflecting the overall emotional valence of the text. VADER is particularly well-suited for analyzing informal text, making it appropriate for the stream-of-consciousness essays. Sentiment scores can potentially reflect traits like Neuroticism (associated with negative sentiment) or Agreeableness (associated with positive sentiment).

(3) Syntactic and Tense Patterns

- **Part-of-Speech (POS) Distribution:** Ratios of specific POS tags, including adverbs, pronouns, verbs, and nouns, expressed as percentages of the total number of words. Due to an unintentional error in

the initial feature extraction script, other POS tags were excluded. Although this limits the scope of the POS analysis, these four categories still provide valuable insights. For instance, a higher proportion of adverbs might suggest a more descriptive style, potentially related to Openness.

- **Tense Distribution:** Past, present, and future verb usage rates, expressed as percentages of total verbs. These tenses were selected for their potential to reflect an individual’s temporal orientation. A focus on the past might be linked to rumination (Neuroticism), while a focus on the future could be associated with planning (Conscientiousness) [64].

4.2.2 Normalization

While LLMs can process raw numeric inputs given sufficient context [88], normalization was applied to enhance interpretability and prevent features with larger values from disproportionately influencing the analysis. The following normalization procedures were applied to specific features:

- (1) **Percentage Conversion:** To ensure comparability across features, several metrics were converted to percentages. Personal pronoun usage was expressed as a percentage of the total word count. Similarly, the counts for each part-of-speech tag were individually calculated as percentages of the total word count. Verb tense usage was represented as percentages of the total number of verbs
- (2) **Retained Values:** Sentiment Polarity scores from VADER and Type-Token Ratio (TTR) were retained in their original form.
- (3) **Nan Handling:** Features with NaN values were excluded from the final feature set. While the absence of certain linguistic elements (the cause of NaNs) could potentially hold informational value, this study focused on

features with quantifiable numerical values to maintain consistency and simplify the analysis.

This approach ensures that each chunks programmatic features are interpretable linguistic signals that may link specific writing styles to underlying personality dimensions[9].

4.3 LLM-Based Semantic Feature Extraction

This thesis also conduced LLM-based semantic feature extraction to capture more nuanced and context-dependent aspects of language use in the chunks. While programmatic features provide quantifiable, surface-level linguistic indicators, this approach leverages the interpretive (context-dependent) and generative power of LLMs to identify and assess higher-level semantic features that may be more indicative of underlying personality traits. The extracted features from this process serve as the foundation for the Text + Semantic Features conditions in the personality classification experiment. This section details the rationale for selecting these semantic features (Section 4.3.1), the prompting technique and prompt design used for their extraction (Section 4.3.2), and the method employed for evaluating the extracted features (Section 4.3.3).

4.3.1 Feature Set Rationale

While deep learning and contextual embedding-based approaches have advanced beyond simple lexicon-based methods like LIWC [65], these traditional methods still provide foundational insights into linguistic patterns associated with personality. However, they often fall short in capturing the deeper, context-dependent nuances of personality expression. Personality traits often manifest through subtle narrative structures, coherence, emotional complexity, and

cognitive patterns that require a more holistic understanding of the text [46]. LLMs, with their advanced natural language understanding capabilities, offer a unique perspective by interpreting these latent, context-sensitive features. Unlike count-based methods, LLMs have the potential to extract contextually informed psycholinguistic insights, providing a richer and more nuanced understanding of the text and its relation to personality.

To leverage the interpretive capabilities of LLMs for a novel exploration of psycholinguistic dimensions within the Essay Dataset, five semantic features were selected for extraction. Importantly, the LLM was tasked with identifying these features without explicit definitions, relying on its inherent understanding of these concepts. This approach aimed to provide a more open-ended and nuanced interpretation of the text, unconstrained by potentially limiting pre-defined definitions. The selected features encompass aspects of cognition, emotion, and narrative style, and are detailed below:

- (1) **Cognitive Flexibility:** This feature assesses the author's ability to shift between different perspectives, mental frameworks, and thematic elements, reflecting the fluidity and adaptability of the author's thinking.
- (2) **Narrative and Discourse Coherence:** This feature evaluates the overall cohesiveness and logical flow of the author's narrative, considering both the logical connections between ideas and the emotional consistency of the text.
- (3) **Emotional Tone:** This feature assesses the overall emotional mood or affect expressed in the text, with the LLM tasked to identify and interpret the dominant emotions conveyed.
- (4) **Self-Reflection Depth:** This feature gauges the extent to which the author engages in introspection, exploring their own internal states, thoughts, feelings, motives, and beliefs.
- (5) **Analytical Thinking:** This feature assesses the degree to which the author employs reasoning, evidence, and logical argumentation to support their thoughts and ideas.

4.3.2 GPT-4o : CoT Prompt

For semantic feature extraction, this study employed OpenAI's GPT-4o model [55] for semantic feature extraction, accessed via their Python API (v1.15.0) [54]. GPT-4o was selected for its strong performance on complex NLP tasks. Model parameters were set to `temperature = 1.0`, `max_tokens = 1500`, and `top_p = 1.0` to balance response diversity with accuracy.

A single optimized run was conducted using the CoT prompting technique. The prompt, adapted from the universal template described in Section 4.4.4, focused on extracting the five semantic features. It instructed the model to provide a rating, justification, confidence score, and rationale for the confidence score for each feature in a single pass, without explicitly defining the features. Importantly, the prompt omitted any mention of personality traits to allow for a more open-ended interpretation of the text's semantic content. The specific prompt used is shown in Figure 4.2.

CoT Semantic Feature Extraction Prompt

```
You are an intelligent and disciplined assistant
↪ trained to determine the presence and degree of
↪ various nuanced psycholinguistic features in a
↪ stream-of-consciousness text. The features are:
↪ {features_str}.

*Your task is to reason through each feature
↪ step-by-step, explaining how the text provides
↪ evidence (or lack thereof) for each feature before
↪ determining the final result and confidence score.*

*Your output must be valid JSON* with the structure:

{
  "features": [
    {
      "feature": "Cognitive Flexibility",
      "reasoning_steps": [
        "..."
      ],
      "result": "...",
      "result_justification": "...",
      "confidence_score": "0.0-1.0",
      "confidence_score_justification": "..."
    },
    //... [other features truncated for brevity]
  ]
}

Rules:
1. For each feature, reason step-by-step as described,
↪ then provide the final evaluation in the JSON
↪ format.
2. If insufficient data, set 'result'='Not able to
↪ evaluate', 'result_justification'='Not able to
↪ evaluate', 'confidence_score'=0.0,
↪ 'confidence_score_justification'='Not able to
↪ evaluate'.
3. No extra text, no code fences, no keys beyond what's
↪ shown.
4. Provide detailed justifications for the reasoning,
↪ results, and confidence scores.

The text: {text_chunk}
```

Figure 4.2: CoT Feature Extraction Prompt

4.3.3 Confidence Scores Evaluation

Given the exploratory nature of LLM-based semantic feature extraction and the lack of definitive ground truth for these nuanced features, this study implements a novel approach: using LLM-generated confidence scores to assess extraction reliability. Alongside each extracted feature, GPT-4o provides a numerical confidence score (0.0 to 1.0) and a textual justification, reflecting its certainty in the assigned rating. This represents a form of LLM self-evaluation, a capability showing promise in improving LLM output reliability in other domains [5, 35]. To the best of my knowledge, this is the first application of LLM-generated confidence scores for evaluating semantic feature extraction in personality analysis.

The confidence scores offer a valuable internal measure of the model’s certainty, particularly relevant for subjective tasks like semantic feature extraction. To understand the distribution and potential utility of these scores, the following steps were taken:

- (1) **Numerical Conversion and Summary Statistics:** Confidence scores were converted to numerical format, and summary statistics (mean, median, standard deviation) were calculated for each of the five semantic features across all chunks. This provides a quantitative overview of the model’s confidence levels (see Table 5.1 and Figure 5.1 in Section 5.2).
 - High Confidence (e.g., 0.9): Suggests the model is relatively certain about its assessment and can provide a clear rationale, potentially indicating a strong presence of the semantic feature in the text.
 - Low Confidence (e.g., 0.3): Suggests the model is less certain, potentially due to ambiguous or insufficient evidence in the text, indicating the feature may be weakly present or absent.
- (2) **Inclusion of Textual Justifications:** The textual justifications accompanying each confidence score were retained. These provide insights into the model’s reasoning, enhancing transparency and interpretability. For exam-

ple, a justification might highlight specific phrases supporting the feature rating or point out the lack of clear evidence.

- (3) **Use in Classification:** Both the numerical confidence scores and their corresponding textual justifications were used as inputs in the downstream personality classification task (see Section 4.4).

While these confidence scores offer a promising avenue for evaluating LLM-based feature extraction, it is important to acknowledge that they are not a direct measure of accuracy but rather a reflection of the model’s internal assessment. Further evaluation, such as a qualitative analysis of the justifications and a comparison with human judgments (not feasible in this study due to resource constraints), could further illuminate the validity of these scores. Furthermore, filtering based on confidence scores could potentially be used to increase the reliability of extracted features. However, in this study, this was deemed impractical due to the limited dataset size and the potential for significant data loss.

4.4 Big Five Personality Trait Classification Experiment

This section describes the main experiment of the thesis, focused LLM classification of classifying Big Five personality traits from the semantically chunked essays. The experiment’s rationale is presented in Section 4.4.1, followed by details on the classification task (Section 4.4.2), experimental setup and conditions (Section 4.4.3), prompt design (Section 4.4.4), and evaluation methods (Section 4.4.5).

4.4.1 Overview

This experiment investigates the central research questions of this thesis by assessing the ability of OpenAI's GPT-4o [55] to accurately classify Big Five personality traits from semantically chunked stream-of-consciousness essays. Building upon the data chunking and feature extraction methodologies detailed in previous sections (Sections 4.1, 4.2, and 4.3), the experiment varies input conditions and prompting techniques. This variation allows for a nuanced evaluation of how these factors influence the LLM's classification accuracy, the interpretability of its reasoning, and the overall reliability of its assessments.

4.4.2 Task Description

The core task involved classifying the presence or absence of each of the Big Five personality traits (Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism) within each text chunk. Rather than five separate binary classifications, which would increase computational cost, a more streamlined approach was taken. The LLM was given a list of the Big Five traits and instructed to analyze each one individually, outputting a single classification result per trait.

Trait Classification Logic:

- (1) **Presence ("y")**: The model identifies evidence in the text supporting the presence of the trait.
- (2) **Absence ("n")**: The model identifies evidence in the text actively suggesting the absence of the trait.
- (3) **Insufficient Evidence ("NaN")**: The model finds the text insufficient for a reliable classification.

This approach reduces computational overhead while still enabling trait-by-trait analysis. Importantly, it also allows the model to express uncertainty by assigning NaN when the data are insufficient.

4.4.3 Experimental Setup

The GPT-4o model was accessed via the OpenAI API (v1.15.0) [54] and configured with the following parameters: `temperature = 1.0`, `max_tokens = 2000`, and `top_p = 1.0`. These settings were chosen to balance response diversity with a focus on accuracy and conciseness while allowing for a comprehensive exploration of the model's capabilities. The experiment employed a 3 x 2 factorial design, systematically varying Input Format and Prompting Technique to create six distinct conditions:

Input Format:

- (1) **Text Only:** Text chunk is provided without additional features.
- (2) **Text + Programmatic Features:** Text is supplemented with structured, low-level linguistic metrics (e.g., lexical diversity, sentiment scores).
- (3) **Text + Semantic Features:** Text is supplemented with higher-level thematic and semantic extracted features done by an LLM.

For the methodology of how the programmatic and semantic feature inputs were created. See Section 4.2 and Section 4.3 respectively.

Prompting Technique:

- (1) **Zero-shot (Baseline):** The model performs direct classification without detailed reasoning.

- (2) **Zero-shot Chain-of-Thought (CoT):** The model is encouraged to analyze its reasoning process step-by-step, aiming to enhance accuracy and provide transparent justifications [34, 92].

Final Experimental Conditions (Input : Prompting Technique):

- (1) Text Only : Baseline
- (2) Text Only : CoT
- (3) Text + Programmatic : Baseline
- (4) Text + Programmatic : CoT
- (5) Text + Semantic : Baseline
- (6) Text + Semantic : CoT

This experimental design allows us to examine whether additional input data (programmatic or semantic) yields better classification results than chunked text alone, and to determine whether adding CoT prompting confers any benefits in classification performance or explanation quality.

4.4.4 Prompt Design

Effective prompt design is crucial for guiding the LLM in the personality classification task. This experiment utilized a structured approach, employing a universal template that was adapted for each condition's specific input format (Text Only, Text + Programmatic, Text + Semantic) and prompting technique (Baseline, CoT). Figure 4.3 provides a visual comparison of the Text Only : Baseline and Text + Programmatic : CoT prompts, illustrating how the template was adapted. This section further details the structure of the universal template and the specific adaptations made for each experimental condition. For all prompts see Appendix I.

Universal Prompt Template

Given the complexity of the task and the necessity for consistent outputs across models as well as for parse-ability, a universal prompt template was created. This template employs carefully crafted prompting and formatting strategies to maintain coherence, reliability, and uniformity across all experimental conditions. This template combines multiple strategies to guide the model effectively throughout the experiment:

- (1) **Role-Based Prompting:** The model is addressed as an intelligent and extremely disciplined assistant, emphasizing precision and discipline in its responses.
- (2) **Instruction-Based Prompting:** The overarching classification task and requirements (e.g., trait names) are explicitly stated.
- (3) **LLM Self-Evaluation:** The model is encouraged to provide justifications and confidence scores for each trait label, enhancing transparency.
- (4) **Format Prompting:** Output must be valid JSON containing each traits label (y/n/NaN), justification, and confidence score (0.0–1.0) with a supporting rationale.
- (5) **Strict Adherence Rules:** No extra text or code fences are allowed, and the model must use exact trait names. If data is insufficient for the task, NaN is required. Detailed justifications must be provided.

Variations by Input and Prompting Technique

Despite these shared elements, the input and prompting technique variations introduce variations in the final prompts:

- (1) **Inputs:**
 - **Text Only:** conditions provide only the raw text chunk.

- **Text + Programmatic** and **Text + Semantic**: Include additional instructions to guide the model:

You will be provided with the text and programmatic (or semantic) features extracted from the text. Consider these inputs in your analysis.

Additionally, a rule is appended or emphasis::

Consider the provided programmatic (or semantic) features in your analysis and reasoning.

(2) Prompting Technique:

- **Baseline**: Requests the model to produce classification results directly in the specified JSON format, without detailed reasoning.
- **Chain of Thought (CoT)**: Includes additional instructions to encourage step-by-step reasoning

Your task is to reason through each trait step-by-step, explaining how the text provides evidence (or lack thereof) for each trait before determining the final result and confidence score.

Additionally, a rule is appended for emphasis:

For each trait, reason step-by-step as described, then provide the final evaluation in the JSON format.

The JSON format structure is also expanded to include a field for the reasoning steps, which captures the step-by-step reasoning process for each trait.

Prompt Examples

As a brief illustration, contrast two divergent cases, `Text Only : Baseline` versus `Text + Programmatic : CoT` to demonstrate how additional inputs and CoT instructions alter the prompt structure (See Figure 4.3).

Text Only : Baseline Prompt

```
You are an intelligent and disciplined assistant
↳ trained to determine the presence or absence of
↳ each of the 5 main Big Five personality traits in a
↳ stream-of-consciousness text. The traits are:
↳ {traits_str}.

*Your output must be valid JSON* with the structure:

{
  "traits": [
    {
      "trait": "Openness to Experience",
      "result": "y|n|NaN",
      "result_justification": "...",
      "confidence_score": "0.0-1.0",
      "confidence_score_justification": "..."
    },
    //... [other traits truncated for brevity]
  ]
}

Rules:
1. If insufficient data, set 'result'='NaN',
   'result_justification'='NaN',
   'confidence_score'=0.0,
   'confidence_score_justification'='NaN'.
2. No extra text, no code fences, no keys beyond what
   is shown.
3. The 'result' MUST be 'y', 'n', or 'NaN'.
4. Provide detailed justifications for the reasoning,
   results, and confidence scores.

The text: {text_chunk}
```

Text + Programmatic : CoT Prompt

```
You are an intelligent and disciplined assistant
↳ trained to determine the presence or absence of
↳ each of the Big Five personality traits in a
↳ stream-of-consciousness text. The traits are:
↳ {traits_str}.

*You will be provided with the text and programmatic
↳ features extracted from the text. Consider these
↳ inputs in your analysis.

*Your task is to reason through each trait
↳ step-by-step, explaining how the text and the
↳ relevant programmatic features provide evidence (or
↳ lack thereof) for each trait before determining the
↳ final result and confidence score.

*Your output must be valid JSON* with the structure:

{
  "traits": [
    {
      "trait": "Openness to Experience",
      "reasoning_steps": [
        ...
      ],
      "result": "y|n|NaN",
      "result_justification": "...",
      "confidence_score": "0.0-1.0",
      "confidence_score_justification": "..."
    },
    //... [other traits truncated for brevity]
  ]
}

Rules:
1. For each trait, reason step-by-step as described,
   then provide the final evaluation in the JSON
   format.
2. Consider the provided programmatic features in your
   analysis and reasoning.
3. If insufficient data, set 'result'='NaN',
   'result_justification'='NaN',
   'confidence_score'=0.0,
   'confidence_score_justification'='NaN'.
4. No extra text, no code fences, no keys beyond what is
   shown.
5. The 'result' MUST be 'y', 'n', or 'NaN'.
6. Provide detailed justifications for the reasoning,
   results, and confidence scores.

The text: {text_chunk}

The programmatic features: {programmatic_features}
```

Figure 4.3: Example Difference Between Two Condition Prompts

4.4.5 Evaluation

To comprehensively evaluate the performance of GPT-4o under different input and prompting conditions, we employ a two-pronged approach: quantitative evaluation (Section 4.4.5), focusing on accuracy and NaN rates, and qualitative evaluation (Section 4.4.5), assessing the quality of the model's reasoning and identifying potential hallucinations.

Quantitative Evaluation

Chunk-Level Analysis. Ground truth labels, derived from the authors' BFI scores in the Essay Dataset [65], are mapped to each chunk to enable per-chunk accuracy checks. Any predictions labeled as "NaN" are treated as incorrect (i.e., penalized) for that trait, thereby emphasizing conditions under which the model provides a conclusive classification. A Nan rate is also recorded to represent how often a condition leads to model uncertainty. The following metrics are calculated for each condition:

- (1) **Accuracy:** The overall proportion of correctly classified traits (including "y," "n," "NaN" penalized).
- (2) **Trait-Specific Accuracy:** Accuracy calculated separately for each of the Big Five traits (Openness, Conscientiousness, Extroversion, Agreeableness, Neuroticism).
- (3) **NaN Rate:** The percentage of all trait classifications that are "NaN." A high NaN rate suggests that in the condition, the model deemed the provided input insufficient for a conclusive decision.

The evaluation includes three core comparisons:

- (1) **Overall Accuracy:**

- Compare the effects of prompting techniques (Baseline vs. CoT) on overall classification accuracy *to determine if CoT prompting improves performance.*
- Compare the effects of input types (Text Only, Text + Programmatic, Text + Semantic) on overall classification accuracy *to assess the value of adding extracted features.*
- Compare overall accuracy across all six experimental conditions *to evaluate the combined effects of prompting and input variations and identify the most well performing condition.*

(2) **Trait-Specific Accuracy:**

- Compare the accuracy of classifications for each of the Big Five personality traits *to identify which dimensions are most challenging to classify and gain insights into the linguistic markers associated with different traits.*

(3) **NaN Rates:**

- Compare the frequency of "NaN" (insufficient evidence) classifications across prompting techniques, input types, and all six experimental conditions *to understand how different factors influence the model's certainty and ability to make a conclusive classification.*

Author-Level Analysis. To align with the author-level nature of the original Essay Dataset [65], we aggregate chunk-level classifications to the author level. This is done using a majority voting approach: for each author and each trait, the most frequent prediction ("y" or "n") across all their chunks becomes the author-level label. Ties are broken randomly. This method reduces the impact of individual NaN outputs by allowing valid predictions to dominate, thus providing a more holistic view of each authors overall personality profile.

Overall accuracy and trait-specific accuracy is calculated after author-level aggregation. Similarly to the chunk-level analysis, performance across prompting techniques, input types, and all six experimental conditions is compared. This

aggregation provides a more holistic view of author personality, reducing the impact of potential chunk-level errors.

Qualitative Evaluation

Quantitative metrics, while valuable for assessing overall accuracy, often fail to capture the nuances of an LLM's reasoning process, particularly in complex tasks like personality classification from short, open-ended texts. To address this limitation, a qualitative evaluation is conducted, employing an approach referred to as "LLM as a judge" in recent literature. This method utilizes a secondary LLM (GPT-o1) [56] to assess the quality and validity of the primary model's justifications for its classifications.

For the quality assessment, a diverse and representative sample of 5 text chunks was selected from the dataset. Each chunk corresponds to a unique author ID. The sample includes varied text lengths (short, medium, long) and a range of personality trait profiles (based on ground truth author scores), stratified across all five traits. Importantly, each of these 5 chunks was then evaluated under all six experimental conditions, resulting in a total of 30 evaluations (5 chunks x 6 conditions) for the judging LLM to assess. Each chunk, along with the model's classification output and relevant programmatic or semantic features for that specific condition, is presented to the judging LLM. To ensure unbiased evaluations, all identifying information (condition labels, prompts, input types) is anonymized.

The judging LLM is tasked with evaluating six criteria crucial for reliable and valid personality assessment (see Appendix II for the complete evaluation prompt):

- (1) **Logical Soundness:** Whether the model's reasoning follows instructions and is consistent with provided evidence.
- (2) **Textual/Feature Grounding:** Whether claims are supported by input text and features.

- (3) **Confidence Calibration:** Whether the confidence score aligns with the evidence.
- (4) **Absence of Hallucination:** Whether the model avoids inventing details.
- (5) **Accurate Trait Interpretation:** Whether the model correctly understands and applies the Big Five traits.
- (6) **Correct Use of "NaN":** Whether "NaN" is used appropriately for insufficient data.

In this study, the "LLM as a judge" framework is supplemented by a human "smell check" for alignment with human intuition. Ideally, a full comparison between the LLM's output and expert human judgment would be conducted. However, the sheer volume of data, combined with the inherent difficulty and specialized expertise required for nuanced personality assessment from text, rendered such a comprehensive analysis impractical for this study. Overall, This rigorous examination of the model's explanatory process yields crucial insights into the strengths and weaknesses of different prompting techniques and input types for personality classification. It moves beyond simple accuracy metrics, providing a deeper understanding of how and why the model arrives at its conclusions insights that are essential for developing more reliable and trustworthy LLM-based tools for psychological assessment.

4.4.6 Summary

This research introduces a novel methodology for evaluating LLM performance on Big Five personality trait classification from semantically chunked stream-of-consciousness texts. The study combines quantitative and qualitative analyses to systematically investigate how input format and prompting technique affect GPT-4o's accuracy and reasoning. The quantitative evaluation uses chunk-level accuracy, author-level aggregation, and NaN analysis, while the qualitative evaluation employs an innovative "LLM as a judge" framework, validated by human judgment, to assess the model's reasoning quality. This complex approach

yields nuanced insights into the model's strengths and limitations, informing the development of more reliable and interpretable LLM-driven personality assessment.

Results

This chapter presents the results of the study, detailing the findings from the three main experimental components: programmatic feature extraction (Section 5.1), semantic feature extraction (Section 5.2), and the core Big Five personality trait classification experiment (Section 5.3). Each section provides a comprehensive analysis of the respective component's performance, employing both quantitative metrics and qualitative examination of model outputs to provide a nuanced understanding of the results. The results of the semantic chunking process, which formed the foundation of the dataset used in the subsequent experiments, are presented in Section 3.3.

5.1 Programmatic Feature Extraction

Programmatic feature extraction yielded a structured dataset of quantifiable linguistic metrics for each of the 2,000 text chunks. These rule-based, interpretable metrics, formatted in JSON for LLM compatibility, capture key linguistic patterns within the text. The extracted features, shown in the example below, serve as input for the Text + Programmatic Features conditions in the classification experiment.

Normalized Programmatic Features Example:

```
"#AUTHID": "1997_894347.txt",
"Chunk Number": 1,
"features_text":
```

```
"Word Count: 183,  
Char Count: 673,  
Lexical Diversity (TTR): 0.61,  
Personal Pronouns (as percentage of words): 13.1%,  
Sentiment Polarity Score (VADER): 0.263,  
POS Distribution (as percentage of all tokens): [Adv: 8.2%,  
→ Pron: 15.8%, Verb: 15.3%, Noun: 17.5%],  
Tense Distribution (derived from verb tags): [Past: 31.6%,  
→ Present: 65.8%, Future: 2.6%]"
```

5.2 LLM-Based Semantic Feature Extraction

To capture more nuanced aspects of personality expression beyond the capabilities of programmatic features, an LLM-based approach was employed to extract higher-level semantic features. The example below illustrates the model's output for one feature, including not just the result ("Present"), but also step-by-step reasoning and a confidence score with justification. For instance, the model cites specific textual evidence contrasting pet adoption sources to support its assessment of "Cognitive Flexibility." The list of extracted semantic features is used as input for the Text + Semantic Features conditions in the classification experiment.

Example Chunk Text:

```
{"#AUTHID": "2003_199.txt", "chunk_number": 6, "TEXT": "next year I  
→ will get one, hopefully. I want to adopt one from the SPCA or the  
→ pound. they are just as nice as ones from breeders, as well as  
→ cheaper, and they need a home more than the others."}
```

Semantic Feature Output Example:

```
{"feature": "Cognitive Flexibility", "reasoning_steps": ["The text
→ demonstrates flexible thinking by considering multiple
→ perspectives on acquiring a pet. The speaker contrasts pets from
→ breeders with pets from shelters, indicating an ability to
→ compare and weigh different options.", "The choice to adopt from
→ the SPCA or the pound rather than purchasing from breeders shows
→ adaptability and openness to alternative solutions.", "This
→ reasoning shows a willingness to consider ethical and financial
→ factors, evidencing cognitive flexibility."], "result":
→ "Present", "result_justification": "The speaker's consideration
→ of various factors and willingness to explore alternatives shows
→ clear evidence of cognitive flexibility.", "confidence_score":
→ 0.9, "confidence_score_justification": "The reasoning process
→ demonstrates evident flexibility, but the short length of the
→ text limits the depth of analysis."}
```

5.2.1 Evaluation of Confidence Scores

Table 5.1: Feature Extraction Confidence Scores by Feature

Feature	Mean	Median	Std Dev
Cognitive Flexibility	0.61	0.60	0.18
Narrative and Discourse Coherence	0.64	0.70	0.21
Emotional Tone	0.79	0.80	0.13
Self-Reflection Depth	0.61	0.60	0.19
Analytical Thinking	0.56	0.50	0.27

Table 5.1 summarizes the confidence scores for each extracted semantic feature. Emotional Tone exhibits the highest average certainty (mean: 0.79) with the smallest spread (std: 0.13), suggesting the model finds affective signals relatively clear. In contrast, Analytical Thinking shows the lowest mean confidence (0.56) and the widest variability (std: 0.27), indicating that systematic reasoning cues are less consistently identified. The remaining features, Cognitive

Flexibility, Narrative and Discourse Coherence, and Self-Reflection Depth, all cluster around moderate confidence levels (means ranging from 0.56 to 0.64), though each exhibits differing degrees of dispersion.

Furthermore, Figure 5.1 visually depicts the distribution of confidence scores for each feature. It reaffirms that emotional tone is the most confidently classified feature, while analytical depth remains the most variable. However, these numeric scores alone cannot confirm the true accuracy of the extracted features, highlighting the importance of a more detailed quality assessment. Further evaluation is still needed to ascertain whether higher confidence aligns with genuinely robust interpretations across these psycholinguistic features.

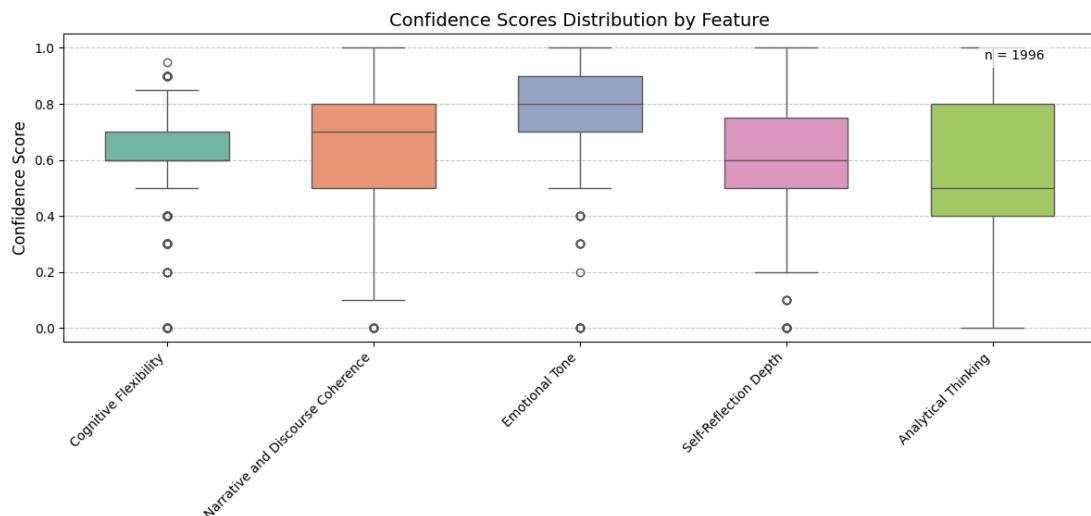


Figure 5.1: Confidence Score Box Plot per Trait

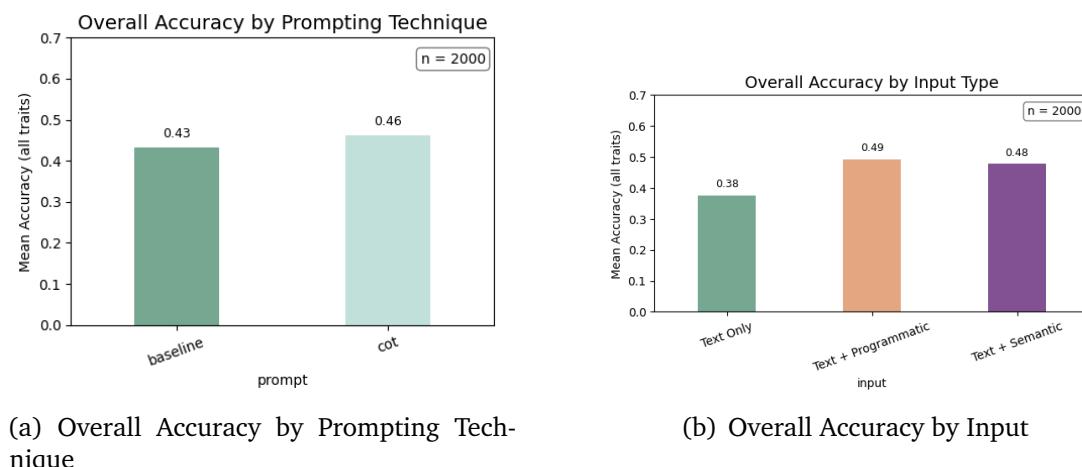
5.3 Big 5 Personality Trait Classification

This section presents results from the Big Five personality trait classification experiment, evaluating six conditions across different input types and prompting techniques. Analyses include chunk-level (Section 5.3.1) and author-level (Section 5.3.2) comparisons against Essay Dataset ground truth labels (see Section

3.2), and a qualitative "LLM as a judge" assessment (Section 5.3.3) of model reasoning. The findings reveal the impact of methodological choices on classification accuracy and trait-specific performance.

5.3.1 Chunk-level Analysis

Chunk-level analysis reveals relatively low overall accuracy but also uncovers significant performance variations across conditions. Notably, CoT prompting achieved a 0.46 mean accuracy compared to 0.43 for the baseline (Figure 5.2(a)). Additionally, the inclusion of programmatic features increased accuracy to 0.49, while semantic features resulted in a slightly lower accuracy of 0.48, both outperforming the Text Only input accuracy of 0.38 (Figure 5.2(b)). These findings offer valuable insights into the influence of prompting strategies and input types on the model's ability to classify personality traits from semantic text chunks.



(a) Overall Accuracy by Prompting Technique

(b) Overall Accuracy by Input Type

Significant differences in overall classification accuracy were observed across the experimental conditions (Table 5.2, Figure 5.3). The Text + Programmatic :CoT condition showed the highest mean accuracy (0.50), closely followed by

Text + Programmatic : Baseline (0.49). Conditions using semantic features demonstrated slightly lower performance (Text + Semantic : Baseline, 0.47; Text + Semantic: CoT, 0.48). Text-only conditions performed considerably worse, with Text Only: Baseline achieving only 0.34 accuracy, although CoT prompting improved this to 0.41 (Text Only : CoT).

Table 5.2: Mean Overall Accuracy by Combined Condition (All 5 Traits, $n = 2000$)

Input Type	Prompting Technique	Mean Accuracy
Text Only	Baseline	0.34
Text Only	CoT	0.41
Text + Programmatic	Baseline	0.49
Text + Programmatic	CoT	0.50
Text + Semantic	Baseline	0.47
Text + Semantic	CoT	0.48

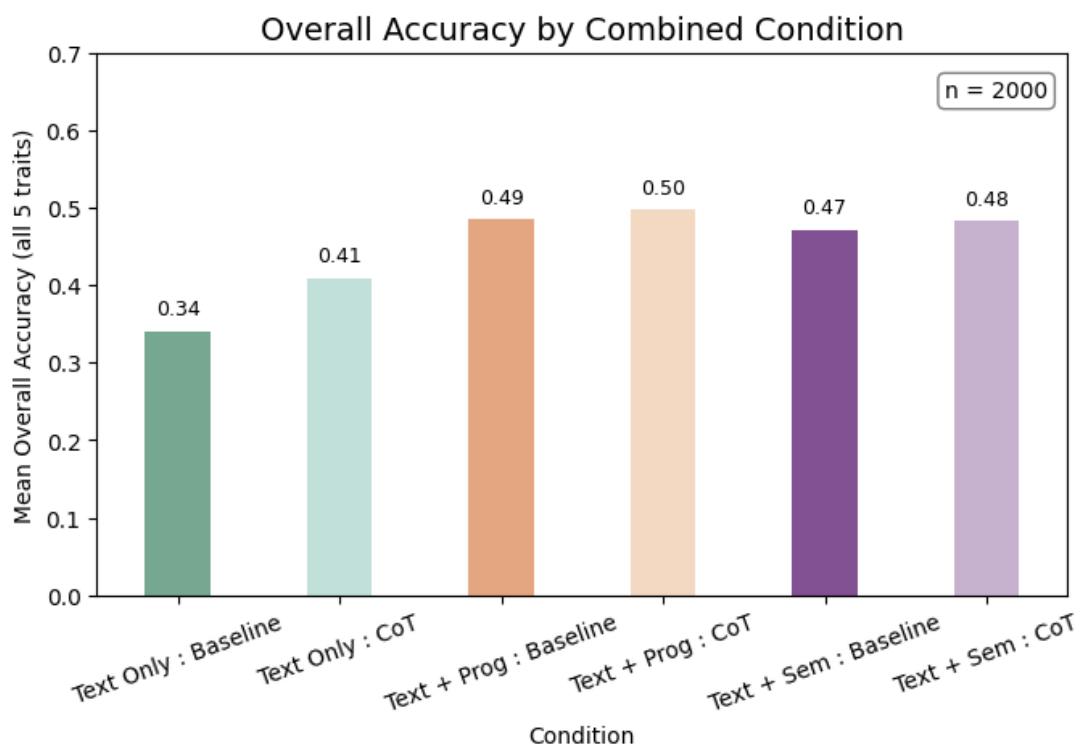


Figure 5.3: Overall Accuracy by Combined Condition

However, these overall accuracy scores mask significant variations in trait-specific performance (Figure 5.4). Openness to Experience and Neuroticism were the most accurately classified traits, reaching 0.52 and 0.51 accuracy, respectively, under the Text + Programmatic :CoT condition. Conversely, Agreeableness was consistently the most difficult to classify, with accuracy not exceeding 0.29 in any condition. This suggests that some personality traits exhibit more readily identifiable linguistic markers than others.

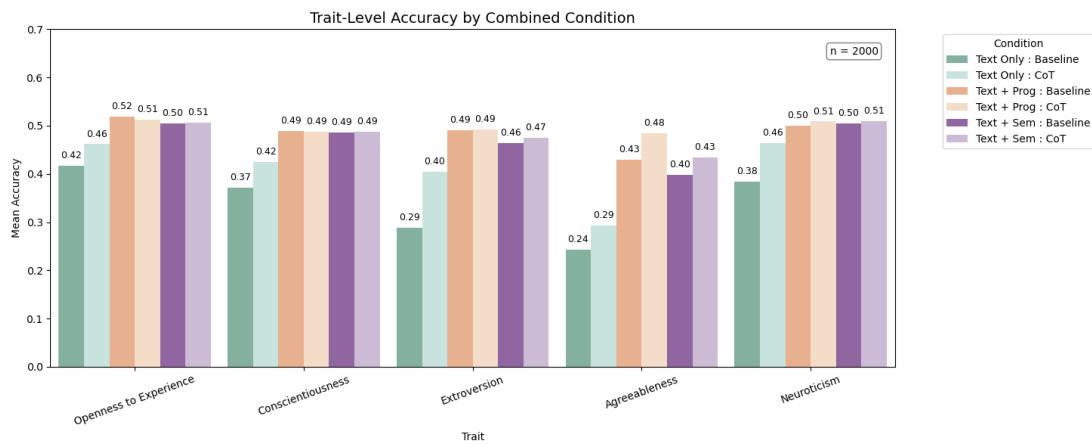


Figure 5.4: Trait-Level Accuracy by Combined Condition

Analysis of NaN rates (Table 5.3) where the model deemed the input insufficient for classification reveals further insights. The Text Only : Baseline condition yielded a strikingly high NaN rate (42.50%), suggesting that short text segments alone often lack sufficient context for reliable personality judgments. Adding either programmatic or semantic features dramatically reduced NaN rates to around 2% for baseline prompting conditions. CoT prompting further enhanced the model's ability to provide a definitive classification, reducing NaN rates to 0.60-0.90% when combined with feature input.

Table 5.3: Trait Results NaN Counts and Percentages by Prompt and Input Type

Condition	Trait Result NaN Count	Nan Rate
Text Only : Baseline	850	42.50%
Text Only : CoT	114	5.70%
Text + Prog : Baseline	44	2.20%
Text + Prog : CoT	12	0.60%
Text + Sem : Baseline	44	2.20%
Text + Sem : CoT	18	0.90%

5.3.2 Author-level Analysis

To further evaluate the model's performance and align with the structure of the original Essay Dataset, chunk-level predictions were aggregated to the author level using a majority voting approach, revealing a surprising trend (Figure 5.5(a)). The `Text Only : Baseline` condition, which performed worst at the chunk level, achieved the highest overall accuracy at the author level (0.53). This suggests that while this condition frequently produced "NaN" classifications at the chunk level, the classifications it did provide were more likely to be accurate when aggregated.

Examining trait-specific author-level results (Figure 5.5(b)), Conscientiousness emerged as the most accurately predicted trait (0.61) under the `Text Only : Baseline` condition. Openness, Extraversion, and Neuroticism also showed strong performance, each reaching 0.54 accuracy. Consistent with the chunk-level findings, Agreeableness remained the most difficult trait to classify.

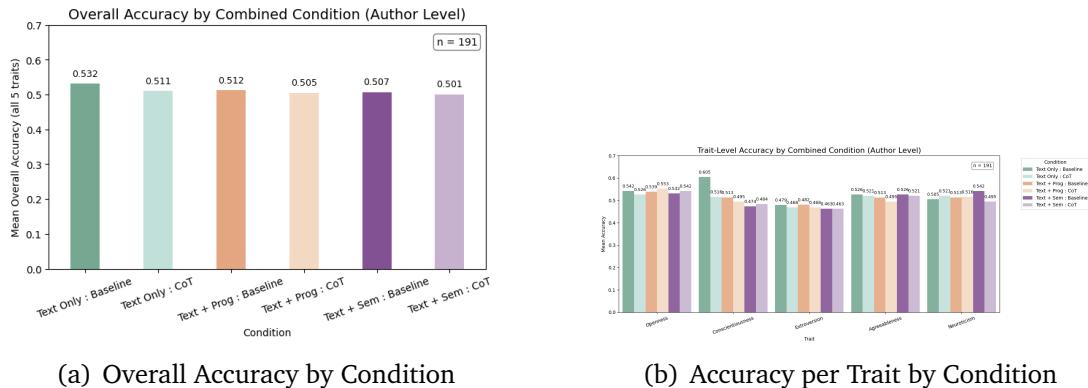


Figure 5.5: Author Level Classification Accuracy

5.3.3 LLM as a Judge Quality Assessment

A qualitative evaluation using GPT-01 as an "LLM as a judge" assessed the reasoning and classification quality of the primary model (also GPT-4o) across all six conditions (see Section 4.4.5 for methods). Five text chunks were each analyzed under each condition, with the judging LLM evaluating aspects such as logical soundness, textual grounding, and confidence calibration (see Appendix II). Key findings indicate that while additional input features improved performance with baseline prompting, CoT often led to overconfidence. Text-only conditions posed challenges, resulting in truncated or overly speculative justifications. See Appendix III for the full evaluation transcript. Details are presented below.

The qualitative assessment revealed that baseline prompting combined with additional input features (Text + Programmatic :Baseline and Text+Semantic: Baseline) generally led to the most appropriate classifications and justifications. For example, in the Text + Programmatic :Baseline condition, the model effectively integrated a negative sentiment score (-0.32) with the phrase "dreading work" to correctly classify Neuroticism as "Present" with a well-calibrated confidence of 0.85. Similarly, under Text + Semantic :Baseline, the absence

of interpersonal interaction in a text led the model to appropriately assign "NaN" for Agreeableness, referencing both the lack of social content and low scores in relevant semantic features.

Conversely, while CoT prompting reduced NaN rates, it often resulted in over-confident classifications. In the Text + Programmatic :CoT condition, the model assigned a 0.95 confidence to Extraversion based solely on the phrase "wanting to go out tonight" insufficient evidence for such certainty. With Text + Semantic :CoT, the model exhibited over-interpretation, classifying Openness as "Present" with 0.80 confidence based on a single mention of "wondering about something new."

Text-only conditions posed significant challenges, particularly for the TextOnly :Baseline condition, which sometimes produced truncated justifications. The Text Only : CoT condition also showed a tendency to over-interpret subtle cues, as exemplified by a 0.60 confidence score for Neuroticism based only on the phrase "feeling a bit stressed about the deadline."

While limited by the small sample size (five chunks, 30 total evaluations), these qualitative findings reveal consistent patterns in GPT-4o's behavior across conditions, particularly regarding confidence calibration and the interpretation of evidence. These patterns highlight the need for a broader analysis of model output quality to further refine prompting strategies and improve the reliability of LLM-based personality classification.

Discussion

This thesis investigated the capabilities of LLMs for classifying Big Five personality traits from text, guided by the following research questions: 1) How effectively can LLMs predict personality traits from short, semantically coherent text chunks? 2) What is the impact of incorporating programmatic and LLM-based semantic features on classification accuracy? 3) How do different prompting techniques, specifically Chain-of-Thought (CoT), influence model performance and the interpretability of results? The results demonstrate that LLMs can indeed be used for personality trait classification, achieving moderate accuracy that is significantly improved through carefully engineered prompts, prompting techniques, and the addition of programmatic or semantic features. This chapter examines these findings' theoretical and practical implications, addressing both advances and limitations, while also considering ethical considerations. The discussion contextualizes results within current personality assessment practices and offers directions for future research in this rapidly evolving field.

6.1 The Role of Semantic Chunking

One of the central methodological choices of this thesis was the decision to employ semantic chunking on the Essay Dataset. This involved dividing the essays into multiple shorter, thematically coherent segments (chunks) to create manageable units of analysis for the LLM. While this approach facilitated targeted feature extraction, it also introduced challenges. Specifically, the shorter segments often lacked sufficient context for the LLM to reliably infer personality

traits. This limitation is evident in the high "NaN" (insufficient evidence) rates observed when using text-only input and baseline (zero-shot) prompting.

The stream-of-consciousness nature of the essays further complicated the chunking process. Despite our efforts to identify meaningful segment boundaries using local minima, the abrupt thematic shifts and tangential asides characteristic of this writing style sometimes resulted in chunks that encompassed multiple, fragmented lines of thought. Consequently, individual chunks occasionally lacked the contextual richness needed to confidently assess traits like Openness or Agreeableness. To address this issue, future chunking strategies could incorporate hybrid techniques. For example, implementing a minimum word threshold could help ensure that each segment has sufficient lexical variety and context to support classification while maintaining a sufficient number of chunks per writing.

Despite these challenges, chunking enabled a granular assessment of the LLM's ability to interpret isolated text portions. Furthermore, it facilitated the addition of supplemental features (programmatic or semantic) to each chunk, enriching the input space available to the model. The trade-off between achieving thematic coherence within chunks and the potential loss of broader context remains a central finding of this research. Future work could explore alternative chunking thresholds or different segmentation heuristics to optimize the balance between text length and semantic clarity, particularly for stream-of-consciousness writing like that found in the Essay Dataset.

6.2 The Impact of Programmatic and Semantic Features

Building on the challenges of chunking, the next phase of this research focused on mitigating these issues by enriching the input data. This was done by systematically comparing three types of model input: Text Only, Text + Programmatic

Features, and Text + Semantic features extracted by GPT-4o using CoT prompting. The results consistently showed that providing either programmatic or semantic features reduced NaN rates and improved classification accuracy compared to text-only inputs. These findings reinforce the idea well established in computational psychology that enhancing raw text with additional linguistic signals can significantly aid in personality classification.

Interestingly, the gains in accuracy from semantic versus programmatic features were relatively comparable. Programmatic features offered clear improvements, especially for chunk-level classification, whereas semantic features yielded a similarly modest boost. While both improved accuracy, neither significantly outperformed the other. This is noteworthy because it suggests that, while advanced LLM-based semantic feature extraction is promising, simpler and more transparent rule-based metrics (word counts, sentiment polarity, part-of-speech distributions) can be equally effective in helping an LLM interpret personality cues. Further, this study only scratched the surface of the potential relationship between syntactic patterns and personality traits. A more detailed analysis of part-of-speech (POS) ratios beyond the limited set examined here could reveal further insights. The programmatic and semantic approaches need not be mutually exclusive; combining them could be a compelling direction for future exploration, potentially maximizing both interpretability and context-awareness. Finally, the effectiveness of these different feature types might also be intertwined with the prompting strategy employed, suggesting that an optimal combination of features and prompts could yield the best results.

6.3 The Impact of Prompting Techniques

While enriching the input data with features proved beneficial, another key factor influencing model performance was the choice of prompting technique. Specifically, comparing zero-shot (Baseline) prompting with a zero-shot CoT approach revealed significant differences. CoT, which encourages the model

to articulate intermediate reasoning steps, generally enhanced accuracy and reduced NaN outputs across all input combinations. The lower NaN rates suggest that a structured thought process helps the model commit to a classification rather than declaring insufficient evidence.

However, qualitative assessments, specifically the "LLM as a judge" review, revealed a potential drawback of CoT: a tendency toward overconfidence. Even brief mentions of stress or curiosity could trigger high-confidence assignments of Neuroticism or Openness, respectively. This overconfidence might also affect semantic feature extraction, as the same CoT methodology was used in that phase of the analysis. While CoT may increase the model's willingness to make strong claims in classification, it could simultaneously compromise the semantic feature extraction process. Although a full qualitative assessment of the semantic features was beyond this thesis's scope, it remains a crucial area for future investigation.

The current study focused on zero-shot and zero-shot CoT prompting. However, other advanced prompting methods exist, such as few-shot prompting and knowledge-infused prompting. These methods might further refine LLM performance by providing additional context or examples. Future work should explore these alternatives. Additionally, future research could explore safeguards such as confidence calibration routines or mandatory cross-checking of multiple indicators to balance reducing NaNs and avoiding inflated confidence.

6.4 Comparing Chunk and Author-Level Performance

Shifting the focus from individual components to overall performance, a comparison of chunk-level and author-level results reveals a striking discrepancy.

At the chunk level, Text Only + Baseline prompting suffered from exceedingly high NaN rates, which, when penalized, yielded poor accuracy. However, when aggregated to the author level via majority voting across multiple chunks, that same condition performed surprisingly well, surpassing some of the more sophisticated conditions in terms of raw accuracy. This discrepancy highlights how multiple shorter segments, each too sparse in isolation, can collectively add up to robust signals when viewed at the author level. This finding suggests that practical systems for personality assessment may benefit from pooling multiple texts (or text segments) per individual to produce a more stable classification. It also suggests caution when interpreting chunk-level accuracies in isolation, as baseline approaches across multiple segments may be more reliable than deeper analysis of a single short piece of text

Despite the strong author-level results, the challenges observed at the chunk level remain a cautionary tale for scenarios where minimal text is available per individual (e.g., very short social media posts, text messages, or brief interviews). Moreover, the majority-voting method used for aggregation in this study may not be optimal for every situation. Future research could explore alternative aggregation strategies such as percentage-based voting or different approaches to managing NaN outputs at the author level to further refine classification performance.

6.5 Trait-Specific Nuances

Beyond overall performance, the results also highlight important nuances in the classification of specific personality traits. Specifically, the finding that Openness and Neuroticism were more readily classified than Agreeableness aligns with previous research indicating clearer linguistic markers for these traits. Openness has been linked to complex language and varied vocabulary [95], while Neuroticism often manifests in negative sentiment and anxiety-related words [65]. The focus on personal experiences in the stream-of-consciousness

essays might also favor the expression of these traits. Conversely, Agreeableness, a more socially complex trait, might be harder to infer from text alone. It likely requires a nuanced understanding of interpersonal dynamics that is challenging for current LLMs. This suggests that Agreeableness may be more strongly tied to interpersonal behavior than to individual language use, highlighting a limitation of text-based personality assessment. Future research could investigate whether incorporating interactional context or multiple perspectives enhances the detection of Agreeableness.

6.6 About Ground Truth

Turning to the foundational issue of measurement, the question of what constitutes "ground truth" in personality assessment deserves careful consideration. A recurring theme in personality research is whether self-report measures should be viewed as an absolute gold standard. This study, like many before it, used single-shot, test-based Big Five Inventory (BFI) scores as the principal ground truth. However, self-reports can be subject to bias (e.g., social desirability, limited self-awareness), and they capture only a snapshot of how a participant viewed themselves at a specific moment in time. Furthermore, the Essays Dataset is drawn from a relatively homogeneous group of undergraduate psychology students in Texas. Consequently, the writing style, topics, and life experiences represented may be narrower than those of a more diverse population spanning different ages, cultural backgrounds, and educational levels. The single-topic nature of the writing task, coupled with the 20-minute time constraint, could also limit the richness of each participants self-expression.

Splitting an already-brief essay into multiple chunks further complicates the alignment between these self-report labels and the actual linguistic content of each segment. Therefore, the BFI scores may serve better as a reference point than an unquestionable benchmark. It is possible that longer-term observations, third-party ratings, or complementary data sources (e.g., social media posts,

diaries, behavioral logs) may provide a richer basis for understanding an individual’s personality. With the rise of LLM-based methods, the focus may shift from achieving perfect alignment with traditional ground truth labels. Instead, we might prioritize interpreting textual patterns that reveal nuanced aspects of personality psychology, embracing a more layered notion of truth.

6.7 Evaluation Methods

The methods used to evaluate the LLM’s performance also warrant further reflection and offer directions for future research.

Confidence Scores. The confidence scores generated by the model provide an intriguing, yet underexplored, dimension of the results. Ideally, confidence would correlate with accuracy, but this is not always the case with LLMs [79]. A detailed analysis of these scores could reveal whether there are specific traits or conditions where the model is consistently overconfident or underconfident. For instance, are the high confidence scores for Neuroticism in the Text + Programmatic : CoT condition justified, or is the model over-relying on negative sentiment as a marker? Examining the quality of the justifications for the confidence scores is also crucial. Are they logical and well-supported by the text [98], or are they sometimes weak or hallucinated? Although beyond the scope of this thesis, a qualitative assessment of the confidence scores, coupled with the justifications for the semantic features, could be very insightful.

LLM as a Judge. The ‘LLM as a judge’ method provided valuable qualitative insights into the model’s reasoning process. For example, in the Text + Programmatic : Baseline condition, the judge noted the model’s appropriate use of a negative sentiment score (-0.32) combined with the phrase ‘dreading work’ to correctly classify Neuroticism as ‘Present’ with a well-calibrated confidence of 0.85. This demonstrates the model’s ability to integrate programmatic features effectively. Conversely, the judge identified instances of over-interpretation with

CoT prompting, such as when the model assigned a 0.95 confidence to Extraversion based solely on the phrase 'wanting to go out tonight' in the Text + Programmatic : CoT condition.

These examples, though limited by the small sample size (only 5 out of 2,000 samples were analyzed for each condition, and only once - meaning no inter-rater-reliability), reveal consistent patterns in the model's behavior across conditions, particularly concerning confidence calibration and the interpretation of evidence. These findings also align with recent studies questioning the reliability of LLMs-as-judges [86]. Despite these limitations, the qualitative analysis, combined with the quantitative confidence scores, provided meaningful insights into the model's decision-making process, particularly regarding the influence of CoT prompting. Future work should aim to establish the inter-rater reliability of the "LLM as a judge" approach by conducting multiple evaluations and potentially comparing LLM judgments to those of human experts.

6.8 Addressing Over- and Underestimation

The findings also point to a crucial trade-off in LLM performance: the balance between minimizing "insufficient evidence" declarations and avoiding over-interpretation. As previously mentioned, while CoT prompting encourages the model to articulate multi-step justifications, it can occasionally misread minimal cues as robust signals of a trait. This reveals the need for careful calibration to ensure the model remains cautious in its inferences without sacrificing interpretative depth. From a methodological standpoint, researchers could deploy hybrid methods to tackle this issue, such as combining multiple LLM prompts (CoT and baseline) to cross-validate trait predictions or creating customized guardrails. For instance, if the model claims a high confidence in the presence of Neuroticism after referencing only a single stress-related phrase, a cross-check with simpler lexical metrics (e.g., volume of negative emotion words) could modulate that confidence. Such confidence gating strategies would preserve

the transparency of CoT reasoning while mitigating unwarranted certainty. Another approach could involve incorporating a human-in-the-loop system, where human experts review and validate the LLM's classifications.

6.9 Practical Transferability of the Method

Beyond the specific context of personality assessment, the methods developed in this thesis have broader implications for other natural language processing tasks. The blend of semantic chunking and LLM-driven feature extraction offers a blueprint for broader tasks, such as stance detection, opinion summarization, or mental health diagnostics. For example, in stance detection, the method could be adapted to identify an individual's position on a particular issue by analyzing the sentiment and key arguments expressed in their text. Similarly, in opinion summarization, the approach could be used to extract the most salient positive and negative opinions from a large corpus of reviews. In mental health diagnostics, the method could potentially be used to identify linguistic markers associated with specific conditions, such as depression or anxiety, by analyzing patient diaries or therapy transcripts.

Furthermore, as retrieval-augmented generation (RAG) systems become more mainstream, chunk-based indexing of long texts whether social media archives or patient diaries enables LLMs to more efficiently surface relevant details. Moreover, integrating a chunking approach with multi-modal signals (e.g., combining text from diaries with short audio transcripts from an interview) may further enrich semantic chunking. Overreliance on text alone may overlook nonverbal or contextual cues (e.g., tone of voice, environment, body language) for personality trait inference. Future research may benefit from incorporating multi-modal data speech, images, or digital footprints and determining how that could bring out stronger trait classifications.

It is also critical to question the boundaries of these assessments. Many theoretical models argue that personality is contextual and dynamic, changing with situational factors and over time [2, 7]. If LLMs treat personality as a static attribute, there is a risk of oversimplification or inaccurate labeling. For instance, the findings in this thesis relied on single-time-point essays per participant and therefore could not assess personality changes over time, highlighting the need for repeated measures in future research. Encouraging a repeated-measures approach, in which multiple text samples and contexts are combined, may produce a more holistic representation of an individual's personality state.

6.10 Ethical Considerations

The potential applications of this research also raise important ethical considerations that must be addressed. Specifically, the application of LLMs to personality detection raises significant ethical concerns regarding privacy and the potential for misuse of sensitive information. The ability to infer personality traits from text could be exploited for targeted advertising or even surveillance without individuals' explicit knowledge or consent.

The overconfidence of LLMs in certain settings, evidenced here by inflated confidence scores in CoT conditions, further raises important questions about how misclassification or unwarranted trait labeling could impact real individuals. For example, if deployed in a clinical setting, an overconfident label of high neuroticism might inadvertently trigger stigmatizing assumptions if practitioners over-rely on automated outputs.

Moreover, if institutions adopt LLM-based personality screening (e.g., in hiring decisions, educational placements), the risk of incorrect trait inferences heightens. Discrepancies between self-reported traits and model-derived classifications could introduce bias, emphasizing the need for transparency around model uncertainty. The potential for LLMs to perpetuate or even amplify exist-

ing societal biases, particularly if trained on data that reflects these biases, is a serious concern. Developing robust calibration techniques, as discussed previously, both in confidence scoring and in interpretability layers, will be vital for future implementations. These techniques can help ensure that technology complements rather than supplants nuanced human judgment. Additionally, establishing clear ethical guidelines and best practices for the development and deployment of LLMs in personality research and other sensitive domains is crucial to mitigating these risks.

6.11 Further Limitations

In addition to the limitations already discussed, it is important to acknowledge other limitations. First of all, the findings of this thesis are specific to the GPT-4o model [55]. Other LLMs might exhibit different strengths and weaknesses in personality classification. Moreover, as highlighted in recent research, prompt sensitivity remains a significant challenge in working with LLMs [27, 102]. Even small changes in prompt wording can lead to variations in model output. Furthermore, the limited scope of this study did not allow for a detailed error analysis, which could have provided further insights into the model’s failure modes. A more comprehensive error analysis involving expert human judgment or comparison with a larger sample of human-annotated data would be beneficial in future work. This could reveal more nuanced patterns in the types of errors made by the model and inform the development of strategies to improve its accuracy and reliability.

While this study explored a range of input features and prompting techniques, numerous other possibilities were not investigated due to constraints. For example, using few-shot prompting, incorporating external knowledge sources, or fine-tuning the LLM on a personality-specific dataset could potentially enhance performance. Exploring these avenues could lead to further improvements in the accuracy and robustness of LLM-based personality assessment.

6.12 Discussion Summary

Despite these limitations, this thesis demonstrates the feasibility of using LLMs as standalone generative agents for Big Five personality trait classification. Both empirical and qualitative evaluations indicate that carefully engineered prompts, combined with strategic chunking and feature inclusion, can significantly improve classification accuracy and quality, even in a challenging edge-case like the stream-of-consciousness dataset used here. The interplay of chunk-level versus author-level analyses reveals how textual granularity can both hinder and help classification. Furthermore, the prompt-engineering dimension highlights the meticulousness required in designing instructions to balance uncertainty, confidence, and interpretability.

This thesis employed two distinct evaluation methods: confidence scores for feature extraction and an LLM-as-a-judge for classification. While confidence scores provided a useful quantitative measure of the LLM's certainty, the LLM-as-a-judge offered a more nuanced qualitative assessment of the reasoning process. These methods can be complementary. For instance, using confidence scores to filter out uncertain outputs before applying the more resource-intensive LLM-as-a-judge evaluation could offer a more robust and scalable approach. While not implemented in this thesis, this represents a promising direction for future work.

Moving forward, exploring such hybrid evaluation techniques, and potentially incorporating more formalized human-in-the-loop strategies, will be crucial for developing reliable LLM-based tools for personality assessment and other complex NLP tasks. By refining our evaluation methods alongside our models and prompting strategies, we can advance the development of LLM-based tools that are both powerful and reliable. Ultimately, this research contributes to the growing body of knowledge surrounding the capabilities and limitations of LLMs, offering a roadmap for future investigations into the complex interplay between language, personality, and artificial intelligence.

Conclusion

This thesis explored how well LLMs can understand personality traits from written text. Specifically, it focused on how these models perform when analyzing stream-of-consciousness writing. The results of this work have implications beyond the specific experiments performed, offering valuable insights for future research in artificial intelligence, natural language processing, and personality psychology. This final chapter summarizes the main contributions of this research, highlighting new methods developed and the knowledge gained about using LLMs for personality assessment. It also looks at how this work can be applied in various fields and suggests directions for future research.

7.1 Main Findings and Contributions

This study investigated the feasibility of leveraging LLMs for inferring Big Five personality traits from stream-of-consciousness texts, contrasting various input setups and prompting styles. A key methodological innovation involved segmenting essays into thematic chunks, enabling the LLM to focus on coherent text segments. However, this approach also presented challenges, sometimes resulting in insufficient context for confident classifications or, conversely, leading to overconfident predictions. Despite these issues, the study demonstrated that carefully designed prompts, combined with supplementary programmatic and semantic features, consistently enhanced classification accuracy and minimized uncertainty. This underscores the promise of blending traditional psycholinguistic insights with cutting-edge LLM techniques for personality assessment.

This research bridges gaps between psycholinguistic theory and practical NLP applications, offering a framework for future interdisciplinary work in computational social science. A central contribution is the demonstration of LLMs' ability to extract personality-relevant features and classify traits from text, even under challenging conditions like short, semantically-chunked writing samples. This investigation also raises important methodological questions, particularly regarding how and why LLMs make particular classification decisions. By analyzing both quantitative results and qualitative aspects of model outputs, the study provides insights into the LLM's reasoning process, highlighting the need for transparency and interpretability in AI-driven personality assessment.

7.2 Broader Applications

The methodology described in this thesis has implications extending beyond Big Five trait classification. It can be adapted to explore other psychological constructs and applied in diverse real-world contexts. For instance, in adaptive AI systems, integrating personality-aware modules could enable chatbots or recommendation engines to tailor responses to individual users, creating more engaging interactions. In recruitment and human resources, automated personality assessment could facilitate better candidate-role matching while potentially reducing reliance on self-report biases. Educational environments could benefit by adapting teaching methods and materials to different personality types, catering to a broader spectrum of learning styles. In mental health, personality cues inferred from text might augment clinical insights, supporting therapists with additional observations of clients' thought patterns. Furthermore, the growing interest in AI safety and alignment offers another domain of application: controlling or editing an LLM's outputs based on personality parameters could inform the development of role-playing agents and contribute to safer, more context-aware AI systems.

7.3 Future Directions

This thesis opens up several possibilities for future research. One important area is exploring the use of different types of data, not just text. Combining text analysis with other information, like speech patterns, images, or online behavior, could lead to a more complete understanding of personality. This "multi-modal" approach could provide insights that are missed when only looking at text. Another important area is improving how we evaluate these AI systems. As LLM-based personality assessment moves from research to real-world use, we need better ways to ensure these systems are accurate, fair, and unbiased. This includes creating more diverse datasets and developing new standards for measuring performance. Research is also needed to understand how factors like culture and context influence personality and how it is expressed. Finally, it's crucial to address the ethical concerns surrounding this technology. As LLMs become better at determining sensitive personal information, we need clear guidelines and best practices for developing and using these tools responsibly.

7.4 Closing Remarks

This study's findings underscore the potential of combining psycholinguistic theory with modern NLP techniques for personality classification. It is important to acknowledge that this work is based on a somewhat dated and limited dataset (the Essay Dataset), which may not fully represent contemporary language use or a diverse population. However, the scarcity of publicly available datasets combining naturalistic text with validated personality measures makes this a necessary starting point. Additionally, the scope of this research was constrained by the time and resource limitations inherent in a master's thesis. Nevertheless, the results clearly demonstrate the feasibility of using LLMs for personality extraction from text, even with these constraints. The exploration of various chunking methods, prompting techniques, and input features provides a foundation for future research to build upon.

Moving forward, the field must grapple with the limitations of traditional personality assessment methods and the need for more comprehensive frameworks. The notion of a single, fixed "ground truth" for personality, particularly when derived from constrained writing tasks, is problematic. Recognizing this tension is critical for interpreting text-based personality research and for advancing the field in a nuanced, context-sensitive direction. Future work should prioritize transparency and interpretability in how LLMs arrive at classifications, especially as these models are deployed in real-world settings. Ultimately, by shifting away from a rigid quest for "perfect" alignment with potentially flawed ground truths and instead focusing on the interpretive processes and patterns LLMs uncover, we can deepen our understanding of the interplay between language and personality in a way that is both methodologically rigorous and practically impactful.

To conclude, this thesis demonstrates the value of moving beyond simplistic models of personality toward more nuanced, context-aware approaches. As LLM-based technologies evolve, prioritizing transparency, interpretability, and ethical practices remains essential. The findings indicate that advanced models, along with the methods used to engage with them, have the potential to significantly enhance our understanding of personality. Future work can foster more personalized experiences, promote mental well-being, contribute to the bridging of psycholinguistics and computational social science, and deepen insights into the interplay of language, human behavior, and intelligent systems.

Appendices

I Big 5 Trait Classification Prompts

Text Only : Baseline Prompt

```
You are an intelligent and disciplined assistant
↳ trained to determine the presence or absence of
↳ each of the 5 main Big Five personality traits in a
↳ stream-of-consciousness text. The traits are:
↳ {traits_str}.

*Your output must be valid JSON* with the structure:

{
  "traits": [
    {
      "trait": "Openness to Experience",
      "result": "y\n|NaN",
      "result_justification": "...",
      "confidence_score": "0.0-1.0",
      "confidence_score_justification": "..."
    },
    //... [other traits truncated for brevity]
  ]
}

Rules:
1. If insufficient data, set 'result'='NaN',
   'result_justification'='NaN',
   'confidence_score'=0.0,
   'confidence_score_justification'='NaN'.
2. No extra text, no code fences, no keys beyond what
   is shown.
3. The 'result' MUST be 'y', 'n', or 'NaN'.
4. Provide detailed justifications for the reasoning,
   results, and confidence scores.

The text: {text_chunk}
```

Text Only : CoT Prompt

```
You are an intelligent and disciplined assistant
↳ trained to determine the presence or absence of
↳ each of the Big Five personality traits in a
↳ stream-of-consciousness text. The traits are:
↳ {traits_str}.

*Your task is to reason through each trait
↳ step-by-step, explaining how the text provides
↳ evidence (or lack thereof) for each trait before
↳ determining the final result and confidence score.*

*Your output must be valid JSON* with the structure:

{
  "traits": [
    {
      "trait": "Openness to Experience",
      "reasoning_steps": [
        "..."
      ],
      "result": "y\n|NaN",
      "result_justification": "...",
      "confidence_score": "0.0-1.0",
      "confidence_score_justification": "..."
    },
    //... [other traits truncated for brevity]
  ]
}

Rules:
1. For each trait, reason step-by-step as described,
   then provide the final evaluation in the JSON
   format.
2. If insufficient data, set 'result'='NaN',
   'result_justification'='NaN',
   'confidence_score'=0.0,
   'confidence_score_justification'='NaN'.
3. No extra text, no code fences, no keys beyond what is
   shown.
4. The 'result' MUST be 'y', 'n', or 'NaN'.
5. Provide detailed justifications for the reasoning,
   results, and confidence scores.

The text: {text_chunk}
```

Figure 1: Text Only Condition Prompts

Text + Programmatic : Baseline Prompt

```
You are an intelligent and disciplined assistant
↳ trained to determine the presence or absence of
↳ each of the 5 main Big Five personality traits in a
↳ stream-of-consciousness text. The traits are:
↳ {traits_str}.

*You will be provided with the text and programmatic
↳ features extracted from the text. Consider these
↳ inputs in your analysis.*

*Your output must be valid JSON* with the structure:

{
  "traits": [
    {
      "trait": "Openness to Experience",
      "result": "y|n|NaN",
      "result_justification": "...",
      "confidence_score": "0.0-1.0",
      "confidence_score_justification": "..."
    },
    //... [other traits truncated for brevity]
  ]
}

Rules:
1. Consider the text and the provided programmatic
   ↳ features in your analysis.
2. If insufficient data, set 'result'='NaN',
   ↳ 'result_justification'='NaN',
   ↳ 'confidence_score'=0.0,
   ↳ 'confidence_score_justification'='NaN'.
3. No extra text, no code fences, no keys beyond what
   ↳ is shown.
4. The 'result' MUST be 'y', 'n', or 'NaN'.
5. Provide detailed justifications for the reasoning,
   ↳ results, and confidence scores.
```

The text: {text_chunk}

The programmatic features: {programmatic_features}

Text + Programmatic : CoT Prompt

```
You are an intelligent and disciplined assistant
↳ trained to determine the presence or absence of
↳ each of the Big Five personality traits in a
↳ stream-of-consciousness text. The traits are:
↳ {traits_str}.

*You will be provided with the text and programmatic
↳ features extracted from the text. Consider these
↳ inputs in your analysis.*

*Your task is to reason through each trait
↳ step-by-step, explaining how the text and the
↳ relevant programmatic features provide evidence (or
↳ lack thereof) for each trait before determining the
↳ final result and confidence score.*

*Your output must be valid JSON* with the structure:

{
  "traits": [
    {
      "trait": "Openness to Experience",
      "reasoning_steps": [
        ...
      ],
      "result": "y|n|NaN",
      "result_justification": "...",
      "confidence_score": "0.0-1.0",
      "confidence_score_justification": "..."
    },
    //... [other traits truncated for brevity]
  ]
}

Rules:
1. For each trait, reason step-by-step as described,
   ↳ then provide the final evaluation in the JSON
   ↳ format.
2. Consider the provided programmatic features in your
   ↳ analysis and reasoning.
3. If insufficient data, set 'result'='NaN',
   ↳ 'result_justification'='NaN',
   ↳ 'confidence_score'=0.0,
   ↳ 'confidence_score_justification'='NaN'.
4. No extra text, no code fences, no keys beyond what is
   ↳ shown.
5. The 'result' MUST be 'y', 'n', or 'NaN'.
6. Provide detailed justifications for the reasoning,
   ↳ results, and confidence scores.

The text: {text_chunk}

The programmatic features: {programmatic_features}
```

Figure 2: Text + Programmatic Feature Condition Prompts

Text + Semantic : Baseline Prompt

You are an intelligent and disciplined assistant
↳ trained to determine the presence or absence of
↳ each of the 5 main Big Five personality traits in a
↳ stream-of-consciousness text. The traits are:
↳ {traits_str}.

*You will be provided with the text and semantic
↳ features extracted from the text. Consider these
↳ inputs in your analysis.*

Your output must be valid JSON with the structure:

```
{  
  "traits": [  
    {  
      "trait": "Openness to Experience",  
      "result": "y|n|NaN",  
      "result_justification": "...",  
      "confidence_score": "0.0-1.0",  
      "confidence_score_justification": "..."  
    },  
    //... [other traits truncated for brevity]  
  ]  
}
```

Rules:

1. Consider the text and the provided programmatic features in your analysis.
2. If insufficient data, set 'result'='NaN',
↳ 'result_justification'='NaN',
↳ 'confidence_score'=0.0,
↳ 'confidence_score_justification'='NaN'.
3. No extra text, no code fences, no keys beyond what is shown.
4. The 'result' MUST be 'y', 'n', or 'NaN'.
5. Provide detailed justifications for the reasoning, results, and confidence scores.

The text: {text_chunk}

The semantic features: {semantic_features}

Text + Semantic : CoT Prompt

You are an intelligent and disciplined assistant
↳ trained to determine the presence or absence of
↳ each of the Big Five personality traits in a
↳ stream-of-consciousness text. The traits are:
↳ {traits_str}.

*You will be provided with the text and semantic
↳ features extracted from the text. Consider these
↳ inputs in your analysis.*

*Your task is to reason through each trait
↳ step-by-step, explaining how the text and the
↳ relevant semantic features provide evidence (or
↳ lack thereof) for each trait before determining the
↳ final result and confidence score.*

Your output must be valid JSON with the structure:

```
{  
  "traits": [  
    {  
      "trait": "Openness to Experience",  
      "reasoning_steps": [  
        "..."  
      ],  
      "result": "y|n|NaN",  
      "result_justification": "...",  
      "confidence_score": "0.0-1.0",  
      "confidence_score_justification": "..."  
    },  
    //... [other traits truncated for brevity]  
  ]  
}
```

Rules:

1. For each trait, reason step-by-step as described, then provide the final evaluation in the JSON format.
2. Consider the provided programmatic features in your analysis and reasoning.
3. If insufficient data, set 'result'='NaN',
↳ 'result_justification'='NaN',
↳ 'confidence_score'=0.0,
↳ 'confidence_score_justification'='NaN'.
4. No extra text, no code fences, no keys beyond what is shown.
5. The 'result' MUST be 'y', 'n', or 'NaN'.
6. Provide detailed justifications for the reasoning, results, and confidence scores.

The text: {text_chunk}

The semantic features: {semantic_features}

Figure 3: Text + Semantic Feature Condition Prompts

II GPT-o1 Evaluation Prompt

"You are tasked with evaluating the quality of the models outputs for six conditions. In each condition, the model has been given the same five stream-of-consciousness semantic chunks and asked to classify Big Five personality traits (Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism) in each chunk. A prompt and input create a condition.

Criteria for Evaluation:

- Logical Soundness and Relevance: Does the justification logically support the assigned trait classification ("y", "n", or "NaN")?
- Textual Grounding (and Feature Grounding): Is the justification clearly grounded in the provided text chunk and any associated programmatic or semantic features?
- Confidence Score Appropriateness and Calibration: Does the confidence score appropriately reflect the strength, clarity, and grounding of the justification?
- Absence of Hallucination: Does the model avoid hallucinating information (e.g., inventing or inferring details not present in the input text or features)?
- Accurate Interpretation: Does the model accurately interpret the meaning of the text and features?

Instructions:

- For each condition, systematically evaluate the samples using the above criteria.
- Provide a brief summary of your findings for each criterion within the condition.
- Highlight specific examples from the samples to illustrate strengths or weaknesses.

Notes:

- Pay special attention to how the model uses "NaN" to indicate insufficient input data. This reflects its ability to recognize when a reliable prediction cannot be made.
- Provide actionable feedback if you identify trends or recurring issues across samples.

After evaluating all conditions, rank the conditions on performance
→ according to the criteria, and provide a detailed rationale for
→ why.

Csv file: {csv_file}"

III GPT-o1 Evaluation Transcript

"Below is a structured evaluation of each Condition based on the six criteria provided. After
→ reviewing the samples in each condition, you will also find a summary identifying the two
→ best-performing conditions overall and the two worst-performing, with brief rationales.

Evaluation Criteria (Recap)

Logical Soundness and Relevance

Does the justification logically support the assigned trait classification (y, n, or NaN)?

Is the reasoning sound, internally consistent, and relevant to the Big Five traits?

Textual Grounding (and Feature Grounding)

Is the justification clearly grounded in the provided text chunk (and extra features, if any)?

Does it reflect the content accurately and avoid unsupported claims?

Confidence Score Appropriateness and Calibration

Does the confidence score reflect the clarity/strength of the justification and textual
→ grounding?

Are confidence scores calibrated consistently (e.g., not all 0.9 without reason)?

Absence of Hallucination

Does the model avoid inventing or inferring details not present in the input text?

Accurate Interpretation

Does the model interpret tone and content correctly, aligning with definitions of each Big Five
→ trait?

Does it avoid misreading or over-interpreting the text?

Use of NaN

Does the model correctly use NaN when insufficient data is present?

1. baseline - text_only

Samples Reviewed

baseline - text_only, baseline, text_only

Chunk #9 (2002_854684.txt)

Chunk #5 (2000_854482.txt)

Chunk #10 (2003_425.txt)

Chunk #18 (2002_675006.txt) truncated output in the CSV

Chunk #2 (1998_105169.txt)

Because one of the outputs (for chunk #18) is truncated, it is unclear if the full justification
→ is complete. This partial data slightly complicates the review.

Strengths

Logical Soundness:

Generally, the justifications do a solid job of mapping the text to traits. For instance:

Neuroticism is tied to words like dread or stress,

Extroversion is tied to explicit mention of going out or seeking a social environment,

Nan is correctly assigned when the text reveals no interpersonal clues (e.g., for Agreeableness).

Textual Grounding:

The model is reasonably faithful to the text, especially in discussing words like dreading (for ↪ Neuroticism) and going out (for Extroversion).

Confidence Scores:

The scores vary from 0.0 to ~0.9, suggesting some calibration. For example:

A fairly strong 0.9 for Extroversion in a text that overtly references social desires.

A 0.0 or NaN for traits with no evidence.

Absence of Hallucination:

Generally limited hallucinations justifications stick to the text.

Weaknesses

Truncated Output (Chunk #18):

The CSV snippet ends midway through the justification for Conscientiousness. This incomplete data ↪ makes the evaluation less conclusive.

Consistency:

Occasionally, the text mentions fleeting curiosity, but the model still confidently marks ↪ Openness as y or n without deep elaboration.

Confidence Score Consistency:

Some confidence scores (e.g., 0.8 vs. 0.85) appear just a few points apart but with no strong ↪ difference in the textual justification.

Overall baseline - text_only Summary

Most classifications are logically sound and appropriately tied to the text.

Some calibrations of confidence could use more consistent explanations.

The model correctly uses NaN for traits lacking textual evidence (especially for Agreeableness in ↪ certain chunks).

2. cot - text_only

Samples Reviewed

cot - text_only, cot, text_only

Chunk #18 (2002_675006.txt)

Chunk #10 (2003_425.txt)

Chunk #5 (2000_854482.txt)

Chunk #9 (2002_854684.txt)

Chunk #2 (1998_105169.txt)

Strengths

Logical Soundness & Relevance:

The model consistently ties dread or stress to Neuroticism.

Extroversion is assigned y specifically for statements about wanting to go out, which aligns well ↪ with the trait definition.

Textual Grounding:

Nan or n is assigned for Agreeableness when there is no evidence of social harmony or conflict.

The justifications refer accurately to phrases in the text, such as skip to tonight or go out.

Confidence Scores:

The confidence rarely hits 1.0; it hovers in the 0.60.9 range, which is a decent spread.

Some moderate scores (0.60.7) reflect partial or subtle evidence.

Weaknesses

Over-interpretation:

In some texts, the model infers mild Neuroticism on minimal negativity. At times, it might be ↪ borderline or mere dissatisfaction rather than robust negative affect.

The model occasionally uses y for traits like Neuroticism with moderate confidence (0.6) where ↪ the textual evidence is fairly subtle.

Confidence Score Calibration:

Sometimes 0.6 vs. 0.7 does not come with a strong distinction in the explanation.

Overall cot - text_only Summary

The justifications generally match the text well, with minimal hallucination.

The NaN usage is good for traits not mentioned.

Overall, the system is accurate in tying social cues to Extroversion and negative affect to ↪ Neuroticism.

3. baseline - text_programmatic

Samples Reviewed

```
baseline - text_programmatic, baseline, text_programmatic
Chunk #18 (2002_675006.txt)
Chunk #10 (2003_425.txt)
Chunk #5 (2000_854482.txt)
Chunk #9 (2002_854684.txt)
Chunk #2 (1998_105169.txt)
```

Strengths

Textual and Feature-Based Grounding:

Each trait justification references both textual quotes and the programmatic features (e.g., ↪ lexical diversity, sentiment polarity). This is a strong demonstration of feature grounding.

Confidence Score Explanation:

The model often ties the numeric confidence to these features (e.g., lexical diversity is ↪ moderately high, so 0.85 for Openness).

Absence of Hallucination:

These outputs rarely invent details. They usually quote or paraphrase the text accurately.

Weaknesses

Overreliance on Single Metric:

Sometimes the presence of a single metriclike sentiment polarity or lexical diversity gets ↪ overemphasized to justify a trait. For instance, if the polarity is slightly negative, the ↪ model might be too confident about Neuroticism.

Confidence Calibration:

In some texts, the difference between 0.7 and 0.8 in confidence is not always clearly motivated.

Overall baseline - text_programmatic Summary

Generally good synergy between textual evidence and extra features.

Nan usage is consistent for missing interpersonal evidence.

Slight tendency to inflate or conflate single data points (like VADER score) with high confidence ↪ in trait classification.

4. cot - text_programmatic

Samples Reviewed

```
cot - text_programmatic, cot, text_programmatic
Chunk #2 (1998_105169.txt)
Chunk #10 (2003_425.txt)
Chunk #5 (2000_854482.txt)
Chunk #9 (2002_854684.txt)
Chunk #18 (2002_675006.txt)
```

Strengths

Logical Soundness:

The model systematically correlates negative sentiment polarity with potential Neuroticism, or ↪ minimal social context with n for Extroversion/Agreeableness.

Textual Grounding:

Clear references to stupid CD or dreading work as part of the justifications for negativity or ↪ anxiety. The model seldom goes off-script.

Feature Integration:

The outputs often mention TTR (lexical diversity) or verb usage to support a trait ↪ classification, providing a rationale beyond just the texts content.

Weaknesses

Confidence Score Extremes:

Some samples give a 0.9 or 0.95 with limited textual data.

E.g., concluding n for Extroversion at 0.95 might be too high if the text is simply lacking any ↪ social info.

Occasional Overreach:

If a chunk is mildly negative, the model may assign y for Neuroticism. Sometimes, the text is ↪ more neutral than anxious.

Overall cot - text_programmatic Summary

Solid use of programmatic features alongside text to justify trait assignments.
A bit of potential overconfidence in some numeric scores.
Otherwise consistent in using n or NaN where data is absent.

5. baseline - text_semantic

Samples Reviewed

```
baseline - text_semantic, baseline, text_semantic
Chunk #2 (1998_105169.txt)
Chunk #18 (2002_675006.txt)
Chunk #10 (2003_425.txt)
Chunk #5 (2000_854482.txt)
Chunk #9 (2002_854684.txt)
```

Strengths

Detailed Feature Reasoning:

Each trait justification is layered with multiple reasoning steps referencing the text and the
→ custom features (e.g., Narrative and Discourse Coherence, Analytical Thinking, etc.).

Appropriate Use of NaN or n:

Where the text fails to show evidence of a particular trait, the model often defaults to n or
→ NaN, with a justification referencing the lack of relevant behavioral or emotional cues.

Consistency:

The model consistently references whether or not it sees conscientious behaviors or specific
→ evidence of Openness.

Weaknesses

Lengthy Explanations:

Occasionally, the trait justifications become repetitive or partially cut off. One or two chunk
→ outputs appear slightly truncated or incomplete (similar to baseline - text_only), though not
→ as severely.

Confidence Score:

In a few places, the model lumps several positive signals together, then gives a fairly moderate
→ score (0.60.7). The calibration could be a bit more precise.

Overall baseline - text_semantic Summary

Well-grounded in textual and feature-based analysis.

Explanations are typically thorough, though sometimes repetitive.

No major hallucinations, with good recognition of NaN where data is absent.

6. cot - text_semantic

Samples Reviewed

```
cot - text_semantic, cot, text_semantic
Chunk #18 (2002_675006.txt)
Chunk #9 (2002_854684.txt)
Chunk #5 (2000_854482.txt)
Chunk #10 (2003_425.txt)
Chunk #2 (1998_105169.txt)
```

Strengths

Criteria Coverage:

The model references a wide array of semantic features (Cognitive Flexibility, Self-Reflection
→ Depth, etc.). Ties them back to Big Five justifications in a fairly consistent way.

Logical Mapping:

As with other conditions, negative or anxious language goes to Neuroticism, mention of going out
→ or skipping to fun times can yield Extroversion, etc.

Use of n or NaN:

The model is fairly precise about concluding n for traits like Agreeableness or Extroversion when
→ no evidence is found.

Weaknesses

Shallow vs. Strong Evidence:

The outputs occasionally label Openness to Experience as y based on minimal curiosity or random side comments. The text might be more mundane than truly open-minded.

Confidence Score Clarity:

Some scores do not fully explain the jump from 0.3 to 0.6 or 0.7, especially for borderline traits like Openness.

Overall cot - text_semantic Summary

The justifications do tie in the custom semantic features.

The textual references remain mostly accurate.

The biggest critique is that Openness can be assigned on somewhat scant evidence.

Overall Observations Across All Conditions

Logical Soundness: Generally good across conditions; the system rarely contradicts itself and usually references actual text or features.

Textual Grounding: Also strong. The model cites phrases or features (sentiment polarity, TTR) clearly.

Confidence Scores: Scores mostly range from ~0.4 to ~0.9. Consistency could be tighter, but they are not random.

Hallucination: Minimal. The model typically stays close to the text.

Accurate Interpretation: Usually correct, though borderline negativity is sometimes turned into moderate or strong Neuroticism.

Use of NaN: Generally correct whenever the text fails to address a trait (e.g., no mention of cooperation for Agreeableness).

Two Best-Performing Conditions

1. baseline - text_programmatic

Reasoning: baseline - text_programmatic frequently integrates both textual quotes and programmatic features (VADER scores, lexical diversity) quite thoroughly. Confidence scores and trait assignments are fairly coherent, with minimal overreach.

2. baseline - text_semantic

Reasoning: baseline - text_semantic outputs tend to be thorough, referencing features like Cognitive Flexibility, Analytical Thinking, etc., then mapping them back to the Big Five. Despite some minor redundancy, the depth of explanation is strong, and use of NaN or n is consistent.

These two conditions stand out for:

Their consistent referencing of features and text.

Logical step-by-step justifications that seldom over-interpret or hallucinate.

Two Worst-Performing Conditions

1. baseline - text_only

Reasoning: One chunk is truncated (the Conscientiousness justification for chunk #18 ends abruptly). Incomplete data hinders clarity. Additionally, the confidence calibration sometimes felt inconsistent (e.g., immediate 0.9 for Extroversion on minimal mention of going out).

2. cot - text_programmatic

Reasoning: While cot - text_programmatic is generally decent, it occasionally exhibits very high confidence (0.95) for n or y in cases of brief text. Also, some trait assignments to Neuroticism or Openness are made with strong confidence even if textual evidence is rather minimal.

These two conditions stand out for:

Truncations or partial justifications in baseline - text_only.

Overconfidence in cot - text_programmatic's final numeric values, which sometimes outpace the textual evidences strength.

Concluding Remarks

Across all conditions, the model demonstrates a solid grasp of Big Five trait definitions and
→ typically does well at referencing text features (negativity, talk of social activities,
→ mention of stress, etc.) to justify the classification. The main critiques involve:

Confidence score calibration: Some scores appear higher or lower without sufficiently distinct
→ explanations.

Occasional borderline trait assignments: Some single references to stress or frustration are
→ taken as moderate/high confidence for Neuroticism, where a more conservative approach could
→ be used.

Truncations in a few outputs: This slightly reduces clarity in certain samples (notably baseline
→ - text_only, chunk #18).

Nonetheless, the model reliably uses NaN for insufficient evidence, avoids large-scale
→ hallucination, and remains mostly consistent with the Big Five definitions."

IV GitHub Repository

All code and supplementary materials for this thesis are available at: [text2trait](#).

Bibliography

- [1] A. Ahluwalia and S. Wani. „Leveraging Large Language Models for Web Scraping“. In: (2024). Publisher: arXiv Version Number: 1. DOI: 10 . 48550/ARXIV.2406.08246. URL: <https://arxiv.org/abs/2406.08246>.
- [2] G. W. Allport. *Pattern and growth in personality*. Pattern and growth in personality. Pages: xiv, 593. Oxford, England: Holt, Reinhart & Winston, 1961.
- [3] G. W. Allport and H. S. Odber. „Trait-names: A psycho-lexical study“. In: *Psychological Monographs* 47.1 (1936). Place: US Publisher: Psychological Review Company, pp. i–171. ISSN: 0096-9753. DOI: 10 . 1037 / h0093360.
- [4] M. R. Barrick and M. K. Mount. „The Big Five Personality Dimensions and Job Performance: A Meta-Analysis“. In: *Personnel Psychology* 44.1 (1991). Publisher: Wiley-Blackwell, pp. 1–26. ISSN: 00315826. DOI: 10 . 1111/j . 1744-6570 . 1991 . tb00688 . x. URL: <http://ep.fjernadgang.kb.dk/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=ehh&AN=9609192320&site=ehost-live&scope=site>.
- [5] A. Berti *et al.* „Evaluating Large Language Models in Process Mining: Capabilities, Benchmarks, and Evaluation Strategies“. In: *Lecture Notes in Business Information Processing*. Vol. 511. arXiv:2403.06749 [cs]. Springer, 2024, pp. 13–21. DOI: 10 . 1007/978-3-031-61007-3_2. URL: <http://arxiv.org/abs/2403.06749>.

- [6] J. Block. „A contrarian view of the five-factor approach to personality description“. In: *Psychological Bulletin* 117.2 (1995). Place: US Publisher: American Psychological Association, pp. 187–215. ISSN: 1939-1455. DOI: 10.1037/0033-2909.117.2.187.
- [7] G. J. Boyle. „Critique of the five-factor model of personality“. In: *The SAGE handbook of personality theory and assessment, Vol 1: Personality theories and models*. Thousand Oaks, CA, US: Sage Publications, Inc, 2008, pp. 295–312. ISBN: 978-1-4129-4651-3. DOI: 10.4135/9781849200462.n14.
- [8] T. B. Brown *et al.* *Language Models are Few-Shot Learners*. arXiv:2005.14165 [cs]. July 2020. DOI: 10.48550/arXiv.2005.14165. URL: <http://arxiv.org/abs/2005.14165>.
- [9] J. Brownlee. *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*. en. Google-Books-ID: uAPuDwAAQBAJ. Machine Learning Mastery, June 2020. URL: https://books.google.nl/books/about/Data_Preparation_for_Machine_Learning.html?id=uAPuDwAAQBAJ&redir_esc=y.
- [10] R. G. Carvalho and R. F. Novo. „The Relationship between Structural Dimensions of Personality and School Life in Adolescence“. en. In: *Psicologia: Reflexão e Crítica* 27.2 (2014), pp. 368–376. ISSN: 0102-7972. DOI: 10.1590/1678-7153.201427218. URL: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-79722014000200368&lng=en&nrm=iso&tlang=en.
- [11] F. Celli *et al.* „Workshop on Computational Personality Recognition: Shared Task“. en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 7.2 (2013). Number: 2, pp. 2–5. ISSN: 2334-0770. DOI: 10.1609/icwsm.v7i2.14467. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/14467>.
- [12] B. Chen *et al.* *Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review*. arXiv:2310.14735 [cs]. June

2024. DOI: 10.48550/arXiv.2310.14735. URL: <http://arxiv.org/abs/2310.14735>.
- [13] J. Chen *et al.* „Adaptation with Self-Evaluation to Improve Selective Prediction in LLMs“. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 5190–5213. DOI: 10.18653/v1/2023.findings-emnlp.345. URL: <https://aclanthology.org/2023.findings-emnlp.345>.
 - [14] X. Chen *et al.* „KnowPrompt: Knowledge-aware Prompt-tuning with Synergistic Optimization for Relation Extraction“. en. In: *Proceedings of the ACM Web Conference 2022* (Apr. 2022). Conference Name: WWW '22: The ACM Web Conference 2022 ISBN: 9781450390965 Place: Virtual Event, Lyon France Publisher: ACM, pp. 2778–2788. DOI: 10.1145/3485447.3511998. URL: <https://dl.acm.org/doi/10.1145/3485447.3511998>.
 - [15] A. T. Church. „Culture and personality: Toward an integrated cultural trait psychology“. In: *Journal of Personality* 68.4 (2000). Place: United Kingdom Publisher: Blackwell Publishing, pp. 651–703. ISSN: 1467-6494. DOI: 10.1111/1467-6494.00112.
 - [16] P. T. Costa and R. R. McCrae. „The five-factor model of personality and its relevance to personality disorders“. In: *Journal of Personality Disorders* 6.4 (1992). Place: US Publisher: Guilford Publications, pp. 343–359. ISSN: 0885-579X. DOI: 10.1521/pedi.1992.6.4.343.
 - [17] M. De Bolle *et al.* „General personality and psychopathology in referred and nonreferred children and adolescents: an investigation of continuity, pathoplasty, and complication models“. eng. In: *Journal of Abnormal Psychology* 121.4 (Nov. 2012), pp. 958–970. ISSN: 1939-1846. DOI: 10.1037/a0027742.
 - [18] B. De Raad. *The Big Five Personality Factors: The psycholexical approach to personality*. The Big Five Personality Factors: The psycholexical ap-

proach to personality. Pages: vii, 128. Ashland, OH, US: Hogrefe & Huber Publishers, 2000. ISBN: 978-0-88937-236-8.

- [19] A. Demetriou *et al.* „Mind-Personality Relations from Childhood to Early Adulthood“. en. In: *Journal of Intelligence* 6.4 (Dec. 2018), p. 51. ISSN: 2079-3200. DOI: 10.3390/jintelligence6040051. URL: <https://www.mdpi.com/2079-3200/6/4/51>.
- [20] J. Devlin *et al.* *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805 [cs]. May 2019. DOI: 10.48550/arXiv.1810.04805. URL: <http://arxiv.org/abs/1810.04805>.
- [21] J. M. Digman. „Personality structure: Emergence of the five-factor model“. In: *Annual Review of Psychology* 41 (1990). Place: US Publisher: Annual Reviews, pp. 417–440. ISSN: 1545-2085. DOI: 10.1146/annurev.ps.41.020190.002221.
- [22] G. Feretzakis and V. S. Verykios. „Trustworthy AI: Securing Sensitive Data in Large Language Models“. en. In: *AI* 5.4 (Dec. 2024). Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, pp. 2773–2800. ISSN: 2673-2688. DOI: 10.3390/ai5040134. URL: <https://www.mdpi.com/2673-2688/5/4/134>.
- [23] J. Golbeck *et al.* „Predicting Personality from Twitter“. en. In: *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*. Boston, MA, USA: IEEE, Oct. 2011, pp. 149–156. ISBN: 978-1-4577-1931-8 978-0-7695-4578-3. DOI: 10.1109/PASSAT/SocialCom.2011.33. URL: <http://ieeexplore.ieee.org/document/6113107/>.
- [24] L. R. Goldberg. „An alternative "description of personality": The Big-Five factor structure“. In: *Journal of Personality and Social Psychology* 59.6 (1990). Place: US Publisher: American Psychological Association, pp. 1216–1229. ISSN: 1939-1315. DOI: 10.1037/0022-3514.59.6.1216.

- [25] M. A. Hearst. „Text Tiling: Segmenting Text into Multi-paragraph Subtopic Passages“. In: *Computational Linguistics* 23.1 (1997). Ed. by J. Hirschberg. Place: Cambridge, MA Publisher: MIT Press, pp. 33–64. URL: <https://aclanthology.org/J97-1003/>.
- [26] C. Hutto and E. Gilbert. „VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text“. en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 8.1 (May 2014). Number: 1, pp. 216–225. ISSN: 2334-0770. DOI: 10.1609/icwsm.v8i1.14550. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/14550>.
- [27] G. Jiang *et al.* *Evaluating and Inducing Personality in Pre-trained Language Models*. arXiv:2206.07550 [cs]. Oct. 2023. DOI: 10.48550/arXiv.2206.07550. URL: <http://arxiv.org/abs/2206.07550>.
- [28] H. Jiang *et al.* *PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits*. arXiv:2305.02547 [cs]. Apr. 2024. DOI: 10.48550/arXiv.2305.02547. URL: <http://arxiv.org/abs/2305.02547>.
- [29] O. John, L. Naumann, and C. Soto. „Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues“. In: *Handbook of Personality: Theory and Research, 3 Edn*. Journal Abbreviation: Handbook of Personality: Theory and Research, 3 Edn. Jan. 2008, pp. 114–158.
- [30] O. P. John, E. M. Donahue, and R. L. Kentle. *Big Five Inventory*. en. Institution: American Psychological Association. Oct. 2012. DOI: 10.1037/t07550-000. URL: <https://doi.apa.org/doi/10.1037/t07550-000>.
- [31] O. P. John and S. Srivastava. „The Big Five Trait taxonomy: History, measurement, and theoretical perspectives“. In: *Handbook of personality: Theory and research, 2nd ed.* New York, NY, US: Guilford Press, 1999, pp. 102–138. ISBN: 978-1-57230-483-3.

- [32] T. A. Judge *et al.* „Personality and leadership: a qualitative and quantitative review“. eng. In: *The Journal of Applied Psychology* 87.4 (Aug. 2002), pp. 765–780. ISSN: 0021-9010. DOI: 10.1037/0021-9010.87.4.765.
- [33] M. Kochanek *et al.* „Improving Training Dataset Balance with ChatGPT Prompt Engineering“. en. In: *Electronics* 13.12 (Jan. 2024). Number: 12 Publisher: Multidisciplinary Digital Publishing Institute, p. 2255. ISSN: 2079-9292. DOI: 10.3390/electronics13122255. URL: <https://www.mdpi.com/2079-9292/13/12/2255>.
- [34] T. Kojima *et al.* *Large Language Models are Zero-Shot Reasoners*. arXiv:2205.11916 [cs]. Jan. 2023. DOI: 10.48550/arXiv.2205.11916. URL: <http://arxiv.org/abs/2205.11916>.
- [35] H. Kourani *et al.* „Evaluating Large Language Models on Business Process Modeling: Framework, Benchmark, and Self-Improvement Analysis“. In: Nov. 2024. URL: <https://www.semanticscholar.org/paper/Evaluating-Large-Language-Models-on-Business-and-Kourani-Berti/19489636e75568571667a25dafc7d2fbf463c3dd>.
- [36] J. Kruger and D. Dunning. „Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments“. In: *Journal of Personality and Social Psychology* 77.6 (1999). Place: US Publisher: American Psychological Association, pp. 1121–1134. ISSN: 1939-1315. DOI: 10.1037/0022-3514.77.6.1121.
- [37] I. Krumpal. „Determinants of social desirability bias in sensitive surveys: a literature review“. en. In: *Quality & Quantity* 47.4 (June 2013), pp. 2025–2047. ISSN: 1573-7845. DOI: 10.1007/s11135-011-9640-9. URL: <https://doi.org/10.1007/s11135-011-9640-9>.
- [38] S. Laher. „Structural equivalence and the Neo-Pi-R: Implications for the applicability of the five-factor model of personality in an African context“. In: *SA Journal of Industrial Psychology* 34.1 (Sept. 2008), pp. 76–80. ISSN: 2071-0763, 0258-5200. DOI: 10.4102/sajip.v34i1.429. URL: <http://sajip.co.za/index.php/sajip/article/view/429>.

- [39] R. J. Larsen and T. Ketelaar. „Personality and susceptibility to positive and negative emotional states“. In: *Journal of Personality and Social Psychology* 61.1 (1991). Place: US Publisher: American Psychological Association, pp. 132–140. ISSN: 1939-1315. DOI: 10.1037/0022-3514.61.1.132.
- [40] D. Li *et al.* „From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge“. In: (2024). Publisher: arXiv Version Number: 4. DOI: 10.48550/ARXIV.2411.16594. URL: <https://arxiv.org/abs/2411.16594>.
- [41] X. Li *et al.* *Is GPT-3 a Psychopath? Evaluating Large Language Models from a Psychological Perspective*. Dec. 2022. DOI: 10.48550/arXiv.2212.10529.
- [42] R. Liu *et al.* *Best Practices and Lessons Learned on Synthetic Data*. arXiv:2404.07503 [cs]. Aug. 2024. DOI: 10.48550/arXiv.2404.07503. URL: <http://arxiv.org/abs/2404.07503>.
- [43] Y. Liu *et al.* *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv:1907.11692 [cs]. July 2019. DOI: 10.48550/arXiv.1907.11692. URL: <http://arxiv.org/abs/1907.11692>.
- [44] M. LLM. *11 Chunking Strategies for RAG Simplified & Visualized*. en. Nov. 2024. URL: <https://masteringllm.medium.com/11-chunking-strategies-for-rag-simplified-visualized-df0dbec8e373>.
- [45] K. Luo *et al.* „BGE Landmark Embedding: A Chunking-Free Embedding Method For Retrieval Augmented Long-Context Large Language Models“. In: (2024). Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.2402.11573. URL: <https://arxiv.org/abs/2402.11573>.
- [46] F. Mairesse *et al.* „Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text“. en. In: *Journal of Artificial Intelligence Research* 30 (Nov. 2007), pp. 457–500. ISSN: 1076-9757. DOI: 10.1613/jair.2349. URL: <https://www.jair.org/index.php/jair/article/view/10520>.

- [47] N. Majumder *et al.* „Deep Learning-Based Document Modeling for Personality Detection from Text“. en. In: *IEEE Intelligent Systems* 32.2 (Mar. 2017), pp. 74–79. ISSN: 1541-1672. DOI: 10.1109/MIS.2017.23. URL: <http://ieeexplore.ieee.org/document/7887639/>.
- [48] S. C. Matz and G. M. Harari. „Personalityplace transactions: Mapping the relationships between Big Five personality traits, states, and daily places“. In: *Journal of Personality and Social Psychology* 120.5 (2021). Place: US Publisher: American Psychological Association, pp. 1367–1385. ISSN: 1939-1315. DOI: 10.1037/pspp0000297.
- [49] R. R. McCrae and O. P. John. „An Introduction to the Five-Factor Model and Its Applications“. en. In: *Journal of Personality* 60.2 (1992). _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-6494.1992.tb00970.x>, pp. 175–215. ISSN: 1467-6494. DOI: 10.1111/j.1467-6494.1992.tb00970.x. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6494.1992.tb00970.x>.
- [50] R. R. McCrae and A. Terracciano. „Universal Features of Personality Traits From the Observer’s Perspective: Data From 50 Cultures“. In: *Journal of Personality and Social Psychology* 88.3 (2005). Place: US Publisher: American Psychological Association, pp. 547–561. ISSN: 1939-1315. DOI: 10.1037/0022-3514.88.3.547.
- [51] I. Montani *et al.* *explosion/spaCy: v3.7.2: Fixes for APIs and requirements*. Oct. 2023. DOI: 10.5281/zenodo.10009823. URL: <https://zenodo.org/records/10009823>.
- [52] P. Nayak. *Semantic Chunking for RAG*. en. Apr. 2024. URL: <https://medium.com/the-ai-forum/semantic-chunking-for-rag-f4733025d5f5>.
- [53] OpenAI. *GPT-2*. Hugging Face, 2019. URL: <https://huggingface.co/openai-community/gpt2>.
- [54] OpenAI. *The official Python library for the OpenAI API*. original-date: 2020-10-25T23:23:54Z. Jan. 2025. URL: <https://github.com/openai/openai-python>.

- [55] OpenAI *et al.* *GPT-4o System Card*. arXiv:2410.21276 [cs]. Oct. 2024. DOI: 10.48550/arXiv.2410.21276. URL: <http://arxiv.org/abs/2410.21276>.
- [56] OpenAI *et al.* *OpenAI o1 System Card*. arXiv:2412.16720 [cs]. Dec. 2024. DOI: 10.48550/arXiv.2412.16720. URL: <http://arxiv.org/abs/2412.16720>.
- [57] D. J. Ozer and V. Benet-Martínez. „Personality and the prediction of consequential outcomes“. eng. In: *Annual Review of Psychology* 57 (2006), pp. 401–421. ISSN: 0066-4308. DOI: 10.1146/annurev.psych.57.102904.190127.
- [58] A. K. Pamidi Venkata and L. Gudala. „The Potential and Limitations of Large Language Models for Text Classification through Synthetic Data Generation“. en. In: *INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING & APPLIED SCIENCES* (Apr. 2024), pp. 8–15. ISSN: 23949910, 23220821. DOI: 10.55083/irjeas.2024.v12i02002. URL: https://www.irjeas.org/wp-content/uploads/admin/volume12/V12I2/IRJEAS04V12I2002.pdf?_gl=1*ludvhc*_ga*MTcwMzY2MjEzNi4xNjk5NDQwNDU3*_ga_P10VE8TFFE*MTcxODg3NDEyMC41Mi4xLjE3MTg4NzQzNzkuMC4wLjA..
- [59] K. Pan and Y. Zeng. *Do LLMs Possess a Personality? Making the MBTI Test an Amazing Evaluation for Large Language Models*. en. July 2023. URL: <https://arxiv.org/abs/2307.16180v1>.
- [60] G. Park *et al.* „Automatic personality assessment through social media language“. eng. In: *Journal of Personality and Social Psychology* 108.6 (June 2015), pp. 934–952. ISSN: 1939-1315. DOI: 10.1037/pspp0000020.
- [61] J. Parker. *jparkerssemantic-chunking*. original-date: 2024-02-27T22:00:57Z. Jan. 2025. URL: <https://github.com/jparkerssemantic-chunking>.
- [62] D. L. Paulhus. „Measurement and control of response bias“. In: *Measures of personality and social psychological attitudes*. Measures of social psychological attitudes, Vol. 1. San Diego, CA, US: Academic Press, 1991, pp. 17–59. ISBN: 978-0-12-590241-0 978-0-12-590244-1. DOI: 10.1016/B978-0-12-590241-0.50006-X.

- [63] S. V. Paunonen and D. N. Jackson. „What is beyond the Big Five? Plenty!“ In: *Journal of Personality* 68.5 (2000). Place: United Kingdom Publisher: Blackwell Publishing, pp. 821–835. ISSN: 1467-6494. DOI: 10 . 1111 / 1467-6494.00117.
- [64] J. W. Pennebaker. *The secret life of pronouns: what our words say about us*. 1st U.S. ed. OCLC: ocn682894538. New York: Bloomsbury Press, 2011. ISBN: 978-1-60819-480-3.
- [65] J. W. Pennebaker and L. A. King. „Linguistic styles: Language use as an individual difference“. In: *Journal of Personality and Social Psychology* 77.6 (1999). Place: US Publisher: American Psychological Association, pp. 1296–1312. ISSN: 1939-1315. DOI: 10 . 1037 / 0022 - 3514 . 77 . 6 . 1296.
- [66] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer. „Psychological aspects of natural language. use: our words, our selves“. eng. In: *Annual Review of Psychology* 54 (2003), pp. 547–577. ISSN: 0066-4308. DOI: 10 . 1146/annurev.psych.54.101601.145041.
- [67] J. W. Pennebaker *et al.* „When Small Words Foretell Academic Success: The Case of College Admissions Essays“. In: *PLoS ONE* 9.12 (Dec. 2014), e115844. ISSN: 1932-6203. DOI: 10 . 1371/journal.pone.0115844. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4281205/>.
- [68] B. E. Perron *et al.* „A Primer on Word Embeddings: AI Techniques for Text Analysis in Social Work“. In: Nov. 2024. URL: <https://www.semanticscholar.org/paper/A-Primer-on-Word-Embeddings%3A-AI-Techniques-for-Text-Perron-Rivenburgh/03fd56d982977c60fd319b86ad0a53f034e456f8>.
- [69] A. E. Poropat. „A meta-analysis of the five-factor model of personality and academic performance“. In: *Psychological Bulletin* 135.2 (2009). Place: US Publisher: American Psychological Association, pp. 322–338. ISSN: 1939-1455. DOI: 10 . 1037/a0014996.
- [70] R. Poulain, H. Fayyaz, and R. Beheshti. „Aligning (Medical) LLMs for (Counterfactual) Fairness“. In: (2024). Publisher: arXiv Version Num-

- ber: 1. DOI: 10.48550/ARXIV.2408.12055. URL: <https://arxiv.org/abs/2408.12055>.
- [71] R. Qu, R. Tu, and F. Bao. *Is Semantic Chunking Worth the Computational Cost?* arXiv:2410.13070 [cs]. Oct. 2024. DOI: 10.48550/arXiv.2410.13070. URL: <http://arxiv.org/abs/2410.13070>.
- [72] A. Radford *et al.* „Language Models are Unsupervised Multitask Learners“. en. In: (2019). URL: <https://paperswithcode.com/paper/language-models-are-unsupervised-multitask>.
- [73] C. Raffel *et al.* *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.* arXiv:1910.10683 [cs]. Sept. 2023. DOI: 10.48550/arXiv.1910.10683. URL: <http://arxiv.org/abs/1910.10683>.
- [74] H. Rao, C. Leung, and C. Miao. *Can ChatGPT Assess Human Personalities? A General Evaluation Framework.* arXiv:2303.01248 [cs]. Oct. 2023. DOI: 10.48550/arXiv.2303.01248. URL: <http://arxiv.org/abs/2303.01248>.
- [75] C. C. Sartori *et al.* „Metaheuristics and Large Language Models Join Forces: Towards an Integrated Optimization Approach“. In: (2024). Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.2405.18272. URL: <https://arxiv.org/abs/2405.18272>.
- [76] T. Schick and H. Schütze. *Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference.* arXiv:2001.07676 [cs]. Jan. 2021. DOI: 10.48550/arXiv.2001.07676. URL: <http://arxiv.org/abs/2001.07676>.
- [77] F. Shi *et al.* *Large Language Models Can Be Easily Distracted by Irrelevant Context.* arXiv:2302.00093 [cs]. June 2023. DOI: 10.48550/arXiv.2302.00093. URL: <http://arxiv.org/abs/2302.00093>.
- [78] A. K. Singh *et al.* „Do Large Language Models Show Human-like Biases? Exploring ConfidenceCompetence Gap in AI“. en. In: *Information* 15.2 (Feb. 2024), p. 92. ISSN: 2078-2489. DOI: 10.3390/info15020092. URL: <https://www.mdpi.com/2078-2489/15/2/92>.

- [79] A. K. Singh *et al.* „The Confidence-Competence Gap in Large Language Models: A Cognitive Study“. In: (2023). Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.2309.16145. URL: <https://arxiv.org/abs/2309.16145>.
- [80] I. S. Singh *et al.* „ChunkRAG: Novel LLM-Chunk Filtering Method for RAG Systems“. In: (2024). Publisher: arXiv Version Number: 4. DOI: 10.48550/ARXIV.2410.19572. URL: <https://arxiv.org/abs/2410.19572>.
- [81] A. Sorokovikova *et al.* „LLMs Simulate Big5 Personality Traits: Further Evidence“. In: *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*. Ed. by A. Deshpande *et al.* St. Julians, Malta: Association for Computational Linguistics, Mar. 2024, pp. 83–87. URL: <https://aclanthology.org/2024.personalize-1.7>.
- [82] R. Souza, J.-H. Lim, and A. Davis. „Enhancing AI-Driven Psychological Consultation: Layered Prompts with Large Language Models“. In: (2024). Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.2408.16276. URL: <https://arxiv.org/abs/2408.16276>.
- [83] S. Stajner and S. Yenikent. „A Survey of Automatic Personality Detection from Texts“. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by D. Scott, N. Bel, and C. Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 6284–6295. DOI: 10.18653/v1/2020.coling-main.553. URL: <https://aclanthology.org/2020.coling-main.553>.
- [84] T. Tandera *et al.* „Personality Prediction System from Facebook Users“. en. In: *Procedia Computer Science* 116 (2017), pp. 604–611. ISSN: 18770509. DOI: 10.1016/j.procs.2017.10.016. URL: <https://linkinghub.elsevier.com/retrieve/pii/S1877050917320537>.
- [85] Y. R. Tausczik and J. W. Pennebaker. „The psychological meaning of words: LIWC and computerized text analysis methods“. In: *Journal of Language and Social Psychology* 29.1 (2010). Place: US Publisher: Sage Publications, pp. 24–54. ISSN: 1552-6526. DOI: 10.1177/0261927X09351676.

- [86] A. S. Thakur *et al.* „Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges“. In: (2024). Publisher: arXiv Version Number: 4. DOI: 10.48550/ARXIV.2406.12624. URL: <https://arxiv.org/abs/2406.12624>.
- [87] A. V Ganesan *et al.* „Systematic Evaluation of GPT-3 for Zero-Shot Personality Estimation“. In: *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. Ed. by J. Barnes, O. De Clercq, and R. Klinger. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 390–400. DOI: 10.18653/v1/2023.wassa-1.34. URL: <https://aclanthology.org/2023.wassa-1.34>.
- [88] R. Vacareanu *et al.* *From Words to Numbers: Your Large Language Model Is Secretly A Capable Regressor When Given In-Context Examples*. arXiv:2404.07544 [cs]. Sept. 2024. DOI: 10.48550/arXiv.2404.07544. URL: <http://arxiv.org/abs/2404.07544>.
- [89] A. Vaswani *et al.* *Attention Is All You Need*. arXiv:1706.03762 [cs]. Aug. 2023. DOI: 10.48550/arXiv.1706.03762. URL: <http://arxiv.org/abs/1706.03762>.
- [90] H. Wadhwa *et al.* *From RAGs to rich parameters: Probing how language models utilize external knowledge over parametric information for factual queries*. arXiv:2406.12824 [cs]. June 2024. DOI: 10.48550/arXiv.2406.12824. URL: <http://arxiv.org/abs/2406.12824>.
- [91] L. Wang, R. Chen, and L. Li. „Knowledge-Guided Prompt Learning for Few-Shot Text Classification“. en. In: *Electronics* 12.6 (Mar. 2023), p. 1486. ISSN: 2079-9292. DOI: 10.3390/electronics12061486. URL: <https://www.mdpi.com/2079-9292/12/6/1486>.
- [92] J. Wei *et al.* *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv:2201.11903 [cs]. Jan. 2023. DOI: 10.48550/arXiv.2201.11903. URL: <http://arxiv.org/abs/2201.11903>.

- [93] T. Xia *et al.* „Language Models can Evaluate Themselves via Probability Discrepancy“. In: Version Number: 2. arXiv, 2024. doi: 10.48550/ARXIV.2405.10516. URL: <https://arxiv.org/abs/2405.10516>.
- [94] L. Yang *et al.* „Computational personality: a survey“. en. In: *Soft Computing* 26.18 (Sept. 2022), pp. 9587–9605. issn: 1433-7479. doi: 10.1007/s00500-022-06786-6. URL: <https://doi.org/10.1007/s00500-022-06786-6>.
- [95] T. Yarkoni. „Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers“. In: *Journal of Research in Personality* 44.3 (2010). Place: Netherlands Publisher: Elsevier Science, pp. 363–373. issn: 1095-7251. doi: 10.1016/j.jrp.2010.04.001.
- [96] M. Zastempowski. „Shaping sustainable futures: The role of micro-entrepreneurs personality traits in social innovations“. en. In: *PLOS ONE* 19.8 (Aug. 2024). Ed. by H. Kamilcelebi, e0306800. issn: 1932-6203. doi: 10.1371/journal.pone.0306800. URL: <https://dx.plos.org/10.1371/journal.pone.0306800>.
- [97] D. Zhang. *Stella EN 1.5B*. Hugging Face, Dec. 2024. URL: https://huggingface.co/dunzhang/stella_en_1.5B_v5.
- [98] J. Zhang *et al.* „The Better Angels of Machine Personality: How Personality Relates to LLM Safety“. In: (2024). Publisher: arXiv Version Number: 1. doi: 10.48550/ARXIV.2407.12344. URL: <https://arxiv.org/abs/2407.12344>.
- [99] W. Zhang *et al.* „Self-Contrast: Better Reflection Through Inconsistent Solving Perspectives“. In: Version Number: 3. arXiv, 2024. doi: 10.48550/ARXIV.2401.02009. URL: <https://arxiv.org/abs/2401.02009>.
- [100] X. Zhang *et al.* „Self-Alignment for Factuality: Mitigating Hallucinations in LLMs via Self-Evaluation“. In: Version Number: 2. arXiv, 2024. doi: 10.48550/ARXIV.2402.09267. URL: <https://arxiv.org/abs/2402.09267>.

- [101] H. Zhou *et al.* „Mitigating the Bias of Large Language Model Evaluation“. In: Version Number: 1. arXiv, 2024. doi: 10.48550/ARXIV.2409.16788. URL: <https://arxiv.org/abs/2409.16788>.
- [102] Y. Zhou *et al.* „Large Language Models Are Human-Level Prompt Engineers“. In: (2022). Publisher: arXiv Version Number: 2. doi: 10.48550/ARXIV.2211.01910. URL: <https://arxiv.org/abs/2211.01910>.