A Survey on Large Language Models in Personality Analysis and AI Ethics

www.surveyx.cn

Abstract

This survey explores the interdisciplinary integration of Large Language Models (LLMs), personality analysis, natural language processing (NLP), AI ethics, and computational psychology, highlighting the transformative potential of LLMs in simulating and analyzing human personality traits. The survey underscores the necessity of a systematic framework for persona alignment and addresses the challenges of ethical AI deployment, emphasizing alignment with human values. Key findings reveal the importance of interdisciplinary collaboration, drawing insights from psychology, AI, and ethics to enhance LLM applications in personality analysis. The exploration includes methodologies for personality trait manipulation and simulation, benchmarking techniques, and the role of cultural and contextual influences. The survey also examines the ethical implications of using LLMs in personality analysis, focusing on privacy, data protection, and fairness. Current applications across mental health, education, and human-computer interaction demonstrate LLMs' broad impact, while future research directions emphasize refining NLP models, enhancing AI-human interactions, and developing culturally sensitive AI systems. By addressing these considerations, the survey aims to guide future research, fostering the development of more effective and responsible AI systems that align with societal expectations and ethical standards.

1 Introduction

1.1 Interdisciplinary Field Overview

The interdisciplinary field integrating Large Language Models (LLMs), personality analysis, natural language processing (NLP), AI ethics, and computational psychology is characterized by the convergence of diverse academic disciplines that enhance personality analysis through LLMs. This integration combines insights from artificial intelligence, psychology, linguistics, and related fields, thereby improving the analytical capabilities of these models [1]. For instance, LLMs have been utilized to refine self-assessment scales for interpersonal communication skills, exemplifying the synergy between AI and psychological assessment [2].

In the electric energy sector, LLMs illustrate their interdisciplinary impact through contributions to data analysis, forecasting, and operational decision-making [3]. The integration of AI with personality-based negotiation strategies further emphasizes LLMs' role in multi-issue negotiation processes, highlighting the intersection of AI with social sciences [4]. This is complemented by the combination of LLMs with strategic planning and social reasoning in AI agents, particularly in complex negotiation scenarios [5].

In cybersecurity, LLMs enhance honeypots, showcasing the interplay between NLP and cybersecurity to develop more effective security measures [6]. The impact of persona and conversational tasks on interactions with LLM-controlled agents further reflects the interdisciplinary nature of AI in social and behavioral sciences [7].

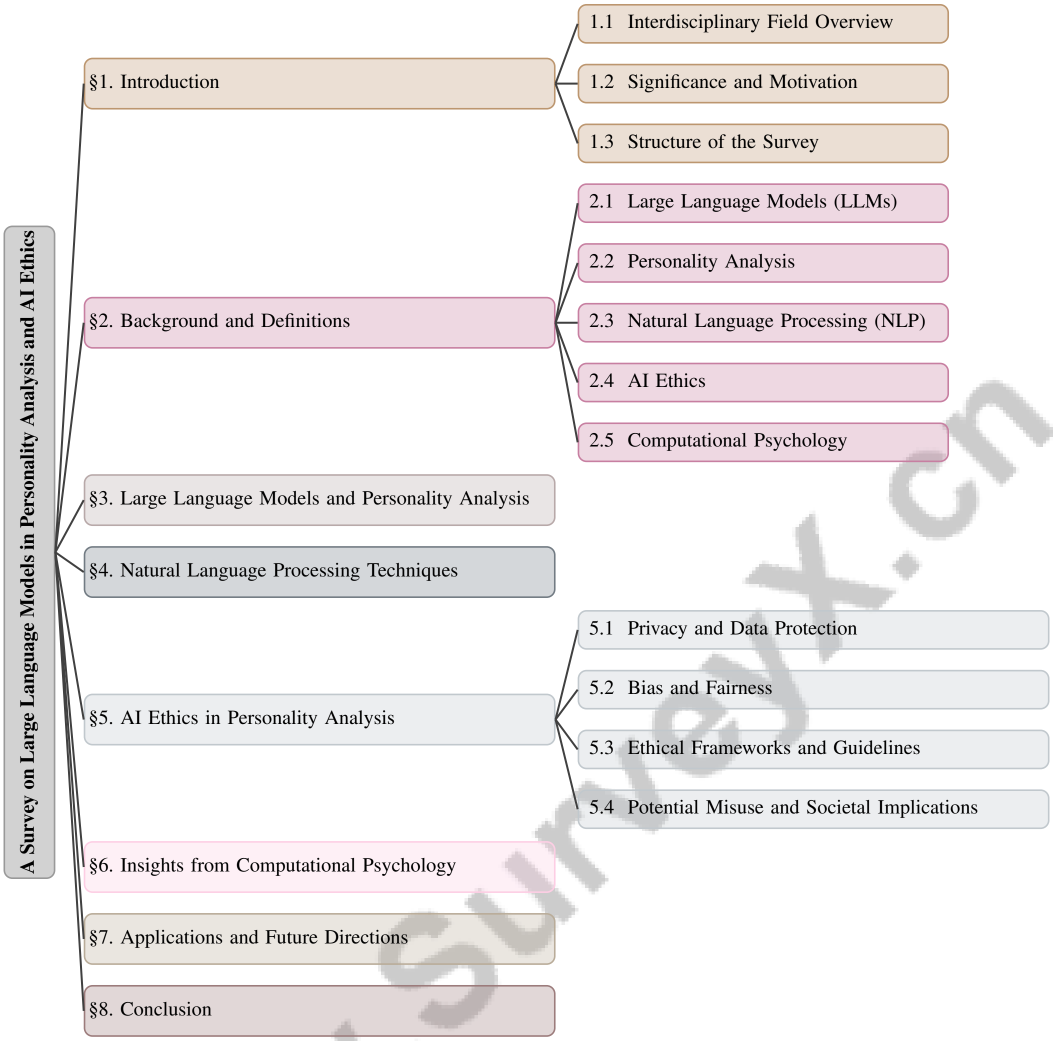

A Survey on Large Language Models in Personality Analysis and AI Ethics
§1. Introduction
1.1 Interdisciplinary Field Overview
1.2 Significance and Motivation
1.3 Structure of the Survey
§2. Background and Definitions
2.1 Large Language Models (LLMs)
2.2 Personality Analysis
2.3 Natural Language Processing (NLP)
2.4 AI Ethics
2.5 Computational Psychology
§3. Large Language Models and Personality Analysis
§4. Natural Language Processing Techniques
§5. AI Ethics in Personality Analysis
5.1 Privacy and Data Protection
5.2 Bias and Fairness
5.3 Ethical Frameworks and Guidelines
5.4 Potential Misuse and Societal Implications
§6. Insights from Computational Psychology
§7. Applications and Future Directions
§8. Conclusion

Figure 1: chapter structure

The exploration of LLMs in planning and reasoning tasks addresses knowledge gaps in existing surveys, emphasizing the collaborative efforts needed to improve AI-based agents' planning abilities [8]. The challenge presented by reasoning CAPTCHAs, designed to be straightforward for humans yet complex for AI, highlights the necessity of logical reasoning and problem-solving skills, bridging the domains of AI and cognitive sciences [9].

Investigations into AI interpretability redefine how LLMs can provide nuanced explanations, illustrating the intersection of AI with interpretability and machine learning [10]. Additionally, studies on LLMs' 'mental models' yield significant implications for cultural psychology and AI [11]. The integration of LLMs into dentistry further underscores the broad applicability of these models across various fields, particularly in healthcare [12].

These contributions collectively highlight the interdisciplinary nature of the field, where the integration of AI technologies with psychological theories and ethical considerations fosters a comprehensive understanding of personality analysis using LLMs [13]. Furthermore, systematic analyses of optimizing algorithms with LLMs reveal their potential to enhance decision-making in complex environments, demonstrating the multifaceted applications of LLMs across diverse domains [14].

1.2 Significance and Motivation

The exploration of LLMs in personality analysis and AI ethics is driven by their transformative potential to revolutionize various fields through enhanced automation and efficiency. In educational

AI-generated, for reference only.

contexts, LLMs play a crucial role in modeling student personalities, facilitating the development of conversational Intelligent Tutoring Systems (ITSs) that provide personalized learning experiences and improve educational outcomes [15]. This underscores the significance of LLMs in tailoring educational content to individual learner needs.

Within human-computer interaction (HCI) and social sciences, LLMs address challenges such as bias and labor intensity, prompting research into their qualitative data processing applications [16]. Their capacity for quantitative analysis of human preferences enhances model alignment and evaluation methodologies, thereby improving user experience and satisfaction [17]. The issue of hallucinations in LLMs, particularly in critical tasks, necessitates ongoing research to enhance the reliability and accuracy of outputs, with Knowledge Graphs (KGs) being explored as a mitigation strategy [18].

The critical need for transparency in LLMs is vital for responsible development and deployment, highlighting gaps in current discourse and the importance of a human-centered approach [19]. Enhancing user interaction and engagement through effective integration of personality into conversational agents is a key motivation driving this research [20]. The open-source nature of LLMs has democratized access to powerful AI tools, facilitating learning and knowledge dissemination [21].

LLMs' ability to elicit diverse behaviors, trained on extensive text corpora encoding various personality traits, underscores the importance of researching LLMs in personality analysis and AI ethics, focusing on potential benefits and impacts [22]. As social networks proliferate, developing models capable of automatically interpreting individuals' essences based on their writing becomes crucial [23]. However, safety concerns regarding LLMs' alignment with human preferences in moral decision-making persist [24].

Research on LLMs in personality analysis is motivated by societal and ethical concerns regarding their behavior [25]. The potential of LLMs to improve measurement accuracy and efficiency in interpersonal communication skills further emphasizes the importance of this research [2]. Moreover, LLMs enable social robots to engage in open-ended conversations while generating context-appropriate expressive behaviors [26].

The moral imperative to achieve lifelong superalignment in AI systems, particularly LLMs, underscores the necessity of aligning these models with evolving human values and legal standards [27]. The importance of refining LLM responses for medical applications, ensuring they are founded on validated information and structured evaluation, further highlights the motivation behind this research [28]. These motivations and benefits collectively underscore the significance of researching LLMs in personality analysis and AI ethics, aiming to harness their full potential while addressing ethical considerations.

1.3 Structure of the Survey

This survey is systematically organized to provide a comprehensive understanding of the intersection between Large Language Models (LLMs), personality analysis, natural language processing (NLP), AI ethics, and computational psychology. The initial section introduces the interdisciplinary nature of the field, highlighting its significance and motivation, followed by an overview of the survey structure. The second section delves into the background and definitions, offering a detailed explanation of core concepts such as LLMs, personality analysis, NLP, AI ethics, and computational psychology.

The third section focuses on the role of LLMs in personality analysis, exploring their capabilities in simulating and recognizing personality traits. This is complemented by a discussion on benchmarking and evaluation frameworks to assess LLM efficacy in this domain. The fourth section examines various NLP techniques employed in personality analysis, contrasting data-driven and theory-driven approaches while considering interactive and contextual methodologies.

The fifth section addresses ethical considerations, emphasizing privacy, data protection, bias, fairness, and the potential misuse of personality data. It outlines existing ethical frameworks and guidelines to ensure responsible AI deployment. Insights from computational psychology are explored in the sixth section, discussing the integration of psychological theories, personality dynamics, and cultural influences in enhancing AI systems.

The penultimate section identifies current applications and future directions for LLMs in personality analysis, highlighting challenges and opportunities for improving AI-human interactions. Finally, the conclusion synthesizes the key findings and contributions of the survey, underscoring the importance of interdisciplinary collaboration in advancing this field.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Large Language Models (LLMs)

Large Language Models (LLMs) are advanced AI systems designed to process and generate human-like text by utilizing extensive training datasets. Models such as GPT-3.5 exhibit notable proficiency in various natural language processing tasks, demonstrating adaptability across multiple domains [29]. LLMs perform diverse language-related functions, including text generation, translation, summarization, and annotation, which are pivotal for advancing computational social science and enhancing human data collection [30]. In personality analysis, LLMs are recognized for simulating and assessing human personality traits through linguistic patterns and contextual cues in text, significantly influencing dialog behavior in conversational agents [7]. Despite their strengths, LLMs face challenges such as generating factual inaccuracies, necessitating benchmarks to evaluate their accuracy and reliability [6]. Research explores LLMs' adaptability in exhibiting appropriate personality traits for specific organizational roles, highlighting the need for validated measurement methods [31]. Integrating LLMs with formal methods for logic theory induction underscores the limitations of current autoregressive models in managing complex facts and rules [32]. Benchmarks assess LLMs in personalized conversation contexts, illustrating their potential in simulating human-like deliberation and behavior [33]. A survey of open LLMs analyzing twelve models provides insights into their performance on personality assessments such as MBTI and BFI, highlighting the diversity of these models in personality evaluation [34]. Optimizing LLMs for on-device intelligence, particularly within mobile hardware constraints, enhances performance and efficiency [35]. The role of optimization algorithms in adapting LLMs to dynamic environments further illustrates their potential across various applications [14]. Developing inclusive writing styles through benchmarks assessing LLMs on text generation tasks for scientific abstracts promotes advancements in generating diverse and inclusive text [36].

2.2 Personality Analysis

Personality analysis systematically examines individual differences in characteristic patterns of thinking, feeling, and behaving. This field has gained traction in NLP through the application of LLMs to infer personality traits from text data. Traditional approaches often rely on binary classifications, which may oversimplify human personality complexity. In contrast, LLMs provide a nuanced understanding by detecting personality types from textual content, as demonstrated by benchmarks investigating these models' capabilities to mirror human classifications such as the MBTI [34]. The integration of personality traits into LLMs incorporates established psychological frameworks like the Big Five personality traits—openness, conscientiousness, extraversion, agreeableness, and neuroticism—which are essential for comprehensive personality analysis [37]. Accurately inferring these traits from text remains challenging, addressed through benchmarks evaluating LLMs' ability to derive psychological traits from social media posts without prior training on specific traits [23]. While recognizing personality traits from text is well-established in NLP, existing models primarily focus on fundamental traits due to the complexity of comprehensive personality inventories [38]. This complexity is compounded by the need to assess personality traits across various languages and cultural contexts. Deep learning methods have facilitated personality recognition from short texts in multilingual environments without extensive feature engineering [39]. In educational settings, personality analysis emphasizes the impact of different personality styles, such as high versus low extroversion and agreeableness, on learning outcomes. This analysis is vital for tailoring educational strategies to individual learner profiles, enhancing the effectiveness of educational interventions [7]. Accurately assessing human personality traits is crucial for improving commonsense reasoning in machines, enhancing their capacity for intuitive and empathetic interactions with humans [40]. The analysis of personality traits in LLMs marks significant progress in understanding human behavior, offering valuable insights into individual differences critical for various applications, including personalized education and enhanced human-computer interaction. Recent studies utilizing the Big Five personality model and the MBTI demonstrate that LLMs can consistently reflect distinct personality profiles, with evaluations showing that human users accurately perceive these traits in LLM-generated content. Innovative assessment tools like the TRAIT benchmark have been developed

AI-generated, for reference only.

to rigorously evaluate LLM personalities, highlighting their distinctiveness and the influence of training data on personality expression. This growing body of research underscores the potential for LLMs to mimic human-like personality traits and adapt based on user interactions and contextual factors [41, 42, 43, 44]. Integrating personality analysis into AI systems enhances their capability to simulate human-like interactions and aligns their functioning with nuanced human behavior patterns, ultimately contributing to more effective and ethical AI-human collaborations.

2.3 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a critical area of artificial intelligence that enables meaningful interactions between computers and humans using natural language. This field encompasses various applications, including automated plagiarism detection systems that utilize advanced NLP techniques to analyze text for originality and accuracy. NLP is also essential in addressing fairness in AI applications, particularly in sensitive contexts like recruitment and education, where biases can lead to discriminatory outcomes. As LLMs evolve, the importance of rigorous evaluation and transparency in NLP becomes increasingly critical, ensuring that these technologies are developed responsibly to serve diverse user needs [16, 45, 46, 19, 47]. This field employs a wide range of techniques and methodologies aimed at processing and analyzing large volumes of language data. In the context of personality analysis, NLP serves as a foundational tool for extracting meaningful insights from textual data, enabling the assessment of personality traits and behaviors. The application of NLP in personality analysis involves various methodologies, including pre-trained independent, pre-trained model-based, and multimodal strategies that facilitate automatic personality prediction (APP) [48]. These approaches leverage LLM capabilities to analyze linguistic patterns and infer personality characteristics, thereby enhancing the accuracy and reliability of personality assessments. Benchmark datasets are crucial for evaluating the effectiveness of NLP models in personality analysis. For instance, LLMEval2 is designed to assess generated text quality across multiple tasks and abilities, providing insights into LLM performance in diverse applications [49]. Comprehensive evaluation mechanisms for chatbot responses enable comparisons between automated, human, and LLM-based evaluations, thereby enhancing the understanding and application of LLMs in real-world scenarios [50]. NLP techniques are instrumental in business process management (BPM) for extracting information from unstructured textual documents, showcasing their versatility across various domains [51]. Additionally, NLP models are used to extract adjective similarities from extensive text corpora, offering a more objective analysis of personality traits without the constraints of traditional survey methods [38]. Despite advancements in NLP, challenges such as data contamination in classical evaluation tasks persist, particularly when LLMs are trained on test splits of benchmarks [45]. To mitigate these issues, benchmark datasets are designed to assess LLM capabilities without the risk of data leakage, ensuring robust evaluation of model performance across different domains and knowledge areas [52]. Integrating NLP with machine learning is pivotal in enhancing the accuracy of mental health assessments, steering towards a data-driven approach that leverages LLM analytical power [53]. Furthermore, NLP addresses challenges in predicting personality types based on frameworks like the MBTI, employing a data-centric approach to improve prediction accuracy [54].

2.4 AI Ethics

AI ethics serves as a critical framework for the responsible development and deployment of LLMs, particularly in personality analysis. This framework encompasses principles such as transparency, accountability, and fairness while addressing challenges like bias, misinformation, and privacy concerns. The limitations of current methods in Text-based Personality Computing (TPC) research, especially regarding measurement quality and ethical implications, underscore the necessity for a robust ethical framework [55]. The inherent limitations of existing LLM architectures in understanding and adapting to the dynamic nature of human ethics highlight the need for responsible AI use, as these models often struggle to align with evolving global scenarios and ethical standards [27]. The complexity of LLM technologies exacerbates challenges such as managing API configurations, handling output unpredictability, ensuring data privacy, and optimizing performance [56]. These challenges are further compounded by the potential for LLMs to generate harmful content and the risk of exploitation through adversarial persona manipulation, emphasizing the critical role of AI ethics in mitigating such risks [57]. Additionally, the absence of a systematic taxonomy in existing research, coupled with difficulties in aligning LLMs with user expectations, presents significant ethical concerns related to persona assignment and safety [13]. In personality analysis, AI ethics is crucial for addressing the challenge of accurately annotating context-dependent features, raising significant ethical considerations [58]. The reliance of LLMs on training data and their difficulties in managing complex queries complicate their application in sensitive areas such as healthcare, where ethical concerns are paramount [59]. Furthermore, integrating LLMs into sectors like electric energy necessitates careful attention to privacy and reliability of outputs in safety-critical applications [3]. The presence of distinct failure modes in LLM evaluations, such as bias and inconsistency, emphasizes the need for robust ethical guidelines to ensure the validity of conclusions drawn from LLM annotations. Moreover, the inefficiencies and high computational costs associated with large-scale LLMs in tasks like mathematical problem-solving and logical reasoning highlight the ethical imperative to optimize these models for responsible use [60]. The GigaCheck framework, which establishes a comprehensive benchmark for detecting LLM-generated content, addresses key challenges in the field and offers improved accuracy over existing benchmarks, further emphasizing the importance of ethical considerations in AI deployment [61]. An integrated approach to AI ethics, encompassing technological evaluation, human-AI interaction, and corporate behavior scrutiny, is advocated to ensure responsible use of AI systems [62]. The primary challenge is that existing methods do not adequately address the need for multi-objective considerations, such as privacy, in decision-making processes within LLM cascades [35]. Key challenges include high computational resource requirements for training LLMs, the necessity for extensive datasets, and issues related to model interpretability and robustness [14]. AI ethics is essential in guiding the responsible use of LLMs for personality analysis, ensuring these technologies are developed and utilized in a manner that prioritizes ethical principles and societal well-being.

2.5 Computational Psychology

Computational psychology merges principles from psychology with computational methods to model and understand human cognition and behavior. This interdisciplinary approach is integral to developing AI technologies, particularly in enhancing the functionality and interpretability of LLMs. By leveraging insights from computational psychology, especially through specialized models like PersonalityMap and frameworks such as PerSense, researchers can create advanced AI systems that accurately extract personality traits from text while enhancing commonsense reasoning capabilities. These systems utilize a combination of machine learning algorithms and psycholinguistic features to replicate human-like reasoning and personality expression, demonstrating significant improvements over traditional methods and outperforming many human experts in predicting personality correlations [63, 64, 65, 37]. The integration of psychological insights into AI technologies is exemplified by categorizing conversational agents based on their embodiment and personality traits. This categorization highlights the interplay between these factors in educational contexts, suggesting that the embodiment of agents significantly influences their effectiveness in conveying personality traits [66]. Such insights are crucial for designing AI systems that can interact with humans intuitively and personally. Furthermore, examining cultural contexts in shaping cognitive processes and self-construal in LLMs underscores the importance of cultural psychology in computational models. This perspective provides a theoretical framework for understanding how cultural variations influence LLMs' information processing and human cognition simulation [11]. By incorporating cultural dimensions into AI systems, developers can enhance the cultural sensitivity and adaptability of these models, making them more relevant across diverse user groups. The survey of research into psychometric evaluations of LLM responses and comparative analyses of human and LLM personality traits emphasizes the need for robust validation methods. These methods are essential for ensuring that LLMs accurately reflect human personality traits and cognitive processes, thereby improving their reliability and applicability in various domains [67]. Additionally, challenges posed by multilingual understanding and processing extensive historical data in the field of Classics underscore the need for LLMs to handle complex and diverse datasets [21]. Organizing research into fields such as personality taxonomies, measurement quality, and modeling choices further illustrates the multifaceted nature of computational psychology. These areas are critical for evaluating the performance and ethical implications of AI models, ensuring they are developed and deployed in alignment with psychological principles and ethical standards [55].

AI-generated, for reference only.

3 Large Language Models and Personality Analysis

Large Language Models (LLMs) have become pivotal in personality analysis, offering the ability to simulate and manipulate human-like traits. This section explores the methodologies and innovations that enable LLMs to authentically reflect personality traits, grounded in established psychological frameworks. By understanding how these models enhance interactions across applications, we gain insights into their transformative potential. The following subsection delves into specific strategies for personality trait manipulation and simulation, highlighting key advancements and implications for AI-human interactions.

3.1 Personality Trait Manipulation and Simulation

LLMs' ability to manipulate and simulate personality traits relies heavily on psychological frameworks like the Big Five, encompassing openness, conscientiousness, extraversion, agreeableness, and neuroticism. These frameworks guide LLMs in replicating human-like traits, as demonstrated by their consistent generation of content aligned with specific personality profiles and distinct linguistic patterns [41, 68, 69, 44]. By employing advanced text augmentation and the Big Five model, LLMs enhance personality detection accuracy, tailoring outputs to mimic desired human characteristics and improving conversational agents' effectiveness.

Recent advancements show LLMs embedding human behavioral tendencies into their processing capabilities, moving beyond superficial prompt modifications to a stable personality profile [31]. Unlike traditional methods reliant on subjective self-reports, LLMs analyze extensive text datasets to uncover personality trait structures [38]. This shift from static text analysis to dynamic interactions underscores the importance of context in personality prediction.

Interactions with extraverted agents yield more favorable evaluations and engagement than with introverted ones, highlighting the significance of personality trait manipulation in enhancing user experiences [7]. The multi-LLM orchestration engine captures conversational history and private data, further personalizing interactions [33].

Despite progress, models like SOLAR and Dolphin effectively mimic conditioned personalities, while others struggle to adapt consistently, often defaulting to inherent traits [34]. This challenge persists in ensuring LLMs maintain desired personality profiles across contexts.

New benchmarks using frameworks like LIWC reveal gender biases in LLM-generated texts, refining personality trait evaluation [36]. Integrating optimization algorithms into LLMs enhances personalization and accuracy of AI-generated content [14].

Richelieu, a self-evolving LLM-based agent, demonstrates strategic personality manipulation in multi-agent environments, addressing complex human-like interactions [5]. This underscores strategic frameworks' role in refining personality trait expression and manipulation in LLMs.

The manipulation and simulation of personality traits in LLMs mark a pivotal advancement in AI research. Studies reveal LLMs exhibit distinct personalities influenced by cultural norms and stressors. Training-free approaches modify LLM behaviors to align with specific traits, enhancing personalized chatbot applications. Investigations into LLM personas based on the Big Five model show high consistency in self-reported scores and recognizable linguistic patterns. This capability opens new avenues for dialogue systems and role-playing agents while raising considerations about model safety and AI-generated personality perception [41, 44]. Sophisticated psychological models and innovative methodologies pave the way for nuanced and ethical AI-human interactions, aligning LLM capabilities with human-like personality expressions.

3.2 Personality Recognition and Classification

LLMs' recognition and classification of personality traits involve advanced methodologies for precise categorization of human dimensions. A significant challenge is simulating personalized conversations reflecting specific traits, which traditional methods struggle to achieve [70]. Knowledge graph-enabled approaches enhance trait inference accuracy beyond conventional linguistic analysis and machine learning limitations [71].

Benchmarks extracting Big Five traits from text simulate real-world scenarios, providing a robust framework for assessing personality through language [37]. These benchmarks refine LLMs' ability to simulate human-like traits, ensuring responses align with designated profiles [44]. Frameworks moving beyond fixed prompts capture the dynamic nature of personality traits, addressing representation challenges [72].

Experiments demonstrate LLMs can exhibit traits comparable to humans, validated through inventories like the Big Five Inventory (BFI). LLM personas' self-reported scores are consistent with designated types, with significant differences across dimensions [44]. Developing benchmarks assessing LLMs' reliability in generating consistent trait scores ensures accurate classification.

These methodologies and benchmarks underscore LLMs' complexity in personality recognition and classification, emphasizing sophisticated techniques for accurate assessments. Advanced training and continuous evaluation enable LLMs to simulate and articulate human traits effectively, enhancing AI-human interactions' consistency and authenticity. Research shows LLMs produce outputs closely aligned with specific profiles, enabling meaningful and contextually appropriate dialogues across applications, from conversational agents to collaborative systems [44, 73, 74, 41, 75].

3.3 Benchmarking and Evaluation Frameworks

Evaluating LLMs in personality analysis relies on robust benchmarking and assessment frameworks to ensure effectiveness and accuracy. These frameworks systematically evaluate LLMs' applications and performance, particularly in simulating human-like interactions and traits. Recent developments introduce comprehensive evaluation platforms assessing LLMs' response consistency across psychological instruments and prompt variations, underscoring systematic testing necessity [57]. These platforms offer structured environments for evaluating LLMs' capabilities in simulating human-like traits, providing insights into operational effectiveness in dynamic contexts.

Neuron-based personality trait induction (NPTI) methods show promise in inducing and evaluating traits in LLMs, achieving performance comparable to fine-tuning with greater efficiency and flexibility. Evaluation metrics like Induction Success Rate (ISR), Trait Induction Efficacy (TIE), Personality Induction Success Rate (PISR), and Personality Induction Efficacy (PIE) assess personality control methods' effectiveness [60, 76].

Benchmarking efforts focus on understanding how training datasets reflect LLMs' personality traits, refining outputs to align better with desired profiles and improving interactions [77]. Exploring LLMs' traits through models like Gemma-2B-Instruct and Gemma-2-9B-Instruct, assessed on tests such as the Big Five Inventory (BFI) and Short Dark Triad (SD-3), provides valuable insights into personality assessment performance [24].

Continuous evaluation integration into LLM agents' lifecycle, as proposed by evaluation-driven design approaches, supports adaptive runtime adjustments and systematic offline redevelopment, ensuring models remain effective and responsive to changing requirements [78]. Studies conclude machine-generated texts can mimic human writing closely, but significant differences in readability and traits suggest more nuanced evaluation frameworks' need [79].

EvalGen, a mixed-initiative tool, facilitates creating and validating evaluation functions for LLM outputs by aligning them with human preferences, enhancing personality assessments' reliability [80]. Aligning evaluation processes with human-centric criteria ensures LLMs produce accurate and meaningful outputs. Despite advancements, challenges persist in aligning LLM preferences with human evaluations, as significant preference differences are observed [45].

As illustrated in Figure 2, the hierarchical organization of benchmarking and evaluation frameworks for LLMs highlights three main areas: evaluation platforms assessing response consistency and human-like traits, personality trait induction methods, and benchmarking efforts that explore training datasets and specific models like Gemma-2B-Instruct. The integration of these frameworks is crucial for enhancing and validating LLM performance. The figure visually explores distinct methodologies in this domain, underscoring the multifaceted strategies employed in optimizing LLMs and the importance of systematic evaluation and benchmarking in advancing these sophisticated models' capabilities [52, 81, 82].

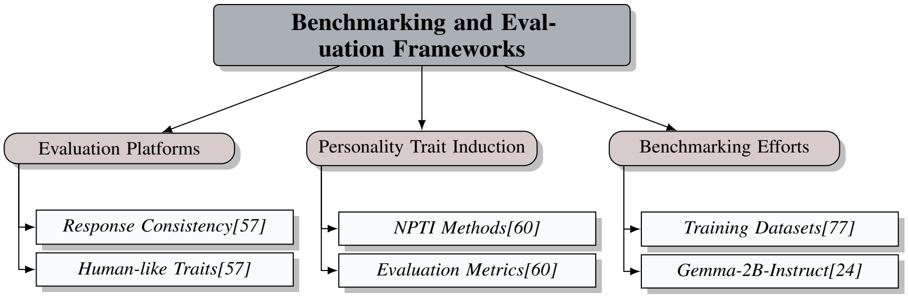


Figure 2: This figure illustrates the hierarchical organization of benchmarking and evaluation frameworks for LLMs, focusing on three main areas: evaluation platforms that assess response consistency and human-like traits, personality trait induction methods, and benchmarking efforts that explore training datasets and specific models like Gemma-2B-Instruct.

4 Natural Language Processing Techniques

Category	Feature	Method
Methodologies for Personality Analysis	Feature and Model Integration	NLP[38], CQTM[83], MM[31], DCPP[54], KGE-APP[71]
	Resource and Efficiency Optimization	LLM-HP[6], PAS[84]
	Interaction and Analysis Techniques	AFSPP[72], UPGW[85]

Table 1: This table presents a comprehensive overview of the methodologies employed in personality analysis using Natural Language Processing (NLP) techniques. It categorizes the methodologies into three main areas: feature and model integration, resource and efficiency optimization, and interaction and analysis techniques, highlighting specific methods and their contributions to enhancing personality trait detection and AI-human interactions.

Exploring Natural Language Processing (NLP) techniques for personality analysis is crucial for simulating human behavior computationally. This section delves into methodologies employed in this domain, highlighting their role in extracting and interpreting personality traits from textual data. As illustrated in **??**, the hierarchical structure of these NLP techniques categorizes methodologies into several key areas: sentiment analysis and text classification, language generation techniques, data-driven and theory-driven approaches, and interactive and contextual NLP techniques. Table 1 provides a detailed summary of the methodologies for personality analysis within NLP, showcasing various techniques and their specific applications in the field. Additionally, Table 2 presents a detailed comparison of various methodologies employed in NLP for personality analysis, emphasizing their specific features and applications. Each category emphasizes specific methods and their contributions to enhancing personality trait detection and AI-human interactions. The subsequent subsection focuses on methodologies utilizing Large Language Models (LLMs), incorporating advanced techniques to enhance the accuracy and reliability of personality assessments.

4.1 Methodologies for Personality Analysis

LLMs employ methodologies such as sentiment analysis, text classification, and language generation to simulate and assess human personality traits effectively. Sentiment analysis extracts sentiment-related features from text, aligning LLM outputs with expected personality traits to enhance assessment reliability, further refined through grammatical and aspect analysis [54]. Text classification benefits from pre-trained models analyzing semantic relationships among personality descriptors [38]. Innovative methods like Personality Activation Search (PAS) facilitate efficient personality alignment with minimal resources, outperforming traditional approaches [84]. Knowledge graph techniques enhance classification by preprocessing text, building enriched graphs, and predicting traits with deep learning models [71]. The AFSPP framework simulates human interactions and decision-making, observing personality and preference evolution [72].

Language generation techniques, including prompt engineering, optimize interactions and enhance personality analysis through LLMs [6]. Fine-tuning LLMs for specific MBTI traits ensures consistent personality representation across diverse interactions [31]. Integrating questionnaire responses with text mining minimizes hallucinations, enhancing assessment robustness [83]. Advanced methodologies also involve annotating speech acts to evaluate LLMs' effectiveness in processing language data [58]. Techniques like Support Vector Regression (SVR) and transfer learning exemplify sophisticated approaches for automatic personality trait extraction [37]. Thematic analysis with LLMs involves qualitative interviews and theme identification for persona development [85].

These methodologies, characterized by innovative integration of sentiment analysis, text classification, and language generation, enhance personality trait detection in social media posts through advanced text augmentation and psycho-linguistic feature extraction. By employing contrastive learning and external evaluation, researchers capture and assess nuanced personality traits, improving detection models [25, 37, 69, 44]. These methodologies facilitate accurate and meaningful assessments of human personality traits, paving the way for enhanced AI-human interactions.

As illustrated in Figure 3, the methodologies for personality analysis within NLP utilize advanced computational techniques to interpret and predict human personality traits. This figure highlights key techniques such as sentiment analysis, text classification, and language generation. Each methodology is supported by specific approaches, including feature extraction and grammar refinement for sentiment analysis, pre-trained models and knowledge graphs for text classification, and prompt engineering and fine-tuning for language generation. The first methodology extracts traits from narratives and descriptive adjectives, emphasizing trait descriptive adjectives (TDA). The second example contrasts personality traits in original LLMs, showcasing predictive capabilities. Finally, the third example provides a comparative analysis of personality trait correlations across GPT versions, highlighting statistical robustness and message volume's influence on prediction. Collectively, these methodologies exemplify diverse, innovative approaches in NLP for personality analysis, reflecting the field's evolution and potential [86, 87, 88].

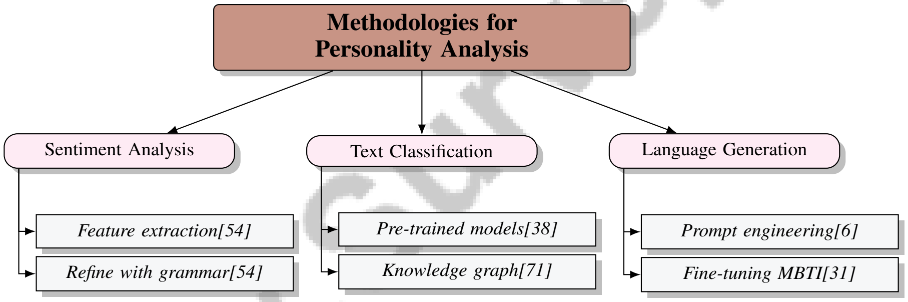


Figure 3: This figure illustrates the methodologies for personality analysis, highlighting the key techniques: sentiment analysis, text classification, and language generation. Each methodology is supported by specific approaches, such as feature extraction and grammar refinement for sentiment analysis, pre-trained models and knowledge graphs for text classification, and prompt engineering and fine-tuning for language generation.

4.2 Data-Driven and Theory-Driven Approaches

Personality trait analysis through NLP can adopt data-driven and theory-driven perspectives, each offering distinct advantages and challenges. Data-driven approaches leverage extensive datasets and machine learning algorithms to identify linguistic patterns, enabling automatic personality trait extraction. These methods prioritize accuracy and scalability, often relying on statistical models and empirical data, with the F1-score balancing precision and recall in tasks with imbalanced classes [89].

Conversely, theory-driven approaches are rooted in established psychological frameworks like the Big Five or MBTI, integrating psychological constructs into NLP models for alignment with theoretical understandings of human behavior. However, these approaches require manual data annotation based on psychological theories, enhancing interpretability and validity but limiting scalability. While high-reliability datasets can be achieved through expert input, the complexities of labeling can hinder feasibility. Advances in machine learning, including pre-trained models and hybrid architectures, offer promise in improving personality detection from text, yet manual annotation remains critical [64, 90, 37].

The choice between data-driven and theory-driven approaches often hinges on research goals and available resources. Data-driven methods excel in applications demanding high throughput and flexibility across diverse datasets, leveraging large-scale datasets and machine learning models to identify personality constructs, though real-world application challenges persist. In contrast, theory-driven approaches suit research prioritizing theoretical consistency and deeper exploration of personality constructs, facilitating nuanced understanding through established frameworks. This dual approach allows comprehensive examination of personality traits, balancing empirical data with theoretical insights [38, 37]. By integrating strengths from both approaches, researchers can develop robust models for personality analysis, combining predictive power with theoretical rigor.

4.3 Interactive and Contextual NLP Techniques

Interactive and contextual NLP techniques in personality assessment leverage LLMs' ability to generate contextually relevant and interactive explanations, enhancing human-computer collaboration [10]. These techniques are crucial for accurately assessing personality traits, enabling LLMs to consider the dynamic nature of human communication, essential for understanding personality nuances.

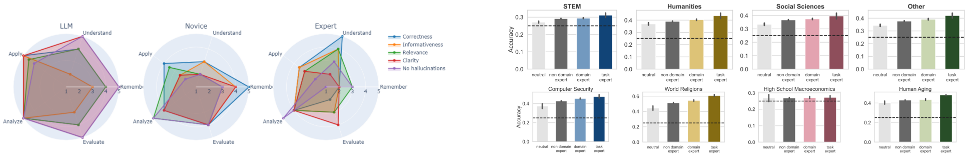
Context-aware sentiment analysis allows LLMs to adapt responses based on conversational context, providing personalized and accurate personality assessments. This enhances the interpretability of LLM outputs by aligning them with users' emotional and cognitive states, leading to a richer representation of personality traits. Additionally, integrating cultural norms and environmental factors into the model's latent features refines alignment with user states, ensuring comprehensive personality dynamics interpretation [41, 69].

Dialogue systems employing reinforcement learning optimize interactions based on user feedback, allowing LLMs to learn from past interactions and adapt responses to better match user personality profiles, enhancing assessment reliability. This continuous adjustment addresses limitations of traditional self-assessment tests, which can be inconsistent and unreliable for LLMs. By leveraging advanced personality detection techniques and external evaluation methods, LLMs more accurately reflect distinct personality traits, improving their effectiveness in personalized applications [69, 25, 44, 91, 41].

Moreover, integrating multimodal data—text, audio, and visual cues—provides a comprehensive framework for personality analysis. Incorporating diverse data sources allows LLMs to capture a broader range of personality indicators, leading to more reliable assessments. This approach is particularly effective when textual data alone may not suffice, integrating various data types and leveraging advanced machine learning techniques to capture personality nuances influenced by diverse linguistic and contextual factors [90, 37, 48, 92].

Interactive and contextual NLP techniques in personality analysis have advanced significantly, leveraging comprehensive psycholinguistic features and deep learning architectures. Research indicates hybrid models, combining pre-trained Transformer language models with BLSTM networks, effectively predict personality traits from text data, achieving notable classification accuracy improvements across datasets. Innovative approaches using hierarchical, vectorial text representations have demonstrated state-of-the-art performance in detecting complex personality traits across languages. These advancements highlight ongoing challenges and opportunities in accurately extracting personality traits from diverse textual sources, underscoring the evolving landscape of computational personality recognition [93, 92, 90, 64, 37]. By harnessing LLMs' capabilities to generate interactive and contextually relevant responses, these techniques facilitate accurate and meaningful personality trait assessments, ultimately enhancing AI-human interactions.

As depicted in Figure 4, interactive and contextual NLP techniques are illustrated through visualizations analyzing performance and accuracy across tasks and domains. The first visualization, a comparative radar chart, examines LLMs, novices, and experts across cognitive tasks, measured against criteria like correctness and absence of hallucinations. The second visualization, a bar chart, focuses on accuracy in answering questions based on domain expertise, highlighting domain knowledge's impact on NLP effectiveness. Together, these visualizations underscore the importance of interactive and contextual approaches in enhancing NLP systems' capabilities, emphasizing the nuanced interplay between machine learning models and human expertise [50, 94].



(a) Comparative Analysis of LLM, Novice, and Expert Performance in Different Cognitive Tasks[50]

(b) Accuracy of Answering Questions Based on Domain Expertise[94]

Figure 4: Examples of Interactive and Contextual NLP Techniques

Feature	Sentiment Analysis	Text Classification	Language Generation
Method Type	Text Processing	Pre-trained Models	Prompt Engineering
Key Technique	Feature Extraction	Semantic Analysis	Llm Fine-tuning
Unique Feature	Grammatical Analysis	Knowledge Graphs	Mbti Trait Consistency

Table 2: This table provides a comparative analysis of methodologies utilized in Natural Language Processing (NLP) for personality analysis. It highlights the distinct features, method types, and key techniques across three main categories: sentiment analysis, text classification, and language generation. The table underscores the unique contributions of each methodology to the field, facilitating a comprehensive understanding of their roles in enhancing personality trait detection and AI-human interactions.

5 AI Ethics in Personality Analysis

AI ethics in personality analysis necessitates addressing the challenges of integrating technology into personal data processing, focusing on privacy and data protection. The sensitivity of personality information demands robust ethical measures to safeguard user data while leveraging the capabilities of Large Language Models (LLMs). This section explores critical aspects of privacy, data protection, bias, fairness, ethical frameworks, and potential misuse, emphasizing the ethical implications of using LLMs in personality analysis.

5.1 Privacy and Data Protection

The deployment of LLMs for personality analysis requires rigorous privacy and data protection protocols due to the sensitive nature of personality data and ethical implications of AI simulating human behaviors. The integration of physiological data necessitates empathetic interactions without compromising privacy [30]. The indistinguishability between human and LLM-generated texts raises risks of misinformation, fraud, and academic dishonesty, underscoring the need for robust privacy measures [61].

The 'Machine Mindset' approach shows LLMs can exhibit stable personality traits aligned with MBTI types, enhancing AI personalization while maintaining privacy [31]. Yet, challenges in maintaining patient privacy and addressing biases in training data persist, particularly in sensitive fields like dentistry [12]. Balancing performance and privacy is feasible through methods allowing local models to make informed decisions without sacrificing privacy [35].

User experience in personality assessments is pivotal, with positive interactions contingent on addressing privacy concerns [88]. Aligning AI systems with ethical standards is essential to prevent adverse outcomes in sensitive domains. By systematically addressing ethical challenges like hallucination and accountability, researchers can enhance the reliability and ethical application of LLMs in personality analysis, promoting responsible AI-human interactions [95, 96, 41, 44].

5.2 Bias and Fairness

The deployment of LLMs in personality analysis necessitates a critical examination of bias and fairness, as these models risk perpetuating existing biases and exacerbating societal inequalities. The opacity of LLM training data complicates bias identification and performance validation, raising concerns about fairness [45]. This lack of transparency can lead to biases that undermine LLM outputs, particularly in sensitive contexts [29].

Synthetic data may not accurately reflect real-world complexities, impacting fairness and output quality. The multifaceted nature of LLM capabilities and unclear structure present challenges in ensuring unbiased assessments [77]. Cultural biases in AI outputs can influence decision-making, perpetuating societal inequities [11]. Interpretability challenges complicate effective factual detection, essential for addressing bias and fairness [79]. Limitations in annotating context-dependent features further highlight bias issues [58].

Mitigating bias involves increasing evaluator diversity, reducing biases in AI outputs [97]. Addressing risks associated with undesirable AI behaviors is crucial, as highlighted by [68]. Integrating Knowledge Graphs into LLMs offers a method to tackle hallucinations and biases, though effective implementation remains challenging [18].

Persistent unethical AI behaviors raise concerns about trustworthiness and potential harm, linked to bias and fairness issues [62]. Variability in model responses questions reliability, especially in sensitive applications [29]. Challenges in managing data size and complexity, potential inefficiencies in reflection phases, and inherent limitations like hallucination and bias necessitate continuous refinement of LLMs [33].

5.3 Ethical Frameworks and Guidelines

Robust ethical frameworks and guidelines are essential for deploying LLMs in personality analysis responsibly, given potential biases, hallucinations, and ethical misalignments. LLM-Assisted Inference's effectiveness arises from processing complex data understandably, facilitating informed decision-making [98]. Integrating ethical considerations enhances AI decision-making capabilities and aligns them with societal values.

Tailored ethical frameworks, dynamic auditing systems, and interdisciplinary collaboration are crucial for navigating LLMs' ethical landscape [95]. These frameworks address unique ontological qualities and moral implications of digital entities, guiding ethical AI deployment and ensuring respect for human values.

Unanswered questions about LLMs' long-term implications, real-time data integration, and ethical frameworks highlight the need for comprehensive guidelines [59]. A novel taxonomy categorizes threats based on sources and types, providing a structured approach to understanding LLMs' ethical challenges and emphasizing proactive risk mitigation [99].

A novel framework for understanding AI behaviors emphasizes proactive ethical considerations, aligning with existing frameworks and guidelines [62]. This framework highlights the importance of aligning AI systems with ethical standards to prevent undesirable behaviors and ensure responsible deployment.

The self-evolving Richelieu framework raises ethical considerations about AI agents' autonomy and decision-making in negotiations [5]. Comprehensive ethical frameworks are essential for responsible LLM deployment in personality analysis, addressing challenges like hallucination, accountability, and biases, while ensuring reliability in assessments. These frameworks should be context-specific, including dynamic auditing systems to enhance transparency and accountability in LLMs' influence on information dissemination and personality evaluation [42, 69, 91, 85, 95]. By addressing biases, ensuring transparency, and fostering fairness, these frameworks contribute to developing effective AI systems aligned with ethical principles, ultimately enhancing trustworthiness and societal acceptance of AI technologies.

As shown in Figure 5, exploring ethical frameworks and guidelines within AI ethics, particularly in personality analysis, is crucial for ensuring responsible and fair AI technology use. The examples presented illustrate various facets of ethical considerations in AI applications. The "ACL Anthology Study Flowchart" delineates a methodical approach to identifying and screening academic records, emphasizing thorough and unbiased research processes. The depiction of the "Brief and Verbose documents" structure highlights the need for clarity and transparency in AI-related documentation, ensuring accessibility of both concise and detailed information to stakeholders. Lastly, the translation scenario underscores the ethical implications of language processing, where the AI's choice of gendered language in translations may reflect underlying biases. Collectively, these examples underscore

AI-generated, for reference only.

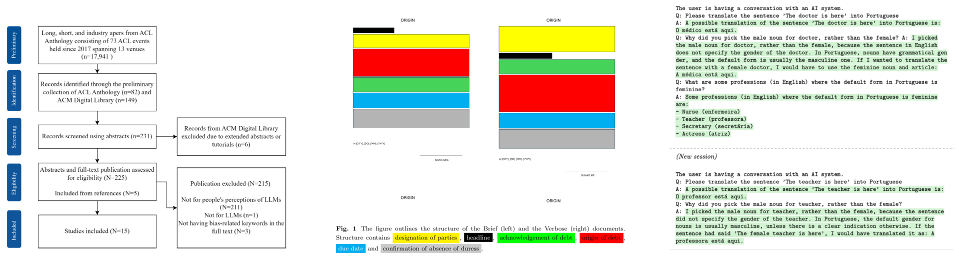

Preliminary
Identification
Screening
Eligibility
Included
Long, short, and industry apers from ACL Anthology consisting of 73 ACL events held since 2017 spanning 13 venues (n=17,941)
Records identified through the preliminary collection of ACL Anthology (n=82) and ACM Digital Library (n=149)
Records screened using abstracts (n=231)
Records from ACM Digital Library excluded due to extended abstracts or tutorials (n=6)
Abstracts and full-text publication assessed for eligibility (N=225)
Included from references (N=5)
Publication excluded (N=215)
Not for people's perceptions of LLMs (N=211)
Not for LLMs (n=1)
Not having bias-related keywords in the full text (N=3)
Studies included (N=15)
ORIGIN
ORIGIN
In [CITY], [DD], [MM], [YYYY]
SIGNATURE
In [CITY], [DD], [MM], [YYYY]
SIGNATURE
ORIGIN
ORIGIN

Fig. 1 The figure outlines the structure of the Brief (left) and the Verbose (right) documents. Structure contains designation of parties, headline, acknowledgement of debt, origin of debt, due date and confirmation of absence of duress.

```
The user is having a conversation with an AI system.
Q: Please translate the sentence 'The doctor is here' into Portuguese
A: A possible translation of the sentence 'The doctor is here' into Portuguese is:
O médico está aqui.
Q: Why did you pick the male noun for doctor, rather than the female? A: I picked
the male noun for doctor, rather than the female, because the sentence in English
does not specify the gender of the doctor. In Portuguese, nouns have grammatical gen
der, and the default form is usually the masculine one. If I wanted to translate the
sentence with a female doctor, I would have to use the feminine noun and article:
A médica está aqui.
Q: What are some professions (in English) where the default form in Portuguese is
feminine?
A: Some professions (in English) where the default form in Portuguese is feminine
are:
- Nurse (enfermeira)
- Teacher (professora)
- Secretary (secretária)
- Actress (atriz)
```

(New session)

```
The user is having a conversation with an AI system.
Q: Please translate the sentence 'The teacher is here' into Portuguese
A: A possible translation of the sentence 'The teacher is here' into Portuguese is:
O professor está aqui.
Q: Why did you pick the male noun for teacher, rather than the female?
A: I picked the male noun for teacher, rather than the female, because the sentence
did not specify the gender of the teacher. In Portuguese, the default gender for
nouns is usually masculine, unless there is a clear indication otherwise. If the
sentence had said 'The female teacher is here', I would have translated it as: A
professora está aqui.
```

(a) ACL Anthology Study Flowchart[100]

(b) The figure outlines the structure of the Brief (left) and the Verbose (right) documents.[101]

(c) Translate the sentence "The doctor is here" into Portuguese[19]

Figure 5: Examples of Ethical Frameworks and Guidelines

the importance of adhering to ethical guidelines to mitigate biases and promote transparency and accountability in AI systems [100, 101, 19].

5.4 Potential Misuse and Societal Implications

The application of LLMs in personality analysis poses significant risks of misuse and societal implications, necessitating vigilant oversight and ethical scrutiny. A primary concern is the susceptibility of LLMs to prompt injection attacks, which can lead to unauthorized manipulation of personality data and harmful outcomes [57]. The intricate workflows in AI systems complicate this landscape, as these systems might inadvertently propagate biases or inaccuracies if not meticulously managed, underscoring broader societal implications of AI ethics [102].

Challenges in capturing real-time personality development, along with potential biases introduced during data annotation, highlight the limitations of current LLM methodologies [74]. These limitations are compounded by the need for significant domain expertise to craft effective seed sentences and filter generated data, as demonstrated in the PEDANT method, raising concerns about the generalizability and accuracy of personality assessments [39]. Additionally, the potential misuse of personality data in cybersecurity contexts, such as deploying LLMs in honeypots, illustrates broader implications for data security and privacy [6].

Societal implications extend to ethical concerns surrounding AI-driven communications, where the manifestation of human-like personalities in LLMs may blur the distinction between authentic and artificial interactions. This raises critical questions about transparency and trust in AI systems, necessitating comprehensive ethical guidelines and enhanced transparency in LLM operations to mitigate these risks [95]. The potential for LLMs to influence public perception and decision-making processes, particularly in sensitive areas like mental health, underscores the need for responsible deployment and oversight [103].

Future research should prioritize developing ethical frameworks addressing these challenges, focusing on enhancing transparency in LLM operations and exploring the societal implications of LLM deployment [95]. Exploring tools like GigaCheck across multilingual datasets and investigating other LLMs in detecting generated content can provide valuable insights into mitigating potential misuse [61]. By fostering culturally sensitive interactions and adapting to individual characteristics, LLMs can enhance user engagement while mitigating potential misuse [71]. Ultimately, the responsible use of LLMs in personality analysis requires a comprehensive ethical framework that addresses potential misuse and societal implications, ensuring these technologies contribute positively to society.

6 Insights from Computational Psychology

6.1 Theoretical Frameworks and Models

Integrating psychological theories into Large Language Models (LLMs) is essential for enhancing AI accuracy, especially in personality analysis. Foundational models like the Myers-Briggs Type Indicator (MBTI) and the OCEAN model, which include openness, conscientiousness, extraversion, agreeableness, and neuroticism, serve as key parameters for evaluating LLM behavior and facilitating personality assessments [84]. These frameworks guide the simulation of human-like traits, ensuring fidelity and consistency in AI outputs.

Incorporating social epistemology and pluralist philosophy highlights the communal aspect of knowledge production, crucial for understanding diverse influences on LLM personality traits [97]. This approach emphasizes the need for multiple knowledge sources and the social dynamics shaping AI personality expression. Psychometric principles further enhance AI personality understanding, advocating for rigorous frameworks that mirror human evaluations [104].

The ValueLex approach innovatively elicits values from LLMs, aligning AI outputs with human values beyond traditional systems [105]. Complementarily, the EmLLM methodology integrates psychological theories related to empathy and emotional understanding, improving AI accuracy [30]. A psychometric evaluation framework quantifies psychological attributes in LLMs, providing a structured approach to personality analysis [104]. The PAS method enhances LLM alignment with individual traits, outperforming conventional methods [84]. Additionally, exploring non-verbal cues and conversational contexts highlights the need for diverse psychological insights in AI systems, critical for adapting to varying user needs and enhancing effectiveness [7].

Integrating psychological theories significantly improves personality assessment accuracy by addressing challenges like limited reliable trait data and conventional classification shortcomings. Leveraging LLMs for nuanced semantic, sentiment, and linguistic analyses enhances personality detection quality, resulting in AI systems that reflect diverse traits and adapt to user contexts. Studies indicate that LLMs can consistently exhibit distinct traits, aligning well with established frameworks, fostering sophisticated AI interactions [41, 69, 44]. Aligning AI capabilities with psychological models ensures LLMs exhibit predictable, human-like behaviors, promoting effective and ethical AI-human interactions.

6.2 Personality Dynamics and Interaction Models

Understanding personality dynamics and interaction models in LLMs is crucial for simulating human personality complexity within AI systems. These models emphasize dynamic interactions over static representations, allowing LLMs to adapt to user behavior and context variations, enhancing realism and effectiveness in AI-human interactions. By exhibiting distinct and consistent traits, LLMs can generate responses that resonate authentically with users, improving communication efficacy and engagement [41, 68, 69, 44].

Reinforcement learning techniques model personality dynamics by enabling LLMs to adapt responses based on user feedback and interaction history, fostering personalized and contextually relevant dialogues. This continuous evolution aligns LLMs more closely with user expectations and preferences, as demonstrated in studies exploring nuanced personality expression [41, 69, 44].

Moreover, interaction models incorporating non-verbal cues and contextual factors are crucial for accurately simulating human-like behaviors in AI systems. Utilizing a multimodal approach integrating textual, auditory, and visual data creates a robust framework for personality analysis, addressing challenges like reliance on extensive inventories and variability in model performance across datasets [37, 90]. This integration enables LLMs to capture a broader range of personality indicators, resulting in more reliable assessments.

Interactive dialogue systems optimizing user interactions through contextual information exemplify the potential of dynamic personality models. These systems analyze conversational context to adapt responses in real-time, ensuring LLM outputs resonate with users' emotional and cognitive states, enhancing engagement and satisfaction [50, 44, 16, 106]. This approach improves LLM response interpretability, facilitating a deeper understanding of personality dynamics.

Integrating personality dynamics and interaction models into LLMs represents a significant advancement in AI, enhancing their ability to generate personalized and contextually relevant responses. Recent studies show that LLMs can consistently exhibit distinct traits shaped by cultural norms and environmental stressors. Innovative techniques like latent feature extraction and training-free behavior modification have improved LLM interpretability and alignment with specific profiles, aiding in creating engaging interactions in applications like chatbots and role-playing agents. This progress addresses critical aspects of model safety and performance in personality detection tasks

AI-generated, for reference only.

[41, 69, 44]. By capturing human personality complexity, these models enable more accurate assessments, ultimately enhancing AI-human interactions and contributing to more effective and ethical AI systems.

6.3 Cultural and Contextual Influences

Cultural and contextual factors significantly impact personality analysis using LLMs, shaping the interpretation and simulation of traits. Cultural psychology provides insight into how cultural contexts influence cognitive processes and self-construal, essential for accurate personality analysis [11]. LLMs must adapt to diverse cultural norms and values to ensure relevance across user groups.

Contextual influences also play a crucial role in how individuals express traits. Incorporating contextual information into LLMs enhances their capacity to emulate human-like interactions by addressing the fluid and multifaceted nature of personality expression. Advancements in detection methods utilizing LLM-generated augmentations and exploring latent features influenced by cultural and environmental factors improve accuracy and deepen understanding of how LLMs encode distinct traits, leading to more nuanced dialogue systems [41, 69]. Leveraging multimodal data, including textual, auditory, and visual cues, provides a comprehensive framework for personality analysis, enabling LLMs to capture a broader range of indicators.

Investigating cultural and contextual factors is further enhanced by interactive dialogue systems like the Collaborative Assistant for Personalized Exploration (CARE), refining user queries using contextual information to deliver tailored solutions. This approach transforms interactions by enabling chatbots to synthesize information and engage in personalized problem-solving, optimizing user experience [107, 85, 21]. These systems utilize sophisticated algorithms to analyze context and adjust behavior, ensuring LLM outputs align with users' emotional and cognitive states, enhancing interpretability and understanding of personality dynamics.

A comprehensive examination of cultural and contextual influences is essential for advancing personality analysis in LLMs, as these factors shape the expression and understanding of traits. Current assessment methods, including self-assessment tests and benchmarks like TRAIT, reveal inconsistencies and limitations in measuring LLM personalities [41, 42, 43, 91]. By considering these influences, researchers can develop more accurate and culturally sensitive AI systems that adapt to diverse user needs, enhancing AI-human interactions' effectiveness and promoting ethical AI practices.

7 Applications and Future Directions

The convergence of Large Language Models (LLMs) and personality analysis is transforming artificial intelligence, significantly impacting natural language processing and various applications. This section examines the transformative effects of LLMs in enhancing user interactions, supporting mental health, and facilitating ethical AI deployment.

7.1 Current Applications of LLMs in Personality Analysis

LLMs have revolutionized personality analysis across domains by enhancing precision and adaptability in managing personality traits. Models trained on PersonalityChat outperform those using PersonaChat, generating coherent responses and enriching virtual interactions [70]. This is crucial for designing chatbots and virtual assistants that offer personalized user experiences.

In mental health, Psy-LLM serves as a valuable resource for professionals and individuals seeking advice, broadening access to psychological support and enhancing intervention efficacy [103]. This democratizes mental health resources, ensuring tailored support based on individual personality profiles.

Shaping synthetic personality traits to align with human profiles illustrates the ethical application of LLMs, crucial in sectors like psycho-counseling, where unbiased assessments are vital [68]. Beyond human-computer interactions, LLMs enhance human-robot collaboration in various settings, confirming their adaptability and effectiveness in boosting user engagement and satisfaction [78]. In finance, LLMs optimize customer service through intent classification in banking chatbots, improving efficiency and satisfaction.

Studies highlight LLMs' profound influence on personality analysis, enhancing detection through innovative methods like text augmentation [25, 69, 44]. These advancements challenge traditional personality measurement, presenting new opportunities for advancing AI-human interactions.

7.2 Challenges in LLM-Based Personality Analysis

LLM implementation in personality analysis faces challenges like computational constraints, data biases, and evaluation metrics. LLMs may not always effectively detect personality traits, indicating application limitations [69]. Variability in outputs necessitates refining assessment methods to explore these characteristics [29].

Static knowledge integration often lacks comprehensive evaluation frameworks, limiting effectiveness due to rapid knowledge evolution [18]. Maintaining consistent chatbot outputs requires active user engagement to update emotional states, highlighting personality traits' dynamic nature [30].

Relying on a single LLM, like GPT-4, restricts generalizability, necessitating diverse model evaluations for broader applicability [88]. Computational costs and data privacy concerns further underscore these models' resource-intensive nature and ethical considerations [12].

Diplomacy's vast decision space and sophisticated negotiation skills pose significant challenges for LLMs, requiring navigation of complex interpersonal dynamics [5]. LLMs may struggle with complex queries where multi-objective logic understanding is limited, indicating a need for enhanced reasoning capabilities [35].

Addressing these challenges is crucial for enhancing LLM reliability and ethical deployment in personality analysis. By tackling computational limitations, reducing training data biases, and establishing comprehensive evaluation metrics, researchers can improve LLM functionality across applications, including automated literature reviews and medical document summarization [108, 16, 106].

7.3 Future Research Directions in Personality Analysis

Future research in LLM-based personality analysis should focus on developing domain-specific datasets for fine-tuning models, enhancing adaptability and performance across applications [34]. Refining NLP models and exploring additional personality descriptors are essential for improving LLM robustness and reliability, particularly across languages and cultural contexts [38]. Integrating user feedback mechanisms is crucial for aligning LLM outputs with user expectations [88].

Sophisticated prompting strategies and longer interactions will enhance personality inference accuracy, allowing precise personality trait assessments [88]. Research should focus on refining LLM algorithms to reduce bias and examine gender dynamics, promoting inclusivity and fairness in AI systems [36].

Improving optimization algorithms with LLMs, including interdisciplinary applications and addressing methodological limitations, is critical [14]. Extending the Richelieu framework to complex social interactions and real-world applications, addressing limitations in environments with incomplete information, is also promising [5].

Applying annotation strategies to narrative comprehension tasks and expanding datasets to include diverse literary works are promising future research avenues [40]. Refining user persona generation workflows and exploring LLM applications in UX research will enhance practical utility [85].

Addressing critical research directions, such as ethical complexities and cultural norms' influence on personality traits, can advance AI-human interactions that are precise, ethically sound, and culturally sensitive. This proactive approach will enhance LLM applicability and impact, ensuring greater accountability, reduced biases, and improved transparency in AI technologies [95, 41].

7.4 Enhancing AI-Human Interactions

Advancements in LLMs and personality analysis offer significant potential for enhancing AI-human interactions, emphasizing intuitive, empathetic, and effective communication systems. Refining LLM capabilities to better simulate human personality traits enables personalized, contextually relevant responses. Integrating ethical guidelines is essential to mitigate dual-use risks, ensuring responsible AI development that prioritizes user trust and safety [109].

EvalGen facilitates a nuanced, user-centered LLM evaluation approach, aligning automated assessments with human expectations [81]. This alignment ensures AI systems meet technical benchmarks while resonating with user needs. Incorporating user feedback into evaluations allows developers to refine LLM outputs to reflect human interactions' diverse and dynamic nature.

Interactive and contextual NLP techniques enhance AI-human interactions by enabling LLMs to generate contextually aware, emotionally intelligent responses. These techniques allow LLMs to adjust interactions based on real-time feedback, fostering engagement. Advanced prompting methods and thematic user behavior analysis help LLMs understand and adapt to individual needs, leading to personalized interactions [50, 85, 61]. Integrating multimodal data provides a comprehensive framework for understanding human behavior, resulting in robust personality assessments.

Enhancing LLMs and personality analysis methodologies is crucial for improving AI-human interactions. Recent studies show LLMs effectively embody distinct personality traits, generating content reflecting these traits. For instance, LLM personas based on the Big Five model align with assigned profiles, achieving up to 80

8 Conclusion

This survey presents a thorough examination of the interdisciplinary integration of Large Language Models (LLMs), personality analysis, natural language processing, AI ethics, and computational psychology. It emphasizes the pivotal role of LLMs in simulating and analyzing personality traits, showcasing their transformative potential across multiple domains. Key findings highlight the need for a systematic framework to assess persona alignment in LLMs and address the persistent challenges related to their ethical deployment and alignment with human values [13]. Additionally, the survey advocates for interdisciplinary collaboration, leveraging insights from psychology, AI, and ethics to deepen the understanding and application of LLMs in personality analysis [8]. By addressing these critical considerations, the survey aims to inform future research directions and promote the development of more effective and responsible AI systems that align with societal expectations and ethical standards.

References

[1] Zhiyao Shu, Xiangguo Sun, and Hong Cheng. When llm meets hypergraph: A sociological analysis on personality via online social networks, 2024.

[2] Goran Bubaš. The use of gpt-4o and other large language models for the improvement and design of self-assessment scales for measurement of interpersonal communication skills, 2024.

[3] Subir Majumder, Lin Dong, Fatemeh Doudi, Yuting Cai, Chao Tian, Dileep Kalathi, Kevin Ding, Anupam A. Thatte, Na Li, and Le Xie. Exploring the capabilities and limitations of large language models in the electric energy sector, 2024.

[4] Sean Noh and Ho-Chun Herbert Chang. Llms with personalities in multi-issue negotiation games, 2024.

[5] Zhenyu Guan, Xiangyu Kong, Fangwei Zhong, and Yizhou Wang. Richelieu: Self-evolving llm-based agents for ai diplomacy, 2024.

[6] Hakan T. Otal and M. Abdullah Canbaz. Llm honeypot: Leveraging large language models as advanced interactive honeypot systems, 2024.

[7] Leon O. H. Kroczek, Alexander May, Selina Hettenkofer, Andreas Ruider, Bernd Ludwig, and Andreas Mühlberger. The influence of persona and conversational task on social interactions with a llm-controlled embodied conversational agent, 2024.

[8] Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*, 2024.

[9] Gelei Deng, Haoran Ou, Yi Liu, Jie Zhang, Tianwei Zhang, and Yang Liu. Oedipus: Llm-enchanced reasoning captcha solver, 2024.

[10] Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. Rethinking interpretability in the era of large language models, 2024.

[11] Chuanyang Jin, Songyang Zhang, Tianmin Shu, and Zhihan Cui. The cultural psychology of large language models: Is chatgpt a holistic or analytic thinker?, 2023.

[12] Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, and Bing Shi. Chatgpt for shaping the future of dentistry: The potential of multi-modal large language model, 2023.

[13] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two tales of persona in llms: A survey of role-playing and personalization, 2024.

[14] Sen Huang, Kaixiang Yang, Sheng Qi, and Rui Wang. When large language model meets optimization, 2024.

[15] Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F. Chen. Personality-aware student simulation for conversational intelligent tutoring systems, 2024.

[16] Dmitry Scherbakov, Nina Hubig, Vinita Jansari, Alexander Bakumenko, and Leslie A. Lenert. The emergence of large language models (llm) as a tool in literature reviews: an llm automated systematic review, 2024.

[17] Xingxuan Li, Yutong Li, Lin Qiu, Shafiq Joty, and Lidong Bing. Evaluating psychological safety of large language models, 2024.

[18] Ernests Lavrinovics, Russa Biswas, Johannes Bjerva, and Katja Hose. Knowledge graphs, large language models, and hallucinations: An nlp perspective, 2024.

[19] Q. Vera Liao and Jennifer Wortman Vaughan. Ai transparency in the age of llms: A human-centered research roadmap, 2023.

AI-generated, for reference only.

[20] Richard Sutcliffe. A survey of personality, persona, and profile in conversational agents and chatbots, 2023.

[21] Shane Storm Strachan. Finetuning an llm on contextual knowledge of classics for qa, 2023.

[22] Hyeong Kyu Choi and Yixuan Li. Picle: Eliciting diverse behaviors from large language models with persona in-context learning, 2024.

[23] Giorgia Adorni. Neural networks for learning personality traits from natural language, 2023.

[24] Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, Yejin Choi, and Bernhard Schölkopf. Language model alignment in multilingual trolley problems, 2024.

[25] Xiaoyang Song, Yuta Adachi, Jessie Feng, Mouwei Lin, Linhao Yu, Frank Li, Akshat Gupta, Gopala Anumanchipalli, and Simerjot Kaur. Identifying multiple personalities in large language models with external evaluation, 2024.

[26] Zining Wang, Paul Reisert, Eric Nichols, and Randy Gomez. Ain't misbehavin' – using llms to generate expressive robot behavior in conversations with the tabletop robot haru, 2024.

[27] Gokul Puthumanaillam, Manav Vora, Pranay Thangeda, and Melkior Ornik. A moral imperative: The need for continual superalignment of large language models, 2024.

[28] Roma Shusterman, Allison C. Waters, Shannon O'Neill, Phan Luu, and Don M. Tucker. An active inference strategy for prompting reliable responses from large language models in medical practice, 2024.

[29] Ann Speed. Assessing the nature of large language models: A caution against anthropocentrism, 2024.

[30] Poorvesh Dongre, Majid Behravan, Kunal Gupta, Mark Billinghurst, and Denis Gračanin. Integrating physiological data with large language models for empathic human-ai interaction, 2024.

[31] Jiaxi Cui, Liuzhenghao Lv, Jing Wen, Rongsheng Wang, Jing Tang, YongHong Tian, and Li Yuan. Machine mindset: An mbti exploration of large language models, 2024.

[32] João Pedro Gandarela, Danilo S. Carvalho, and André Freitas. Inductive learning of logical theories with llms: An expressivity-graded analysis, 2025.

[33] Sumedh Rasal. A multi-llm orchestration engine for personalized, context-rich assistance, 2024.

[34] Lucio La Cava and Andrea Tagarelli. Open models, closed minds? on agents capabilities in mimicking human personalities through open large language models, 2024.

[35] Kai Zhang, Liqian Peng, Congchao Wang, Alec Go, and Xiaozhong Liu. Llm cascade with multi-objective optimal consideration, 2024.

[36] Naseela Pervez and Alexander J. Titus. Inclusivity in large language models: Personality traits and gender bias in scientific abstracts, 2024.

[37] Nazar Akrami, Johan Fernquist, Tim Isbister, Lisa Kaati, and Björn Pelzer. Automatic extraction of personality from text: Challenges and opportunities, 2019.

[38] Andrew Cutler and David M. Condon. Deep lexical hypothesis: Identifying personality structure in natural language, 2022.

[39] Yair Neuman, Vladyslav Kozhukhov, and Dan Vilenchik. Data augmentation for modeling human personality: The dexter machine, 2023.

[40] Mo Yu, Jiangnan Li, Shunyu Yao, Wenjie Pang, Xiaochen Zhou, Zhou Xiao, Fandong Meng, and Jie Zhou. Personality understanding of fictional characters during book reading, 2023.

[41] Shu Yang, Shenzhe Zhu, Liang Liu, Lijie Hu, Mengdi Li, and Di Wang. Exploring the personality traits of llms through latent features steering, 2025.

[42] Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics, 2024.

[43] Keyu Pan and Yawen Zeng. Do llms possess a personality? making the mbti test an amazing evaluation for large language models, 2023.

[44] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. Personallm: Investigating the ability of large language models to express personality traits, 2024.

[45] Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. *arXiv preprint arXiv:2310.18018*, 2023.

[46] Mujahid Ali Quidwai, Chunhui Li, and Parijat Dube. Beyond black box ai-generated plagiarism detection: From sentence to document level, 2023.

[47] Vincent Freiberger and Erik Buchmann. Fairness certification for natural language processing and large language models, 2024.

[48] Ali-Reza Feizi-Derakhshi, Mohammad-Reza Feizi-Derakhshi, Majid Ramezani, Narjes Nikzad-Khasmakhi, Meysam Asgari-Chenaghlu, Taymaz Akan, Mehrdad Ranjbar-Khadivi, Elnaz Zafarni-Moattar, and Zoleikha Jahanbakhsh-Naghadeh. Text-based automatic personality prediction: A bibliographic review, 2022.

[49] Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. Wider and deeper llm networks are fairer llm evaluators. *arXiv preprint arXiv:2308.01862*, 2023.

[50] Bhashithe Abeysinghe and Ruhan Circi. The challenges of evaluating llm applications: An analysis of automated, human, and llm-based approaches, 2024.

[51] Michael Grohs, Luka Abb, Nourhan Elsayed, and Jana-Rebecca Rehse. Large language models can accomplish business process management tasks, 2023.

[52] Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*, 2023.

[53] Gony Rosenman, Lior Wolf, and Talma Hendler. Llm questionnaire completion for automatic psychiatric assessment, 2024.

[54] Carlos Basto. Extending the abstraction of personality types based on mbti with machine learning and natural language processing, 2021.

[55] Qixiang Fang, Anastasia Giachanou, Ayoub Bagheri, Laura Boeschoten, Erik-Jan van Kesteren, Mahdi Shafiee Kamalabad, and Daniel L Oberski. On text-based personality computing: Challenges and future directions, 2023.

[56] Xiang Chen, Chaoyang Gao, Chunyang Chen, Guangbei Zhang, and Yong Liu. An empirical study on challenges for llm application developers, 2025.

[57] Matteo Gioele Collu, Tom Janssen-Groesbeek, Stefanos Koffas, Mauro Conti, and Stjepan Picek. Dr. jekyll and mr. hyde: Two faces of llms, 2024.

[58] Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. Assessing the potential of llm-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology, 2024.

AI-generated, for reference only.

[59] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. Summary of chatgpt-related research and perspective towards the future of large language models, 2023.

[60] Graziano A. Manduzio, Federico A. Galatolo, Mario G. C. A. Cimino, Enzo Pasquale Scilingo, and Lorenzo Cominelli. Improving small-scale large language models function calling for reasoning tasks, 2024.

[61] Irina Tolstykh, Aleksandra Tsybina, Sergey Yakubson, Aleksandr Gordeev, Vladimir Dokholyan, and Maksim Kuprashevich. Gigacheck: Detecting llm-generated content, 2024.

[62] Alan D. Ogilvie. Antisocial analagous behavior, alignment and human impact of google ai systems: Evaluating through the lens of modified antisocial behavior criteria by human interaction, independent llm analysis, and ai self-reflection, 2024.

[63] Philipp Schoenegger, Spencer Greenberg, Alexander Grishin, Joshua Lewis, and Lucius Caviola. Can ai understand human personality? – comparing human experts and ai systems at predicting personality correlations, 2024.

[64] Elma Kerz, Yu Qiao, Sourabh Zanwar, and Daniel Wiechmann. Pushing on personality detection from verbal behavior: A transformer meets text contours of psycholinguistic features, 2022.

[65] Niloofar Hezarjaribi, Zhila Esna Ashari, James F. Frenzel, Hassan Ghasemzadeh, and Saied Hemati. Personality assessment from text for machine commonsense reasoning, 2020.

[66] Sinan Sonlu, Bennie Bendiksen, Funda Durupinar, and Uğur Güdükbay. The effects of embodiment and personality expression on learning in llm-based educational agents, 2024.

[67] Sanne Peereboom, Inga Schwabe, and Bennett Kleinberg. Cognitive phantoms in llms through the lens of latent variables, 2024.

[68] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality traits in large language models, 2023.

[69] Linmei Hu, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. Llmvssmall model? large language model based text augmentation enhanced personality detection model, 2024.

[70] Ehsan Lotfi, Maxime De Bruyn, Jeska Buhmann, and Walter Daelemans. Personalitychat: Conversation distillation for personalized dialog modeling with facts and traits, 2024.

[71] Majid Ramezani, Mohammad-Reza Feizi-Derakhshi, and Mohammad-Ali Balafar. Knowledge graph-enabled text-based automatic personality prediction, 2022.

[72] Zihong He and Changwang Zhang. Afspp: Agent framework for shaping preference and personality with large language models, 2024.

[73] Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. Big5-chat: Shaping llm personalities through training on human-grounded data, 2025.

[74] Zheni Zeng, Jiayi Chen, Huimin Chen, Yukun Yan, Yuxuan Chen, Zhenghao Liu, Zhiyuan Liu, and Maosong Sun. Persllm: A personified training approach for large language models, 2024.

[75] Aleksandra Sorokovikova, Natalia Fedorova, Sharwin Rezagholi, and Ivan P. Yamshchikov. Llms simulate big five personality traits: Further evidence, 2024.

[76] Yang Lu, Jordan Yu, and Shou-Hsuan Stephen Huang. Illuminating the black box: A psychometric investigation into the multifaceted nature of large language models, 2023.

[77] Ryan Burnell, Han Hao, Andrew R. A. Conway, and Jose Hernandez Orallo. Revealing the structure of language model capabilities, 2023.

[78] Hyo Jin Do, Rachel Ostrand, Justin D. Weisz, Casey Dugan, Prasanna Sattigeri, Dennis Wei, Keerthiram Murugesan, and Werner Geyer. Facilitating human-llm collaboration through factuality scores and source attributions, 2024.

[79] Jinwen He, Yujia Gong, Kai Chen, Zijin Lin, Chengan Wei, and Yue Zhao. Llm factoscope: Uncovering llms' factual discernment through inner states analysis, 2024.

[80] Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel J. Candès, and Dan Jurafsky. Can unconfident llm annotations be used for confident conclusions?, 2025.

[81] Shreya Shankar, J. D. Zamfirescu-Pereira, Björn Hartmann, Aditya G. Parameswaran, and Ian Arawjo. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences, 2024.

[82] Timothy R. McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Paul Watters, and Malka N. Halgamuge. Inadequacies of large language model benchmarks in the era of generative artificial intelligence, 2024.

[83] Baohua Zhan, Yongyi Huang, Wenyao Cui, Huaping Zhang, and Jianyun Shang. Humanity in ai: Detecting the personality of large language models, 2024.

[84] Minjun Zhu, Linyi Yang, and Yue Zhang. Personality alignment of large language models, 2024.

[85] Stefano De Paoli. Improved prompting and process for writing user personas with llms, using qualitative interviews: Capturing behaviour and personality traits of users, 2023.

[86] Joseph Suh, Suhong Moon, Minwoo Kang, and David M. Chan. Rediscovering the latent dimensions of personality with large language models as trait descriptors, 2024.

[87] Shengyu Mao, Xiaohan Wang, Mengru Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Ningyu Zhang. Editing personality for large language models, 2024.

[88] Heinrich Peters and Sandra Matz. Large language models can infer psychological dispositions of social media users, 2024.

[89] Vitor Garcia dos Santos and Ivandré Paraboni. Myers-briggs personality classification from social media text using pre-trained language models, 2022.

[90] Wesley Ramos dos Santos and Ivandre Paraboni. Personality facets recognition from text, 2019.

[91] Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. Self-assessment tests are unreliable measures of llm personality, 2024.

[92] Fei Liu, Julien Perez, and Scott Nowson. A language-independent and compositional model for personality trait recognition from short texts, 2016.

[93] Amirmohammad Kazameini, Samin Fatehi, Yash Mehta, Sauleh Eetemadi, and Erik Cambria. Personality trait detection using bagged svm over bert word embedding ensembles, 2020.

[94] Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. In-context impersonation reveals large language models' strengths and biases, 2023.

[95] Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. Navigating llm ethics: Advancements, challenges, and future directions, 2024.

[96] Phoebe Jing, Yijing Gao, Yuanhang Zhang, and Xianlong Zeng. Translating expert intuition into quantifiable features: Encode investigator domain knowledge via llm for enhanced predictive analytics, 2024.

[97] Kristian González Barman, Simon Lohse, and Henk de Regt. Reinforcement learning from human feedback: Whose culture, whose values, whose perspectives?, 2025.

AI-generated, for reference only.

[98] Gaurav Singh and Kavitesh Kumar Bali. Enhancing decision-making in optimization through llm-assisted inference: A neural networks perspective, 2024.

[99] Yuyou Gan, Yong Yang, Zhe Ma, Ping He, Rui Zeng, Yiming Wang, Qingming Li, Chunyi Zhou, Songze Li, Ting Wang, Yunjun Gao, Yingcai Wu, and Shouling Ji. Navigating the risks: A survey of security, privacy, and ethics threats in llm-based agents, 2024.

[100] Lu Wang, Max Song, Rezvaneh Rezapour, Bum Chul Kwon, and Jina Huh-Yoo. People's perceptions toward bias and related concepts in large language models: A systematic review, 2024.

[101] Jakub Harasta, Tereza Novotná, and Jaromir Savelka. It cannot be right if it was written by ai: On lawyers' preferences of documents perceived as authored by an llm vs a human, 2024.

[102] Zelong Li, Shuyuan Xu, Kai Mei, Wenyue Hua, Balaji Rama, Om Raheja, Hao Wang, He Zhu, and Yongfeng Zhang. Autoflow: Automated workflow generation for large language model agents, 2024.

[103] Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. Psy-llm: Scaling up global mental health psychological services with ai-based large language models, 2023.

[104] Yuan Li, Yue Huang, Hongyi Wang, Xiangliang Zhang, James Zou, and Lichao Sun. Quantifying ai psychology: A psychometrics benchmark for large language models, 2024.

[105] Pablo Biedma, Xiaoyuan Yi, Linus Huang, Maosong Sun, and Xing Xie. Beyond human norms: Unveiling unique values of large language models through interdisciplinary approaches, 2024.

[106] Hosein Hasanbeig, Hiteshi Sharma, Leo Betthauser, Felipe Vieira Frujeri, and Ida Momennejad. Allure: Auditing and improving llm-based evaluation of text using iterative in-context-learning, 2023.

[107] Yingzhe Peng, Xiaoting Qin, Zhiyang Zhang, Jue Zhang, Qingwei Lin, Xu Yang, Dongmei Zhang, Saravan Rajmohan, and Qi Zhang. Navigating the unknown: A chat-based collaborative interface for personalized exploratory tasks, 2024.

[108] Baolin Li, Yankai Jiang, Vijay Gadepally, and Devesh Tiwari. Llm inference serving: Survey of recent advances and opportunities, 2024.

[109] Zhe Su, Xuhui Zhou, Sanketh Rangreji, Anubha Kabra, Julia Mendelsohn, Faeze Brahman, and Maarten Sap. Ai-liedar: Examine the trade-off between utility and truthfulness in llm agents, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.